SIGDIAL 2023

The 24th Meeting of the Special Interest Group on Discourse and Dialogue

Proceedings of the Conference

September 11 - 15, 2023
Prague, Czechia

©2023 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-028-8

Preface

We are glad to pen the first few words for the proceedings of SIGDIAL 2023, the 24rd Annual Meeting of the Special Interest Group on Discourse and Dialogue. The SIGDIAL conference is a premier publication venue for research in discourse and dialogue. This year the conference is organized together with the conference on International Natural Language Generation (INLG). The format is hybrid with most participants and presenters in-person. Zoom was used for remote presentations and Discord was used as a communication platform for both remote and local participants.

The joint SIGDIAL-INLG 2023 took place on September 11-15, 2023 in Prague, Czech Republic at OREA Hotel Pyramida. The joint conference was collocated with five full-day and one half-day workshops and one satellite event on September 11-12:

- Taming Large Language Models: Controllability in the era of Interactive Assistants
- Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge
- The 1st Workshop on Counter Speech for Online Abuse
- The Eleventh Dialog System Technology Challenge
- PracticalD2T: 1st Workshop on Practical LLM-assisted Data-to-Text Generation
- Connecting multiple disciplines to AI techniques in interaction-centric autism research and diagnosis (half-day workshop)
- The 2023 Young Researchers' Roundtable on Spoken Dialog Systems (YRRSDS 2023) was held as a satellite event

We thank the organizers of the workshops. We are grateful for their coordination with the main conference.

SIGDIAL received 136 submissions this year, comprising 87 long papers, 42 short papers, and 7 demo descriptions. We have aimed to develop a broad, varied program spanning the many positively-rated papers identified by the review process. We therefore accepted 61 papers in total: 41 long papers (47%), 16 short papers (38%), and 4 demo descriptions, for an overall acceptance rate of 45.5%. In addition, we considered 7 articles from Dialogue and Discourse journal and selected 3 for poster presentations.

SIGDIAL-INLG 2023 jointly featured 4 keynotes (one of them presented remotely), a panel discussion, and a joint virtual poster/demo session. The SIGDIAL 2023 program featured 5 oral sessions with 24 talks. The presented topics included Analysis of discourse and dialogue, LLM for dialogue, Dialogue modeling and evaluation, Language understanding and multimodality, and Topics in open-domain dialogue (arguments, opinions, empathy).

7 of the talks were presented virtually and were evenly distributed among the oral sessions. The format of the talks was a 15 minute presentation with 5 minutes for Q&A. The conference had two in-person poster-demo sessions featuring a total of 27 poster and 2 demo presentations. A virtual joint SIGDIAL-INLG poster/demo session was held on Discord during the conference and featured 8 posters and 2 demos from SIGDIAL submissions.

In organizing this hybrid in-person/ remote conference, we have tried to maintain as much of the spirit of a fully in-person conference as possible, allowing opportunities for questions and discussion both from in-person and remote audiences. Online participants were able to ask questions using the Discord platform which also featured a channel for online discussions.

We had 131 reviewers and 13 Senior Program Committee (SPC) members, who were each responsible for 9-11 papers, leading the discussion process and also contributing with meta-reviews. Each submission was assigned to an SPC member and received at least three reviews. Decisions carefully considered the original reviews, meta-reviews, and discussions among reviewers facilitated by the SPCs. We are immensely grateful to the members of the Program Committee and Senior Program Committee for their efforts in providing excellent, thoughtful reviews of the large number of submissions. Their contributions have been essential to selecting the accepted papers and providing a high-quality technical program for the conference.

A conference of this scale requires the energy, guidance, and contributions of many parties, and we would like to take this opportunity to thank and acknowledge them all. We thank our four keynote speakers, Emmanuel Dupoux (Ecole des Hautes Etudes en Sciences Sociales), Ryan Lowe (OpenAI), Barbara Di Eugenio (University of Illinois Chicago), and Elena Simperl (King's College London) for their inspiring talks on "Textless NLP: towards language processing from raw audio", "Aligning ChatGPT: past, present, and future", "Engaging the Patient in Healthcare: Summarization and Interaction", and "Knowledge graph use cases in natural language generation", respectively.

Ryan Lowe's talk was followed by a panel discussion on 'Social Impact of LLMs'. We thank the panel chair David Traum and the Panelists: Malihe Alikhani, Maria Keet, Ryan Lowe, and Ehud Reiter for engaging discussion on this important topic.

SIGDIAL 2023 was made possible by the dedication and hard work of our community, and we are indebted to many. The hybrid nature (in-person and remote), the collocation with the INLG and seven workshops put additional burden on the organization process. The conference would not have been possible without the advice and support of the SIGDIAL board, particularly Gabriel Skantze and Milica Gasic as well as Emiel van Miltenburg and Dave Howcroft who helped coordination between the collocated events.

The tireless work by the local organizing team led by Ondřej Dušek who was involved in countless discussions prior and during the conference coordinating SIGDIAL, INLG, and collocated workshops. We thank the local team who ensured that the conference ran very smoothly, and was enjoyed greatly by all participants. Without that team, there would not have been a conference.

Special thanks go to Zdenek Kasner and Ondrej Platek for their tireless efforts in managing the website with timely updates, and to the team handling various online aspects of participation: Ondrej Platek, Patricia Schmidtova, Dave Howcroft. We would like to thank Patricia Schmidtova, Mateusz Lango and Simone Balloccu for further help with conference preparation. We are grateful to Souro Mukherjee, Kirill Semenov, Nalin Kumar, and Peter Polák, as well as Zdenek Kasner, Ondrej Platek, Patricia Schmidtova, Simone Balloccu, and Mateusz Lango again, for support with the registration, A/V and all other local organizing tasks. Many thanks also go to Jan Hajič for his support, and especially to Anna Kotěšovcová for making all the local arrangements possible. We would also like to thank the sponsorship chair Ramesh Manuvinakurike, who brought to the conference an impressive panel of conference sponsors. We gratefully acknowledge the support of our sponsors: LivePerson (Platinum), LuxAI (Platinum), Apple (Gold), Furhat Robotics (Silver), AX Semantics (Bronze), and Bloomberg (Bronze). In addition, we thank Malihe Alikhani, the publication chair, and Casey Kennington, the mentoring chair for their dedicated service.

Finally, it was our great pleasure to welcome you physically and remotely to the conference. We hope that we have provided an enriching and productive experience at the joint SIGDIAL-INLG 2023.

Svetlana Stoyanchev, Shafiq Joty, Program Co-Chairs
David Schlangen, General Chair

Organizing Committee

General chair:

David Schlangen, University of Potsdam

Program chair 1:

Svetlana Stoyanchev, Toshiba Cambridge Research Laboratory

Program chair 2:

Shafiq Joty, Nanyang Technological University / Salesforce Research

Sponsorship chair:

Ramesh Manuvinakurike, Intel Labs

Local chair:

Ondrej Dusek, Charles University Prague

Mentoring chair:

Casey Kennington, Boise State University

Publication chair:

Malihe Alikhani, Northeastern University

Program Committee:

Gavin Abercrombie, Heriot Watt University
Tazin Afrin, Educational Testing Service
Ron Artstein, USC Institute for Creative Technologies
Katherine Atwell, Northeastern University
M Saiful Bari, Nanyang Technological University
Timo Baumann, Ostbayerische Technische Hochschule Regensburg
Nitu Bharati, TheOpenUniversity
Nate Blaylock, Canary Speech
Johan Boye, KTH
Kristy Boyer, University of Florida
Chloé Braud, IRIT, CNRS
Hendrik Buschmeier, Bielefeld University
Justine Cassell, Carnegie Mellon University and Inria Paris
Senthil Chandramohan, Staples
Lin Chen, Engineering Manager, Meta Platform Inc.
Derek Chen, Columbia University
Hailin Chen, NTU
Zhiyu Chen, Meta
Jinho D. Choi, Emory University
Paul Crook, Meta
Heriberto Cuayahuitl, University of Lincoln
Vera Demberg, Saarland University
Nina Dethlefs, University of Hull
David DeVault, Anticipant Speech, Inc.

Barbara Di Eugenio, University of Illinois at Chicago
Bosheng Ding, Nanyang Technological University
Rama Sanand Doddipatla, Toshiba Cambridge Research Laboratory
Cecilia Domingo, TheOpenUniversity
Ondrej Dusek, Charles University
Alex Fabbri, Salesforce AI Research
Youmna Farag, University of Cambridge
Shutong Feng, Heinrich-Heine-Universität Düsseldorf
Elisa Ferracane, Abridge AI, Inc.
Milica Gasic, Heinrich Heine University Duesseldorf
Kallirroi Georgila, University of Southern California Institute for Creative Technologies
Alborz Geramifard, Facebook AI
Felix Gervits, US Army Research Laboratory
Tirthankar Ghosal, Oak Ridge National Laboratory
Jonathan Ginzburg, Université Paris Cité
Akhilesh Deepak Gotmare, Salesforce Research
Venkata Subrahmanyan Govindarajan, University of Texas at Austin
Yulia Grishina, Amazon
David Gros, University of California - Davis
Prakhar Gupta, Carnegie Mellon University
Joakim Gustafson, KTH
Dilek Hakkani-Tur, Amazon Alexa AI
Devamanyu Hazarika, Amazon
Jie He, University of Edinburgh
Larry Heck, Georgia Institute of Technology
Behnam Hedayatnia, Amazon
Ryuichiro Higashinaka, Nagoya University/NTT
Yufang Hou, IBM Research
Julian Hough, Swansea University
David M. Howcroft, Edinburgh Napier University
Christine Howes, University of Gothenburg
Ruihong Huang, Texas A&M University
Michimasa Inaba, The University of Electro-Communications
Mert Inan, Northeastern University
Koji Inoue, Kyoto University
Di Jin, Amazon
Prathyusha Jwalapuram, Rakuten
Tatsuya Kawahara, Kyoto University
Simon Keizer, Toshiba Europe Ltd
Seokhwan Kim, Amazon Alexa AI
Kazunori Komatani, Osaka University
Philippe Laban, Salesforce Research
Mateusz Lango, Poznan University of Technology
Staffan Larsson, University of Gothenburg
Kornel Laskowski, Carnegie Mellon University
Fabrice Lefèvre, Avignon Univ.
Hsien-chin Lin, Heinrich Heine University
Zhaojiang Lin, Meta
Pierre Lison, Norwegian Computing Centre
Yang Janet Liu, Georgetown University
Linlin Liu, Nanyang Technological University

Eduardo Lleida Solano, University of Zaragoza
Nurul Lubis, Heinrich Heine University
Ramesh Manuvinakurike, Intel labs
Alessandro Mazzei, Università degli Studi di Torino
Teruhisa Misu, Honda Research Institute USA
Anhad Mohananey, Google
Tasnim Mohiuddin, Nanyang Technological University
Han Cheol Moon, Nanyang Technological University
Roger Moore, University of Sheffield
Sourabrata Mukherjee, Charles University
Philippe Muller, IRIT, University of Toulouse
Satoshi Nakamura, Nara Institute of Science and Technology
Mikio Nakano, C4A Research Institute, Inc.
Anna Nedoluzhko, Charles University in Prague
Douglas O'Shaughnessy, INRS-EMT (Univ. of Quebec)
Suraj Pandey, The Open University
Rebecca Passonneau, The Pennsylvania State University
Paul Piwek, The Open University
Ondrej Platek, Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Massimo Poesio, Queen Mary University of London
Stephen Pulman, Apple Inc.
Matthew Purver, Queen Mary University of London
Kun Qian, Columbia University
Chengwei Qin, Nanyang Technological University
Liang Qiu, Amazon Alexa AI
Vikram Ramanarayanan, University of California, San Francisco
Hannah Rashkin, Google Research
Mathieu Ravaut, Nanyang Technological University
Ehud Reiter, University of Aberdeen
Giuseppe Riccardi, University of Trento
Antonio Roque, Soar Technology
Saurav Sahay, Intel Labs
Sakriani Sakti, Japan Advanced Institute of Science and Technology / Nara Institute of Science and Technology (JAIST/NAIST)
Chinnadhurai Sankar, SliceX AI
Ruhi Sarikaya, Amazon
David Schlangen, University of Potsdam
Ethan Selfridge, LivePerson
Samira Shaikh, University of North Carolina at Charlotte
Anthony Sicilia, Northeastern University
Gabriel Skantze, KTH Speech Music and Hearing
Georg Stemmer, Intel Corp.
Amanda Stent, Colby College
Matthew Stone, Rutgers University
Carl Strathearn, Edinburgh Napier University
Kristina Striegnitz, Union College
Hiroaki Sugiyama, NTT Communication Science Labs.
Alessandro Suglia, Heriot-Watt University
António Teixeira, DETI/IEETA, University of Aveiro
Megh Thakkar, BITS Pilani

Takenobu Tokunaga, Tokyo Institute of Technology
Stefan Ultes, University of Bamberg
Carel van Niekerk, Heinrich Heine University
David Vandyke, Apple
Renato Vukovic, Heinrich Heine University Düsseldorf
Hsin-Min Wang, Academia Sinica
Yi-Chia Wang, Facebook AI
Nigel Ward, University of Texas at El Paso
Michael White, The Ohio State University
Qingyang Wu, Columbia University
Chien-Sheng Wu, Salesforce
Deyi Xiong, Tianjin University
Yuxiao Ye, University of Cambridge
Koichiro Yoshino, Guardian Robot Project (GRP) RIKEN, Nara Institute of Science and Technology
Yanchao Yu, Edinburgh Napier University
Hongli Zhan, University of Texas at Austin
Tiancheng Zhao, Binjiang Institute of Zhejiang University
Ruochen Zhao, Nanyang Technological University
Mingyang Zhou, Post Doctoral Research Scientist at Columbia University
Ingrid Zukerman, Monash University

Invited Speakers:

Barbara Di Eugenio, University of Illinois Chicago
Emmanuel Dupoux, Ecole des Hautes Etudes en Sciences Sociales (EHESS) / Meta AI Labs
Elena Simperl, King's College London
Ryan Lowe, OpenAI

Table of Contents

Conference Program

Wednesday September 13, 2023

09:00–09:15 Opening

09:15–10:15 Keynote: Textless NLP: towards language processing from raw audio

10:15–10:45 Coffee Break

10:45–12:30 Oral Session 1: Analysis of discourse and dialogue

Sources of Noise in Dialogue and How to Deal with Them
Derek Chen and Zhou Yu

Investigating Explicitation of Discourse Connectives in Translation using Automatic Annotations
Frances Yung, Merel Scholman, Ekaterina Lapshinova-Koltunski, Christina Pollkläsener and Vera Demberg

What's Hard in English RST Parsing? Predictive Models for Error Analysis
Yang Janet Liu, Tatsuya Aoyama and Amir Zeldes

Grounded Complex Task Segmentation for Conversational Assistants
Rafael Ferreira, David Semedo and Joao Magalhaes

A Statistical Approach for Quantifying Group Difference in Topic Distributions Using Clinical Discourse Samples
Grace Lawley, Peter A. Heeman, Jill K. Dolata, Eric Fombonne and Steven Bedrick

Wednesday September 13, 2023 (continued)

12:30–13:30 Lunch

13:30–15:10 Poster Session 1

OpinionConv: Conversational Product Search with Grounded Opinions
Vahid Sadiri Javadi, Martin Potthast and Lucie Flek

Dial-M: A Masking-based Framework for Dialogue Evaluation
Suvodip Dey and Maunendra Sankar Desarkar

From Chatter to Matter: Addressing Critical Steps of Emotion Recognition Learning in Task-oriented Dialogue
Shutong Feng, Nurul Lubis, Benjamin Ruppik, Christian Geishauser, Michael Heck, Hsien-chin Lin, Carel van Niekerk, Renato Vukovic and Milica Gasic

Analyzing Differences in Subjective Annotations by Participants and Third-party Annotators in Multimodal Dialogue Corpus
Kazunori Komatani, Ryu Takeda and Shogo Okada

Frame-oriented Summarization of Argumentative Discussions
Shahbaz Syed, Timon Ziegenbein, Philipp Heinisch, Henning Wachsmuth and Martin Potthast

Towards Multilingual Automatic Open-Domain Dialogue Evaluation
John Mendonca, Alon Lavie and Isabel Trancoso

Dialog Action-Aware Transformer for Dialog Policy Learning
Huimin Wang, Wai Chung Kwan and Kam-Fai Wong

The Wizard of Curiosities: Enriching Dialogues with Fun Facts
Frederico Vicente, Rafael Ferreira, David Semedo and Joao Magalhaes

The Road to Quality is Paved with Good Revisions: A Detailed Evaluation Methodology for Revision Policies in Incremental Sequence Labelling
Brielen Madureira, Patrick Kahardipraja and David Schlangen

The effect of conversation type on entrainment: Evidence from laughter
Bogdan Ludusan and Petra Wagner

'What are you referring to?' Evaluating the Ability of Multi-Modal Dialogue Models to Process Clarificational Exchanges
Javier Chiyah-Garcia, Alessandro Suglia, Arash Eshghi and Helen Hastie

PGTask: Introducing the Task of Profile Generation from Dialogues
Rui Ribeiro, Joao Paulo Carvalho and Luisa Coheur

Question Generation to Elicit Users' Food Preferences Considering the Semantic Content
Jie Zeng, Yukiko Nakano and Tatsuya Sakato

Roll Up Your Sleeves: Working with a Collaborative and Engaging Task-Oriented Dialogue System
Lingbo Mo, Shijie Chen, Ziru Chen, Xiang Deng, Ashley Lewis, Sunit Singh, Samuel Stevens, Chang-You Tai, Zhen Wang, Xiang Yue, Tianshu Zhang, Yu Su and Huan Sun

15:10–15:40 Coffee Break

15:40–17:00 Oral Session 2: LLM for dialogue

Leveraging Large Language Models for Automated Dialogue Analysis
Sarah E. Finch, Ellie S. Paek and Jinho D. Choi

Are Large Language Models All You Need for Task-Oriented Dialogue?
Vojtěch Hudeček and Ondrej Dusek

Multi-party Goal Tracking with LLMs: Comparing Pre-training, Fine-tuning, and Prompt Engineering
Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernandez Garcia, Christian Dondrup and Oliver Lemon

ChatGPT vs. Crowdsourcing vs. Experts: Annotating Open-Domain Conversations with Speech Functions
Lidiia Ostyakova, Veronika Smilga, Kseniia Petukhova, Maria Molchanova and Daniel Kornev

Wednesday September 13, 2023 (continued)

17:10–18:10 Keynote: Aligning ChatGPT: past, present, and future

18:10–18:45 Panel Discussion: Social Impact of LLMs

19:00–20:00 Welcome Reception

Thursday September 14, 2023

09:00–10:00 Keynote: Engaging the Patient in Healthcare: Summarization and Interaction

10:00–10:30 Coffee Break

10:30–12:10 Oral Session 3: Dialogue modeling and evaluation

DiactTOD: Learning Generalizable Latent Dialogue Acts for Controllable Task-Oriented Dialogue Systems
Qingyang Wu, James Gung, Raphael Shu and Yi Zhang

Approximating Online Human Evaluation of Social Chatbots with Prompting
Ekaterina Svikhnushina and Pearl Pu

Dialogue Response Generation Using Completion of Omitted Predicate Arguments Based on Zero Anaphora Resolution
Ayaka Ueyama and Yoshinobu Kano

Syndicom: Improving Conversational Commonsense with Error-Injection and Natural Language Feedback
Christopher Richardson and Larry Heck

"What do others think?": Task-Oriented Conversational Modeling with Subjective Knowledge
Chao Zhao, Spandana Gella, Seokhwan Kim, Di Jin, Devamanyu Hazarika, Alexandros Papangelis, Behnam Hedayatnia, Mahdi Namazifar, Yang Liu and Dilek Hakkani-Tur

Thursday September 14, 2023 (continued)

12:10–13:00 SIGDIAL Business meeting

12:10–13:00 SIGGEN Business meeting

13:00–14:00 Lunch

14:00–15:40 Poster Session 2

UD_Japanese-CEJC: Dependency Relation Annotation on Corpus of Everyday Japanese Conversation
Mai Omura, Hiroshi Matsuda, Masayuki Asahara and Aya Wakasa

Unravelling Indirect Answers to Wh-Questions: Corpus Construction, Analysis, and Generation
Zulipiye Yusupujiang and Jonathan Ginzburg

A New Dataset for Causality Identification in Argumentative Texts
Khalid Al Khatib, Michael Voelske, Anh Le, Shahbaz Syed, Martin Potthast and Benno Stein

Controllable Generation of Dialogue Acts for Dialogue Systems via Few-Shot Response Generation and Ranking
Angela Ramirez, Kartik Agarwal, Juraj Juraska, Utkarsh Garg and Marilyn Walker

Reference Resolution and New Entities in Exploratory Data Visualization: From Controlled to Unconstrained Interactions with a Conversational Assistant
Abari Bhattacharya, Abhinav Kumar, Barbara Di Eugenio, Roderick Tabalba, Jillian Aurisano, Veronica Grosso, Andrew Johnson, Jason Leigh and Moira Zellner

CONVERSER: Few-shot Conversational Dense Retrieval with Synthetic Data Generation
Chao-Wei Huang, Chen-Yu Hsu, Tsu-Yuan Hsu, Chen-An Li and Yun-Nung Chen

Speaker Role Identification in Call Centre Dialogues: Leveraging Opening Sentences and Large Language Models
Minh-Quoc Nghiem, Nichola Roberts and Dmitry Sityaev

Synthesising Personality with Neural Speech Synthesis
Shilin Gao, Matthew P. Aylett, David A. Braude and Catherine Lai

Prompting, Retrieval, Training: An exploration of different approaches for task-oriented dialogue generation
Gonçalo Raposo, Luisa Coheur and Bruno Martins

Bootstrapping a Conversational Guide for Colonoscopy Prep
Pulkit Arya, Madeleine Bloomquist, SUBHANKAR CHAKRABORTY, Andrew Perrault, William Schuler, Eric Fosler-Lussier and Michael White

Applying Item Response Theory to Task-oriented Dialogue Systems for Accurately Determining User's Task Success Ability
Ryu Hirai, Ao Guo and Ryuichiro Higashinaka

An Open-Domain Avatar Chatbot by Exploiting a Large Language Model
Takato Yamazaki, Tomoya Mizumoto, Katsumasa Yoshikawa, Masaya Ohagi, Toshiki Kawamoto and Toshinori Sato

15:40–16:10 Coffee Break

16:10–17:50 Oral Session 3: Language understanding and multimodality

Learning Multimodal Cues of Children's Uncertainty
Qi Cheng, Mert Inan, Rahma Mbarki, Grace Grmek, Theresa Choi, Yiming Sun, Kimele Persaud, Jenny Wang and Malihe Alikhani

Grounding Description-Driven Dialogue State Trackers with Knowledge-Seeking Turns
Alexandru Coca, Bo-Hsiang Tseng, Jinghong Chen, Weizhe Lin, Weixuan Zhang, Tisha Anders and Bill Byrne

Resolving References in Visually-Grounded Dialogue via Text Generation
Bram Willemsen, Livia Qian and Gabriel Skantze

Slot Induction via Pre-trained Language Model Probing and Multi-level Contrastive Learning
Hoang Nguyen, Chenwei Zhang, Ye Liu and Philip Yu

The timing bottleneck: Why timing and overlap are mission-critical for conversational user interfaces, speech recognition and dialogue systems
Andreas Liesenfeld, Alianda Lopez and Mark Dingemanse

Thursday September 14, 2023 (continued)

17:35–18:30 GenChal Poster Session + demos

19:00–22:00 Conference Dinner

Friday September 15, 2023

09:00–10:00 Virtual Poster Session

Enhancing Task Bot Engagement with Synthesized Open-Domain Dialog
Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao and Zhu (Drew) Zhang

Enhancing Performance on Seen and Unseen Dialogue Scenarios using Retrieval-Augmented End-to-End Task-Oriented System
Jianguo Zhang, Stephen Roller, Kun Qian, Zhiwei Liu, Rui Meng, Shelby Heinecke, Huan Wang, silvio savarese and Caiming Xiong

Transformer-based Multi-Party Conversation Generation using Dialogue Discourse Acts Planning
Alexander Chernyavskiy and Dmitry Ilvovsky

Incorporating Annotator Uncertainty into Representations of Discourse Relations
S. Magalí López Cortez and Cassandra L. Jacobs

Investigating the Representation of Open Domain Dialogue Context for Transformer Models
Vishakh Padmakumar, Behnam Hedayatnia, Di Jin, Patrick Lange, Seokhwan Kim, Nanyun Peng, Yang Liu and Dilek Hakkani-Tur

C3: Compositional Counterfactual Contrastive Learning for Video-grounded Dialogues
Hung Le, Nancy Chen and Steven C.H. Hoi

No that's not what I meant: Handling Third Position Repair in Conversational Question Answering
Vevake Balaraman, Arash Eshghi, Ioannis Konstas and Ioannis Papaioannou

When to generate hedges in peer-tutoring interactions
Alafate Abulimiti, Chloé Clavel and Justine Cassell

PaperPersiChat: Scientific Paper Discussion Chatbot using Transformers and Discourse Flow Management
Alexander Chernyavskiy, Max Bregeda and Maria Nikiforova

Friday September 15, 2023 (continued)

FurChat: An Embodied Conversational Agent using LLMs, Combining Open and Closed-Domain Dialogue with Facial Expressions
Neeraj Cherakara, Finny Varghese, Sheena Shabana, Nivan Nelson, Abhiram Karukayil, Rohith Kulothungan, Mohammed Afil Farhan, Birthe Nesset, Meriam Moujahid, Tanvi Dinkar, Verena Rieser and Oliver Lemon

10:00–11:00 Keynote: Knowledge graph use cases in natural language generation

11:00–11:30 Coffee Break

11:30–13:10 Oral Session 5: Topics in open-domain dialogue

Towards Breaking the Self-imposed Filter Bubble in Argumentative Dialogues
Annalena Aicher, Daniel Kornmueller, Yuki Matsuda, Stefan Ultes, Wolfgang Minker and Keiichi Yasumoto

The Open-domain Paradox for Chatbots: Common Ground as the Basis for Human-like Dialogue
Gabriel Skantze and A. Seza Doğruöz

MERCY: Multiple Response Ranking Concurrently in Realistic Open-Domain Conversational Systems
Sarik Ghazarian, Behnam Hedayatnia, Di Jin, Sijia Liu, Nanyun Peng, Yang Liu and Dilek Hakkani-Tur

Empathetic Response Generation for Distress Support
Anuradha Welivita, Chun-Hung Yeh and Pearl Pu

Reasoning before Responding: Integrating Commonsense-based Causality Explanation for Empathetic Response Generation
Yahui Fu, Koji Inoue, Chenhui Chu and Tatsuya Kawahara

Friday September 15, 2023 (continued)

13:10–14:10 Lunch

14:10–14:40 Sponsors

14:40–15:00 Closing

15:00–15:30 Birds-of-Feather

15:00–15:30 Coffee Break

Keynote Abstracts

Keynote 1: Engaging the Patient in Healthcare: Summarization and Interaction
Barbara Di Eugenio
University of Illinois, Chicago

Abstract: Effective and compassionate communication with patients is becoming central to healthcare. I will discuss the results of and lessons learned from three ongoing projects in this space. The first, MyPHA, aims to provide patients with a clear and understandable summary of their hospital stay, which is informed by doctors' and nurses' perspectives, and by the strengths and concerns of the patients themselves. The second, SMART-SMS, models health coaching interactions via text exchanges that encourage patients to adopt specific and realistic physical activity goals. The third, HFChat, envisions an always-on-call conversational assistant for heart failure patients, that they can ask for information about lifestyle issues such as food and exercise. All our work is characterized by: large interdisciplinary groups of investigators who bring different perspectives to the research; grounding computational models in ecologically valid data, which is small by its own nature; the need for culturally valid interventions, since our UI Health system predominantly serves underprivileged, minority populations; and the challenges that arise when dealing with the healthcare enterprise.

Bio: Barbara Di Eugenio is a Professor and Director of Graduate Studies in the Computer Science department at the University of Illinois Chicago. There she leads the NLP laboratory (http://nlp.cs.uic.edu/). She obtained her PhD in Computer Science from the University of Pennsylvania (1993). Her research has always focused on the pragmatics and computational modeling of discourse and dialogue, grounded in authentic data collection on the one hand, and in user studies on the other. The applications of her work run the gamut from educational technology to human-robot interaction, from data visualization to health care. Dr. Di Eugenio is an NSF CAREER awardee (2002); a UIC University Scholar (2018-2020); and a Zenith Award recipient from AWIS, the Association for Women in Science (2022). She has been the editor-in-chief for the Journal of Discourse and Dialogue since 2019. She is very proud to have graduated 15 PhD and 32 Master's students.

Keynote 2: Textless NLP: towards language processing from raw audio

Emmanuel Dupoux
Ecole des Hautes Etudes en Sciences Sociales (EHESS)

Abstract: The oral (or gestural) modality is the most natural channel for human language interactions. Yet, language technology (Natural Language Processing, NLP) is primarily based on the written modality, and requires massive amounts of textual resources for the training of useful language models. As a result, even fundamentally speech-first applications like speech-to-speech translation or spoken assistants like Alexa, or Siri, are constructed in a Frankenstein way, with text as an intermediate representation between the signal and language models. Besides this being inefficient, This has two unfortunate consequences: first, only a small fraction of the world's languages that have massive textual repositories can be addressed by current technology. Second, even for text-rich languages, the oral form mismatches the written form at a variety of levels, including vocabulary and expressions. The oral medium also contains typically unwritten linguistic features like rhythm and intonation (prosody) and rich paralinguistic information (non verbal vocalizations like laughter, cries, clicks, etc, nuances carried through changes in voice qualities) which are therefore inaccessible to language models. But is this a necessity? Could we build language applications directly from the audio stream without using any text? In this talk, we review recent breakthroughs in representation learning and self-supervised techniques which have made it possible to learn latent linguistic units directly from audio which unlock the learning of generative language models without the use of any text. We show that these models can capture heretofore unaddressed nuances of the oral language including in a dialogue context, opening up the possibility of speech-to-speech textless NLP applications. We outline existing technical challenges to achieve this goal, including challenges to build expressive oral language datasets at scale.

Bio: E. Dupoux is professor at the Ecole des Hautes Etudes en Sciences Sociales (EHESS) and Research Scientist at Meta AI Labs. He directs the Cognitive Machine Learning team at the Ecole Normale Supérieure (ENS) in Paris and INRIA. His education includes a PhD in Cognitive Science (EHESS), a MA in Computer Science (Orsay University) and a BA in Applied Mathematics (Pierre & Marie Curie University). His research mixes developmental science, cognitive neuroscience, and machine learning, with a focus on the reverse engineering of infant language and cognitive development using unsupervised or weakly supervised learning. He is the recipient of an Advanced ERC grant, co-organizer of the Zero Ressource Speech Challenge series (2015–2021), the Intuitive Physics Benchmark (2019) and led in 2017 a Jelinek Summer Workshop at CMU on multimodal speech learning. He is a CIFAR LMB and a ELLIS Fellow. He has authored 150 articles in peer reviewed outlets in cognitive science and language technology.

Keynote 3: Knowledge graph use cases in natural language generation

Elena Simperl
King's College London

Abstract: Natural language generation (NLG) makes knowledge graphs (KGs) more accessible. I will present two applications of NLG in this space: in the first one, verbalisations of KG triples feed into downstream KG applications, allowing users with diverse levels of digital literacy to share their knowledge, and contribute to the KG. In the second one, having text representations of KG triples helps us verify the content of a KG against external sources towards more trustworthy KGs. I will present human-in-the-loop solutions to these applications that leverage a range of machine learning techniques to scale to the large, multilingual knowledge graphs modern applications use.

Bio: Elena Simperl is a Professor of Computer Science and Deputy Head of Department for Enterprise and Engagement in the Department of Informatics at King's College London. She is also the Director of Research for the Open Data Institute (ODI) and a Fellow of the British Computer Society and the Royal Society of Arts. Elena features in the top 100 most influential scholars in knowledge engineering of the last decade. She obtained her doctoral degree in Computer Science from the Free University of Berlin, and her diploma in Computer Science from the Technical University of Munich. Prior to joining King's in 2020, she was a Turing Fellow, and held positions in Germany, Austria and at the University of Southampton. Her research is at intersection between AI and social computing, helping designers understand how to build smart sociotechnical systems that combine data and algorithms with human and social capabilities. Elena led 14 European and national research projects, including recently QROWD, ODINE, Data Pitch, Data Stories, and ACTION. She is currently the scientific and technical director of MediaFutures, a Horizon 2020 programme that is using arts-inspired methods to design participatory AI systems that tackle misinformation and disinformation online. Elena's interest in leading initiatives within the scientific community has also taken form through chairing several conferences in her field, including the European and International Semantic Web Conference series, the European Data Forum, and the European Semantic Technologies conference. She is the president of the Semantic Web Science Association.

Keynote 4: Aligning ChatGPT: past, present, and future

Ryan Lowe
OpenAI

Abstract: In this talk I will present different perspectives on the alignment of chatbots like ChatGPT. I'll review reinforcement learning from human feedback (RLHF), the core training technique behind InstructGPT and ChatGPT, including a brief history of how it was developed. I'll discuss some of the pitfalls of RLHF, and what is being done today to address them. I'll then speculate on some of the alignment challenges I expect we'll face with this new generation of powerful personal assistants, how they could reshape society, and some things we'll need to do to make sure these changes are good for humans.

Bio: Ryan is a researcher at OpenAI on the Alignment team. His most recent work involved proving out RLHF on language models, starting with summarization, then moving to InstructGPT and most recently ChatGPT and GPT-4. Previously, he worked on multi-agent RL, emergent communication, and dialogue systems at McGill University.

Sources of Noise in Dialogue and How to Deal with Them

Derek Chen
Columbia University, NY
dc3761@columbia.edu

Zhou Yu
Columbia University, NY
zy2461@columbia.edu

Abstract

Training dialogue systems often entails dealing with noisy training examples and unexpected user inputs. Despite their prevalence, there currently lacks an accurate survey of dialogue noise, nor is there a clear sense of the impact of each noise type on task performance. This paper addresses this gap by first constructing a taxonomy of noise encountered by dialogue systems. In addition, we run a series of experiments to show how different models behave when subjected to varying levels of noise and types of noise. Our results reveal that models are quite robust to label errors commonly tackled by existing denoising algorithms, but that performance suffers from dialogue-specific noise. Driven by these observations, we design a data cleaning algorithm specialized for conversational settings and apply it as a proof-of-concept for targeted dialogue denoising.

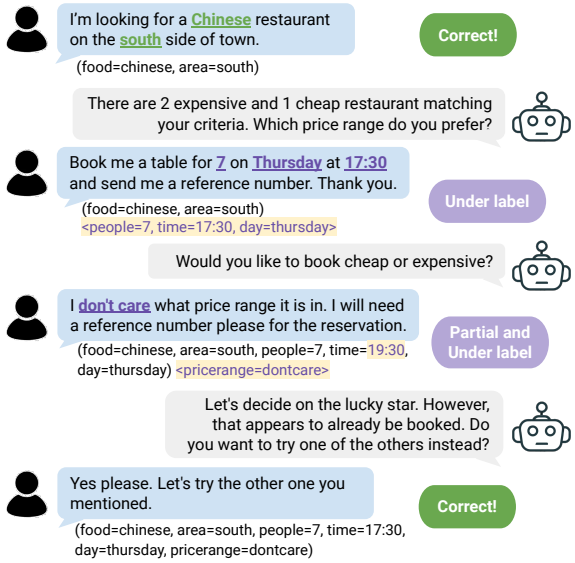


Figure 1: An example of label errors within MultiWoz 2.0 which contains partially filled and missing labels. We categorize this as two types of instance-level noise.

1 Introduction

Quality labeled data is a necessity for properly training deep neural networks. More data often leads to better performance, and dialogue tasks are no exception (Qian and Yu, 2019). However, in the quest for more data, practitioners increasingly rely on crowdsourcing or forms of weak supervision to meet scaling requirements. Even when acting in good faith, crowdworkers are not trained experts which understandably leads to mistakes. This ultimately results in noisy *training inputs* for our conversational agents. Moreover, when dialogue systems are deployed into the real world, they must also deal with noisy *user inputs*. For example, a user might make an ambiguous request or mention an unknown entity. All these sources of noise eventually take their toll on model performance.

Before building noise-robust dialogue systems or denoising dialogue datasets, it would be helpful to know what types of noise exist in the first place. Then our efforts can be spent more wisely tackling the sources of noise that actually make a difference. Prior works have looked into counteracting noisy user interactions (Peng et al., 2021; Liu et al., 2021), but did not study the impact of noisy training data. Moreover, they lack analysis on how noise influences performance across different model types or conversational styles. Other works claim that dialogue agents can be easily biased by offensive language found in noisy training data (Ung et al., 2022; Dinan et al., 2020). Given such a danger, we wonder "How much toxic data actually exists in annotated dialogue data?"

To investigate these concerns, we survey a wide range of popular dialogue datasets and outline the different types of naturally occurring noise. Building on this exercise, we also study the patterns of annotation errors to determine the prevalence of each noise type and identify the most likely causes of noise. Next, we run transformer models through the gamut to find out how well they handle the different types of noise documented in the previous step. In total, we test 3 model types on 7 categories

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 1–20
September 11–15, 2023. ©2023 Association for Computational Linguistics

Dataset	Abbr.	Num. Dialogs	Collection Methodology	Open Domain	Goal Oriented	Synchronous Chat	KB/ Document
Action-Based Conversations Dataset	ABCD	10,042	Expert Live Chat		X	X	X
DailyDialog	DD	13,118	Post-conv Annotation	X			
Empathetic Dialogues	ED	24,850	Live Chat	X		X	
Google Simulated Conversations	GSIM	3,008	Machine to Machine		X		
Key-Value Retrieval for In-Car	KVRET	3,031	Wizard of Oz		X		X
Machine Interaction Dialog Act Schema	MIDAS	468	Live Chat	X		X	
MultiWoz 2.3	MWOZ	10,419	Wizard of Oz		X		
Schema Guided Dialogue	SGD	42,706	Post-conv Annotation		X		
TicketTalk (TaskMaster 3)	TT	23,789	Dialogue Self-Play		X		
Wizard of Wikipedia	WOW	22,311	Wizard of Oz	X		X	X

Table 1: Breakdown of ten dialogue datasets used in constructing the noise taxonomy. The datasets were chosen to span a wide variety of annotation schemes, task specifications and conversation lengths. KB/Document refers to a dataset containing an external knowledge base or document to ground the conversation. (See Appendix A)

of noise across 10 diverse datasets spanning 5 dialogue tasks. We discover that most models are quite robust to the label errors commonly targeted by denoising algorithms (Natarajan et al., 2013; Reed et al., 2015), but perform poorly when subjected to dialogue-specific noise. Finally, to verify we have indeed identified meaningful noise types, we apply our findings to denoise a dataset containing real dialogue noise. As a result, we are able to raise joint goal accuracy on MultiWOZ 2.0 by 42.5% in relative improvement.

In total, our contributions are as follows: (a) Construct a realistic taxonomy of dialogue noise to guide future data collection efforts. (b) Measure the impact of noise on multiple tasks and neural models to aid the development of denoising algorithms. (c) Establish a strong baseline for dealing with noise by resolving dialogue specific concerns, and verify its effectiveness in practice.

2 Dialogue Datasets

A data-driven taxonomy of dialogue noise was designed by manually reviewing thousands of conversations across ten diverse datasets and their accompanying annotations. The datasets were chosen from non-overlapping domains to exhaustively represent all commonly considered dialogue tasks. At a high level, they are divided into six task-oriented dialogue datasets and four open domain chit-chat datasets. The task-oriented datasets are comprised of MultiWoz 2.0 (MWOZ) (Budzianowski et al., 2018), TicketTalk (TT) (Byrne et al., 2019), Schema Guided Dialogue (SGD) (Rastogi et al., 2020), Action Based Conversations Dataset (ABCD) (Chen et al., 2021), Google Simulated Conversations (GSIM) (Shah et al., 2018), and Key-Value Retrieval for In-car Assistant (KVRET) (Eric et al., 2017). The open domain datasets include DailyDialog (DD) (Li et al., 2017), Wizard of Wikipedia (WOW) (Dinan et al., 2019b), Empathetic Dialogues (ED) (Rashkin et al., 2019), and Machine Interaction Dialog Act Schema (MIDAS) (Yu and Yu, 2021). The datasets also span a variety of data collection methodologies, such as M2M or Wizard-of-Oz, which has a close connection to the types of noise produced. We also consider whether the interlocutors engage in real-time vs. non-synchronous chat. Details of each dataset can be found in Table 1 and Appendix A.

The taxonomy creation process starts by uniformly sampling 1% of conversations from each corpus, rounding up as needed to include at least 100 dialogues per dataset. Five expert annotators then conducted two rounds of review per conversation to tally noise counts, with a third round to break ties if needed. The group also cross-referenced each other to merge duplicate categories and resolve disagreements. Notably, the final taxonomy purposely excludes sources of noise that occur less than 0.1% of the time. This active curation supports future denoising research by focusing attention on the most prominent sources of noise.

3 Sources of Noise

Through careful review of the data, we discover that dialogue systems encounter issues either from noisy training inputs during model development or from noisy user inputs during model inference.

3.1 Training Noise

Noisy training data impacts model learning, before any user interaction with the system. The sources of noise are derived from labeling errors, ontology inconsistencies or undesirable discourse attributes.

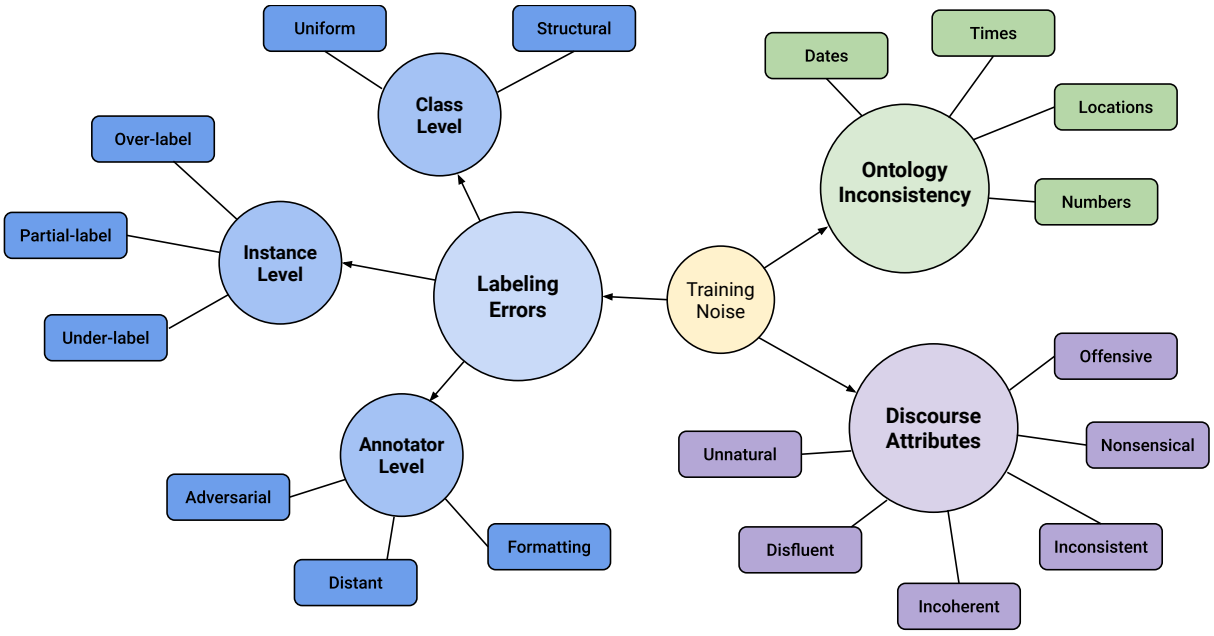


Figure 2: Diagram of the main sources of noise that affect training, based on review of the data. Our taxonomy also includes inference noise which occurs when users interact with the dialogue agent (See Fig 3).

3.1.1 Labeling Errors

For a given dataset of (X, Y) pairs, any occasion when the target label y is labeled incorrectly can be considered a labeling error.

Class Level When noise occurs due to confusion between two classes, this is considered a class-level labeling error. This can be further sub-divided into *Uniform Label Swapping* or *Structured Label Swapping*. In the former, symmetric noise implies all classes have equal likelihood to be confused with any other class, whereas in the latter certain classes are more likely to confused with other related classes. For example, "anger" as a label is more likely to be confused with "frustration" than "joy" when performing emotion detection.

Instance Level Noise comes from the example itself due to the complexity of interpreting natural language, which is especially common within dialogues (Zhang et al., 2021). For example, annotators may carry over the dialogue act from the previous turn, even though it is no longer relevant, resulting in *Over Labeling*. Conversely, *Under Labeling* is when a label is missed. *Partial Labels* occur when some labels are correct, while others are not. This is common in dialogue due to the prevalence of multi-label examples, such as an utterance with two slot-values to fill. (See Figure 1)

Annotation Level Noise arises due to the labeler or data collection process. (Snow et al., 2004). Applying heuristics on a gazetteer to label named entities in NER produces *Distant Supervision* noise. Human annotators are also a source of noise either purposely from *Adversarial Actors* or inadvertently from annotators acting in good faith still leading to *Formatting Mistakes*. (See Table 2)

3.1.2 Ontology Inconsistency

Another source of noise comes from inconsistent formatting when constructing the ontology. The only entities which actually contained issues are (a) **Dates**: tomorrow, Jan 3rd, 1/3/2022, January 3 (b) **Times**: 14:15, 2:15 PM, quarter past 2, 215pm (c) **Locations**: NYC, New York, ny, the big apple (d) **Numbers**: three, 'wife daughter & I', 3, 'Me and my two buddies'. In contrast, inconsistent names (ie. Fred Miyato, Mr. Miyato, fred miyato, my father) only occurred occasionally. Lack of standardization in the ontology was so pronounced in certain datasets that classifying labels becomes untenable, leaving generation or copying as the only viable method of predicting slot-values.

3.1.3 Discourse Attributes

Dialogue agents developed for response generation often mimic the behavior found in the training examples, so one hopes they contain positive discourse attributes while avoiding negative ones. We identify six such attributes by following qualitative metrics commonly used for dialogue evaluation and through our own review of the conversations.

Dialogue	Labels

SGD – [Ontology Inconsistency > Date, Time]
User: I need a rental car in Chicago on the 3rd of this month.
System: When and for how long will you need the car?
User: I'd like it from 12:30 in the afternoon till next Wednesday.
...
System: So you'd like to reserve a standard car from March 3rd at 12:30 pm until March 6th from the O'Hare International Airport location?
User: Yes that'll work
	`GetCarsAvailable(pickup_city=Chicago, pickup_date=3rd of this month)`
	
	`GetCarsAvailable(pickup_time=12:30 in the afternoon, dropoff_date=Wednesday)`
	
	`ReserveCar(dropoff_date=March 6th, pickup_time=12:30 pm)`

MIDAS – [Discourse Attribute > Incoherent] User: one guy Agent: what do you think about christopher nolan's acting User: you can't get a boy	Revised dialog act: `statement` → `nonsense`

TT – [Labeling Error > Annotator Level > Formatting]
User: We would like to see the Rhythm Section. That sounds good.
Assistant: How many tickets will you need today?
User: We will need 4 tickets.
Assistant: Where would you like to see the movie?
User: We would like to see it in San Antonio at Cinemark McCreless Market.
	`(name.movie='the Rhythm Section')`
	
	`(num.tickets=4)`
	`(location='San Antonio', name.theater ='inemark McCreless Market.')`

Table 2: Selected qualitative examples of dialogue noise. Best viewed in color. Many more examples in Appendix I.

(1) **Fluent** utterances flow well, obey proper grammar, and are syntactically valid. (2) **Coherent** dialogues are semantically valid, and make sense such that they are interpretable and understandable by a general audience. (3) **Consistent** models do not contradict what was stated earlier in the conversation, or haphazardly change their stance on a subject. (4) **Sensible** models follow common sense principles and understand basic natural laws (ie. gravity). (5) **Polite** dialogue models avoid toxic language or offensive speech. They should not exhibit overt bias towards certain groups or minorities. (6) **Natural** dialogues reflect how people generally talk in real life. In addition, the speakers should not break the fourth wall by directly or indirectly referring to the data collection process.

3.2 Inference Noise

Inference noise refers to issues that occur in test time, during user interaction with the system after deployment to production. This aligns nicely with the concept of out-of-scope errors (Chen and Yu, 2021), which are made up of two categories: out-of-distribution cases and dialogue breakdowns.

3.2.1 Out-of-Distribution (OOD)

Causes of OOD (Peng et al., 2021) include:

Novel queries The user asks the model to do something it was not trained to do. Example: the customer asks about frequent flyer miles, but the agent is only capable of making or modifying flight reservations. The model fails for these requests since it was never taught to handle such queries.

Unseen entities Facing new entities or values not seen during training. Although difficult, we could still expect a model to understand a portion of such queries by generalizing from the context. For example, "I would like a flight from Miami to Puffville". Even though the model has never heard of 'Puffville', it can infer from context that this is the desired value for the destination slot.

Domain shift The dialogue system must make predictions in a new domain (taxi vs. flight). Commonly tackled in zero-shot settings, we can expect models to occasionally generalize because there may be shared slots across domains (ie. departure time is shared by both taxi and flights queries).

3.2.2 Dialogue Breakdown

In contrast to OOD issues, dialogue breakdowns are situations a model should be able to handle since the scenario is within the bounds (i.e. in-domain) of what the model was trained to understand (Higashinaka et al., 2016). However, due to noise from ambiguous or unclear user input, communication breaks down and the conversation is unable to continue. (Higashinaka et al., 2015).

Ambiguous Meaning Query or statement that the model should be able to handle, but caused confusion, possibly because the model failed to take the dialogue context into consideration. For example, a co-reference issue may cause difficulty in interpreting the user intent. "Yea, let's go with that one" is unclear when viewed in isolation. To resolve this type of noise a model should look at the broader conversational context.

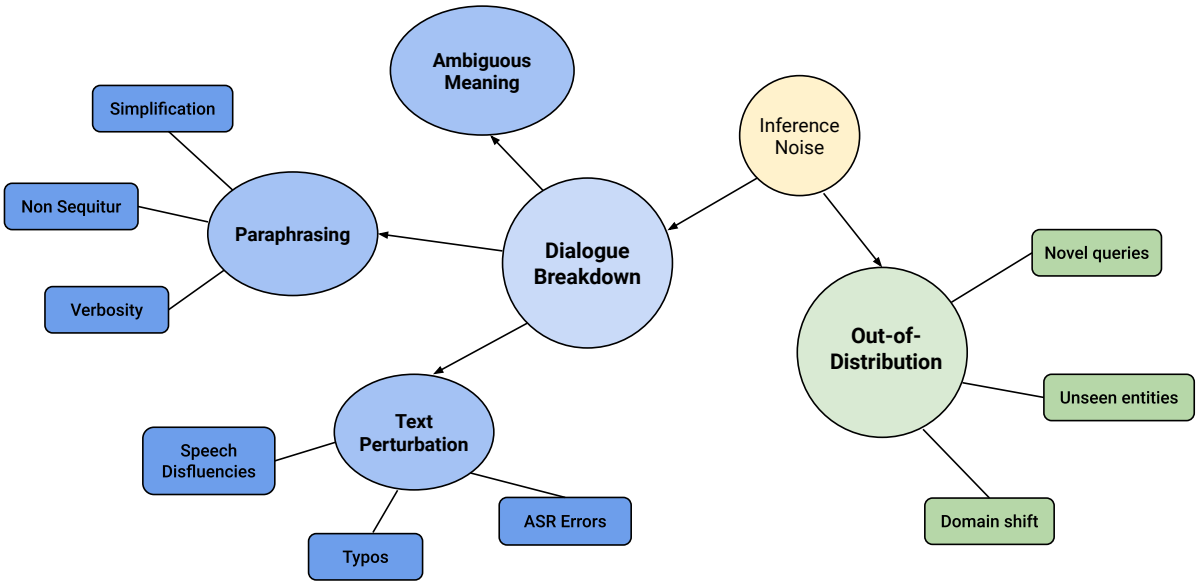


Figure 3: Diagram of the sources of noise that affect dialogue systems during inference.

Paraphrasing The text is rephrased to become: (a) *Simplification*: request may be simplified or shortened that makes it unclear what the user wants. (b) *Non Sequitur*: response is plausibly in-distribution, but does not reasonably answer the question. (c) *Verbosity*: request is so verbose that the underlying request is lost. (See Appendix D)

Text Perturbations Notable instances include (a) *ASR Errors* that fail to "wreck a nice beach" (recognize speech) (b) *Typos* and other syntax errors on the user input. This is distinct from formatting mistakes by annotators, which are errors on the target output. (c) *Speech Disfluencies* such as repeats, corrections, or adding 'umm' to start a utterance (Liu et al., 2021; Peng et al., 2021).

4 Noise Patterns

Beyond categorization, manually reviewing 10K+ utterances also provides unique insights.

How often does noise appear? The percentage of dialogues with at least one instance of noise comes out to an average of 11.2%, a median of 10.6%, with a standard deviation of 3.7%. However, given the approximate nature of sampling, the extra digits may not be significant. Instead, we assert the rate of noise in curated dialogue datasets is usually over 5%, rarely above 20% and typically around 10%. Since these rates are relatively low, denoising techniques aiming to combat extremely high levels of noise may be impractical.

What noise types are most common? While most existing denoising algorithms are designed to resolve class confusion (Sukhbaatar et al., 2015; Patrini et al., 2017; Goldberger and Ben-Reuven, 2017), our analysis reveals that instance-level noise is actually much more common, showing up in nearly 10% of cases compared to just 5% for class-level errors. Class-level noise assumes a latent noise transition matrix stochastically switches labels from one class to another. However, the prevalence of instance-level noise implies that the more likely explanation is that some examples are simply more confusing then others due to the genuinely ambiguous nature of dialogue (Pavlick and Kwiatkowski, 2019; Nie et al., 2020)

From an algorithmic perspective, the upshot is that developing denoising methods to target individual examples rather than class errors are likely to be most effective. Furthermore, we discovered that noise is clustered rather than evenly distributed, so filtering out or relabeling these particularly noisy instances should have an out-sized impact.

Why is X source of noise missing? The expected influence of some sources of noise are greatly exaggerated. Building out the taxonomy not only shows the most likely sources of noise, but equally notable is uncovering the *least* likely noise types. Concretely, the threat of adversarial actors is largely overblown (Dinan et al., 2019a), as spam-like activity appears less than 2% of the time. Offensive speech is the subject of numerous dialogue studies (Khatri et al., 2018; Xu et al., 2021; Sun

et al., 2022), but is practically non-existent in reality (<0.5% of cases). While hate speech may be a problem when training on raw web text (Schmidt and Wiegand, 2017), our empirical review reveals that toxic language is exceedingly rare in curated datasets. Instead, unnatural utterances generated by crowdworkers role-playing as real users occurs much more often. (Full breakdown in Appendix E)

Other types of noise occur so infrequently that they are missing from the taxonomy completely! Noteworthy options include inconsistent *names* or *titles* within the ontology (See Appendix C), as well as *improper reference* texts for dialogue generation tasks. While these noise types are possible, they did not occur in practice. We intentionally exclude all such candidates from the taxonomy since the aim is not to be comprehensive, but rather to highlight where researchers should spend their efforts.

Where does noise come from? Our survey found that each data collection method had a propensity to produce certain kinds of noise. This suggests noise arises as a result of how examples are annotated, rather than other factors such as conversation length (number of utterances) or dialogue style (open-domain vs. task-oriented). For example, positive discourse attributes are most common with Post-conversation Annotation and Live Chat, which involve two human speakers engaging in real dialogue. Wizard-of-Oz datasets are less time-consuming to produce, but contain more label noise. In contrast, dialogues from Machine-to-Machine or Dialog Self-play (ie. starting with the labels to generate the dialogue) contain fewer label errors, but also sound less natural. Separately, annotator and ontology issues can be mitigated with well-written agent guidelines and proactive crowdworker screening. Thus, practitioners should consider these noise trade-offs when collecting dialogue data.

5 Experiments and Results

This section explores to what degree various models and dialogue tasks are impacted by each of the seven different categories of noise outlined in Section 3. To study this, a model is trained on a clean version of the dataset and on a corrupted version with either natural or injected noise. The level of corruption for all trials is held constant at 10% to allow for comparison across noise types. Datasets for each noise type are selected to maximize the overall variety, while always keeping one instance of MultiWOZ 2.3 to aid comparison. Intuitively, sources of noise that induce a larger gap in models trained on cleaned versus corrupted data are more significant, and consequently deserve more attention as targets to denoise.

5.1 Task Setup

All trials are conducted with GPT2-medium as a base model (Brown et al., 2020). The chosen tasks are: (1) *Conversation Level Classification* (CLC) – Choose from a finite list of labels for each conversation. (2) *Turn Level Classification* (TLC) – Make a prediction for each turn that contains a label. (3) *Dialogue State Tracking* (DST) – Predict the overall dialogue state, which may contain multiple slot-values or no new slot-values at all. Individual values come from an enumerable or open-ended ontology. (4) *Response Generation* (RG) – Produce the agent response given the dialogue context so far. (5) *Information Retrieval* (IR) – Find and rank the appropriate information from an external data source, such as a knowledge base (KB) or separate document. Metrics were chosen to adhere to the evaluation procedure introduced with the original dataset or from related follow-up work.

5.2 Noise Injection

For each noise category, we start by independently sampling 10% of the data, adding the corresponding noise and training a model to convergence. For example, consider instance-level label errors applied to MultiWOZ. This dataset contains 113,556 total utterances so 11,356 of them are selected for corruption. Next, one of the three sub-categories of instance noise are chosen uniformly at random. Over-labeling occurs when a label that has recently appeared in previous turns is no longer valid. To match this behavior, we keep a running tally of recent slot-labels and occasionally insert an extra one from this pool into the current training example. Partial-labeling is achieved by replacing a slot-label with a randomly selected one from the recent pool, and under-labeling is achieved by simply dropping a slot-label from the example. Finally, a model is trained with the noisy data applying the same hyper-parameters as the ones used for training the standard, original model. This process is repeated for each other noise type, with details for each source of noise found in Appendix F.

5.3 Main Results

Denoising methods targeting class-level noise may have minimal impact since it turns out such label er-

Noise Source	**MultiWoz**	**Dataset 2**	**Dataset 3**	**Dataset 4**	**Average**
Label Noise by Class	84.1 (0.13%)	75.8 $(0.37\%)^{DD}$	58.1 $(1.15\%)^{ED}$	78.8 $(1.92\%)^{MIDAS}$	0.89%
Label by Instance	59.1 (4.88%)	82.4 $(3.03\%)^{SGD}$	72.9 $(0.96\%)^{TT}$	98.9 $(0.12\%)^{GSIM}$	2.25%
Label by Annotator	58.2 (18.1%)	73.6 $(3.36\%)^{DD}$	90.2 $(1.43\%)^{TT}$	44.7 $(15.9\%)^{WOW}$	9.68%
Discourse Attributes	62.9 (9.31%)	36.8 $(8.42\%)^{WOW}$	25.6 $(5.08\%)^{ABCD}$	39.2 $(10.7\%)^{KVRET}$	8.38%
Ontology Inconsistency	61.9 (3.41%)	98.7 $(0.40\%)^{GSIM}$	58.7 $(26.8\%)^{ED}$	84.9 $(0.94\%)^{SGD}$	7.89%
Out-of-Distribution	48.1 (28.9%)	83.2 $(2.04\%)^{SGD}$	83.3 $(10.5\%)^{ABCD}$	74.6 $(23.6\%)^{SGD}$	**16.3%**
Dialogue Breakdown	61.8 (11.3%)	49.8 $(4.02\%)^{WOW}$	4.07 $(4.44\%)^{ED}$	72.1 $(2.08\%)^{TT}$	5.45%

Table 3: Performance across various datasets when injected with 10% noise. Scores in parentheses are the percent degradation when compared to the clean version of the data. Datasets 2-4 contain a superscript representing the dataset name as described in Table 1. Please see Appendix 5 for the exact task and dataset mapping for each item.

	RoBERTa	**GPT2**	**BART**
Original	45.7	61.9	62.3
Noised	39.4	59.1	61.4

(a) Performance on MultiWOZ for each model

	CLC	**TLC**	**DST**	**RG**	**IR**
Median	3.4%	0.9%	4.0%	10.3%	8.4%
Average	6.5%	4.6%	8.4%	10.5%	8.1%

(b) Change in performance for each task due to noise.

Table 4: Breakdown by dialogue task and model type

rors are not all that damaging with just 0.89% drop in performance. On the other hand, annotator noise is quite powerful causing a 9.7% disturbance and should be mitigated whenever possible. Luckily, our manual review showed that spamming behavior occurs infrequently in reality simply by following some best practices[1]. Negative discourse attributes can also cause major harm leading to a 8.4% gap.

Moving onto inference noise, ontology issues are not only quite common, but also have meaningful impact on performance, causing a 7.9% drop. Dataset creators can ameliorate this by deciding on an ontology upfront, rather than creating one after the fact. Dialogue breakdowns also cause noticeable degradation, but the impact of OOD is most prominent among all noise types. Neural networks are powerful enough to learn from any training signal, even complete random noise (Zhang et al., 2017). However, OOD cases are by definition areas the network has not seen, leading to poor performance. Data augmentation and other robustness methods may serve as a strong tool to cover the unknown space by maximizing the diversity of the examples (Ng et al., 2020; Chen and Yin, 2022).

[1]For example, gold checks insert questions with known labels; timers ensure adequate time is spent on each task.

5.3.1 Task Breakdown

In order to study tasks across noise types, we look at the percentage change between models, rather than absolute difference. Furthermore, to minimize the influence of outliers, we emphasize the median of change, rather than the average. The results in Table 4b show that RG and IR observe the largest drops when noise is added. Somewhat surprisingly, CLC has larger performance shift than TLC despite being an easier task. We hypothesize this is because CLC examples only occur once for each conversation, whereas TLC examples occur at every turn, leading to an order of magnitude less data. Training with the existence of noisy data depends on both the rate of noisy data as well as on a minimum number of clean examples.

5.3.2 Model Robustness

Prior work has suggested that models behave differently when faced with distinct types of noise (Belinkov and Bisk, 2018). In addition to GPT2-medium (345M parameters), we also consider a masked language model in RoBERTa-Large (355M parameters) (Liu et al., 2019) and a sequence-to-sequence model with BART-large (406M parameters) (Lewis et al., 2020). These are selected due to having a comparable number of training parameters. Based on the results in Table 4a, RoBERTa is the weakest performer of the group. We hypothesize this is because many dialogue tasks are generation based, whereas BERT-based models typically perform well on classification. Conversely, BART deals quite well with noise, suggesting encoder-decoder models as reasonable starting points for future dialogue projects.

5.4 Amount of Noise

We simulate increasing levels of noise by adding instance-level label errors and incoherent discourse

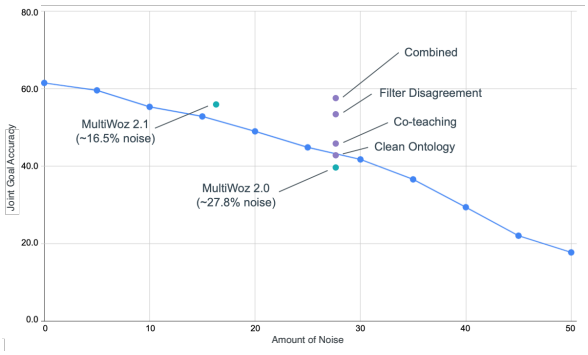


Figure 4: Impact of injecting different amounts of label and discourse noise to MultiWOZ dataset. MultiWOZ 2.3 is defined to be 0% noise. MultiWOZ 2.0 and 2.1 have estimated noise levels based on the ratio of labels that are changed compared to MultiWOZ 2.3 data.

attributes to the MultiWOZ 2.3 dataset (Han et al., 2021), which we define to be noise-free. We additionally plot the performance of models trained on MultiWOZ 2.0 (Budzianowski et al., 2018) and 2.1 (Eric et al., 2020), where all models are evaluated on the MultiWOZ 2.4 test set (Ye et al., 2021). Looking at Fig 4, we first note that scores on naturally noisy data from MWOZ 2.0 and MWOZ 2.1 fall close to the plotted trajectory, lending credence to the overall trend. Furthermore, we notice that as we vary the amount of noise, model performance decreases logarithmically, but surprisingly does not have a tipping point at which it fails to converge.

6 Dialogue Denoising

Informed by our understanding of the sources of dialogue noise, we now design a preliminary denoising algorithm for learning in the presence of noisy labels. We select MultiWOZ to serve as our testbed not only because it is one of the most popular dialogue datasets, but also because it is representative of how noise affects most datasets in general (see Figure 6). While our method produces promising results, our aim is not to declare the noise issue solved, but rather to establish a baseline others can further improve. (More details in Appendix G.)

6.1 Algorithm

Based on analysis in Section 3, MultiWOZ 2.0 is most plagued by three types of errors: ontology inconsistencies, instance label errors and out-of-distribution issues. We now devise three solutions to resolve each source of noise accordingly.

(1) To clean up the *ontology*, we drop values that do not conform to the correct format, and remove the associated examples from training. For example, if `time_of_day` slot expects the HH:MM format, then we remove all values referencing day formats (e.g. Friday). (2) To deal with label errors, we filter out individual *instances* where the predicted label from a pre-trained GPT2-medium model disagrees with the annotator label (Cuendet et al., 2007; Jiang et al., 2018; Chen et al., 2019). We calibrate the model with temperature scaling to prevent it from being over-confident in its predictions (Guo et al., 2017). (3) To counteract issues caused by *OOD*, we augment our training data by pseudo-labeling the examples stripped out in the first two steps. When the model used for filtering is also used for pseudo-labeling, biases may propagate across each iteration. As a result, inspired by co-teaching (Han et al., 2018), we instead use a different BART-base model for pseudo-labeling to force divergence of model parameters and avoid errors from accumulating.

6.2 Denoising Results

We once again evaluate with MultiWOZ 2.4 since this is the cleanest version of test data. As seen in Figure 4, we are able to outperform MultiWOZ 2.0 (39.8) by 16.9% absolute accuracy and 42.5% relative accuracy. Ontology Clean (43.2), Filter Disagree (53.7) and Co-teaching (46.7) all show marked improvement over the original baseline, but Combined (58.6) does the best overall, reaching a score that even surpasses MultiWOZ 2.1 (56.5). These initial efforts show our ability to successfully identify and counteract sources of noise within MultiWOZ, which we encourage others to build upon.

7 Related Works

Our work is related to efforts to categorize noise within speech and dialog. Clark (1996) proposed a theory of miscommunication consisting of channel, signal, intention and conversation where each of the four levels serves as a potential vector for noise. Others have also studied noise in spoken dialogue systems, where they found that the main culprit stems from errors in speech transcription (Paek, 2003; Bohus, 2007). Rather than a high-level framework of general communication, our hierarchical taxonomy focuses on understanding the multiple layers of noise found in written text.

More recent works on dialogue noise discuss robustness to noisy user inputs, whereas we expand this view to also analyze noisy training in-

puts. Peng et al. (2021) introduce RADDLE as a platform which covers OOD due to paraphrasing, verbosity, simplification, and unseen entities, as well as general typos and speech errors. Liu et al. (2021) create a robustness benchmark which considers paraphrasing through word perturbations as well as speech disfluencies. Lastly, Krone et al. (2021) considers noise from abbreviations, casing, misspellings, paraphrasing, and synonyms.

7.1 Survey of Denoising Methods

Most prior works exploring learning with noisy labels were originally developed for the computer vision domain (Smyth et al., 1994; Mnih and Hinton, 2012; Sukhbaatar et al., 2015). Some methods model the noise within a dataset in order to remove it, often through the use of a noise transition matrix (Dawid and Skene, 1979; Goldberger and Ben-Reuven, 2017). Others have designed noise-insensitive training schemes by modifying the loss function (van Rooyen et al., 2015; Ghosh et al., 2017; Patrini et al., 2017), while a final set of options manipulate noisy examples by either reweighting or relabeling them. (Reed et al., 2015; Jiang et al., 2018; Li et al., 2020). While denoising work certainly exists for NLP (Snow et al., 2008; Raykar et al., 2009; Wang et al., 2019), none of them specifically touch upon the dialogue scenario.

7.2 Denoising by Source of Noise

To support the effort of designing improved algorithms for combating dialogue-specific noise, we highlight potential methods that can be adapted to deal with the noise categories identified by our taxonomy in Section 3. To start, a common technique for dealing with *class-level* errors is to learn a noise adaptation layer to recognize label noise (Goldberger and Ben-Reuven, 2017). For *instance-level* noise, besides filtering by disagreement, core-set selection (Mirzasoleiman et al., 2020) or the Shapley algorithm (Liang et al., 2021) can be used to identify important datapoints and thereby remove the noisy ones. Modeling the likelihood of *annotator-level* error in order to reverse its impact is also worth considering (Welinder et al., 2010; Hovy et al., 2013; Guan et al., 2018). Next, a model trained on NLI data can be used to screen out inconsistent *discourse* examples (Welleck et al., 2019). A model trained on Prosocial Dialogue data can learn to reduce toxicity (Kim et al., 2022). In terms of *discourse fluency*, one can train a student model to reweight its logits during inference based on a large language model (Brown et al., 2020) to improve the fluency of the student. Another method is to create an *ontology* upfront which defines the allowed entities before data collection and enforcing this by having checks upon label submission. *Out-of-Domain* issues can be handled with the use of more examples to increase the coverage and diversity of the solution space to limit OOD errors. This can be tackled by performing data augmentation on the in-domain (Feng et al., 2021) or out-of-domain examples (Chen and Yu, 2021). Lastly, *dialogue breakdown* can be mitigated by screening for annotators through minimum acceptance rates, language filters, and pre-qualifications quizzes (ie. quals).

8 Conclusion

This paper categorizes the different sources of noise found in dialogue data and studies how models react to them. We find that dialogue noise is divided into issues that occur during training and during inference. We also find that conversations pose unique challenges not found in other NLP corpora, such as discourse naturalness and dialogue breakdowns. Our study further reveals that the most common sources of noise are actually based on the ambiguity of individual instances, rather than systematic noise across classes or adversarial annotators actively harming data collection efforts.

Despite being surprisingly resilient, dialogue models nonetheless experience a notable drop in performance when exposed to high levels of noise. To combat this, we design a proof-of-concept denoising algorithm to serve as a strong foundation for others to compare against. We apply this algorithm successfully to the MultiWOZ 2.0 dataset, raising the accuracy by 42.5% over the original baseline. We hope our survey informs the collection of cleaner dialogue datasets and the development of advanced denoising algorithms targeting the true sources of dialogue noise.

9 Limitations

The main limitation of the taxonomy is only considering natural language text within dialogue. It could be useful to conduct a detailed breakdown of speech noise or multi-modal noise that occurs in conversations grounded by images. Our survey also does not include all theoretically possible sources of noise and instead is limited to actual sources of noise saw occuring in the data. We argue this type of taxonomy serves a more practical purpose.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Dan Bohus. 2007. *Error awareness and recovery in conversational spoken language interfaces*. Ph.D. thesis, Carnegie Mellon University.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4515–4524. Association for Computational Linguistics.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3002–3017. Association for Computational Linguistics.

Derek Chen and Claire Yin. 2022. Data augmentation for intent classification. *CoRR*, abs/2206.05790.

Derek Chen and Zhou Yu. 2021. GOLD: improving out-of-scope detection in dialogues using data augmentation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 429–442. Association for Computational Linguistics.

Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. 2019. Understanding and utilizing deep neural networks trained with noisy labels. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1062–1070. PMLR.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Sébastien Cuendet, Dilek Hakkani-Tür, and Elizabeth Shriberg. 2007. Automatic labeling inconsistencies detection and correction for sentence unit segmentation in conversational speech. In *Machine Learning for Multimodal Interaction , 4th International Workshop, MLMI 2007, Brno, Czech Republic, June 28-30, 2007, Revised Selected Papers*, volume 4892 of *Lecture Notes in Computer Science*, pages 144–155. Springer.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019a. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Kumar Goyal, Peter Ku, and Dilek Hakkani-Tür. 2020. Multiwoz 2.1: A consolidated multi-domain

dialogue dataset with state corrections and state tracking baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 422–428. European Language Resources Association.

Mihail Eric, Lakshmi Krishnan, François Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 37–49. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 968–988. Association for Computational Linguistics.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1919–1925. AAAI Press.

Jacob Goldberger and Ehud Ben-Reuven. 2017. Training deep neural-networks using a noise adaptation layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Melody Y. Guan, Varun Gulshan, Andrew M. Dai, and Geoffrey E. Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3109–3118. AAAI Press.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8536–8546.

Ting Han, Ximing Liu, Ryuichi Takanobu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang. 2021. Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. In *Natural Language Processing and Chinese Computing - 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13-17, 2021, Proceedings, Part II*, volume 13029 of *Lecture Notes in Computer Science*, pages 206–218. Springer.

Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 87–95. The Association for Computer Linguistics.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning whom to trust with MACE. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 1120–1130. The Association for Computational Linguistics.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2309–2318. PMLR.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5574–5584.

Chandra Khatri, Behnam Hedayatnia, Rahul Goel, Anushree Venkatesh, Raefer Gabriel, and Arindam Mandal. 2018. Detecting offensive content in open-domain conversations using two stage semi-supervision. *NeurIPS Workshop on ConvAI*, abs/1811.12900.

Ashish Khetan, Zachary C. Lipton, and Animashree Anandkumar. 2018. Learning from noisy singly-labeled data. In *6th International Conference on*

Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. In *EMNLP*.

Jason Krone, Sailik Sengupta, and Saab Mansoor. 2021. On the robustness of goal oriented dialogue systems to real-world noise. In *Robust ML Workshop at ICLR 2021*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Junnan Li, Richard Socher, and Steven C. H. Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Weixin Liang, Kaihui Liang, and Zhou Yu. 2021. HERALD: an annotation efficient method to detect user disengagement in social conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3652–3665. Association for Computational Linguistics.

Jiexi Liu, Ryuichi Takanobu, Jiaxin Wen, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, and Minlie Huang. 2021. Robustness testing of language understanding in task-oriented dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2467–2480. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Baharan Mirzasoleiman, Kaidi Cao, and Jure Leskovec. 2020. Coresets for robust training of deep neural networks against noisy labels. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Volodymyr Mnih and Geoffrey E. Hinton. 2012. Learning to label aerial images from noisy data. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 1196–1204.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283, Online. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9131–9143. Association for Computational Linguistics.

Tim Paek. 2003. Toward a taxonomy of communication errors. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, page 53–58.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2233–2241. IEEE Computer Society.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Trans. Assoc. Comput. Linguistics*, 7:677–694.

Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2021. RADDLE: an evaluation benchmark and analysis platform for robust task-oriented dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4418–4429. Association for Computational Linguistics.

Kun Qian and Zhou Yu. 2019. Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2639–2649. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8689–8696. AAAI Press.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna K. Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. 2009. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 889–896. ACM.

Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2015. Training deep neural networks on noisy labels with bootstrapping. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gökhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 3 (Industry Papers)*, pages 41–51. Association for Computational Linguistics.

Padhraic Smyth, Usama M. Fayyad, Michael C. Burl, Pietro Perona, and Pierre Baldi. 1994. Inferring ground truth from subjective labelling of venus images. In *Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994]*, pages 1085–1092. MIT Press.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 1297–1304.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 254–263. ACL.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2015. Training convolutional neural networks with noisy labels. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 843–852. IEEE Computer Society.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, Dublin, Ireland. Association for Computational Linguistics.

Megan Ung, Jing Xu, and Y-Lan Boureau. 2022. SaFeR-Dialogues: Taking feedback gracefully after conversational safety failures. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6462–6481, Dublin, Ireland. Association for Computational Linguistics.

Brendan van Rooyen, Aditya Krishna Menon, and Robert C. Williamson. 2015. Learning with symmetric label noise: The importance of being unhinged. *CoRR*, abs/1505.07634.

Hao Wang, Bing Liu, Chaozhuo Li, Yan Yang, and Tianrui Li. 2019. Learning with noisy labels for sentence-level sentiment classification. In *Proceedings of the 2019 Conference on Empirical Methods*

in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 6285–6291. Association for Computational Linguistics.

Peter Welinder, Steve Branson, Serge J. Belongie, and Pietro Perona. 2010. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 2424–2432. Curran Associates, Inc.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3731–3741. Association for Computational Linguistics.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online. Association for Computational Linguistics.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021. Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *CoRR*, abs/2104.00773.

Dian Yu and Zhou Yu. 2021. MIDAS: A dialog act annotation scheme for open domain humanmachine spoken conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1103–1120. Association for Computational Linguistics.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7620–7632. Association for Computational Linguistics.

A Dataset Descriptions

In no particular order, the datasets we study are:

1. DailyDialog (DD) - a collection of conversations from the web about everyday events, curated for teaching English. (Li et al., 2017)

2. Wizard of Wikipedia (WoW) - a wizard reads an article on Wikipedia and then talks about it with their partner (Dinan et al., 2019b)

3. Google Simulated Dialogue (GSIM) - a large scale Machine-to-Machine (M2M) dataset build through paraphrasing, covers movie and restaurant domains. (Shah et al., 2018)

4. Action Based Conversations Dataset (ABCD) - customer service conversations that mimic agents in real-life where actions are taken to resolve customer issues based on instructions in agent guidelines (Chen et al., 2021)

5. MultiWoz 2.0 (MWoz) - a multi-domain dialogue dataset. Note that we use the original version for initial analysis because it contains true noise, before any additional cleaning. (Budzianowski et al., 2018)

6. TicketTalk (TT) - As part of the third installment of TaskMaster, this dataset also uses the M2M style, but focuses on the single vertical of movie ticket booking. (Byrne et al., 2019)

7. Empathetic Dialogues (ED) - a set of dialogues that aim to teach models to be empathetic by being more attuned to what a user is feeling. (Rashkin et al., 2019)

8. Machine Interaction Dialog Act Schema (MIDAS) - created for the Amazon Alexa challenge with Gunrock. Transcribed conversations are with actual Alexa users, and not crowdworkers. (Yu and Yu, 2021)

9. Schema Guided Dialogue (SGD) - the most comprehensive DST dataset to date, with a heavy focus on slot-filling for API calls. Contains natural OOD splits. (Rastogi et al., 2020)

10. Key-Value Retrieval for In-Car Assistant (KVRET) - Task oriented dataset with a knowledge base for querying items. Covers navigation, weather and scheduling domains. (Eric et al., 2017)

B Label Error Details

Class Level Examples are labeled incorrectly due to confusion with another class.

- *Uniform Label Swapping*: symmetric noise where all classes have equal likelihood to be confused with any other class. The assumption is that noise is injected through a randomly initialized noise transition matrix.
- *Structured Label Swapping*: asymmetric noise where certain classes are more likely to confused with other related classes. For example, a cheetah is more likely to be confused with leopard than a refrigerator when performing image recognition. Alternatively, dogs and wolves are likely to be confused for each other much more often than with horses since those animals are similar to each other.

Instance Level Noise comes from the example itself due to the complexity of interpreting natural language. This is the realization that even when annotators act in good-faith, mistakes are still made since the instances themselves are difficult to label. Errors must be determined on a case-by-case basis.

- *Over Labeling*: annotator added a label, but should be removed since it is unnecessary. Example: carrying over a slot-value from the previous turn to the current dialogue state when it is not warranted.
- *Under Labeling*: annotator missed the label, when most people would include it. Example: failing to notice a newly mentioned criteria in the dialogue state. This also includes cases where a better label could have been used, but the option is missing from the ontology and consequently prevents the example from being properly labeled.
- *Partial Labeling*: part of the label is correct, but other parts are not. For multi-intent utterances, the annotator may have captured one intent, but not the other. For slot-filling tasks, the annotator may have selected the appropriate value, but assigned it to the wrong slot.

Annotation Level Noise arises due to the labeler or data collection process. (Snow et al., 2004)

- *Distant Supervision*: the noise results from the fact that the label is not from a human, but rather weakly labeled from distant supervision (Sun et al., 2017). For example, using a

gazetteer for labeling named entities in NER. As another example, you use the SQL results to train a semantic parser, rather than an annotated SQL query.

- *Adversarial Actors*: meant to mimic spammers, this is characterized by repeating patterns or irrational behavior. For example, the annotator selects "greeting" dialogue act as the label for every single utterance regardless of the underlying text. (Raykar et al., 2009; Hovy et al., 2013; Khetan et al., 2018) Other examples include bad actors in social media who provoke chatbots into producing unsafe content or labelers who mark every review as possessing positive sentiment without actually reading the passage.
- *Formatting Mistakes*: Caused by non-experts making human mistakes, which are independent of the dialogue context. For example, typos or off-by-one errors, such as when the labeler failed to highlight the entire phrase during span selection. (See Table 2)

C Ontology Inconsistency Details

Another source of noise comes from inconsistent formatting when constructing the ontology. More specifically, the creators of the dataset did not set a canonical format for each type of slot being tracked. While we can imagine many other slot-types causing issues, the types of errors which actually occurred in practice include:

- **Dates**: tomorrow, Jan 3rd, 1/3/2022, Monday, January 3, mon
- **Times**: 14:15, 2:15 PM, quarter past 2, 215pm
- **Locations**: NYC, New York, ny, the big apple
- **Numbers**: three, 'wife daughter & I', 3, 'Me and my two buddies'.

Other ontology issues which we thought might occur more often, turn out to happen very rarely. For example, naming inconsistency such as [Fred Miyato, Mr. Miyato, fred miyato, my father] did not really occur. Titles of people or places [Macdonalds, MickeyD's, McDonald's, mcdonalds] also were not present. To minimize the amount of noise from ontology inconsistency, a recommendation is to declare the allowable slot-values upfront before data collection begins.

D Paraphrasing Examples

Paraphrasing can take on three general forms:

1. **Simplification** – the request may be simplified so much that it becomes unclear what the user wants. For a restaurant scenario:

   ```
   Agent: What part of town
   would you like to eat?
   User: W
   ```
   (as a shorthand for West side)

2. **Non Sequitur** – response is plausibly in-distribution, but does not reasonably answer the question.

   ```
   Agent: What part of town
   would you like to eat?
   User: I would like Italian
   food.
   ```

   Note that the user's response is still in distribution since it could have been a reasonable answer to "What cuisine do you prefer?". However, in this instance, this type of response is very noisy because it fails to answer the agent's question.

3. **Verbosity** – the request contains extra words or entities, which makes it confusing as to exactly what the answer may be.

   ```
   Agent: What part of town
   would you like to eat?
   User: I prefer food in the
   East, but I live in the South
   right now.
   ```

   In this case, the user's response is not necessarily long, but it is verbose enough to make it unclear whether the user wants food in the east side of town or the south side of town.

True paraphrasing noise should alter the text without altering the user's underlying intent. If the text has changed so much that the user's intent has also shifted, then it should be considered adversarial behavior beyond the scope of typical dialogue noise.

```
Agent: What part of town would
you like to eat?
User: The Northern Lights are
beautiful this time of year.
```

The example above displays positive sentiment, but the user has completely ignored the agent's request. This case borders on being incoherent and fails to move the dialogue forward.

E Results Breakdown

Aggregated amounts of noise by each sub-category:

	Average	Median	Std. Dev.
Class-level	4.9%	3.8%	0.7%
Instance-level	9.7%	6.9%	5.4%
Annotator-level	1.8%	0.7%	2.1%
Dates	3.6%	0.5%	6.3%
Times	1.1%	<0.1%	2.0%
Locations	1.3%	0.3%	2.1%
Numbers	2.3%	0.2%	4.6%
Incoherent	3.4%	3.8%	1.9%
Disfluent	2.6%	2.4%	2.0%
Inconsistent	1.7%	1.3%	1.5%
Nonsensical	2.0%	2.6%	1.1%
Offensive	0.2%	<0.1%	0.9%
Unnatural	4.8%	5.8%	1.6%
Overall	11.2%	10.6%	3.7%

Table 5: Breakdown across noise sub-categories

F Noise Injection Methods

Class-level Label Errors We create a noise transition matrix to mimic structured confusion. Specifically, given a certain class label, we want to determine what is likely to be confused with it so we can substitute the current label for that other class. To fill the noise transition matrix, we embed all class labels into bag-of-word GloVe embeddings and measure their similarity to other classes by cosine distance. Then, for 10% of examples, we sample an incorrect label given the original class according to the likelihood in the transition matrix.

Instance-level Label Errors To match the behavior of over-labeling, we keep a running tally of recent labels and occasionally insert an extra one from this pool into the example. Partial-labeling is achieved by replacing a label from the recent pool, and under-labeling is achieved by simply dropping a random label from the example.

Annotator-level Label Errors We mimic spammers who apply preset answers to every occasion without considering the actual dialogue. For the classification tasks, we assume a spammer randomly picks from one of the three most common labels for that task as the noisy target label. For response generation tasks, we assume a spammer randomly responds with one of three generic phrases.

Undesirable Discourse Attributes We replace a subset of the utterances with noisy versions 10% of the time. Incoherent utterances are randomly selected sentences from other dialogues within the dataset. Disfluent utterances are generated by shuffling the tokens within the current utterance. Unnatural utterances are generated by selecting from a list of awkward sentences referencing the task.

Ontology Inconsistency To clean the data, we manually remove entries that do not comply to the proper format. We also merge similar categories to create more compact ontologies. Training examples that are covered by the remaining entries are considered the clean version, while the full, original dataset is considered the noisy version.

Out-of-Distribution Multi-domain data is divided such that training data contains a subset of domains while the test set includes examples from all domains. Choosing the domains to exclude was straightforward for ABCD and SGD since they are given by the task design. Rather than choosing an arbitrary domain to leave out for MWOZ, we instead run the experiment once for each domain, and report the average of the five results.

Dialogue Breakdown We reproduce this behavior by pre-training a paraphrase model and applying it to perturb 10% of utterances. Paraphrase model is trained on QQP, MRPC and PAWS corpora.

G Denoising Procedure for MultiWOZ

We identify the highest likelihood sources of noise for any given dataset and dealing with each one accordingly. MultiWOZ in particular has (1) ontology issues, (2) instance level label errors and (c) out-of-distribution examples caused by low coverage in the training set. In turn, we proceed to deal with each of these issues as follows:

(1) To clean up the *ontology*, we drop values that do not conform to the correct format for their given slots, and remove the associated examples from training. For example, if the slot is a time of day expecting the HH:MM format, then we remove all values referencing 'Friday' or 'afternoon' which are incorrectly formatted.

(2) To deal with possible label errors, we filter out individual *instances* where the predicted label from a pre-trained GPT2-medium model disagrees with the annotator label (Cuendet et al., 2007; Jiang et al., 2018; Chen et al., 2019).

(3) Lastly, we augment our training data to counteract issues caused by *OOD* cases. In order to augment, we pseudo-labeling the datapoints that have been stripped out in the first two steps. However,

Noise Source	**MultiWoz 2.3**	**Dataset 2**	**Dataset 3**	**Dataset 4**
Label Noise by Class	MWOZ (TLC on intents)	DD (CLC on topics)	ED (CLC on emotions)	MIDAS (TLC on dialog acts)
Label by Instance	MWOZ (DST on slot-values)	SGD (DST w/ slot-values)	TT (DST w/ slot-values)	GSIM (TLC on user acts)
Label by Annotator	MWOZ (RG of agent utt)	DD (CLC on topics)	TT (TLC on APIs)	WOW (RG on wizard utt)
Discourse Attributes	MWOZ (RG of agent utt)	WOW (IR on wizard utt)	ABCD (IR on agent utt)	KVRET (IR on KB entries)
Ontology Inconsistency	MWOZ (DST on slot-values)	GSIM (TLC on user acts)	ED (CLC on emotions)	SGD (DST on slot-values)
Out-of-Distribution	MWOZ (DST on slot-values)	SGD (DST on slot-values)	ABCD (CLC on subflows)	SGD (TLC on intents)
Dialogue Breakdown	MWOZ (RG of agent utt)	WOW (RG on wizard utt)	ED (RG on agent utt)	TT (DST on slot-values)

Figure 5: Mapping of model performance to datasets and dialogue tasks. Parentheses also includes the target of the task. For example, 'CLC on topics' means that the task is to classify the associated topic label at a conversation level, while 'TLC on intents' means the task is to classify the intent of each user turn.

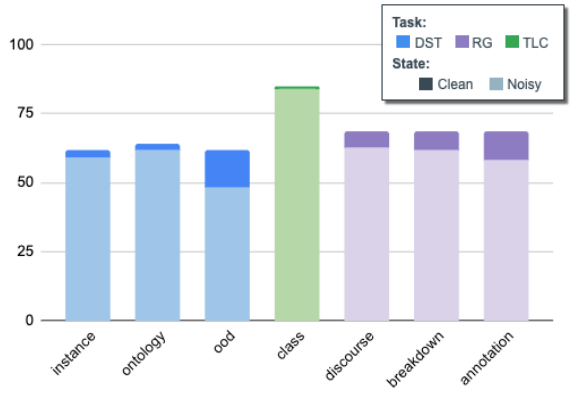


Figure 6: Impact of the different noise types on the MultiWoz2.3 dataset. DST is dialogue state tracking, RG is response generation and TLC is turn level classification.

the pretrained model's predictions are unlikely to be all correct, so rather than keep all the new labels, we only keep the examples where the probability of the max value crosses the 0.5 threshold. Then, since neural networks are often over-confident, we perform calibration with temperature scaling using a λ parameter (Guo et al., 2017). However, pseudo-labeling with the same model that is used to perform filtering causes errors to propagate which hinders performance gains. As a result, inspired by co-teaching (Han et al., 2018), we instead use a different model to force divergence of model parameters and avoid the existing biases. In more detail, we rely on a BART-base model rather than the original GPT-2 medium, which works even though BART-base has much fewer parameters.

H Noise as Uncertainty

An interesting way to view the impact of noise is through the lens of Bayesian uncertainty. In particular, aleatoric and epistemic uncertainty can be seen caused by different types of noise. Kendall and Gal (2017) describe aleatoric uncertainty as uncertainty which "captures noise inherent in the observations." In contrast "epistemic uncertainty accounts for uncertainty in the model parameters which can be explained away given enough data."

Roughly speaking, labeling errors cause epistemic uncertainty since these errors produce uncertainty in the model parameters. If given enough clean data to train a model, the issues caused by the noisy labels should largely be erased. In other words, epistemic uncertainty describes what the model does not know because training data was not appropriate, so by resolving the labeling errors, the training data is now appropriate and the dialogue system can be trained successfully.

On the other hand, ontology inconsistencies cause aleatoric uncertainty since they can lead to situations where it is impossible to fix the problem by altering the training data alone. Suppose we want the dialogue model to predict the desired time for a restaurant reservation (such as 11 AM, 6PM or 8PM), but options such as 'Sunday' or 'afternoon' keep appearing, which are never correct. This would make it harder for a classifier to choose the correct time. In the degenerate case, suppose the ontology only consisted of days of the week such as 'Monday', 'Wednesday' or 'Friday', such that the classifier would only have the ability to choose from seven incorrect options. In this case, adding any amount of extra data (even those labeled in the correct format) would do nothing to resolve the issue since the problem itself has been modeled incorrectly.

Accordingly, a model developer should focus on eliminating certain types of noise based on the type of uncertainty they are seeing in their dialogue system. If the model is consistently making a handful of random mistakes, then relabeling some data or collecting new data may resolve the issue. Alternatively, if the model is a making systematic errors then looking into the ontology or data collection procedure might be a better route.

I Additional Noise Examples

Examples were carefully selected to give good coverage of the different types of noise that occurred frequently within the data.

Dataset	Noise Type	Dialogue	Comments
WOW	Labeling Error ↪ Instance level ↪ Under	*Apprentice* I have visited the United States. To New York City, Los Angeles, and Seattle for work and vacation. Every city was unique with its own culture and loved every one of them. *Wizard* I haven't been to the East coast yet, but I have been to Los Angeles, which is Spanish for "The Angels" *Apprentice* Oh I never knew. The East coast always felt busier, the West coast felt more relaxed. *Wizard* Agreed! I grew up in Hawaii, where the life expectancy is amongst the highest in the nation. Do you like large cities or smaller towns?	*Correct label*: {topic: 'Los Angeles'} *Possible: missing labels:* {topic:'Hawaii' 'longevity'}
DD	Labeling Error ↪ Class level ↪ Uniform	*A* Hello Mike ! Would you like a drink ? *B* No, thank you. I had too much to drink yesterday evening. I had a bad hangover this morning. My head felt terrible. (happiness) *A* Were you celebrating something ? *B* Yes. It was a friend's birthday party. We drank all kinds of things - beer, wine and spirits. After midnight, we even drank cocktails! *A* It's a bad idea to drink a combination of alcoholic drinks. You should stick with one for the whole evening.	*Revise label*: happiness → disgust
ED	Discourse Attribute ↪ Disfluent	*A* I've got popcorn kernels to last me through retirement. I wonder how long they keep for. *B* That is nice. *A* Yea, it is. Do you like popcorn? *B* Yes. Why did you bought that many popcorn kernels?	*grammar mistake*

MWOZ	Ontology Inconsistency
	↪ Location
	
	Labeling Error
	↪ Annotator
	↪ Formatting
		User I'm looking for a special place, can you help?
		`attraction(type=Special)`
		System I need just a little more information to help. I think all places in Cambridge are special ...
		User I am looking specifically for Saint John's College. `attraction(type=special, name=Saint John's College.)`
		System sorry i dont have that in our list. is there something else i can do for you?
		User Okay, well I also need a train departing for Cambridge on Wednesday.
		`train(dest=cambridge,day=wednesday)`
		System I have several trains headed to Cambridge on Wednesday. Where will you be departing from? <*truncated*>
			ontology: uppercase
			ontology: lowercase
			formatting: added an extra period

ABCD	Labeling Error ↪ Class level ↪ Structured	*Agent* Thank you for contacting acmebrands. how can I help you? *Customer* I need to check the status of an order *Agent* Okay. Could I have your name or Account ID please? *Customer* alessandro phoenix *Action* Account has been pulled up for alessandro phoenix. `[Pull-up Account]` *Agent* Thank you. Do you just need to know the order's shipping status? *Customer* The email that I received showed it is being shipped to the wrong address. is it being shipped to <street_address>, jacksonville, mi	*Revision:* The agent followed the 'manage shipping' flow when instead they should have followed the 'shipping status' flow
ED	Labeling Error ↪ Instance level ↪ Partial	*A* I liked the way Lebron was snatched up by the Lakers *B* Will be a game changer. *A* I hope so. I hope they can crush the warriors *B* Did the Warriors get a replacement?	*Emotion label*: impressed *Missing labels*: hopeful, anticipating
SGD	Discourse Attribute ↪Inconsistent, Nonsensical	*User* I want to check in on March 2nd, and stay in 1 room until the 12th. *System* Okay, please confirm: You want to stay in New York at 1 hotel central park, checking in tomorrow, and checking out on March 12th. you also want 1 room. *User* Actually 3 rooms please. *System* Okay, confirm: You want 3 rooms and you check in tomorrow? *User* No. I want to check out on the 13th. *System* Please confirm: You want to check out of your hotel in New York on March 13th. *User* Yes, works for me. *System* THe reservation is made. *User* Does it allow pets? How much does it cost? *System* It is pet friendly, and costs $4,488. *User* Thank you, i'm all set.	*Inconsistent* since the user wants 1 room than arbitrarily changes to 3. *Nonsensical* since you would want to know the price of a hotel room before booking, not afterwards.
MIDAS	Labeling Error ↪ Class level ↪ Structured	*User* actually *Agent* what was that experience like *User* i have not	*Revise label*: statement → neg answer
GSIM	Ontology Inconsistency ↪ General	*System* please confirm the date is next monday at 8 pm for ephesus restaurant for 3 people *User* yes i confirm the date is next monday at 8pm for ephesus restaurant for 3 people *(multi-part label not needed)*	*Original label:* `AFFIRM_AFFIRM_AFFIRM_AFFIRM` *Revised label:* `AFFIRM`
KVRET	Ontology Inconsistency ↪ Date	*User* what is the weather like in the city *Assistant* Which city do you want the weather for? *User* Will it hail in Alameda today? *Assistant* Today in Alameda is gonna be drizzle, with low temperature of 70F, and high of 90F *User* Thank you! *Assistant* You are welcome.	`request - {date:today, loc:Alameda, weather:hail}` Date is annotated as "today", but should be a day of the week such as Monday, Tuesday

Investigating Explicitation of Discourse Connectives in Translation using Automatic Annotations

Frances Yung[1] **Merel C.J. Scholman**[1,2]
Ekaterina Lapshinova-Koltunski[3] **Christina Pollkläsener** [3] **Vera Demberg** [1]
[1]Saarland University, Saarbrücken, Germany
[2] Utrecht University, Utrecht, Netherlands
[3]University of Hildesheim, Hildesheim, Germany
{frances,m.c.j.scholman,vera}@coli.uni-saarland.de
{lapshinovakoltun, christina.pollklaesene}@uni-hildesheim.de

Abstract

Discourse relations have different patterns of marking across different languages. As a result, discourse connectives are often added, omitted, or rephrased in translation. Prior work has shown a tendency for explicitation of discourse connectives, but such work was conducted using restricted sample sizes due to difficulty of connective identification and alignment. The current study exploits automatic methods to facilitate a large-scale study of connectives in English and German parallel texts. Our results based on over 300 types and 18000 instances of aligned connectives and an empirical approach to compare the cross-lingual specificity gap provide strong evidence of the *Explicitation Hypothesis*. We conclude that discourse relations are indeed more explicit in translation than texts written originally in the same language. Automatic annotations allow us to carry out translation studies of discourse relations on a large scale. Our methodology using relative entropy to study the specificity of connectives also provides more fine-grained insights into translation patterns.

1 Introduction

Discourse connectives such as *because* and *however* are considered volatile items in translation: translators often add, rephrase or remove them (e.g. Zufferey and Cartoni, 2014). Prior studies have often focused specifically on whether connectives are added (i.e. the relation sense is *explicitated*) or removed (i.e. *implicitated*), and have shown that there is a tendency for explicitation in translation (but this also depends on various other factors, see e.g., Hoek et al., 2015, 2017; Lapshinova-Koltunski et al., 2022; Zufferey, 2016). The current work focuses on an understudied aspect of connectives in translation, namely when they are underspecified (e.g. connectives like "and" or "but" are compatible with many different types of discourse relations) or highly specific (e.g. the connective "nevertheless" can only mark concessive relations). The question we address is whether we can see a similar pattern of explicitation of connectives in translation for connectives that are already explicit (but possibly unspecific) in the source text.

One factor that impedes a comprehensive study of DCs in translation is the (manual) annotation effort that is required for this task. Consequently, many studies are restricted to limited samples and a subset of DCs. To facilitate a more comprehensive investigation, we explore an automatic approach to identify and align connectives. Specifically, we use language-specific discourse parsers (Bourgonje, 2021; Knaebel, 2021) and a neural word alignment model (Dou and Neubig, 2021) to link a large range of connectives and their translations in English and German parallel texts. We test the feasibility of this approach by replicating the well-established explicitation results in our newly created dataset. Using an empirical measure of cross-lingual specificity gap, we identify all the cases of (under)specifications instead of a subjectively defined subset.

Our contributions are: 1) We demonstrate that automatic word alignments and discourse parsers facilitate a comprehensive study of discourse connectives and relations in translation. 2) We show evidence for explicitation in translation, in terms of both insertion and specification of DCs; 3) We compare the cross-lingual specificity of English and German DCs; 4) The automatically aligned and annotated data are publicly available[1].

2 Background

2.1 Explicitation Hypothesis

Previous studies show that the translation of discourse connectives depends on various factors. One of the most well-known accounts, the Explicitation Hypothesis, suggests that translations tend to be

[1]https://osf.io/ybfxp/?view_only=8ef5f7a591064b7ea3334f706e544118

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 21–30
September 11–15, 2023. ©2023 Association for Computational Linguistics

more explicit than the source texts (Blum-Kulka, 1986). However, this does not mean that discourse relations are always explicitated in translation, or that explicitation of the relations is always due to the translation effect. Klaudy (1998) more specifically distinguishes between *obligatory explicitations* and *translation-inherent explicitations*. Obligatory explicitation results from grammatical and stylistic differences between the source and target languages, as well as pragmatic and cultural preferences of the source and target readers. For example, Becher (2010) found that over 50% of *damit* instances in German translated texts are the result of explicitation, but all except a few are explicitations that address the cross-lingual contrast.

By contrast, translation-inherent explicitations are language-independent and depend on the nature of the translation process. This type of explicitation is separate from structural, formal or stylistic differences between the two languages, and with culture-specific textual elements. Klaudy (2009) argues that, in order to identify any translation-inherent explicitations, corresponding *implicitation in the opposite translation direction* should be taken into account. That is to say, explicitation due to the contrast in the explicitness of the source and target languages (with some languages being more prone to expressing discourse relations through explicit connectives than others), should be counter-balanced by the degree of implicitation when translating in the other direction. Becher (2011b) found that the insertions of discourse connectives in English to German translation are in fact more than the number of omissions in German to English translation, but still, most of the insertions can be qualitatively explained by the known observation that German is more explicit than English (Hawkins, 1986; House, 2014; Becher, 2011a).

Various other factors have also been found to affect the explicitation of connectives, such as the type of the coherence relations and the connectives involved (Zufferey and Cartoni, 2014; Crible et al., 2019), the identity of the source and target languages (Zufferey, 2016), register and translator expertise (Dupont and Zufferey, 2017), contrast between the constraints and communicative norms of the source and target languages (Marco, 2018), the cognitive interpretability and expectedness of the relations in context (Hoek et al., 2015, 2017), information density and the mode of translation (Lapshinova-Koltunski et al., 2022).

2.2 Explicitation of DCs in translation

Much of the earlier work on explicitation of DCs focused largely on cases where connectives are inserted or omitted in translation or they provided qualitative estimations of specificity without basing it on a quantitative method (Crible et al., 2019; Lapshinova-Koltunski et al., 2022). In the current work, we propose a score to quantify the specificity gap between a connective and its translation, such as cases where a stronger connective is used in translation (e.g. "*and*" translated as "*außerdem*" in German). While previous works only study a limited subset of subjectively defined specification, our empirical approach allows us to identify all cases where a more specified connective verbalizes the relation to a greater degree.

The specificity of connectives likely differs between languages due to the contrast between the connective lexicons and discourse marking of these languages. This means that the entropy of English *and* might differ from the precise value of the entropy of German *und*. One connective could therefore appear to be more specific than another connective in a different language due to differences between the lexicons, even though both connectives express a similar range of relation senses. Previous studies found that the explicitation pattern of a given connective in a target language is directly related to the alternative options available in that language (Becher, 2011b; Zufferey and Cartoni, 2014). To address the issue of cross-lingual correspondence, we derive estimates of a connective's specificity empirically by normalizing connectives' entropy value within a language (see Section 3.3).

2.3 Identification and alignment of discourse connectives

Prior work is often based on a restricted selection of connectives. This can be attributed to the fact that connective identification on a large scale can be difficult, because many discourse connectives can also be used in non-connective contexts (e.g., *indeed* is not always used as a DC). Consequently, prior corpus studies have mostly focused on a handful of connectives and senses. For example, Zufferey and Cartoni (2014) analyzed 200 occurrences each of the English causal connectives *since*, *because* and *given that* in Europarl. The frequent causal connective *as* was excluded because it is often used in a non-connective usage. A more comprehensive analysis that takes into account a larger range of

connectives and coherence relation senses in the same text is critical to be able to get more insight into the general translation patterns of connectives. The current study explores the feasibility of using automatic methods to identify and align discourse connectives.

Automatic word alignment was an essential step in statistical machine translation (Och and Ney, 2000). In the era of neural machine translation, word alignment is often used for annotation projection, including the projection of English discourse annotations (Versley, 2010; Laali, 2017; Sluyter-Gäthje et al., 2020). The focus of these works is to associate discourse sense labels annotated for the DCs in English with the DCs in the human or machine-translated texts, in order to create discourse-annotated resources in the other languages. In contrast, we use word alignments to examine where the DC marking differs between source and target languages, when DCs are inserted, omitted or their specificity is changed.

Another line of work uses automatic word alignments to generate cross-lingual lexica of connectives. For example, Bourgonje et al. (2017) extract alignments between German and Italian adversative connectives that are identified based on connective lexicons of both languages. Özer et al. (2022) link the multilingual annotation of the TED-MDB corpus (Zeyrek et al., 2019) to induce multilingual connective lexicons. Robledo and Nazar (2023) examine the mapping of English and Spanish connectives in order to identify possible new categories of relation senses. In this work, we use a similar technique to investigate whether connectives are explicitated by insertion or specification. In contrast to existing work, we also use language-specific discourse parsers to identify connectives and exclude tokens of non-discourse usage in English and German texts. We then use a neural word aligner which has reported lower error rates compared with statistical aligners.

3 Methodology

3.1 Data

We analyze the parallel texts taken from the Europarl Direct Corpus (Cartoni and Meyer, 2012), which are proceedings from the European Parliament. A total of 33 proceedings are used in the analyses.[2] The data contains 171k tokens of English texts and their German translation from 18 proceedings, and 95k tokens of German texts and their English translation from 15 proceedings.

3.2 Identification and alignment of DCs in English and German texts

We use two language-specific parsers to identify and annotate the discourse relations in the English and German texts. We use the Discopy parser (Knaebel, 2021) to identify and classify DCs in the English original and translated texts. This parser considers the semantic representation of a connective token and its contexts. The classifier distinguishes discourse and non-discourse usage of the connective and labels each with a sense label based on the PDTB 2.0 framework (Prasad et al., 2008). The reported accuracies are 97.20% for connective identification, and 92.12% / 86.26% respectively for 4-way coarse-grained / 14-way fine-grained classification of the relation sense.

For the German texts, we use the German Shallow Discourse Parser (Bourgonje and Stede, 2018; Bourgonje, 2021) to identify and classify DCs in the German original and translated texts. The parser is based on a BERT architecture with additional syntactic features and ambiguity knowledge from the DimLex lexicon (Stede, 2002). It has been trained on the Potsdam Commentary Corpus (PCC) 2.2 (Bourgonje and Stede, 2020) to predict a sense labels defined in the PDTB 3.0 hierarchy (Webber et al., 2019). The reported results on the accuracy of this German parser regarding discourse-usage identification is 87.57% and 85.63% / 80.57% respectively for 4-way coarse-grained / 16-way fine-grained classification of the relation sense.

We align the identified connectives cross-lingually using the Awesome Align word alignment model (Dou and Neubig, 2021), which extracts corresponding tokens (including m:n mappings and "null" alignments) in a pair of bilingual sentences based on multilingual embeddings of the tokens and fine-tuned on parallel texts. An error rate of 15.1% is reported evaluating against human annotation of English-German word alignments (of all words, not just DCs), which out-performs statistical alignment models such as GIZA++ (Och and Ney, 2000) and eflomal (Östling and Tiedemann, 2016).

To ensure that the annotation tools produce reliable output for our data, we manually analyzed the

[2]These 33 proceedings are selected because they overlap with instances included in the discourse-annotated DiscoGeM corpus (Scholman et al., 2022), which could be used in future contrastive studies.

automatic annotations of 200 randomly extracted connective pairs each from the English-German and German-English translation data. The accuracy (precision) of connective identification and 4-way sense classification are 85% and 92% for English and 83% and 90% for German. The alignment accuracy is 90%. Taking into account error-propagation, in our analysis, we annotate DCs only on one side and analyze their alignment to the other side without considering whether the aligned words are also identified as DCs. In addition, we improve the automatic annotations by syntactic rules that remove unlikely DC candidates (e.g. *damit...,..zu..* is not a DC) and "unalign" tokens that cannot mark connectives, such as *'power'* or *'reading'*). We analyze the alignments of the source/target English and German texts respectively, in order to identify explicitation and implicitation in both translation directions.

3.3 Quantifying specificity of connectives

We determine the specificity level of each English and German connective based on their manual annotation in existing discourse-annotated resources. For English connectives, we extract the distribution of sense labels (after removing the *speechact* and *belief* tags) assigned to the *explicit* connectives in PDTB3.0. We extract the sense distribution of each German connective similarly based on their sense annotation in the PCC2.0 corpus (Bourgonje and Stede, 2020).

It is possible that the corpora from which we extract the specificity information differ in domain or aspects of how the annotation schemes were applied, such that in one language, a wider variety of relations was annotated than in the other. In order to remove such effects, we define the specificity of each connective by the entropy of its sense distribution in relation to the entropy of all explicit relations in the corresponding corpus. We further round the values to 1 decimal place. We call this measure *relative entropy*.

Overall, we assign *relative entropy* to 173 English and 126 German connective types. The average relative entropy of the English and German connectives are 0.122 and 0.065 respectively.

Connectives that are aligned to "null" in the target text are considered *omissions*, and connectives that are aligned to "null" in the source text are considered *insertions*. Similarly, connectives in the source and target texts that are aligned to a *less specific* connective are identified as *under-specification* and *specification* respectively.

4 Results

We first look at how connectives are implicitated and explicitated in English and translations, and then we will take a closer look at how the English and German connectives correspond to each other.

4.1 Implicitation & explicitation of DCs

A total of 8058 English and 9739 German connectives have been identified and annotated by the discourse parsers and aligned. Table 1 shows the proportions of automatically identified connectives that are aligned to *"null"* or a DC of higher entropy in the other language, grouped by four categories of relations as identified by the discourse parsers. Alignments of connectives in the **source** texts to *"null"* or a higher entropy DC means **omission** and **under-specification**, while the corresponding alignments of connectives in the **target** texts would mean **insertion** and **specification** in translation.[3]

It can be observed that, when translating from English to German (top sub-table), more DCs are added than removed (26.1% vs 13.8%). The reverse is observed in German to English translation (bottom sub-table), where more DCs are removed than added (21.6% vs 12.3%). The same tendency is observed for under-specification and specification. This confirms the previous qualitative conclusion that German is more explicit in terms of discourse relation marking (Becher, 2011b,a).

Zufferey and Cartoni (2014) and Zufferey (2016) found that, based on the analysis of the translation of a subset of connectives, explicitation is not a general phenomenon. The roles of the source and target languages, the type of relations, and the specific DCs all have influences. We also see different patterns of explicitation depending on the translation directions and category of relations, e.g., CONTINGENCY relations are explicitated more often in English than in German.

Moreover, our analysis of connectives typically expressing all types of relation senses provides a

[3]The implicitation and explicitation proportions do not add up to 100%, because: 1) the proportions are normalized against the total connective counts of the each source/target language; and 2) overall, 58.0% of the connectives have been aligned to a connective of the same specificity level, and the specificity scores of 22.7% of the identified connectives or the aligned tokens is unknown (i.e. those tokens are not annotated in PDTB3.0 or PCC2.0).

EN →DE	**EN original** (171K tokens)				**DE translation** (164K tokens)			
	ttl. DC count	align to 'null' (omission)	align to a DC of higher rel. ent. (under-specif.)	impl. total	ttl. DC count	align to 'null' (insertion)	align to a DC of higher rel. ent. (specification)	expl. total
EXPANSION	2329	13.1%	9.2%	22.4%	2821	**20.6%**	**3.1%**	**23.7%**
CONTINGENCY	906	16.8%	6.8%	23.6%	1383	**33.0%**	**18.7%**	**51.8%**
COMPARISON	978	7.5%	13.3%	20.8%	979	**24.9%**	**35.4%**	**60.4%**
TEMPORAL	426	25.6%	13.8%	39.4%	505	**40.2%**	**16.6%**	**56.8%**
Total	4639	13.8%	10.0%	23.8%	5688	**26.1%**	**13.7%**	**39.8%**

DE →EN	**DE original** (95K tokens)				**EN translation** (107K)			
	ttl. DC count	align to 'null' (omission)	align to a DC of higher rel. ent. (under-specif.)	impl. total	ttl. DC count	align to 'null' (insertion)	align to a DC of higher rel. ent. (specification)	expl. total
EXPANSION	1876	17.6%	3.0%	20.7%	1605	**13.8%**	**20.1%**	**33.9%**
CONTINGENCY	1146	24.5%	16.8%	41.3%	831	10.5%	**7.8%**	18.3%
COMPARISON	638	21.2%	32.1%	53.3%	673	**9.5%**	**15.9%**	**25.4%**
TEMPORAL	391	32.7%	6.4%	39.1%	310	15.8%	**41.9%**	**57.7%**
Total	4051	21.6%	11.8%	33.4%	3419	12.3%	**18.3%**	**30.6%**

Table 1: Proportions of connectives that are not aligned to any words in the target text (*omission*) or the source text (*insertion*); and connectives that are aligned to a connective of higher relative entropy (rel. ent.) in the target text (*under-specification*) or the source text (*specification*). *Impl.* and *expl.* totals are based on the sum of *omission/insertion* and *under-specification/specification* respectively. Bolded proportions refer to proportions of explicitation exceeding the proportions of implicitation of the same type in the opposite translation direction (compared against the sub-table in diagonal).

Implicitation	**Explicitation**
EN→DE omission	EN→DE insertion
and (177), also (69), when (62), if (49), but (43), so (41)	und (287), dann (121), wenn (88), also (61) damit (57), aber (52)
DE→EN omission	DE→EN insertion
und (105), dann (105), aber (78), sondern (68), wenn (52), deshalb (49)	and (158), also (26), but (26), if (25) when (25), so (13)
EN→DE under-specif.	EN→DE specification
also → auch (173), but → sondern (113), then → dann (54), because → da (22), so that → damit (16)	but→ jedoch (89), however→ jedoch (82), but→ doch (70), when→ wenn (67), although→ obwohl (26)
DE→EN under-specif.	DE→EN specification
aber → however (80), wenn → when (67), jedoch → however (32), denn → for (30), allerdings → however (12)	auch → also (281), dann→ then (126), sondern - but (86), damit → so that (25), sondern → rather (13)

Table 2: The most freqent connective omissions, insertions, under-specifications and specifications (counts in brackets) in both translation directions.

more comprehensive picture. The results show that the explicitation strategy also differs across different relation senses and translation directions. For example, relations are explicitated more by insertion, while more relations in German translation, in particular temporal relations, are explicitated by specification in English. Within German translation, many CONTINGENCY (33.0%) and TEMPORAL connectives (40.2%) are inserted, while comparatively, COMPARISON relations are explicitated more by specification (35.4%).

To find out whether these patterns can be explained by obligatory explicitations or translation-inherent explicitions, we look at the connectives that are most frequently omitted/inserted and (under-)specified, see Table 2. It can be seen that connectives that are most frequently added in the translation, are also those that are most frequently omitted in the opposite translation direction, consistent with reports by Hoek et al. (2015) and supporting the findings of Becher (2011b) that most explicitations are obligatory due to the cross-lingual contrast of English and German.

Taking into account obligatory translation effects, we still find more explicitation in the translation than would have expected (see bolded figures in Table 1). In other words, the *Explicitation Hypothesis* is quantitatively confirmed for both explicitation strategies, translation directions and all categories of relations, save two exceptions: CONTINGENCY and TEMPORAL connectives are frequently dropped in English to German translation and they are not counter-balanced by the insertion in German to English translation. Table 2 suggests that the high rate of these omissions could be attributed to the dropping of *when*, *if* and *so* in English to German translation. Previous work has found that CAUSAL DCs like *so* are often omitted

due to processing ease (Hoek et al., 2017).

In addition, many of the explicitated COMPARISON relations come from the translation of *but* and *however*, which are ambiguous because they can signal both CONTRAST and CONCESSION relations. The German translation often specifically signals CONCESSION, such as *jedoch* and *allerdings*. We will analyze some of these cases in Section 5 to see if such explicitation is obligatory or translation-inherent.

4.2 Cross-lingual correspondence of DCs

Next, we look into the mutual correspondence between English and German connectives. Figure 1 shows the normalized distribution of the alignment between each source connective (x-axis) and their translation (y-axis; at least the top two most common translations are displayed). Higher numbers / darker colors represent more frequent translation alignments.

It can be observed that some connectives have one or two dominating translations (e.g. English: *also, and, if, then*; German: *auch, und, weil*), while others can have an even distribution of various translations (e.g. English: *so, but*; German: *deshalb*). While many of the correspondences in the two translation directions are asymmetrical (e.g. 82% of *auch* is translated to *also*, but only 45% of *also* is translated to *auch*), some correspondences are symmetrical, indicating that the pair of connectives are of mutual correspondence (e.g. *and* is frequently translated as *und* and vice versa; the same goes for *then* and *dann*).

Figure 1 also suggests a general trend that English connectives are translated to a wider range of German connectives, while German connectives more often have one dominating English translation (more darker color cells in the bottom figure). It is to be expected that English connectives are more ambiguous than German, as English is less explicit in terms of discourse markedness (House, 1997; Becher, 2011a). We quantify this observation by considering the *cross-lingual specificity* of English and German connectives based on the diversity of their translations. This is calculated as the entropy of the distribution of alignments of each unique connective in the source texts (i.e. the entropy of the distribution per column in Figure 1). Figure 2 shows the distribution of connectives grouped by the entropy of their translation alignments. Connectives with less than 20 occurrences

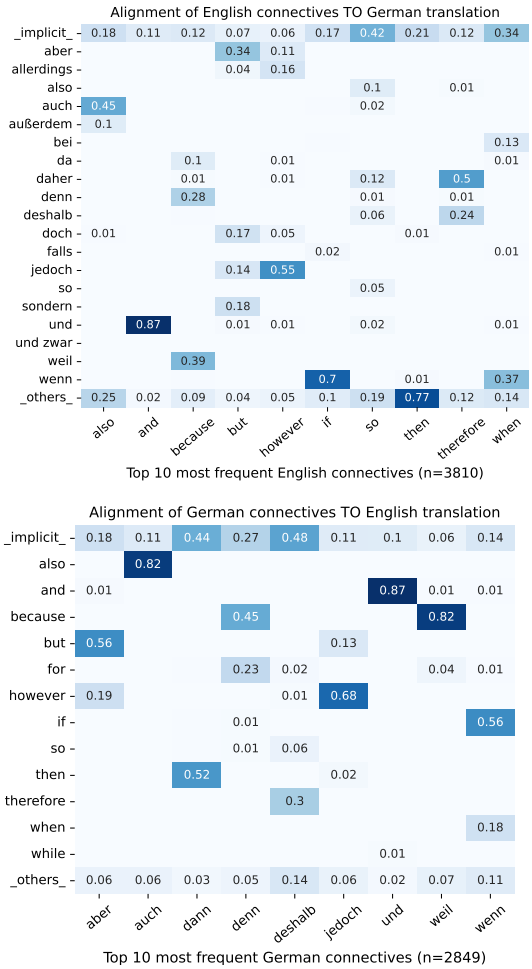


Figure 1: Alignment between connectives in the source texts (x-axis) with their corresponding tokens in translation (y-axis); the first row *_implicit_* means the connective is not aligned to any words in the target sentence, and the last row _others_ refers to the proportions of alignments to tokens that are not displayed on the y-axis.

are not included since the alignment distributions may divert from the actual distribution due to their sample size. It can be seen that most English DCs are more versatile and correspond to a wide range of German DCs, while a normal distribution is observed for the German DCs: some DCs have more correspondences and some have less.

To summarize, the automatic connective annotation and alignment procedure allows us to extract the complex mapping between connectives empirically and instantly. This enables us to identify systematic patterns such as the overall specificity of English connectives in terms of English-German translation. We found empirical evidence that explicitations counter-balance and exceed opposite implicitation.

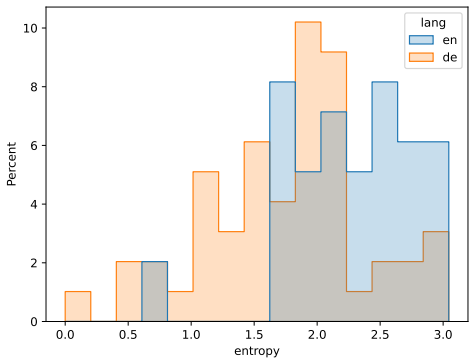


Figure 2: Distribution of connectives grouped by the entropy of their translation alignment.

We however also note that one needs to consider the effect of possible annotation errors using such an automatic approach. Based on our manual inspection of the 400 alignments, most of the error comes from the over-identification[4] of English *and* and German *und*: these often did not function as connectives, but were identified as such by the parsers. In most of these cases, *and* and *und* were aligned, which means that they were not counted as explicitation nor implicitation. Consequently, our reported explicitation / implicitation rate of EXPANSION could actually be higher, because the sample size should be smaller. Regarding errors specific to the alignment of connectives, we found that most alignment errors were false positives (i.e. a connective was aligned to a non-connective word, when in fact it was supposed to align to *null*), meaning the insertion / omission rates could actually be higher.[5] Therefore, manual qualitative analysis is still necessary to confirm the findings. This will also be demonstrated in the next section.

5 Qualitative analysis

The qualitative results show that there are more explicitations in translation after counter-balancing implicitation in the other translation direction. Now the question is, are these explicitations actually coming from the nature of the translation process, or are they due to the contrast between the two languages or other reasons? We try to gain some insights through a qualitative analysis.

[4]Note that the manual inspection did not include cases where a connective was missed by the parsers.

[5]The relative entropy of the falsely aligned words would most likely be "unknown", so they are not counted as (under-)specification.

We analyze the alignment instances to see if the explicitated translations are **obligatory** or **translation-inherent** (see Sec. 2.1). This analysis revealed various cases of obligatory explicitation. First, Table 1 shows that TEMPORAL relations are often specified in German to English translation. Table 2 suggests that the high explicitation rate of German TEMPORALs can be attributed to the frequent specification of *dann* (which can signal both TEMPORAL and CONDITIONAL according to PCC2.0) to *then* (which dominantly signals TEMPORAL in PDTB3.0). These explicitations are likely to belong to obligatory explicitations, because *then* is the only English DC that signals a PRECEDENCE relation like *dann* does, and has a similar level of markedness.

Second, for German translation, Table 1 also reveals that COMPARISON relations are often specified. The high specification rate of English COMPARISONs comes from the frequent translation of *but* to *jedoch* or *doch*, and *however* to *jedoch*, as seen in Table 2. The translation of *however* to *jedoch* might also be categorized as obligatory explicitation. The two connectives are very similar in their meaning and usage (both are predominantly be used to mark CONTRAST and CONCESSION), but English *however* is also occasionally used to mark SYNCHRONOUS relations among its many annotations in PDTB3.0 – this sense did not occur for *jedoch* in the PCC2.0. Similarly, the frequent specification of *wenn* to *when* belongs to this case. *Wenn*, which can ambiguously signal a CONDITION or SYNCHRONOUS relation, often has to be translated to the less specific *when* to mark a SYNCHRONOUS relation naturally because of a lack of other suitable DCs in English.

The translation of *but* to *doch/jedoch* differs from the previously discussed obligatory explicitations and might actually be translation-inherent: translators could have translated *but* to *aber*, which matches *but* semantically and also in terms of strength and specificity, instead of specifying the relation with *jedoch* or *doch*. To gain further insight into the reason for these explicitations, a trained translator manually analyzed these cases using a "substitution test": we produced an alternative translation using *aber*, making necessary grammatical changes. If the resulting translation is equally acceptable, then it could be a case of translation-inherent explicitation.

We found that in 35% of the *but*-instances that

were translated into *doch/jedoch*, these more specific could have been chosen because the resulting syntactic or stylistic structure is preferred; that is, they do actually appear to be cases of obligatory explicitation. For example:

> It is important to have EU and national targets, **but** it is also important to have a European directive...
> Es ist zwar wichtig, Ziele auf EU- und einzelstaatlicher Ebene zu setzen, **doch** ist es ebenso wichtig, eine europäische Richtlinie zu schaffen...

In this case, having chosen *zwar* in the previous clause, the translator likely used *doch*, because they often occur together. But in 65% of the cases, the use of *aber* is equally acceptable, and thus these cases appear to represent translation-inherent explicitation. For example:

> Its starting point is the European Year Against Racism 1997 **but** the context has moved on significantly.
> Ausgangspunkt war das Europäische Jahr gegen Rassismus 1997, **doch/aber** der Kontext wurde seither beträchtlich weiterentwickelt.

Among these acceptable cases, in 38% of the total cases, *doch* or *jedoch* sometimes fit better to the formality of a parliament discussion, while *but* is a lighter DC typical in spoken English, for example:

> **But** as has been pointed out, the adoption of a rigorous definition of the precautionary principle is crucial.
> ***Doch**, wie bereits festgestellt wurde, ist dabei die Verabschiedung einer strikten Definition des Vorbeugeprinzips von entscheidender Bedeutung.*

One possible explanation is the *domain gap* between the source and target texts. The source texts of the Europarl corpus are prepared speeches of the parliament, while the target texts are the published translation of these scripts. In other words, the source texts are prepared to be spoken while the target texts are for reading. This could be a reason that the discourse relations in the translated texts of the Europarl corpus are more specified than the original texts, corresponding to the *situational* and *translation-task* variables as discussed in House (2004). Analysis on data from another genre could confirm this domain and genre effect.

6 Discussion

The current study investigated explicitation and implicitation of discourse connectives in English-German parallel texts. To gain a comprehensive insight of the patterns underlying explicitation, we exploited an automatic approach to connective identification and alignment, which allowed us to study a large variety of connectives (173 English and 126 German connective types) and many samples per language (8058 English and 9739 German connectives were identified in our dataset). We evaluated the feasibility of this approach by first studying whether we could replicate the established effect of explicitation in translation between English and German texts. We furthermore extended existing findings by defining explicitation in a more fine-grained sense as specification of the relation sense, and investigating whether we can see a similar pattern of explicitation of connectives for those connectives that were already explicit in the source text.

Our quantitative results provide strong evidence for the *Explicitation Hypothesis*: taking into account the counter-balance of implicitation in the opposite translation direction, there is still considerable more explicitation in translation. Manual qualitative analysis suggests that a domain effect may have played a role. These findings are in line with already established effects in prior work, and thus support the reliability of the insights that the automatic approach can provide.

We also propose a novel method of studying explicitation in translation, namely by considering the relative entropy of corresponding connectives in parallel text. Our results showed that the general pattern of explicitation in translation replicates to specification of connectives. Furthermore, we found that English connectives are generally less specific than German ones, considering all types of connectives and their translation in our data. The large-scale alignments provide additional insights, such as the fine-grained interaction between relation type and explicitation strategy across different languages. Such analyses would not have been possible without taking into account how all types of DCs are translated within the same span of text and a well-defined measure to identify cross-lingual specificity gap.

We conclude that discourse relations indeed tend to be explicitated in translation. Our proposed automatic approach is feasible for studying translation of connectives in parallel text. We were able to replicate known effects for German-English translations and extend these findings to specification of connectives using relative entropy. The cross-lingual analysis in large scale allows us to identify

language-specific patterns in discourse production, which is useful for the generation of multi-lingual discourses. Future work will focus on applying a similar methodolgy to less studied language pairings to gain further insight into the generalizability of DC translation and production patterns.

Acknowledgements

This project is supported by the German Research Foundation (DFG) under Grant SFB 1102 ("Information Density and Linguistic Encoding", Project-ID 232722074).

References

Viktor Becher. 2010. Towards a more rigorous treatment of the explicitation hypothesis in translation studies. *Trans-kom*, 3(1):1–25.

Viktor Becher. 2011a. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts*. Ph.D. thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky.

Viktor Becher. 2011b. When and why do translators add connectives?: A corpus-based study. *Target. International Journal of Translation Studies*, 23(1):26–47.

Sh Blum-Kulka. 1986. Shifts of cohesion and coherence in translation. *Interlingual and Intercultural Communication. Discourse and Cognition in Translation and Second Language Acquisition Studies*, pages 17–35.

Peter Bourgonje. 2021. *Shallow discourse parsing for German*, volume 351. IOS Press.

Peter Bourgonje, Yulia Grishina, and Manfred Stede. 2017. Toward a bilingual lexical database on connectives: Exploiting a german/italian parallel corpus.

Peter Bourgonje and Manfred Stede. 2018. Identifying explicit discourse connectives in german. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 327–331.

Peter Bourgonje and Manfred Stede. 2020. The potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1061–1066.

Bruno Cartoni and Thomas Meyer. 2012. Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, CONF, pages 2132–2137.

Ludivine Crible, Ágnes Abuczki, Nijolė Burkšaitienė, Péter Furkó, Anna Nedoluzhko, Sigita Rackevičienė, Giedrė Valūnaitė Oleškevičienė, and Šárka Zikánová. 2019. Functions and translations of discourse markers in ted talks: A parallel corpus study of underspecification in five languages. *Journal of Pragmatics*, 142:139–155.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*.

Maïté Dupont and Sandrine Zufferey. 2017. Methodological issues in the use of directional parallel corpora: A case study of english and french concessive connectives. *International journal of corpus linguistics*, 22(2):270–297.

John A Hawkins. 1986. *A comparative typology of English and German: Unifying the contrasts*. London Sydney: Croom Helm.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted JM Sanders. 2015. The role of expectedness in the implicitation and explicitation of discourse relations. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 41–46.

Jet Hoek, Sandrine Zufferey, Jacqueline Evers-Vermeul, and Ted JM Sanders. 2017. Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of pragmatics*, 121:113–131.

Juliane House. 1997. *Translation quality assessment: A model revisited*. Gunter Narr Verlag.

Juliane House. 2004. *Explicitness in discourse across languages*. Bochum: AKS.

Juliane House. 2014. *Translation quality assessment: Past and present*. Springer.

Kinga Klaudy. 1998. Explicitation. *Routledge encyclopedia of translation studies*, pages 80–84.

Kinga Klaudy. 2009. The asymmetry hypothesis in translation research. *Translators and their readers. In Homage to Eugene A. Nida. Brussels: Les Editions du Hazard*, 283:303.

René Knaebel. 2021. discopy: A neural system for shallow discourse parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133.

Majid Laali. 2017. *Inducing discourse resources using annotation projection*. Ph.D. thesis, Concordia University.

Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Heike Przybyl. 2022. Exploring explicitation and implicitation in parallel interpreting and translation corpora. *The Prague Bulletin of Mathematical Linguistics*, (119):5–22.

Josep Marco. 2018. Connectives as indicators of explicitation in literary translation: A study based on a comparable and parallel corpus. *Target. International Journal of Translation Studies*, 30(1):87–111.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 440–447.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Sibel Özer, Murathan Kurfalı, Deniz Zeyrek, Amália Mendes, and Giedrė Valūnaitė Oleškevičienė. 2022. Linking discourse-level information and the induction of bilingual discourse connective lexicons. *Semantic Web*, (Preprint):1–22.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*.

Hernán Robledo and Rogelio Nazar. 2023. A proposal for the inductive categorisation of parenthetical discourse markers in spanish using parallel corpora. *International Journal of Corpus Linguistics*.

Merel Scholman, Dong Tianai, Frances Yung, and Vera Demberg. 2022. Discogem: A crowdsourced corpus of genre-mixed implicit discourse relations. In *the 13th Language Resources and Evaluation Conference (LREC 2022)*, pages 3281–3290. European Language Resources Association.

Henny Sluyter-Gäthje, Peter Bourgonje, and Manfred Stede. 2020. Shallow discourse parsing for under-resourced languages: Combining machine translation and annotation projection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1044–1050.

Manfred Stede. 2002. Dimlex: A lexical approach to discourse markers. *A. Lenci and*, 501:1–15.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pages 83–82.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Deniz Zeyrek, Amalia Mendes, Yulia Grishina, Murathan Kurfali, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (ted-mdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–38.

Sandrine Zufferey. 2016. Discourse connectives across languages: Factors influencing their explicit or implicit translation. *Languages in Contrast. International Journal for Contrastive Linguistics*, 16(2):264–279.

Sandrine Zufferey and Bruno Cartoni. 2014. A multifactorial analysis of explicitation in translation. *Target. International Journal of Translation Studies*, 26(3):361–384.

What's Hard in English RST Parsing? Predictive Models for Error Analysis

Yang Janet Liu and **Tatsuya Aoyama** and **Amir Zeldes**
Department of Linguistics
Georgetown University
{yl879, ta571, amir.zeldes}@georgetown.edu

Abstract

Despite recent advances in Natural Language Processing (NLP), hierarchical discourse parsing in the framework of Rhetorical Structure Theory remains challenging, and our understanding of the reasons for this are as yet limited. In this paper, we examine and model some of the factors associated with parsing difficulties in previous work: the existence of implicit discourse relations, challenges in identifying long-distance relations, out-of-vocabulary items, and more. In order to assess the relative importance of these variables, we also release two annotated English test-sets with explicit correct and distracting discourse markers associated with gold standard RST relations. Our results show that as in shallow discourse parsing, the explicit/implicit distinction plays a role, but that long-distance dependencies are the main challenge, while lack of lexical overlap is less of a problem, at least for in-domain parsing. Our final model is able to predict where errors will occur with an accuracy of 76.3% for the bottom-up parser and 76.6% for the top-down parser.

1 Introduction

Powered by pretrained language models, recent advancements in NLP have led to rising scores on a myriad of language understanding tasks, especially at the sentence level. However, at the discourse level, where analyses require reasoning over multiple sentences, progress has been slower, with generalization to unseen domains remaining a persistent problem for tasks such as coreference resolution (Zhu et al., 2021) and entity linking (Lin and Zeldes, 2021).

One task which remains particularly challenging is hierarchical discourse parsing, which aims to reveal the structure of documents (e.g. where parts begin and end, which parts are more important than others) and make explicit the relationship between clauses, sentences, and larger parts of the text, by labeling them as expressing a type of e.g. CAUSAL, ELABORATION, etc. More specifically, hierarchical discourse parses identify connections between elementary discourse units (EDUs, usually equated with propositions) in a text or conversation, classify their functions using a closed tag set, and form a recursive tree structure, which indicates the locally most prominent EDU in each tree or subtree. Figure 1 shows an example tree in the most popular hierarchical discourse formalism, Rhetorical Structure Theory (RST, Mann and Thompson 1988), in which the list of units 37–38 is the most prominent (being pointed to by other units directly or indirectly), and discourse relation labels such as CAUSE are identified using edge labels, whose definitions in RST are based on the rhetorical effect which the writer (or speaker) is thought to be conveying to the reader (or hearer).

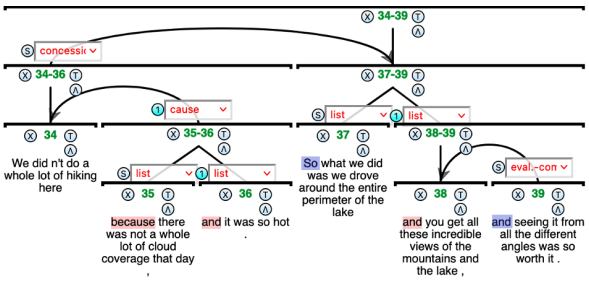


Figure 1: An RST analysis of a *vlog* excerpt. Tokens highlighted in red are discourse markers associated with relations in the tree, while tokens highlighted in blue are distractors, with no corresponding relation.

There is by now substantial evidence showing that even for a high resource language like English, state-of-the-art (SOTA) neural RST discourse parsers, whether employing a top-down or a bottom-up architecture, do not perform well across domains (Atwell et al., 2021, 2022; Yu et al., 2022; Aoyama et al., 2023), with some crucial tasks, such as predicting the most prominent Central Discourse Unit (CDU) of each document, performing at just 50% (Liu and Zeldes, 2023). At the same time,

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 31–42
September 11–15, 2023. ©2023 Association for Computational Linguistics

we do not have a good understanding of what exactly prevents good performance—is it the fact that some relations are **well-marked** (for example, most CONTINGENCY relations are marked by the discourse marker (DM) *if*, but most EVALUATION relations lack a common marker)? Conversely, is the **presence of distracting markers** not associated with the correct relation (e.g. an additional temporal marker such as *then* inside a unit with a non-temporal function)? Alternatively, is it the difficulty in identifying high-level relations, between groups of multiple sentences or paragraphs, compared to less tricky intra-sentential relations between clauses? Or is it just the prevalence of out-of-vocabulary (OOV) items in test data?

In this paper, we would like to systematically evaluate the role of these and other factors contributing to errors in English RST discourse parsing. Our contributions include:

- Annotation and evaluation of the `dev`/`test` sets of the English RST-DT (Carlson et al., 2003) and GUM datasets (Zeldes, 2017), for explicit relation markers, as well as distracting markers not signaling the correct relation;
- Parsing experiments with two different SOTA architectures to examine where degradation happens;
- Development and analysis of multifactorial models predicting where errors will occur and ranking importance for different variables;
- Qualitative and quantitative error analysis.

Our results reveal that while explicit markers and distractors do play a role, the most significant predictor of difficulty is inter-sentential status and the specific relation involved. At the same time, our error analysis indicates that distractors often correspond to true discourse relations which are not included in the gold-standard tree, but may be included in alternative trees produced by other annotators. In addition, we find that OOV rate plays only a minor role, that architecture choice is presently not very important, and that genre continues to matter even when all other factors are known. All code and data are available at `https://github.com/janetlauyeung/NLPErrors4RST`.

2 Related Work

2.1 Discourse Structure in Discourse Parsing

Discourse parsing is the task of identifying the coherence relations that hold between different parts of a text. Regardless of discourse frameworks or formalisms, identifying intra-sentential, inter-sentential, or inter-paragraph discourse relations may pose different levels of difficulty to parsers due to their various characteristics and levels of explicitness (e.g. Zhao and Webber 2021; Dai and Huang 2018; Muller et al. 2012). Intuitively, this becomes increasingly important for discourse parsing in a hierarchical framework such as RST, where long-distance relations are more frequent.

Researchers have therefore been considering ways of dealing with long-distance relations for nearly twenty years, starting with the structure-informed model proposed by Sporleder and Lascarides (2004) to tackle local and global discourse structures such as paragraphs. Other multi-stage parsing models, for example, as developed by Joty et al. (2013, 2015), have taken into account the distribution and associated features of intra-sentential and inter-sentential relations, achieving competitive results for English document-level parsing.

Later models expanded on these approaches by incorporating paragraph information to better capture high-level document structures. For instance, Liu and Lapata (2017) proposed a neural model leveraging global context, enabling it to capture long-distance dependencies and achieving SOTA performance. Yu et al. (2018) used implicit syntactic features in a hierarchical RNN architecture. Active research continues on developing multi-stage parsing algorithms aiming at capitalizing on structural information at the sentence or paragraph-levels (Wang et al., 2017; Lin et al., 2019; Kobayashi et al., 2020; Nishida and Nakayama, 2020; Nguyen et al., 2021).

2.2 Explicit and Implicit Relations in RST

Unlike in hierarchical RST parsing, work on shallow discourse parsing in the framework of the Penn Discourse Treebank (PDTB, Prasad et al. 2014), in which relations apply between spans of text without forming a tree, has long distinguished explicitly and implicitly marked discourse relations. Explicit relations are signaled by connectives such as 'but' or 'on the other hand', while implicit ones lack such marking. It is well-established that shallow parsing of explicit discourse relations is substantially easier due to the availability of connective signals, which, although not unambiguous, narrow down likely senses for relations. For example, the best systems from Knaebel (2021) achieved an F1 score

of 62.75 on explicit relations and an F1 score of 40.71 on implicit relations for Section 23 of WSJ using PDTB v2 (Prasad et al., 2008). The DISRPT shared task created a relation classification task in 2021 (Zeldes et al., 2021), and the 2023 edition (Braud et al., 2023) reported separate mean accuracy scores for explicit (79.32) and implicit (50.85) relations across six datasets in 4 languages.

RST datasets used in hierarchical discourse parsing do not make such a distinction, in part because RST trees include very high-level relations between entire sections of documents, which are less likely to be marked by such items. As a result, such a distinction is not available, meaning that we are in the dark regarding the prevalence and importance of such markers for RST parsing.

We are aware of two prior works analyzing connectives for RST data: the RST Signalling Corpus (RST-SC, Das et al. 2019) analyzes each relation in the English RST-DT dataset, indicating which relations were signaled by a DM (DMs roughly include the same items as PDTB connectives; see Webber et al. (2019) and Das and Taboada (2014) for complete inventories of markers). However, the data is limited to newswire material and does not provide an alignment of analyses to actual tokens, limiting the possibilities for model building (i.e. we only know whether a DM was present somewhere, but not which token in the text it was or in which exact EDU it appeared). It also does not indicate whether DMs were present which ***did not*** signal the relation in the tree (i.e. distractors). Although previous efforts targeted DM tokens in RST-DT (Liu and Zeldes, 2019) as well as such DM tokens in non-newswire texts (Liu, 2019), no previous study has examined the role of DMs in RST parsing.

Stede and Neumann (2014) enriched an RST corpus of German with token-aligned connectives and the relations they signal, allowing investigation of their positions and the presence of distracting connectives. However, the annotations were not mapped to the RST relations in the corpus, making exact inferences again tricky, and the size of the corpus (32K tokens) precludes training high quality models. This corpus too is limited to the newspaper domain, which also motivates us to annotate genre-rich data, described in the next section.

Finally we note that data in other frameworks, including not only PDTB but also SDRT (Segmented Discourse Representation Theory, Asher and Lascarides 2003), contains multiple concurrent discourse relations, providing information about the presence of competing or distracting relations. However, SDRT data does not include connective annotations, and apart from the coverage of RST-SC's overlapping data with the Wall Street Journal (WSJ) in PDTB, there is no way to extract a mapping between connectives and RST relations in any existing dataset (for attempts at aligning PDTB and RST-DT, see Demberg et al. 2019).

In this paper, we therefore begin by creating hand-annotated data (using rstWeb, Gessler et al. 2019) associating exact DM tokens with RST-style relations, or indicating their status as distractors, not associated with any relation in the gold tree. These latter DMs are especially interesting, since they could indicate that some parser errors are not exactly errors, instead corresponding to concurrent relations not present in the gold trees.

3 Data

To examine the role of explicit vs. implicit relations in parsing errors, we first need to know which relations were explicitly signaled. To that end, we use PDTB's methodology to define explicit connectives. Note that RST papers often use the term DM without clear inventories; from this point on we will use 'DM' for brevity, but strictly adhere to the PDTB English inventory. Specifically, we annotate data from the two largest RST corpora for English, covering the `test` set of RST-DT[1] (Carlson et al., 2003) and the `test` and `dev` sets of GUM (Zeldes, 2017) , with 1) **discourse markers** (including 'distractor' DMs) and 2) **associated relations**, thereby attaching DMs to each relation they signal, or no relation. Table 1 gives an overview of the data.

	RST-DT	**GUM v9**
# of docs	385	213
train/dev/test	347 / – / 38	165 / 24 / 24
# of toks	203,352	203,780
# of EDUs	21,789	26,310
# of genres	1	12
# of relation labels	78	32
# of relation classes	17	15
# of relation instances	18,630	23,451

Table 1: Overview of the Largest English RST Corpora.

Inter-Annotator Agreement To assess the reliability and quality of the human annotations, we conduct an inter-annotator agreement study on the `test` set of RST-DT and report average mutual F1

[1]RST-DT has no established separate `dev` set.

scores. The use of RST-DT can also facilitate some comparisons between the PDTB and RST frameworks as a number of documents from the WSJ section of the Penn Treebank (Marcus et al., 1993) were annotated in both PDTB v3 and RST-DT. In total, we double-annotated 38 documents, divided to overlap among three annotators. For DMs, the average F1 score was 95.2, and for associated relations, the average F1 score given a DM was 96.7. These scores indicate a high agreement between annotators for both tasks.

Automatic Parses In order to examine parsing errors from different architectures, we select two SOTA-performing parsers to obtain automatic parses: a BOTTOM-UP one from Guz and Carenini (2020), using their best `SpanBERT-NoCoref` setting, and a TOP-DOWN one from Liu et al. (2021) using XLM-RoBERTa-base (Conneau et al., 2020). Following recommendations by Morey et al. (2017), we use the more stringent original Parseval metric on binary trees. Table 2 shows reproduced 5-run average scores on both `test` sets.[2] It is clear that scores of both architectures are neck and neck, which raises questions on whether, beyond numeric scores, they find similar or different data difficult.

corpora	GUM v9			RST-DT		
metrics	**S**	**N**	**R**	**S**	**N**	**R**
BOTTOM-UP Guz and Carenini (2020)	70.4	57.7	49.9	76.5	65.9	54.8
TOP-DOWN Liu et al. (2021)	71.9	58.9	51.7	76.5	65.8	54.8

Table 2: Parsing Performance on GUM v9 and RST-DT `test` with Gold EDU Segmentation (5 run average). **S=S**pan (whether subtrees span the right EDUs); **N=N**uclearity (whether edges point the right way); **R=R**elation (whether labels are correct).

4 Analysis

Strictly speaking, the types of errors that top-down and bottom-up parsers make are not identical: while bottom-up, and in particular shift-reduce parsers see analyzed preceding discourse units, grouped in a stack, and remaining discourse units in an upcoming queue, top-down parsers analyze a domain of ungrouped tokens to be split and determine the optimal split point and label for each decision. Because we want to analyze what promotes errors both across and for each architecture, we adopt an output-centric view, analyzing EDUs at which parsers do and do not make errors based on their properties in the completed gold vs. predicted tree. At the same time, we do not want our results to be swayed by coincidental variations in neural models, which can have far-reaching consequences due to cascading errors. Instead, we train five models in each architecture, i.e. five training runs, each with a different random seed producing a different initialization for the parser: if only one model fails to predict a relation, it may not be very hard, while 4–5 errors would be indicative of genuinely hard relations.

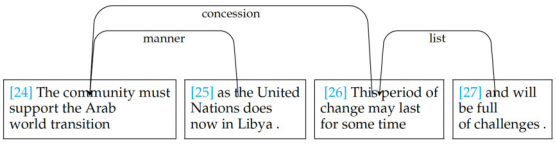


Figure 2: An Example of an RST Constituent Fragment converted into the Discourse Dependency Structure following Li et al. (2014).

Additionally, since models ultimately confront different inputs as a result of such cascaded decisions, we will use a dependency representation of both the gold and predicted RST trees, following the dependency conversion as defined by Li et al. (2014),[3] as exemplified in Figure 2. Although RST uses constituent discourse trees, focusing on each EDU and its dependencies will make it possible to make meaningful comparisons across models, and to intuitively understand how challenging EDUs are at any point in each document, regardless of whether or not they head large constituent structures. In Section 4.2 we will also incorporate the spanned domain of each head EDU's constituent block as an additional feature to assess the role of block size in predicting errors.

4.1 Explicit vs. Implict Relations

Table 3 shows the distribution of explicit or unmarked relations across the genres in the `dev+test` sets of GUM v9 and in comparison to RST-DT's `test` set, for each relation class and overall. The results for RST-DT are consistent with previous work, with 17.0% of test data relations being marked, similarly to the 18.2% identified by Das and Taboada (2017) for the entire corpus (but not anchored to specific tokens). An exami-

[2] Validation performance of each parser on both corpora is provided in Appendix A.

[3] The conversion code is available at `https://github.com/amir-zeldes/rst2dep`.

	# of explicit	explicit prop.	# of implicit	implicit prop.	# of distractor	distractor prop.
RST-DT	398	17.0%	1948	83.0%	81	3.5%
GUM v9	1198	21.7%	4332	78.3%	174	3.1%
academic	73	16.1%	380	83.9%	13	2.9%
bio	66	18.4%	292	81.6%	11	3.1%
conversation	100	12.9%	674	87.1%	23	3.0%
fiction	116	23.7%	374	76.3%	15	3.1%
interview	80	20.2%	317	79.8%	8	2.0%
news	73	18.1%	331	81.9%	7	1.7%
reddit	147	28.3%	373	71.7%	20	3.8%
speech	84	19.1%	356	80.9%	9	2.0%
textbook	95	21.3%	352	78.7%	9	2.0%
vlog	180	35.8%	323	64.2%	38	7.6%
voyage	69	22.4%	239	77.6%	9	2.9%
whow	115	26.4%	321	73.6%	12	2.8%
mean	99.8	21.9%	361	78.1%	14.5	3.1%

Table 3: Distribution of Explicit and Implicit Relations as well as EDUs with Distracting DMs in RST-DT `test` and `dev+test` of GUM v9.

relation class	# of explicit	explicit prop.	# of implicit	implicit prop.
ROOT	0	0.0%	48	100.0%
ADVERSATIVE	222	55.5%	178	44.5%
ATTRIBUTION	0	0.0%	292	100.0%
CAUSAL	131	53.5%	114	46.5%
CONTEXT	143	31.8%	306	68.2%
CONTINGENCY	99	91.7%	9	8.3%
ELABORATION	64	5.8%	1049	94.2%
EVALUATION	4	1.7%	231	98.3%
EXPLANATION	44	12.5%	308	87.5%
JOINT	409	37.2%	689	62.8%
MODE	52	45.2%	63	54.8%
ORGANIZATION	0	0.0%	331	**100.0%**
PURPOSE	21	10.7%	176	89.3%
RESTATEMENT	6	3.8%	150	96.2%
SAME-UNIT	1	0.3%	289	99.7%
TOPIC	2	2.0%	99	98.0%

Table 4: Distribution of Explicit and Implicit Relations across Relation Classes in `dev+test` of GUM v9.

nation of distributions by genre in GUM reveals some differences, highlighted in Table 3, with ***vlog*** exhibiting the most explicit relations, and ***conversation*** the fewest, raising the possibility that it may be more challenging for parsers. And in fact, Liu and Zeldes (2023) pointed to *conversation* as the worst-performing genre at all metric levels using an older version of the corpus (v8), which had less *conversation* data compared to GUM v9.

Looking at the presence of 'distractor' connectives, which are not associated with one of the gold relations in the tree, we see that ***vlog*** is the most prone to such cases, again raising the question of whether these may pose a problem for parsers, which may identify **a possibly correct relation that is not prioritized by the gold tree**. This situation appears to be infrequent in the WSJ data from RST-DT, which has only 81 such cases (3.5%). Taking a closer look at the types of distractors across genres in GUM, we see that the most frequent types are 'and', 'but', and 'so', which are highly ambiguous and common in conversational data such as *vlog* and *conversation*.

Regarding the most and least explicitly signaled relation classes in GUM v9, Table 4 reveals that CONTINGENCY is the most explicitly marked class due to the use of the DM 'if', and that the least explicitly signaled classes are ATTRIBUTION and ORGANIZATION. The former is almost always signaled by speech verbs (a verb such as 'say' or 'argue') and the latter mostly by document layout and graphical features in written texts, or by back-channeling in conversation data. It is also worth noting that instances of EVALUATION, RESTATEMENT, and TOPIC (used predominantly for question-answer pairs) are mostly ***not*** signaled by a discourse marker.

With these descriptive statistics in hand, we can examine each parser's performance on explicit/implicit relations, as well as on EDUs with a distracting DM in either the source or target of the relation (we must consider both ends, since many DMs can mark either a source or target such as 'but' and 'so'). Figure 3 shows the density of relations incurring between 0 and 5 attachment errors (disregarding labels) in each architecture for GUM, broken down by whether a DM marks the relation (top) and whether a distracting DM is present (bottom). The figure reveals several important facts: firstly, DMs are unsurprisingly associated with fewer errors (t=−7.29, D=0.23, p<0.0001), with lack of connectives affecting top-down models slightly more severely (χ^2=3.95, ϕ=0.14, p<0.05). Secondly, lack of distractors is associated with having fewer errors (t=5.0718, D=0.37, p<0.0001), and this is more pronounced for the bottom-up architecture, but the difference between architectures is not significant here.[4] Figure 4 shows the same kind of density plots for RST-DT.

Although it seems obvious that explicitness will facilitate parsing and that distractors should be harmful, it is an open question whether such markers will remain important once we know about other factors known to cause problems, such as OOV items, EDU text length, and intra-sentential status. To compare these, we construct several regression models predicting the number of errors. Because the distribution of error numbers is U-

[4]That said, we recognize that there are also more differences between these parsers than just the top-down/bottom-up distinction, so it is possible that with a broader sample of parsers, more differences would emerge.

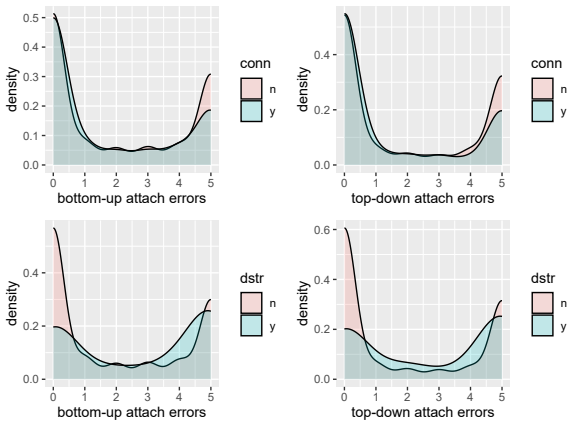


Figure 3: Attachment Error Count Density with and without DMs or Distractors for Each Architecture in `dev+test` of GUM v9.

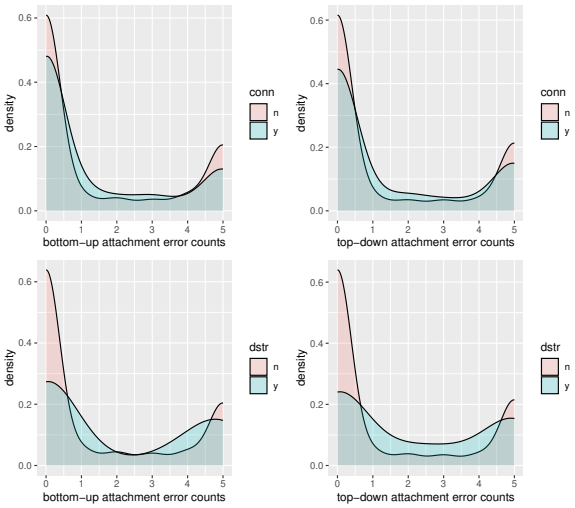


Figure 4: Attachment Error Count Density with and without DMs or Distractors for Each Architecture in `test` of RST-DT.

shaped (many cases with zero or five errors, few in the middle), as shown in Figures 3–4, we cannot use traditional gaussian models, which assume a roughly normal distribution of the data. Instead, we use mixed effects Beta regression, which is suited to U-shaped data, with a random effect for document identity, and re-scale the number of **attachment or relation errors** to the range 0–1, where 1 means the max 5 model errors. Table 5 shows significance for each predictor in each model.[5]

Looking first at GUM on the left, Table 5 shows that, when given only DMs and distractors, both features are significant in predicting errors above a per-document random effect baseline, for both architectures. In other words, predicting implicit relations is unsurprisingly harder in RST, just as it is for PDTB-style shallow discourse parsing, and distractors make things even harder.

However, adding the subordination feature (the second and third pairs of models from the left for GUM v9), which indicates whether an EDU is in a subordinate clause (and therefore likely to have an intra-sentential relation), removes the significance of the presence of a DM (but not of distractors). This suggests DMs are less important in predicting errors (or lack thereof) than intra-sentential status. Adding some more predictors, a fuller model with EDU length, OOV rate (the percentage of lexical items not seen during training per EDU), and genre does not remove the significance of subordination status, and shows that OOV rate is not a significant predictor in this setting. The more complex models with 6 features also restore some significance for DMs, albeit to a lesser degree than other predictors.

Moving to RST-DT, we see a similar pattern, except for a surprising difference between architectures: in the mixed effects model, presence of a DM is ***not*** a significant predictor for the bottom-up architecture, while it is significant for top-down. This pattern is repeated across all sets of features on the right side of Table 5. For RST-DT, since we do not have gold syntactic dependency trees, we use gold intra-sentential relation status to represent the `subord` feature. This feature remains highly significant in all models across architectures. Finally, adding all the features to the right-most models (excluding `genre`, since RST-DT is all newswire), OOV rate again fails to reach significance, while all other features are significant, except for DMs for the bottom-up architecture models.

These numbers suggest several things: first and most important, while DMs may be somewhat important, some representation of intra-sentential status is the more robust predictor of parsing errors. This effect persists even if we know about other plausible features, such as EDU length and OOV rate. This observation fits with the line of work mentioned above on multi-stage models for RST parsing, which attempt to learn separate models for intra-sentential and inter-sentential or inter-paragraph models (e.g. Kobayashi et al. 2020). Although joint models can perform well on all levels regardless, we can confirm that there are substantial differences between these types.

In terms of architecture differences, results for

[5]Significance for `genre`, a multi-nominal feature, is computed via a likelihood ratio test comparing the model with and without this predictor.

corpus	GUM v9						RST-DT					
architecture	bot-up	top-down	bot-up	top-down	bot-up	top-down	bot-up	top-down	bot-up	top-down	bot-up	top-down
dm	<.001***	<.001***	0.059	0.074	0.003**	0.005**	0.988	0.002**	0.244	<.001***	0.445	<.001***
distractor	<.001***	<.001***	<.001***	<.001***	<.001***	<.001***	<.001***	<.001***	<.001***	<.001***	<.001***	<.001***
subord			<.001***	<.001***	<.001***	<.001***			<.001***	<.001***	<.001***	<.001***
length					<.001***	<.001***					<.001***	<.001***
oov					0.115	0.262					0.944	0.563
genre					<.001***	<.001***						

Table 5: Results of the Regression Models for GUM v9 and RST-DT from both Architectures.

RST-DT suggest more sensitivity to DMs for top-down models, but this result is not reproduced in GUM. Finally, all models are sensitive to distractors, which raises questions about the nature of this sensitivity—what kinds of errors are parsers making, and more specifically are they predicting relations corresponding to distractor DMs? We address these questions in the next sections.

4.2 Predicting Parsing Errors

The results in the previous section quantify the importance of different characteristics of discourse relations in promoting errors, and the relative difficulty of implicit relations in SOTA English RST parsing.

However, the linear model comparing the significance of explicit DMs, distractors, and features such as EDU length or OOV rate is rather naive and leaves out a variety of potentially relevant properties of subtrees, such as total number of attached discourse units (which could contribute to ambiguity), or the gold relation to be predicted—some relations are easier to recognize or are less ambiguous, and some relations have high prior likelihood, making guessing them a safe bet. Although these properties may not be useful for realistic prediction of errors when we do not have a gold parse, they can be of interest for understanding tree properties which are difficult for parsers to get right.

To make matters even more complex, the factors mentioned above interact in subtle ways with each other and with explicit marking status. For example, CONTINGENCY relations are easy to recognize thanks to the reliable DM 'if' as in (1), but this is not always the case, as in (2) which uses subject-verb inversion to mark a conditional. Some relations are almost never marked by DMs, but may still be easy, such as ATTRIBUTION, which can be identified via speech verbs, as in (3).

(1) [Um **if** you don't want to do a tour of Pittock Mansion,] $\xrightarrow{\text{gold:CONTINGENCY}}$ [I'd still recommend like taking the trail up there]$_{\text{GUM\_vlog\_portland}}$

(2) ["**Had it happened** an hour later] $\xrightarrow{\text{gold:CONTINGENCY}}$ [It would have been much worse]$_{\text{GUM\_news\_crane}}$

(3) [Any judge in this country would **agree**] $\xrightarrow{\text{gold:ATTRIBUTION}}$ [that opening and closing statements along are not a trial.]$_{\text{GUM\_speech\_impeachment}}$

This complexity means that a realistic model of difficult parsing environments may need to consider more variables, and the interactions mean that a simple linear model cannot capture the rich patterns in the data. In this section, we therefore use XGBoost (Chen and Guestrin, 2016), a highly accurate ensemble gradient boosting framework which is able to harness arbitrary interactions between features and is highly regularized to prevent overfitting, meaning it can be expected to find a near-optimal mapping of our variables to parser error occurrences. For this experiment, we will attempt to predict 'hard' EDUs, which we define as EDUs which most models predict incorrectly.

However, it is not immediately clear what kinds of features we should allow the model to use: on the one hand, we would like to know what constellations in gold RST trees are difficult, including the gold relation label or the relative importance of being a leaf node vs. a hub with many dependents, as well as the contributions of DMs and distractors. On the other hand, in a realistic scenario we would not be able to know whether a DM is a distractor without knowing the gold relation, and we would not know how many dependents a node really has.

We thus construct two models: the **REALISTIC** model only has access to features that can reasonably be predicted without the gold parse, including EDU length in tokens, presence of DMs (whether helpful or distracting), the incoming syntactic dependency relation (which can be predicted by a syntax parser), the OOV rate, and genre. The **FULL** model, by contrast, has access to all gold features, including the gold relation class, intra-/inter-sentential status, DM vs. distractor presence etc. The first model is more relevant for realistic scenarios in which we want to diagnose where parser er-

rors are more likely (or how many we might incur), while the second is more helpful for understanding what is hard in an RST graph given the gold graph itself. Note that neither model is fed features from any outputs of the parser models above: the parsers are only used to compute the number of errors at each point, which the XGBoost model attempts to predict. Figure 5 gives an analysis of feature importances using classification gain[6] for both the REALISTIC and the FULL models, which score 67.3% and 76.3% respectively over a majority baseline score of 58.3%, which predicts that RST parsers will never be wrong, for the bottom-up architecture. For top-down, the scores of the two models are 65.3% and 76.6% respectively.

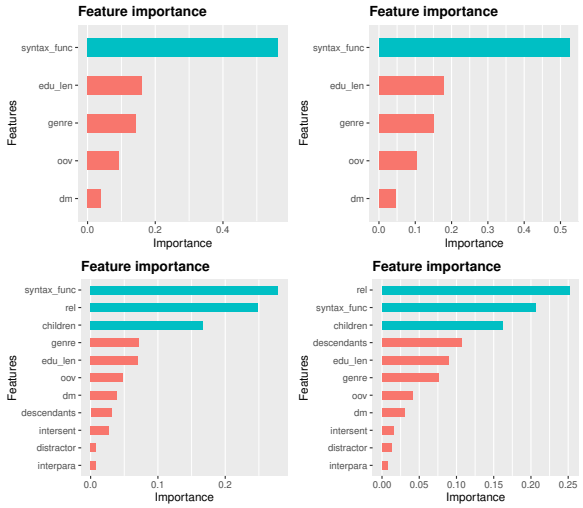


Figure 5: Feature Importances for the **REALISTIC** (top) and **FULL** (bottom) XGBoost Models for GUM from both BOTTOM-UP (left) and TOP-DOWN (right) Architectures. Very important features are highlighted in **teal**.

The XGBoost library's plots automatically highlight the most important features for both parser architectures, which for the **REALISTIC** model is only **the syntactic function of the EDU**. This likely indicates the overwhelming importance of knowing whether an EDU has a typical intra-sentential role, such as a relative or adverbial clause, which is likely to be predicted correctly. The next features begin with length (short EDUs are likely to have similar ones attested in training data compared to long ones), then genre (since some genres are harder), and only then the typical NLP difficulty predictor, the OOV rate (which is slightly less useful when EDU length is also known, since the two correlate). The last feature, presence of DMs, is still useful but less so, especially since it folds in occurrences of helpful and distracting DMs. There are no substantial differences between top-down and bottom-up here for GUM v9.

Turning to the **FULL** model, we see that syntactic function is still very important: it beats gold label for bottom-up models and follows it for top-down. Some relations are easier than others, or different subsequent conditions apply to them, and this matters about as much as the syntactic attachment type. Number of children (a measure of tree centrality vs. leaf status) is third, only then followed by length and genre, which are still quite helpful. Number of descendants (which is correlated with children) follows for top-down, but is far lower for bottom-up parsers. We then see OOV rate outranking DMs, which outrank less important features, such as the no longer crucial inter-sentential/inter-paragraph status, which are also highly correlated with some of the features above (syntax for the former, number of children for the latter, since many children are typical of paragraph head units). Finally distractors are second to last, far below DMs, also because they are rare.

These models indicate that predicting errors without knowing the gold tree is challenging, but a gain of 7–9% over baseline is still possible, mainly by looking at syntactic structure, which indicates inter-/intra-sentential status—a predictor much more valuable than DM marking. By contrast, when looking at gold trees, hard parts can most easily be associated with hard relations and syntactic environments, but combining all of the available features leads to an impressive ability to predict where parser models will likely go wrong, with ∼18% gain over baseline.

4.3 The Nature and Meaning of Distractors

Although the previous results suggest distractors play a minor role, their independent correlation with errors and the fact that DMs are generally relevant to discourse relations, raise questions regarding their very existence: why do they appear and how exactly do they affect parsers?

To begin with the second question, we examined the 174 distractors in GUM. For most bottom-up models, 108/174 (62.1%) were still erroneous, and 107/174 (62.1%) instances from the top-down models were erroneous. We then decided to manually

[6]Because XGBoost relies on gradient boosting with tree-based learners, the effect of variable interactions is computed within the classification gain metric, which is often used to estimate feature importance (see e.g. Shang et al. 2019).

label whether the majority model-predicted label was consistent with the distractor: if the gold relation is ELABORATION, the distractor is *but*, and the prediction is ADVERSATIVE, then prediction is consistent with the distractor, but if the prediction is CONTINGENCY, then it is not. We use PDTB's mapping of connectives to classes to match DMs to relations.

For 74/108 cases (68.5%) from the bottom-up models and 68/107 cases (63.6%) from the top-down models, the majority label was consistent with the distractor—in other words, the parser may be predicting based on a DM which would normally signal a competing relation. This brings us to the second question: if the relations signaled by distractors are incorrect, why are the distractors present? As an example, we consider two such cases from GUM, shown in (4)–(5).

(4) [**if** Steven didn't see it as weird] $\xrightarrow[\text{pred:CONTINGENCY}]{\text{gold:EXPLANATION}}$ [why should it bother us?]$_{\text{GUM\_fiction\_teeth}}$

(5) [so the reason seems to be that there are things out there that put even these kaiju to shame] $\xleftarrow[\text{pred:ADVERSATIVE}]{\text{gold:EVALUATION}}$ [**But** even this presents a problem]$_{\text{GUM\_reddit\_monsters}}$

In (4), the gold tree has the 'if'-clause as a justification for why it 'shouldn't bother us', which makes sense pragmatically; but formally, the clause seems like a legitimate conditional marked by *if*, and parsers predict CONTINGENCY. In (5), the annotation focuses on the evaluative meaning of the words 'a problem', while parsers, probably provoked by *But*, predict ADVERSATIVE.

We thus suspect that multiple, concurrent relations may actually hold in data where distractors appear, which is a standard possibility in frameworks like PDTB, where relations are identified based on the presence of DMs. If this applies in RST as well, then in a sense, such parser errors are not really errors at all. Because RST enforces a strict tree constraint, the only way to find out would be to look at alternative RST trees.

In order to do just this, we utilize RST-DT's official double-annotated subset, which has trees from a second annotator for 53 documents. This subset overlaps only 5 documents in the RST-DT `test` set, which contain only 12 distractors, meaning that the scope of this last analysis is limited; however, in examining these 12 distractors, we discovered that 75% (9/12) actually corresponded to relations **selected as the primary RST relations by the second annotator** in the double annotated data. In other words, the double annotated data confirms that, at least in the case of the RST-DT `test` set, a large majority of distractors do in fact correspond to multiple concurrent relations, which were identified by an experienced RST annotator.

5 Conclusion

This study has several important implications. Firstly and unsurprisingly, the explicit/implicit distinction from shallow discourse parsing is mirrored in RST parsing difficulty, and the dataset released in this paper can help study it further. However, explicit marking is clearly less consequential than intra-sentential status, with which explicitness it correlated. Secondly, OOV rate plays a less important role than we initially suspected, while genre effects remain robust, suggesting that diverse genres may matter more than subject matter. Our results also indicate that current architectures do not differ substantially in what they get right or wrong, and with scores being so similar, differences reduce to computational efficiency and personal preference.

Finally, the study of distractors suggest that RST's tree constraint may mix some cases of multiple concurrent relations with parsing errors, when parsers are actually identifying viable relations. This suggests that we may want to consider ways of allowing and adding concurrent relations to RST parses.

We also note that although the error prediction models evaluated in Section 4.2 were primarily developed in order to gain a greater understanding of the issues in discourse parsing, they could have some practical applications.[7] Predicting regions of low certainty in discourse parses can: 1) assist by highlighting low confidence regions in user-facing downstream applications; 2) flag potential problems during annotation of resources, especially when relying on NLP (Gessler et al., 2020) or less trained annotators/crowd workers (Scholman et al., 2022; Pyatkin et al., 2023); and 3) help guide additional resource acquisition, either automatically using active learning (to prioritize documents predicted to have parsing problems for manual annotation, cf. Gessler et al. 2022) or using qualitative evaluation in deciding what data to collect in terms of the relative importance of genres, presence of OOV items, etc.

[7] We thank an anonymous reviewer for noting this.

References

Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. GENTLE: A genre-diverse multilayer challenge set for English NLP and linguistic evaluation. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 166–178, Toronto, Canada. Association for Computational Linguistics.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press, Cambridge.

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where are we in discourse relation recognition? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325, Singapore and Online. Association for Computational Linguistics.

Katherine Atwell, Anthony Sicilia, Seong Jae Hwang, and Malihe Alikhani. 2022. The change that matters in discourse parsing: Estimating the impact of domain shift on parser error. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 824–845, Dublin, Ireland. Association for Computational Linguistics.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current and New Directions in Discourse and Dialogue*, Text, Speech and Language Technology 22, pages 85–112. Kluwer, Dordrecht.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. Association for Computational Linguistics.

Debopam Das and Maite Taboada. 2014. RST Signalling Corpus Annotation Manual. Technical report, Simon Fraser University.

Debopam Das and Maite Taboada. 2017. Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes*, 55(8):743–770.

Debopam Das, Maite Taboada, and Paul McFetridge. 2019. RST Signalling Corpus. LDC2015T10.

Vera Demberg, Fatemeh Torabi Asr, and Merel Scholman. 2019. How compatible are our discourse annotation frameworks? insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10(1):87—135.

Luke Gessler, Lauren Levine, and Amir Zeldes. 2022. Midas loop: A prioritized human-in-the-loop annotation for large scale multilayer data. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 103–110, Marseille, France. European Language Resources Association.

Luke Gessler, Yang Liu, and Amir Zeldes. 2019. A discourse signal annotation system for RST trees. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 56–61, Minneapolis, MN. Association for Computational Linguistics.

Luke Gessler, Siyao Peng, Yang Liu, Yilun Zhu, Shabnam Behzad, and Amir Zeldes. 2020. AMALGUM – a free, balanced, multilayer English web corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5267–5275, Marseille, France. European Language Resources Association.

Grigorii Guz and Giuseppe Carenini. 2020. Coreference for discourse parsing: A neural approach. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167, Online. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496, Sofia, Bulgaria. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

René Knaebel. 2021. discopy: A neural system for shallow discourse parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Top-down RST parsing utilizing granularity levels in documents. In *AAAI Conference on Artificial Intelligence*.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland. Association for Computational Linguistics.

Jessica Lin and Amir Zeldes. 2021. WikiGUM: Exhaustive entity linking for wikification in 12 genres. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 170–175, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, and M Saiful Bari. 2019. A unified linear-time framework for sentence-level discourse parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4200, Florence, Italy. Association for Computational Linguistics.

Yang Liu. 2019. Beyond the Wall Street Journal: Anchoring and comparing discourse signals across genres. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 72–81, Minneapolis, MN. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2017. Learning contextually informed representations for linear-time discourse parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1289–1298, Copenhagen, Denmark. Association for Computational Linguistics.

Yang Liu and Amir Zeldes. 2019. Discourse relations and signaling information: Anchoring discourse signals in RST-DT. *Proceedings of the Society for Computation in Linguistics*, 2(35):314–317.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? A thorough investigation of the impact of data diversity. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3104–3122, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Special Issue on Using Large Corpora, Computational Linguistics*, 19(2):313–330.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324, Copenhagen, Denmark. Association for Computational Linguistics.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India. The COLING 2012 Organizing Committee.

Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2021. RST parsing from scratch. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1613–1625, Online. Association for Computational Linguistics.

Noriki Nishida and Hideki Nakayama. 2020. Unsupervised discourse constituency parsing using Viterbi EM. *Transactions of the Association for Computational Linguistics*, 8:215–230.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse TreeBank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950.

Valentina Pyatkin, Frances Yung, Merel C. J. Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. 2023. Design choices for crowdsourcing implicit discourse relations: Revealing the biases introduced by task design.

Merel Scholman, Valentina Pyatkin, Frances Yung, Ido Dagan, Reut Tsarfaty, and Vera Demberg. 2022. Design choices in crowdsourcing discourse relation annotations: The effect of worker selection and training.

In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2148–2156, Marseille, France. European Language Resources Association.

Erbo Shang, Xiaohua Liu, Hailong Wang, Yangfeng Rong, and Yuerong Liu. 2019. Research on the application of artificial intelligence and distributed parallel computing in archives classification. In *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 1267–1271.

Caroline Sporleder and Alex Lascarides. 2004. Combining hierarchical clustering and machine learning to predict high-level discourse structure. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 43–49, Geneva, Switzerland. COLING.

Manfred Stede and Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of the Language Resources and Evaluation Conference (LREC '14)*, pages 925–929, Reykjavik.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse TreeBank 3.0 annotation manual. Technical report, University of Edinburgh, Interactions, LLC, University of Pennsylvania.

Nan Yu, Meishan Zhang, and Guohong Fu. 2018. Transition-based neural RST parsing with implicit syntax features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 559–570, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. RST discourse parsing with second-stage EDU-level pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280, Dublin, Ireland. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM Corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zheng Zhao and Bonnie Webber. 2021. Revisiting shallow discourse parsing in the PDTB-3: Handling intra-sentential implicits. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 107–121, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 461–467, Online. Association for Computational Linguistics.

A Validation Performance

Table 6 shows our reproduced 5-run average parsing performance on the `dev` partition of each corpus. GUM v9 has an established `dev` partition following the UD English GUM treebank. While RST-DT does not have an established `dev` partition, we followed previous work by taking 10% of training data stratified by the number of EDUs in each document (Guz and Carenini, 2020), which remained the same in the training for both parsers. The list of document names used as development data can be found in the repository of the paper for reproducibility purposes.

corpora	**GUM v9**			**RST-DT**		
metrics	**S**	**N**	**R**	**S**	**N**	**R**
BOTTOM-UP Guz and Carenini (2020)	67.9	64.8	46.8	76.0	64.9	55.2
TOP-DOWN Liu et al. (2021)	69.3	56.3	48.1	75.0	64.6	55.7

Table 6: Validation Performance on GUM v9 and RST-DT with Gold Segmentation (5 run average).

Grounded Complex Task Segmentation for Conversational Assistants

Rafal Ferreira, David Semedo, João Magalhães
NOVA LINCS, NOVA School of Science and Technology, Portugal
{rah.ferreira}@campus.fct.unl.pt, {df.semedo, jmag}@fct.unl.pt

Abstract

Following complex instructions in conversational assistants can be quite daunting due to the shorter attention and memory spans when compared to reading the same instructions. Hence, when conversational assistants walk users through the steps of complex tasks, there is a need to structure the task into manageable pieces of information of the right length and complexity. In this paper, we tackle the recipes domain and convert reading structured instructions into conversational structured ones. We annotated the structure of instructions according to a conversational scenario, which provided insights into what is expected in this setting. To computationally model the conversational step's characteristics, we tested various Transformer-based architectures, showing that a token-based approach delivers the best results. A further user study showed that users tend to favor steps of manageable complexity and length, and that the proposed methodology can improve the original web-based instructional text. Specifically, 86% of the evaluated tasks were improved from a conversational suitability point of view.[1]

1 Introduction

Voice-based assistants can guide users through everyday complex tasks, such as cooking, crafts, and home repairs. These conversational assistants need to understand the users' intention, find a specific recipe, and communicate it in a structured and well-paced manner. Supporting this type of task-guiding interaction is a recent topic (Gottardi et al., 2022; Choi et al., 2022; Strathearn and Gkatzia, 2022), where conversational assistants must work hand-in-hand with users in order to guide them throughout the task execution.

We argue that most instructional texts found online are structured in a non-optimal way for conversational assistants, due to the inherent differences between screen and voice-based interfaces. Recipes are a great example where the decomposition of the recipe's text into dialogue-suited steps is critical (example in Appendix A) – as Gottardi et al. (2022) observed, the user is dividing attention through various and possibly parallel actions. Hence, and following previous knowledge (Miller, 1956; Cowan, 2001), we aim for steps that are structured and presented to the user in ordered pieces of information, while dosing complexity, with the aim of achieving an efficient task completion.

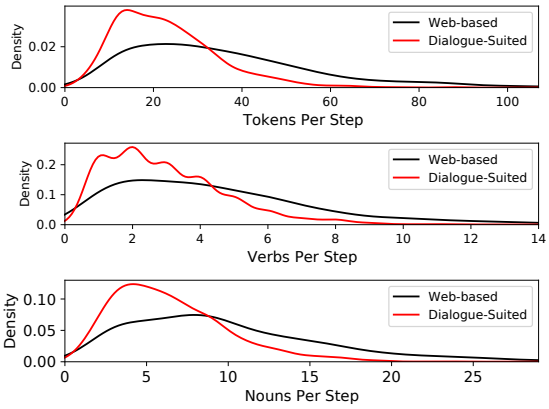


Figure 1: Differences between web-based and dialogue-suited recipes (i.e., ConvRecipes corpus) in terms of the density distribution of tokens, verbs, and nouns per step.

To tackle this new problem, we part ways with topic-based segmentation methods (Koshorek et al., 2018; Choi, 2000; Arnold et al., 2019) and propose a novel human-focused methodology to convert reading-structured instructions into conversational ones. Figure 1 offers a clear view of the differences between the original web-based recipes and their dialogue-suited counterparts. The distribution of linguistic characteristics such as length, verbs (which cover actions), and nouns (covering ingredients, tools, etc.), confirms that dialog-suited instructions should avoid overwhelming users' short-term memory (Miller, 1956; Cowan, 2001).

[1] https://github.com/rafaelhferreira/grounded_task_segmentation_cta

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 43–54
September 11–15, 2023. ©2023 Association for Computational Linguistics

Our proposed methodology starts with the curation of a corpus which we call *ConvRecipes*, where online recipes are segmented into recipes with steps more suited to a conversational agent. Moreover, we identify the key traits of a conversational step. An example of this can be seen in Figure 2. This example shows the need for models that can tackle our task, in specific, in "Step 2" of the task, where it is noticeable that the step would be very difficult to follow in a conversational assistant due to the long sentence and the inherent complexity of its actions.

To tackle this task computationally, we propose the Dialogue-Task Segmenter Transformer (*DTS-Transformer*), which follows state-of-the-art approaches in text-segmentation (Lukasik et al., 2020; Lo et al., 2021; Solbiati et al., 2021) and adopts a Transformer-based backbone (Vaswani et al., 2017). Distinct from previous work, we follow a token-level approach which by modeling steps' text at a finer granularity, is capable of better modeling the inherent structural characteristics of conversational tasks. Note that, we did not follow generative approaches and ground our task segmentation task on the recipes' original text. We do this to avoid the risk of introducing hallucinations or mistakes in step-by-step procedures (Choi et al., 2022).

Finally, we validated the proposed methodology with automatic experiments, and, more interestingly, with a user evaluation. We observed that the best *DTS* model, a *T5-3B* Encoder backbone, trained on the proposed ConvRecipes corpus, was able to improve the conversational structure of 86% of the evaluated tasks. This evidences both the conversational characteristics of the ConvRecipes corpus and the effectiveness of the model's approach to the grounded conversational task segmentation task.

Next, we will relate our contributions to previous corpus and methods. In Section 3, we carefully detail the proposed methodology. Experimental validation and user evaluations are presented in Section 4, and we conclude with the final takeaways and future work.

2 Related Work

Related Corpora. While conversational-suited task segmentation is a novel task, multiple datasets have been created to address article-based text segmentation, with the earliest ones being the Choi Dataset (Choi, 2000), where each document is represented by the concatenation of 10 random passages from a large corpus, and the RST-DT dataset (Carlson et al., 2001), which focuses on intra-sentente granularity on Wall Street Journal articles. Topic and document-section-oriented segmentation datasets such as Wiki-727 (Koshorek et al., 2018) and WikiSection (Arnold et al., 2019) are comprised of Wikipedia articles and focus on topic and section-based text segmentation. Closer to our domain, we highlight works with instructional text such as Task2Dial (Strathearn and Gkatzia, 2022) and the Wizard of Tasks (Choi et al., 2022), which rewrite the tasks' text into dialogue-suited steps. Our approach focuses on grounded structuring of task instructions for dialog while avoiding hallucination problems common in generative/re-writing approaches. We also take a step further by identifying the fundamental traits of conversational-suited tasks, in a principled manner.

Title: Baked Bananas Recipe

Web-based Recipe

Step 1: Preheat oven to 190 degrees C. Spray a baking dish with cooking spray.
Step 2: Arrange banana halves in the prepared baking dish. Drizzle maple syrup over bananas and top with cinnamon. Bake in the oven until heated through, 10-15 minutes.

Dialogue-suited Recipe

Step 1: Preheat oven to 190 degrees C.
Step 2: Spray a baking dish with cooking spray.
Step 3: Arrange banana halves in the prepared baking dish. Drizzle maple syrup over bananas and top with cinnamon.
Step 4: Bake in the oven until heated through 10-15 minutes.

Figure 2: Example of conversion from web/reading-based format to a dialogue-suited format. In blue and orange, we highlight the verbs and nouns, respectively.

Methods and Models. Initial works for text segmentation were mostly based on statistical and unsupervised approaches, such as TextTiling (Hearst, 1997) and C99 (Choi, 2000). After these, supervised neural methods, particularly with the use of RNNs were utilized. In (Badjatiya et al., 2018), a CNN is used to generate sentence embeddings in conjunction with an LSTM to keep sequential information. Li et al. (2018) also presents an RNN-based model with an additional pointing mechanism and in (Koshorek et al., 2018) it is used a hierarchical Bi-LSTM model.

Currently, the state-of-the-art is based on supervised Transformer-based approaches (Lukasik et al., 2020; Lo et al., 2021; Solbiati et al., 2021). In (Lukasik et al., 2020), cross-segment and hierarchical models are proposed, where predictions are

made based on consecutive segments or sentence-based representations of the segments. Lo et al. (2021) presented a hierarchical approach combining sentence and cross-segment embeddings. Xing et al. (2020) proposed a hierarchical BiLSTM to complement BERT's (Devlin et al., 2019) sentence representations, aided by a coherence-related auxiliary task. Some approaches such as (Zhang et al., 2021), tackle task structuring as a generation task, where an end-to-end pipeline is proposed to generate day-to-day tasks. In a dialogue setting, Solbiati et al. (2021) applied a BERT model for transcript-based meetings segmentation (Janin et al., 2003; McCowan et al., 2005) by calculating the similarity between segment embeddings given by a pre-trained model. Given the particular intricacies of conversational-suited task structuring, while we also adopt a Transformer backbone, we propose a task segmentation model that makes decisions at a token-level being able to consider the global task's structure.

3 Structuring Conversational Tasks

Our hypothesis is that the recipe instructions found online are not suited for conversational assistants, motivating both the task and the dataset collection efforts. To convert instructions from a reading structure into a conversationally structured format, we followed a human-focused methodology. First, we collected task instructions and ran a user study to curate them as conversational instructions. Second, we ask users to annotate the relevance of various conversational instructions traits. Third, we analyzed the linguistic characteristics of reading instructions compared to conversational task instructions. Finally, we modeled conversational-steps computationally with various Transformer-based (Vaswani et al., 2017) architectures.

3.1 A Conversational-Tasks Corpus

Currently, there are no explicit corpora for studying the grounded segmentation of a recipe into conversational-suited steps. The closest examples are either section-based document segmentation (Koshorek et al., 2018; Arnold et al., 2019) or rewriting/generative approaches (Choi et al., 2022; Strathearn and Gkatzia, 2022) which are prone to hallucinations. In this section, we introduce the methodology used to create the *ConvRecipes* corpus, consisting of recipes segmented into conversational-suited steps.

3.1.1 Tasks Collection and Annotation

To create the ConvRecipes corpus, we collected recipes from a popular recipes website , where each recipe is self-contained and composed of various steps in English with arbitrary lengths. We started by filtering out recipes with fewer than three steps due to having a structure that is too simple. After this, near-duplicate recipes were identified with SimHash (Charikar, 2002) and removed.

Conversation-Steps Annotation. Even though recipes are human-edited, we argue that they are written for a reading-based setting, making them ill-suited to be used in a conversational setting. Hence, to create grounded conversational instructions, we conducted a user study. In total, we had 8 annotators, 6 male and 2 female all Computer Science MSc. and or Ph.D. students. All annotators had experience with both conversational assistants and cooking applications, making them particularly suited for this annotation task.

The annotators were shown the original recipes and asked to propose (or not) changes to make the recipes dialog-suited, either by adding and/or removing steps. Figure 2 illustrates the annotation process process: given a recipe formatted for the Web, the goal is for the annotator to identify the structure that is better suited for a conversational setting. This approach makes the segmentation grounded on the original task, avoiding the introduction of mistakes prone to happen when using rewriting approaches.

3.1.2 On the Traits of a Conversational Step

After the annotation process described in the previous section, the annotators were asked to quantify, on a Likert scale of 1 to 5, the importance of various conversational traits. In particular, we considered: *Complexity*, *Clarity*, *Length and #Steps*, *Ability to Parallelize Tasks*, and *Naturalness*. For the exact description of these traits refer to Appendix B. This evaluation of the traits aims to further inform us what users value in this conversational task-guiding assistance setting (Gottardi et al., 2022).

Table 1 shows the results of the analysis of the traits. The results reveal that although all traits have some importance, users mostly focus on the complexity and length of the steps, which are generally connected with each other. This means that managing complexity and ensuring a balance in the information given to the user is paramount. On the other hand, the naturalness and the ability to per-

Conversational-Step Trait	Importance
(1) Complexity	4.5
(2) Step Length & #Steps	4.2
(3) Clarity	3.8
(4) Naturalness	3.6
(5) Ability to Parallelize Tasks	3.4

Table 1: Trait importance on a 1 to 5 scale. A higher value represents higher importance.

	Reading	Dialog
Avg. # Tokens	135	
Avg. # Sentences	9.3	
Avg. # Steps	3.80	5.85
Avg. # Tokens step	35.44	23.03
Avg. # Sents. step	2.44	1.59
Avg. # Verbs step	4.23	2.75
Avg. # Nouns step	9.92	6.44

Table 2: Comparison between the 300 original reading-based recipes and the manually annotated set.

form parallel tasks were considered less important traits, which seem to indicate that users are not so concerned with language naturalness given that the step should be short and not too complex.

3.2 ConvRecipes Corpus Analysis

After preparing and curating the task instructions, we analyzed and compared the original to the curated data in order to understand how the language differs from a web/reading setting to a conversational setting.

3.2.1 Reading-suited vs Dialog-suited

In total, 300 recipes were annotated, where 59 recipes were left without changes, and the remaining 241 (80.3%), had at least one new step added, with one, two, and three or more breaks added 75, 66, and 47 times, respectively. Only one recipe was annotated with fewer steps than the original. This result shows that the reading-based instructions are not optimal for a conversational setting, generally missing critical segmentations. Table 2 further evidences the difference between the original and the conversational-suited instructions, where it is clear that there is a preference for shorter segments with fewer actions. These results correlate with the importance of the conversational traits (Table 1), which showed that the complexity and number of steps are particularly important in this setting.

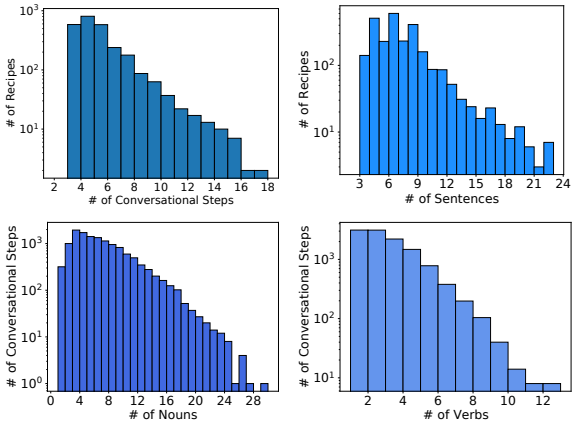


Figure 3: ConvRecipes statistics: conversational-steps per task (top-left), sentences per task (top-right), nouns per conversational-step (bottom-left), and verbs (bottom-right) per conversational-step.

Thus, ConvRecipes presents a step forward in discovering the optimal structure for instructional text in a conversational scenario.

3.2.2 Linguistic Style of Conversational-Steps

Figure 3 shows the corpus's distribution of conversational steps, sentences, nouns, and verbs. The figure indicates that there is a lot of variability that needs to be correctly addressed, due to each recipe having a particular structure.

In contrast to other corpora (Koshorek et al., 2018; Choi, 2000), ConvRecipes is written in an instructional/imperative format, using actionable verbs mostly related to the cooking domain such as "stir", "bake" and "mix". Analyzing how steps start and end, can also bring some insights into the segmentation behavior, so we examined the most common starting and ending n-grams of each step. The top-20 starting and ending tri-grams are available in Figure 4. This showed that many of the steps have temperature mentions, e.g. "preheat oven to", or time aware mentions *e.g.*, "for [N] minutes" ([N] is a placeholder replacing the number). These indicate a start/end of an action which in turn reflects a new step. It is important to note that the majority of both bi-grams (65%) and tri-grams (80%) are only used once, which shows the diversity of actions available, creating a more complex challenge for data-driven approaches.

3.2.3 Corpus Processing.

Annotating a large number of recipes is labor-intensive and expensive. Thus, we use the 300 manually annotated recipes as the test set and cre-

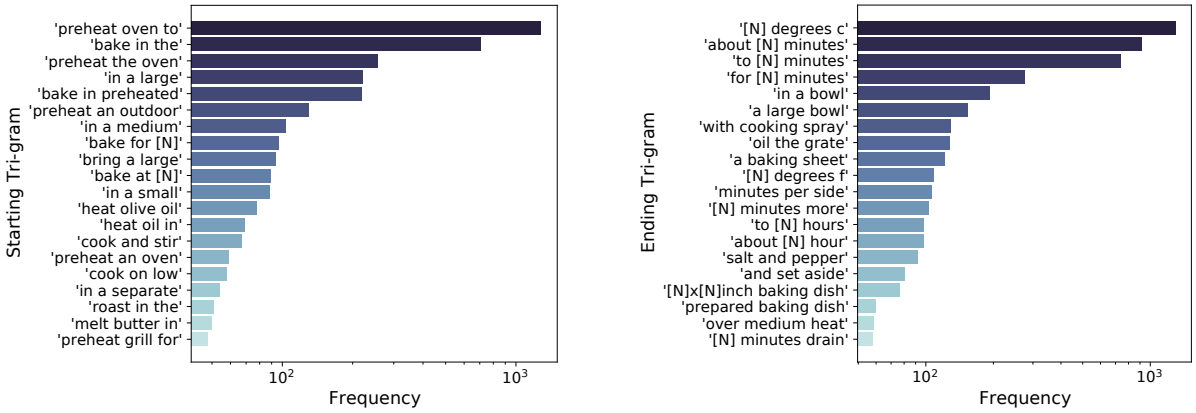


Figure 4: Distribution of the top-20 most frequent starting (left) and ending (right) tri-grams.

ate the training and validation splits automatically, by using the average number of sentences per step of the annotated set (1.59, Table 2) as a maximum threshold to choose recipes from a non-annotated set. These non-annotated recipes use the original step information as the ground-truth step labels, in a similar way as the section markers in the Wiki-727 dataset (Koshorek et al., 2018). This resulted in a dataset with 1930, and 424 recipes for training and validation, respectively. As mentioned before, the test set is composed of the 300 annotated recipes.

To conclude, the aim of this corpus is to create models that learn how to identify steps and segment a task into grounded dialogue-suited steps. Hence, we concatenate all the steps together, resulting in an unstructured text with no segment-identifying structure. Step annotations are then used as labels to train and evaluate the models.

3.3 Dialog-Task Structuring Transformer

To learn the structure of a conversational task, we processed the entire task's text as a whole. By explicitly receiving the entire input sequence, we aim to take into account the size and position of each segment token relative to all the other tokens. Consequently, with a single pass over the input, this approach is able to output all segment predictions, making it more efficient than sentence-based embedding models that output a prediction per sentence (Lukasik et al., 2020).

Given the characteristics of the Transformer model (Vaswani et al., 2017), we use it as the basis for our *Dialogue-Task Structuring* (DTS) model. In particular, we feed the model with the complete recipe, allowing it to create contextualized token representations of the entire recipe. This allows the model to consider all of the tokens in the recipe via the self-attention mechanism.

After the input has been processed by the Transformer, we apply a binary segment-break prediction head, i.e., a linear layer followed by a *softmax* to the embedding of each segment identifying token (emb_t), outputting the probability of a token (t) being a *segmentation token*:

$$P_{seg}(t_i) = softmax(FFNN(emb_{[t_i]})), \quad (1)$$

To identify these end-of-segment tokens, generally punctuation marks (*e.g.* ".", "!", ";"), we use Spacy (Honnibal and Johnson, 2015) to perform basic sentence segmentation over the recipe's text. Finally, we train the model using the cross-entropy loss between the model predictions, $\hat{y}$, and the binary segmentation labels, y, as the following:

$$L_{CE} = y \cdot \log \hat{y} + (1 - y) \cdot \log (1 - \hat{y}), \quad (2)$$

4 Experiments

In this section, we demonstrate how the proposed framework tackles the challenge of structuring task instructions in a conversational setting. Experimental validation was done with both automatic metrics and human evaluation.

4.1 Metrics

We use Precision, Recall, and F-score to measure the detection of the correct location of a conversation step following Li et al. (2018). Moreover, we followed (Koshorek et al., 2018; Arnold et al., 2019), and used the text segmentation metric P_k (Beeferman et al., 1999), which compares the predicted segmentation with the ground-truth labels, where a lower value represents a better model.

4.2 DTS Backbones and Implementation Details

As the basis for our models, we used pre-trained Transformer models. We tested with the encoder-only model BERT (Devlin et al., 2019), the encoder-decoder model T5 (Raffel et al., 2020) in both an encoder-decoder (Enc-Dec) setting and in an encoder-only (Enc-only), i.e. decoder is not used. When using an E-D model, the input sequence of the decoder is the same as the encoder as in (Lewis et al., 2020) for the extractive QA task (i.e. there is no actual decoding). To identify the candidate segments, we used Spacy (Honnibal and Johnson, 2015), to be more robust than a simple punctuation-based approach. We evaluated in the test set the model with the best performance in the validation set in terms of F-Score. Additional information about model training is provided in Appendix C.

4.3 Baselines

As baselines, we considered random and uniform approaches, a classic method (Hearst, 1997), and a strong baseline based on a cross-encoder (Lukasik et al., 2020):

Rand$_p$ and **Every**$_n$: unsupervised methods which use Spacy (Honnibal and Johnson, 2015) to identify sentences. p is the probability of breaking at each sentence, and n is the number of consecutive sentences to break.

TextTiling (Hearst, 1997): one of the earliest text segmentation methods based on lexical co-occurrence.

Cross-Segment (CrossSeg) (Lukasik et al., 2020): BERT-Base (Devlin et al., 2019) model with a classification head that predicts if a pair of input sentences should be segmented.

4.4 Results and Discussion

In Table 3, we present the results of the baselines, along with the results of the proposed DTS models.

Importance of Conversational-Aware Corpora. We trained the same DTS model with a BERT-Base backbone: one on all crawled raw recipes (20.000 recipes), identified as *(All*)*, and one on the ConvRecipes training set (*BERT-Base*). The results on Table 3, show that *BERT-Based (All*)* obtained the highest Precision (93.4), since its training samples have fewer breaks, the model makes less, but correct, break predictions. On the other hand, it achieved the lowest Recall of all supervised methods. More importantly, the results clearly show the importance of training models with suited data, yielding a $\boldsymbol{P}_k$ relative improvement of 15% (*Bert-Base*). This result indicates that the ConvRecipes dataset is constructed in a way that embeds the traits of conversational task instructions (Section 3.1.2).

General results. In Table 3, we observe that the baselines *Rand*$_p$, *Every*$_n$ and *TextTiling* do not generally break the steps at the correct locations as indicated by their low precision ($\leq 62\%$). However, since it implicitly enforces a step distribution that resembles the dataset, *Every*$_1$ achieves a fairly good $\boldsymbol{P}_k$, while also achieving a recall close to 100% due to breaking at every sentence (it is $\neq 100\%$ due to errors in Spacy's sentence identification algorithm). *TextTiling*, which decides the task structure through lexical overlap, performs poorly and does not appear to be a good option for this task. This is because recipe steps are not structured based on overlap, but rather in a sequence of sub-actions, which *TextTiling* overlooks.

The *CrossSeg* achieved a P_k of 19.5 which is already a significant improvement over the best unsupervised baseline which achieved 23.3. This translates into an F1 score improvement from 73.8% to 76.5%. Regarding the *DTS* models, the most solid fact that emerges from Table 3 is that, regardless of the backbone, our *DTS* approach consistently outperforms all the baselines.

In general, the results of the baselines illustrate the difficulty of the problem we are trying to solve. Moreover, there is a clear divide in terms of P_k between previous baselines and the proposed *DTS* framework, which is consistently below 20 P_k, highlighting the importance of capturing the relations between the task and the conversational steps.

Encoder vs Encoder-Decoder Backbones. Comparing the encoder-only model BERT (Devlin et al., 2019) with T5 (Raffel et al., 2020) (Enc-only) or the full encoder-decoder (Enc-Dec) model, in situations with a comparable number of parameters, we see that *T5* outperforms *BERT*. This might be explained by the different pre-training approaches used in T5 (Raffel et al., 2020), which are better suited for our task. Comparing the encoder-only (Enc-only) with the encoder-decoder (Enc-Dec) in the same models, we see an improvement in *T5-Large*, but a decrease in performance in

	Model	# Params	P_k↓	Precision↑	Recall↑	F1↑
Baselines	$Rand_{0.5}$	-	35.4 ± 0.3	59.9 ± 0.5	49.7 ± 0.8	51.7 ± 0.6
	$Rand_{0.75}$	-	28.3 ± 0.5	61.2 ± 0.4	75.0 ± 0.9	65.2 ± 0.6
	$Every_1$	-	23.3	60.9	**98.8**	73.8
	$Every_2$	-	37.9	59.6	37.9	44.9
	TextTiling	-	28.4	58.7	67.7	61.4
	CrossSeg	110 M	19.5 ± 0.4	77.5 ± 0.9	79.5 ± 1.6	76.5 ± 0.4
Dialogue Task Segmenter (DTS)	BERT-Base (All*)	110 M	22.5 ± 0.3	**93.4 ± 0.1**	58.7 ± 0.4	69.6 ± 0.4
	BERT-Base	110 M	19.1 ± 0.4	75.8 ± 0.7	83.6 ± 0.7	77.5 ± 0.4
	BERT-Large	340 M	18.4 ± 0.2	77.0 ± 1.7	83.6 ± 2.8	78.1 ± 0.5
	T5-Base (Enc-only)	110 M	17.7 ± 0.2	77.9 ± 0.7	84.2 ± 0.5	79.0 ± 0.1
	T5-Base (Enc-Dec)	220 M	18.1 ± 0.6	77.9 ± 0.3	82.9 ± 1.6	78.5 ± 0.8
	T5-Large (Enc-only)	335 M	18.1 ± 0.2	77.4 ± 0.4	84.1 ± 0.4	78.6 ± 0.3
	T5-Large (Enc-Dec)	770 M	17.7 ± 0.2	79.1 ± 0.8	81.9 ± 0.9	78.5 ± 0.2
	T5-3B (Enc-only)	1.5 B	**17.0 ± 0.4**	78.3 ± 1.0	85.9 ± 0.9	**80.0 ± 0.2**

Table 3: Results on the ConvRecipes's test set from an average of 3 runs per model. *All** indicates that the model was trained on the set of all recipes crawled, in their original form.

T5-Base. This result implies that the use of the decoder part of *T5* might not be necessarily needed for this particular task.

DTS Model Size Influence. Having established the performance range of *DTS*, we examined the relationship between model size and performance. Results show that increasing the model size can bring improvements, in particular, from *BERT-Base* to *BERT-Large*, however, in the case of *T5-Base* for *T5-Large*, we notice an improvement in the Enc-Dec model and a decrease in performance in the encoder-only (Enc-only) model. Nonetheless, the best results by a significant margin in P_k and F1 are obtained with the largest model *T5-3B* (Enc-only), showing that the use of larger models can bring an improvement as evidenced in (Raffel et al., 2020; Chowdhery et al., 2022).

Main Takeaways. Results indicate that the proposed models are capable of capturing the intrinsic relations of the steps and extract them correctly when trained on high-quality conversationally structured task instructions. We also observe that our *DTS*-Transformer approach gives the best results in this setting. A fundamental difference between *DTS* and other supervised approaches is that it tackles the conversational recipe structuring task at a token-level granularity. As a consequence, it abstracts less information than previous approaches (Lukasik et al., 2020; Lo et al., 2021), such that at each Transformer layer, intermediate token embeddings are contextualized on the full-task sequence. Despite working at a finer granularity (token-level), *DTS* is both faster to train and perform inferences. This makes it highly suited to be applied in a real setting, to structure tasks into conversational steps.

4.4.1 Conversational Tasks Statistics

The recipe task structuring results led us to further examine the resulting conversational steps statistics. These are shown in Table 4, where we contrast the # Steps and # Tokens statistics with the human-annotated set. Specifically, we observe that Exact Match segmentation is higher (17.0%) in the *DTS T5-3B (Enc-only)* model due to its greater ability to capture the segmentation patterns. It is also interesting to note that all methods have a tendency to overestimate the number of steps. Finally, for ΔSteps≤ 1 – the percentage of examples where the model predicts less than one step of difference with the test set – we see an equivalent performance within the supervised baselines.

Overall, by examining these task structuring statistics, we observe that although the average number of steps (# Steps column) is acceptable for most methods, when we look at the finer-grain statistics, we see that there is a non-trivial balance between step length, number of steps, and content of each step. Hence, it is not a sufficient condition to optimize a single statistic but rather a combination of these.

4.4.2 User Evaluation

To compare the model's performance to the original web-based instructions, we asked 6 annotators from the same pool of Section 3 to annotate which segmentation was the best considering a conversational setting.

In total, 50 recipes were randomly selected from

		# Steps	# Tokens	Exact Match	= # Steps	+ # Steps	− # Steps	ΔSteps≤ 1
	Human Annotation	5.86	19.21	-	-	-	-	-
Method	Every_1	9.29	12.11	5.00%	5.33%	94.67%	0.00%	24.00%
	Text Tiling	6.32	17.80	7.00%	24.00%	49.33%	26.67%	58.67%
	CrossSeg	6.08	18.53	13.33%	30.67%	36.22%	33.11%	68.11%
	DTS T5-3B (Enc-only)	6.48	17.37	17.00%	27.56%	46.44%	26.00%	68.44%

Table 4: Detailed conversational task structuring statistics for the ConvRecipes test set (human annotated). Exact Matches is the percentage of predictions exactly matching the ground-truth. (=, + and -) # Steps represent the percentage of predictions that have equal, more, or less steps than the ground-truth. ΔSteps≤ 1 indicates the percentage of times the difference between the # Steps predicted and the ground-truth is ≤ 1 step.

	Web-based	T5-3B (E-only)
Rating 1	18.0%	3.3%
Rating 2	36.0%	12.7%
Rating 3	18.7%	20.7%
Rating 4	20.0%	35.3%
Rating 5	7.3%	28.0%
Best	14.0%	86.0%
Conv. Suitability	2.63	3.72

Table 5: User study results comparing the original web-based segmentations with T5-3B (Enc-only) model predictions. (Conversation Suitability is given on a 1 to 5 scale.

the test set in their original web-based format (i.e., without human annotations). These recipes were then compared to the predictions of the best model *DTS T5-3B* (Enc-only). Examples can be seen in Appendix D. For each recipe, we collected 3 annotations, resulting in an inter-rater agreement of 73% w.r.t. binary preference. Additionally, the annotators were also asked to grade each segmentation (web and model) on a 1 to 5 Likert scale according to the suitability for a conversational agent.

Table 5 shows the results of the user evaluation. We observe a preference for the model's segmentation (86%) since it was trained on a conversational-based data distribution which more accurately reflects the user's preference in this setting. We also analyzed that the annotators had a preference for recipes with more segments 88% of the time. Notwithstanding, it is important to note that breaking too often may result in a sub-optimal experience and in incomplete steps, as shown by the $\textit{Every}_1$ baseline of Table 3.

Considering the 1 to 5 rating of suitability for a conversational agent, the model's prediction scores were much higher (3.72) than the original recipes (2.63). These ratings further reinforce our hypothesis that the original recipes are not dialogue suited, and that the model is able to greatly increase the suitability of a recipe to a conversational-friendly format.

To conclude, these results indicate that the model is able to capture segmentation patterns, showing an ability to improve the suitability of a recipe for a conversational assistant. This, in turn, brings advantages to the user experience by providing a grounded conversationally-suited segmentation.

5 Conclusions

In this paper, we proposed a methodology to tackle the problem of converting web/reading structured instructions into conversationally structured ones, using a task-grounded segmentation by considering the original task's steps. In summary, the key contributions are as follows:

ConvRecipes Corpus. This corpus enables a better understanding of the problem. Its analysis showed that instructional text as it is presented online is not optimal for a conversational setting.

Dialogue-Task Structurer (DTS). We proposed several methods that can effectively capture segmentation linguistic patterns. The best-performing method was a T5-3B (Enc-only) model, a token-level Transformer.

Real-World Improvement. The user evaluation showcased the model's ability to improve over the original segmentation (86%), which brings advantages in user experience in a conversational-assistant scenario.

For future work, we intend to assess how segmentation influences the user's perception of a recipe's quality and generalize our experiments to different domains such as DIY tasks and tutorials.

Acknowledgments

This work has been partially funded by the FCT project NOVA LINCS Ref. UIDP/04516/2020, by the Amazon Science - TaskBot Prize Challenge and the CMU|Portugal projects iFetch CMUP LISBOA-01-0247-FEDER-045920), and by the FCT Ph.D. scholarship grant UI/BD/151261/2021. Any opinions, findings, and conclusions in this paper are the authors' and do not necessarily reflect those of the sponsors.

References

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A neural model for coherent topic segmentation and classification. *Trans. Assoc. Comput. Linguistics*, 7:169–184.

Pinkesh Badjatiya, Litton J. Kurisinkel, Manish Gupta, and Vasudeva Varma. 2018. Attention-based neural text segmentation. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, volume 10772 of *Lecture Notes in Computer Science*, pages 180–193. Springer.

Doug Beeferman, Adam L. Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Mach. Learn.*, 34(1-3):177–210.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the SIGDIAL 2001 Workshop, The 2nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Saturday, September 1, 2001 to Sunday, September 2, 2001, Aalborg, Denmark*. The Association for Computer Linguistics.

Moses Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 380–388. ACM.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *6th Applied Natural Language Processing Conference, ANLP 2000, Seattle, Washington, USA, April 29 - May 4, 2000*, pages 26–33. ACL.

Jason Ingyu Choi, Saar Kuzi, Nikhita Vedula, Jie Zhao, Giuseppe Castellucci, Marcus Collins, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. 2022. Wizard of tasks: A novel conversational dataset for solving real-world tasks in conversational settings. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 3514–3529. International Committee on Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Anna Gottardi, Osman Ipek, Giuseppe Castellucci, Shui Hu, Lavina Vaz, Yao Lu, Anju Khatri, Anjali Chadha, Desheng Zhang, Sattvik Sahai, Prerna Dwivedi, Hangjie Shi, Lucy Hu, Andy Huang, Luke Dai, Bofei Yang, Varun Somani, Pankaj Rajan, Ron Rezac, Michael Johnston, Savanna Stiff, Leslie Ball, David Carmel, Yang Liu, Dilek Hakkani-Tur, Oleg Rokhlenko, Kate Bland, Eugene Agichtein, Reza Ghanadan, and Yoelle Maarek. 2022. Alexa, let's work together: Introducing the first alexa prize taskbot challenge on conversational task assistance. In *Alexa Prize TaskBot Challenge Proceedings*.

Marti A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguistics*, 23(1):33–64.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003*, pages 364–367. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 469–473. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jing Li, Aixin Sun, and Shafiq R. Joty. 2018. Segbot: A generic neural text segmentation model with pointer network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4166–4172. ijcai.org.

Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray L. Buntine. 2021. Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3334–3340. Association for Computational Linguistics.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4707–4716. Association for Computational Linguistics.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of the 5th international conference on methods and techniques in behavioral research*, volume 88, page 100. Citeseer.

George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Calì. 2021. Unsupervised topic segmentation of meetings with BERT embeddings. *CoRR*, abs/2106.12978.

Carl Strathearn and Dimitra Gkatzia. 2022. Task2dial: A novel task and dataset for commonsense-enhanced task-based dialogue grounded in documents. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, DialDoc@ACL 2022, Dublin, Ireland, May 26, 2022*, pages 187–196. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Linzi Xing, Brad Hackinen, Giuseppe Carenini, and Francesco Trebbi. 2020. Improving context modeling in neural topic segmentation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 626–636. Association for Computational Linguistics.

Yi Zhang, Sujay Kumar Jauhar, Julia Kiseleva, Ryen White, and Dan Roth. 2021. Learning to decompose and organize complex tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2726–2735. Association for Computational Linguistics.

A Task Grounded Dialogue Segmentation Example in Conversation

In Table 6, we show an example dialog between a user and a conversational assistant. It is clear that segmenting a task is critical to managing the user's cognitive load, and allows for a better understanding of the recipe's multiple steps.

Table 6: Example of a dialogue-suited segmented recipe in the context of a conversational task assistant.

System	User
Hello! I am your cooking assistant! Try asking me things like: How to make lasagna? or How to bake a cake?	I want to make a Tiramisu.
Sure, I found this very tasty: "Classic Tiramisu". Do you want to try it?	Sure.
Step 1. Separate eggs. Whisk egg yolks with confectioner's sugar until pale and fluffy.	How many eggs should I use?
This recipe uses 2 eggs.	Next.
Step 2. Sure, let's keep going. Add Amaretto and mascarpone. Whisk until smooth.	Next Step.
Step 3. Beat egg whites with a pinch of salt until stiff peaks form. Slowly whisk in the remaining confectioner's sugar.	Go ahead.
Step 4. We are halfway through the task. Now, carefully fold beaten egg whites into mascarpone cream.	Next Step.
Step 5. Combine espresso and remaining Amaretto in a shallow dish. Dip ladyfingers in espresso mixture and then layer soaked ladyfingers in bottom of the serving dish.	What are ladyfingers?
Glad you ask! Ladyfingers are low-density, dry, egg-based, sweet sponge cake biscuits roughly shaped like large fingers.	Next Step.
Step 6. Cover with a layer of mascarpone cream and top with a fine layer of grated chocolate.	I'm done.
Step 7. Repeat layering process until all ingredients are used up. Finish up with a layer of mascarpone cream and dust with unsweetened cocoa powder.	Next.
Step 8. This is the last step of this recipe.Refrigerate for at least 3 hours before serving.	Next.
Excellent work, we just finished Classic tiramisu together! Hope you had fun with me today!	Goodbye!

B Conversational Steps Traits Description

The following traits where classified in a 1 to 5 Likert scale considering their importance in a task-guiding setting:

- **Step Complexity** - Was the complexity of the step an important factor to segment?
- **Step Clarity** - Was the clarity of a step important, i.e. information was clear, organized and well delimited in each step?
- **Step Length and Number of Steps** - Did the length of the steps and the total number of steps influence your decision?
- **Ability to Parallelize Tasks** - Were you concerned that the segmentation should allow the user to parallelize multiple tasks/steps?
- **Naturalness** - Was the naturalness of each step important for your decision? i.e., sounded close to a person explaining the task?

C Model Training and Hyperparameters

We trained every model for 20 epochs and evaluated in the test set the model with the best performance in the validation set in terms of F-Score. We used a batch size of 16, a learning rate of 5^{-5}, and the Adam optimizer (Kingma and Ba, 2015). All models were trained on a single NVIDIA A100 GPU, except for T5-3B (Enc-only), which used 4 NVIDIA A100 GPUs. We also highlight that these models are faster to train compared to the cross-segment model (Lukasik et al., 2020), due to each training sample only predicting one segmentation, instead of all segments as in DTS.

D Model Output Examples

Table 7 shows examples comparing the original web version of a recipe with the output of the T5-3B (Enc-only) DTS model. As we can observe, the model outputs a larger number of breaks complementing our findings in this conversational scenario. The model is also able to keep a notion of the subtask being performed, for instance in Example-3 Step-4, the model does not segment the sentences into multiple steps.

Example 1 - *Soy Garlic Steak* (Web)	**Model Output**
1. In a small bowl, mix vegetable oil, soy sauce, vinegar, ketchup, and crushed garlic. Place flank steak in a large resealable plastic bag. Pour the marinade over steak. Seal, and marinate in the refrigerator at least 3 hours. **2.** Preheat grill for high heat. **3.** Oil the grill grate. Place steaks on the grill, and discard marinade. Cook for 5 minutes per side, or to desired doneness.	**1.** In a small bowl, mix vegetable oil, soy sauce, vinegar, ketchup, and crushed garlic. **2.** Place flank steak in a large resealable plastic bag. Pour the marinade over steak. Seal, and marinate in the refrigerator at least 3 hours. **3.** Preheat grill for high heat. **4.** Oil the grill grate. Place steaks on the grill, and discard marinade. Cook for 5 minutes per side, or to desired doneness.
Example 2 - *Blueberry Yogurt Pops* (Web)	**Model Output**
1. Combine all ingredients in blender. Cover; blend on high speed 15 to 20 seconds or until smooth. **2.** Pour into 8 frozen pop molds (2.5 ounces to 3 ounces each). Cover, insert craft sticks, and freeze for 2 hours or until completely firm. **3.** To serve, dip outsides of molds into warm water to loosen.	**1.** Combine all ingredients in blender. Cover; blend on high speed 15 to 20 seconds or until smooth. **2.** Pour into 8 frozen pop molds (2.5 ounces to 3 ounces each). **3.** Cover, insert craft sticks, and freeze for 2 hours or until completely firm. **4.** To serve, dip outsides of molds into warm water to loosen.
Example 3 - *Quinoa Salad with Roasted Yams* (Web)	**Model Output**
1. Preheat oven to 350 degrees F (175 degrees C). Line a baking sheet with aluminum foil; add yams. **2.** Bake in the preheated oven until yams are tender and wrinkled at the edges, about 20 minutes. Cool to room temperature, about 15 minutes **3.** Bring water to a boil in a large saucepan. Add quinoa, stirring once; return to boil. Cook uncovered until water is absorbed, 10 to 12 minutes. Strain, shaking the sieve well to remove all moisture. Transfer to a mixing bowl. **4.** Stir cucumbers, yams, parsley, olive oil, onion, lemon juice, red wine vinegar, salt, and pepper into the quinoa. Garnish with endive spears.	**1.** Preheat oven to 350 degrees F (175 degrees C). **2.** Line a baking sheet with aluminum foil; add yams. **3.** Bake in the preheated oven until yams are tender and wrinkled at the edges, about 20 minutes. Cool to room temperature, about 15 minutes **4.** Bring water to a boil in a large saucepan. Add quinoa, stirring once; return to boil. Cook uncovered until water is absorbed, 10 to 12 minutes. **5.** Strain, shaking the sieve well to remove all moisture. Transfer to a mixing bowl. **6.** Stir cucumbers, yams, parsley, olive oil, onion, lemon juice, red wine vinegar, salt, and pepper into the quinoa. Garnish with endive spears.

Table 7: Examples comparing original web recipe and the T5-3B (Enc-only) DTS model outputs.

A Statistical Approach for Quantifying Group Difference in Topic Distributions Using Clinical Discourse Samples

Grace O. Lawley[1], **Peter A. Heeman**[1], **Jill K. Dolata**[2], **Eric Fombonne**[3], **Steven Bedrick**[4]
[1]*Computer Science and Engineering*
[2]*Department of Pediatrics*, [3]*Department of Psychiatry*
[4]*Department of Medical Informatics and Clinical Epidemiology*
Oregon Health & Science University, Portland, OR, USA

Abstract

Topic distribution matrices created by topic models are typically used for document classification or as features in a separate machine learning algorithm. Existing methods for evaluating these topic distributions include metrics such as coherence and perplexity; however, there is a lack of statistically grounded evaluation tools. We present a statistical method for investigating group difference in the document-topic distribution vectors created by latent Dirichlet allocation (LDA). After transforming the vectors using Aitchison geometry, we use multivariate analysis of variance (MANOVA) to compare sample means and calculate effect size using partial eta-squared. We report the results of validating this method on a subset of the *20Newsgroup* corpus. We also apply this method to a corpus of dialogues between Autistic and Typically Developing (TD) children and trained examiners. We found that the topic distributions of Autistic children differed from those of TD children when responding to questions about social difficulties. Furthermore, the examiners' topic distributions differed between the Autistic and TD groups when discussing emotions and social difficulties. These results support the use of topic modeling in studying clinically relevant features of social communication such as topic maintenance.

1 Introduction

Throughout the course of a dialogue many different topics are traversed with varying frequencies, and many analytical tasks depend on the ability to meaningfully quantify or otherwise characterize these patterns. For example, a system designed to automatically summarize meetings might need to detect when a new topic has been introduced; in a clinical context, we might wish to characterize the topics discussed during a patient visit to facilitate some sort of downstream analysis involving clustering or classification.

Topic modeling techniques such as latent Dirichlet allocation (LDA; Blei et al., 2003) allow us to capture and quantify the topic distributions across a collection of language samples. Typical methods for evaluating the resulting topic distributions use intrinsic metrics such as within-topic coherence; however, to our knowledge there remains a shortage of methods for statistically comparing the topic distributions produced by a model.

The application of topic modeling methods in clinical research has become more common in recent years (Hagg et al., 2022; Boyd-Graber et al., 2017; Jelodar et al., 2019). While topic modeling approaches have advanced significantly over the last twenty years, evaluation methods have lagged behind (see Hoyle et al., 2021 for a recent survey of methods). Current metrics tend to focus on intrinsically assessing model performance (via perplexity on held-out data) or on attempting to measure the quality of the topics that a model produces using metrics based on constructs such as human interpretability of the topics themselves (sometimes referred to as "coherence"). In a clinical research setting, however, the topic distributions produced by a model are themselves often meant for use in meaningfully quantifying differences between clinical populations. In such a scenario, usefully evaluating the quality of a topic model's "fit", or comparing that "fit" to that of another model (perhaps trained via a different algorithm, or with a different choice of hyperparameters) becomes a question of *extrinsic* evaluation, as intrinsic metrics such as perplexity or coherence are unlikely to be sufficient. Additionally, in clinical research, topic models are typically one piece of a larger analytical puzzle, one which often depends on traditional hypothesis-driven inferential statistical approaches (rather than stand-alone evaluation or use, as is more typical with topic models in machine learning scenarios).

In this paper, we outline a statistical approach to explore and quantify group differences in topic

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 55–65
September 11–15, 2023. ©2023 Association for Computational Linguistics

distributions captured by topic models and demonstrate its application using LDA and two different corpora. First, we validate our method on the *20Newsgroup* corpus, a widely-used reference corpus for developing and evaluating topic modeling algorithms (Mitchell, 1997), by comparing topic distributions between groups of documents that we expect to be similar and groups that we expect to be different. Second, we use our method on a corpus of language samples of Autistic[1] and Typically Developing (TD) children. Based on previous clinical evidence, we expect the topic distribution vectors of Autistic children to differ from those of the TD children. Our proposed method allows for a robust and statistically meaningful evaluation of the output of a topic model in both clinical and non-clinical contexts.

1.1 Topic Maintenance in ASD

Autism Spectrum Disorder (ASD) is a developmental disorder that is characterized by difficulties with social communication and restricted repetitive behavior (RRB) (American Psychiatric Association, 2013). These social communication difficulties sometimes include problems with topic maintenance (Baltaxe and D'Angiola, 1992; Paul et al., 2009), with Autistic children having more difficulty staying on topic than TD children. This difference may result in a signal that could be captured by a topic model as TD and ASD children would have different proportions of their speech assigned to different topics. In an effort to investigate this difference, we applied our statistical approach using LDA and a corpus of transcribed conversations between Autistic and TD children and trained examiners that were recorded during administration of a standard clinical assessment tool, the Autism Diagnosis Observation Schedule (ADOS, described further in section 3.2.1). Previous work with ADOS language samples (Salem et al., 2021; Lawley et al., 2023; MacFarlane et al., 2023) has shown that computational methods are able to capture a variety of differences in the language used by Autistic children from such dialogue samples, but to date have not focused on topic-level features. Our hypotheses for this experiment are two fold: (1) Autistic children will have different topic distributions than the TD children (i.e., talk about different topics than the TD children); (2) examiners will have similar topic distributions regardless of whether they are talking with Autistic children or TD children, as the ADOS task is designed (and examiners are trained) so as to ensure uniformity of delivery on the part of the examiner irrespective of the child's diagnostic status.

[1]We are using identity-first language (i.e., Autistic children) here instead of person-first language (i.e., children with Autism) as the former is the current preference among many Autistic individuals (Brown, n.d.).

2 Statistical Motivation

LDA is a unsupervised, generative probabilistic model that is used on a corpus of text documents to model each document as a finite mixture over k topics (Blei et al., 2003). Each document is treated as a bag-of-words (i.e., order does not matter) and is represented as a set of words and their associated frequencies. Given M documents and an integer k, LDA produces a $M \times k$ document-topic matrix (θ). LDA also produces a $k \times V$ topic-word matrix (β), where V is the total number of unique words across the entire corpus of documents. Since we will not be using the topic-word matrix in this analysis, from this point forward, we will use the phrases "LDA model" and "document-topic matrix" interchangeably.

In the document-topic matrix, each row represents a single document and each column represents one topic. The elements $(\theta_{1,1}, \ldots, \theta_{i,j}, \ldots, \theta_{M,k})$ are the estimated proportion of words in a document that were generated by a topic. From this matrix, each document can now be represented as a k-dimensional topic distribution vector.

These LDA-derived topic distribution vectors often serve as useful document representations for downstream analyses, such as a feature vectors for documentation classification or clustering. They are also commonly used as proxies for document content in more qualitative analyses of the composition of text corpora. To our knowledge, a statistical method for comparing topic distribution vectors between groups of documents has not yet been proposed.

One reason for this is due to the numerical properties of the resulting topic distribution vectors (each component θ_i is bounded between $\{0, 1\}$ with the further constraint of $\sum_{i=1}^{k} \theta_i = 1$), which render them unsuitable for use with many parametric statistical methods. This is an important limitation, because as previously mentioned, as the applications of topic modeling methods expand in clinical and behavioral research, the need for statis-

tically based evaluation tools grows.

We realized that since the components in a topic distribution vector are proportions and all sum to one, they meet the definition of "compositional" data as formalized by Aitchison (1982), who also proposed a family of statistical approaches for such data. Compositional data are vectors of positive numbers that together represent parts of some whole: e.g., the demographic profile of a city or the mineral compositions of rocks.

There are three linear transformations that can be performed on compositional data: additive logratio (ALR), center logratio (CLR), and isometric logratio (ILR) transformation. The ILR transformation was introduced by Egozcue et al. (2003) in an effort to broaden the range of statistical methods that can be applied to compositional data by mapping compositonal data into real space. This transformation maps a composition from its original sample space (the D-part simplex) to the $D-1$ Euclidean space (ILR: $S^D \rightarrow \mathbb{R}^{D-1}$) with all metric properties preserved. Once the compositions are in $\mathbb{R}^{D-1}$, we are able to use classical multivariate analysis tools such as multivariate analysis of variance (MANOVA) to explore group differences (Egozcue et al., 2003; van den Boogaart et al., 2023).[2]

MANOVA is used to compare multivariate sample means and examines the effect of one discrete, independent variable on multiple continuous, dependent variables. For the analyses described in this paper, the independent variable is topic label when using the *20Newsgroup* corpus and diagnosis (ASD, TD) when using the clinical corpus. The dependent variables in both analyses are the various topic distribution probabilities in the document-topic matrix created by LDA: $\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,k-1}$ where $i = 1, 2 \ldots, M$. It is important to note that a different discrete variable can be used as the independent variable, as long as it separates the documents into groups (e.g., author if modeling a corpus of newspaper articles); if one wished to incorporate multiple independent variables, one could could instead use MANCOVA. Since we used a k of 20 in both of our analyses and one dimension is removed during the ILR transformation, there are a total of 19 dependent variables.

In the case that we do find a significant group difference, the next step is to find out the magnitude of the effect. After MANOVA, we can use partial eta-squared (η^2) to calculate effect size. Partial η^2 tell us what proportion of variance of the linear combination of the topics can be explained by the independent variable (Tabachnick and Fidell, 2013).

MANOVA is a compelling choice for this analysis for several reasons. As detailed above, it enables us to statistically determine whether the topic distributions learned by our topic model are significantly associated with our other variables of interest (group membership, etc.) under a conventional hypothesis-testing framework. Second, MANOVA allows us to calculate interpretable measurements of effect size, which in turn facilitate comparison between different models (even if they are trained using different modeling algorithms). Third, this framework enables us to incorporate additional covariates as independent variables (via upgrading to MANCOVA), in a way that a more traditional classification-centric downstream task would not. Lastly, MANOVA is a well-characterized and well-established statistical method and as such has numerous useful extensions; for example, it can be combined with post-hoc Roy–Bargmann stepdown procedure (Tabachnick and Fidell, 2013) which enables detailed statistical analysis of the relationship between individual topics (or combinations of topics) and our independent variable, thereby facilitating a far richer quantitative interpretation of our topic model's output than other methods. Note, however, that this would be slightly complicated under our protocol due to our use of ILR, which results in the loss of a dimension into a new feature space that is decoupled from the original topics learned by the model (but which preserves important semantic properties of the original feature space). In this work, we explore only the first two points mentioned, leaving the rest for future work.

3 Corpora

We demonstrate our approach on two separate corpora: a subset of the *20Newsgroup* corpus and a corpus of transcribed natural language samples of ASD and TD children.

3.1 *20Newsgroup* corpus

The *20Newsgroups* corpus is a collection of approximately 18,000 posts from twenty different Usenet

[2]Our ability to use MANOVA here is contingent on statistical assumptions that must be met before proceeding. These assumptions are discussed in more in detail in section 4.3.

newsgroups,[3] and is a classic and widely-used dataset for text classification and analysis (Mitchell, 1997). We used the version of the *20Newsgroups* corpus that is available through the Python library `scikit-learn` (Pedregosa et al., 2011). For this analysis, we used documents from the following topic labels: *comp.sys.ibm.pc.hardware*, *comp.sys.mac.hardware*, *rec.sport.baseball*, and *rec.sport.hockey*. Documents that contained less than 500 characters were omitted. All utterances were tokenized, converted to lowercase, and lemmatized (e.g., "troubling" and "troubles" both become "trouble"). Stop words and fillers (e.g., "uh-huh", "mmhmm", "hmm", etc.) were dropped.[4]

3.2 Clinical corpus

The data used to in our second analysis consists of transcribed natural language samples of 117 ASD children and 65 TD children between the ages of 4 and 15 years old. All participants were native English speakers and had an IQ of ≥ 70. Sample characteristics for all 182 participants are summarized in Table 1. Intellectual level was estimated using the Wechsler Preschool and Primary Scale of Intelligence, third edition (WPPSI-III; Wechsler, 2002), for children younger than 7 years old. For children 7 years and older, the Wechsler Intelligence Scale for Children, fourth edition (WISC-IV; Wechsler, 2003), was used. Language ability and pragmatic and structural language skills were estimated using the Children's Communication Checklist, version 2 (CCC-2; Bishop, 2003).

3.2.1 Language samples

The language samples are transcribed dialogues between the child and an examiner during the conversation activities in the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2000). The ADOS is a semi-structured interview that is designed to provide opportunities to observe speech and behavior that are characteristic of ASD as defined by the DSM-IV-TR (American Psychiatric Association, 2000). All participants were administered the ADOS-2, Module 3, which is designed for children and adolescents with fluent speech. Sessions were scored using the revised algorithms (Gotham et al., 2009).

Audio files were transcribed by a team of trained transcribers who were blind to participants' diagnostic status and intellectual abilities. Transcription was completed following modified Systematic Analysis of Language Transcripts (SALT) guidelines (Miller and Iglesias, 2012). Both the child and examiner speech were transcribed.

For this analysis, we used the transcribed dialogues from the four ADOS conversation activities: *Emotions*; *Social Difficulties and Annoyance*; *Friends, Relationships, and Marriage*; *Loneliness*. These activities were chosen for this analysis because of their conversational structure and naturalistic dialogue. Other ADOS activities, such as *Description of a Picture* and *Telling a Story From a Book*, were omitted. For each conversation activity, examiners are instructed to ask the child a series of questions, such as "What do you like doing that makes you feel happy and cheerful?" and "Do you have some friends? Can you tell me about them?". We followed same text preprocessing steps as described in section 3.1.

4 Methods

Figure 1 shows an example workflow for our method using LDA and a k of 5. All analyses were completed using the statistical programming language R (R Core Team, 2020). LDA models were estimated using the the `topicmodels` package (Grün and Hornik, 2011). The ILR transformation was performed using the `compositions` package (van den Boogaart et al., 2023). Box's M Test was performed using the `heplots` package (Friendly et al., 2022) and partial eta-squared was calculated using the `effectsize` package (Ben-Shachar et al., 2020). Our code for the *20Newsgroup* analysis is available online.[5]

4.1 *20Newsgroup*

Using the documents from four different topics, we fit a single LDA model with a k value of 20. After transforming the topic distribution vectors using the ILR transformation, we performed seven MANOVA tests. First, we compared the topic distributions between the broader *comp.sys.** and *rec.sport.** categories, where the former is composed of the documents from *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware* and the latter of those from *rec.sport.baseball* and *rec.sport.hockey*.

[3] Usenet was an early internet-based network of hierarchically-organized discussion groups where users could post messages about a given topic.

[4] We used the lexicon of stop words provided in the tidytext package (Silge and Robinson, 2016).

[5] https://github.com/gracelawley/lawley-sigdial-2023

	ASD *(n = 117, 98 males)*				TD *(n = 65, 37 males)*				
	min	*max*	*mean*	*s.d.*	*min*	*max*	*mean*	*s.d.*	*p*
Age in years	4.54	15.6	10.03	2.82	4.21	14.5	8.22	2.83	<.001
IQ	72	138	102.19	15.77	90	147	116.94	12.37	<.001
ADOS SA	3	19	9.18	3.48	0	8	0.95	1.47	<.001
ADOS RRB	0	8	3.59	1.53	0	2	0.45	0.64	<.001
ADOS Total	7	24	12.77	3.73	0	10	1.40	1.79	<.001
CCC-2 Pragmatic	1.5	10.8	4.96	1.69	7.5	15.8	12.05	1.73	<.001
CCC-2 Structural	1	12	7.01	2.29	8.5	15	11.73	1.57	<.001
CCC-2 GCC	45	103	75.13	11.0	87	143	115.18	12.09	<.001

Table 1: Demographic and clinical sample characteristics. Abbreviations: ADOS = Autism Diagnostic Observation Schedule; SA = Social Affect; RRB = Restricted and Repetitive Behavior; CCC-2 = Children's Communication Checklist, version 2; GCC = Global Communication Composite.

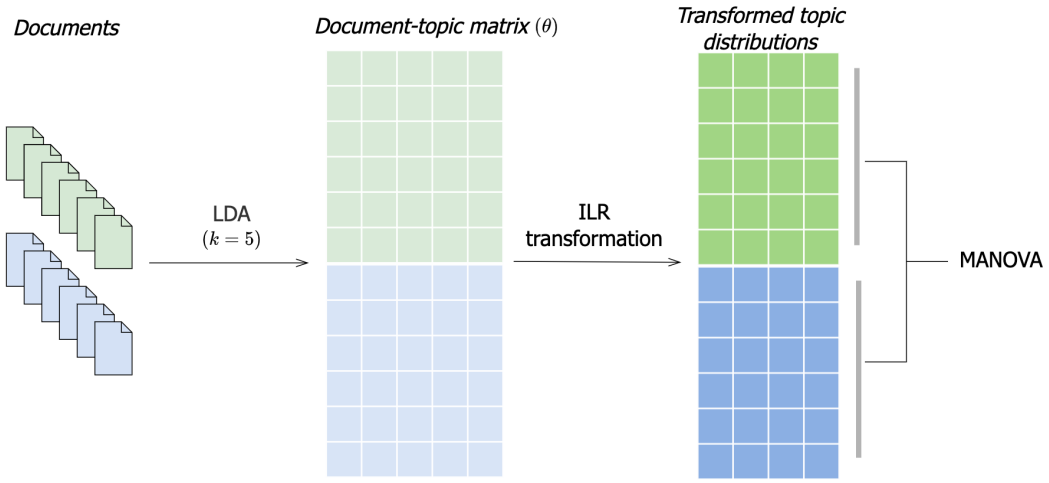


Figure 1: Example workflow for the described statistical approach described to explore and quantify group differences in topic distributions captured by topic models.

We hypothesize that the topic distributions between these groups will be very different. Second, we compared topic distributions between subcategories: *comp.sys.ibm.pc.hardware* vs. *comp.sys.mac.hardware*; *rec.sport.baseball* vs. *rec.sport.hockey*. We hypothesize that these groups will also be different, but not as different as the previous comparison. Third, we compared the topic distributions within each of the four topics by randomly splitting each topic into two groups (e.g., *rec.sport.baseball.1* vs. *rec.sport.baseball.2*). Since the documents are from the same topic, we hypothesize that there will be no difference between the topic distributions. For all of the above MANOVA tests, the independent variable is the topic label and the dependent variables are the topic probability values from the document-topic vectors.

4.2 Clinical corpus

Since our plan involves analyzing the child and examiner speech separately, we created two separate LDA models: one containing only the child speech and one containing only the examiner speech. In both models, we define a document as all words said by a speaker during a single ADOS conversation activity. Since there are four activity types, within each model each child-examiner conversation is associated with four, distinct documents.

We used a k value of 20 for both models. This decision was informed by prior knowledge of the type and quantity of questions the examiners are

instructed to ask during the ADOS conversation activities. Hyperparameter estimation was done using the variational expectation-maximization (VEM) algorithm with a starting α value of $50/k$ (Grün and Hornik, 2011; Griffiths and Steyvers, 2004).

For each of our MANOVA tests, the independent variable is diagnosis (either ASD or TD) and the dependent variables are the topic probability values from the document-topic vectors. Since we used a k of 20 in our analysis and one dimension was lost during the ILR transformation there are 19 dependent variables. The null hypothesis is that the multivariate means of the ASD and TD groups are equal.

4.3 MANOVA assumptions

Before proceeding further with MANOVA, there are multiple assumptions that must be met (Tabachnick and Fidell, 2013). First, each combination of independent and dependent variables should be multivariate normally distributed. Since there are more than 20 observations for each dependent $\times$ independent variable combination the Multivariate Central Limit Theorem holds so we can assume the multivariate normality assumption holds.

Second, dependent variables should have a linear relationship with each group of the independent variable. This assumption was initially not met since each topic distribution vector summed to 1. However after performing the ILR transformation described in section 2, this is no longer the case.

Third, variance-covariance matrices for dependent variables should be equal across groups. This can be tested using Box's M test (Box, 1949), which tests the null hypothesis that the matrices are equal. For our data, Box's M test yielded p-values of $p < 0.001$ for each topic for the *20Newsgroups* documents and also for each conversation activity for both child and examiner speech, and thus this assumption (of equal covariance matrices) was not met. However, MANOVA is robust to unequal covariance matrices when Pillai's criterion is used (Tabachnick and Fidell, 2013; Pillai, 1955), and as such we are able to proceed .

Lastly, there should be no extreme outliers in the dependent variables. Extreme outliers can be identified by calculating the Mahalanobis distance for each observation and then performing a chi-squared test (using $df = k - 1$) to calculate the corresponding p-values. The null hypothesis is that the observation is not an outlier. We repeated analyses with identified outliers excluded and saw no difference in results. The results presented here are with these outliers included.

5 Results

The first part of our analysis was to demonstrate the application of our approach on the *20Newsgroup* corpus, a popular corpus for topic modeling. The results for the MANOVA tests are reported in Table 2. There was a significant difference between the topic distributions from the *comp.sys.** and *rec.sport.** categories, $F(19, 1710) = 414.240$, $p < 0.001$, with a large effect size, partial $\eta^2 = 0.82$. Between the *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware* subcategories, topic distributions were significantly different, $F(19, 795) = 15.008$, $p < 0.001$, with a large effect size, partial $\eta^2 = 0.26$. Topic distributions were also significantly different between the *rec.sport.baseball* and *rec.sport.hockey* subcategories, $F(19, 895) = 15.008$, $p < 0.001$, with a large effect size, partial $\eta^2 = 0.57$. When comparing topic distributions within each topic (by randomly splitting the documents into two groups), there were no significant differences found.

For the second part of our analysis, we compared the children's topic distribution vectors between diagnostic groups (ASD, TD). The results of the MANOVA tests for each ADOS conversation activity for child speech are reported in Table 3. The children's topic distributions were significantly different between the Autistic and TD children within the *Social Difficulties and Annoyance* activity, $F(19, 169) = 2.055$, $p = 0.0083$, with a large effect size, partial $\eta^2 = 0.19$. There was no significant group difference in topic distributions within the other three conversation activities (*Emotions*; *Friends, Relationships, and Marriage*; *Loneliness*). To address potential Type I error from multiple comparisons, p-values can be evaluated using a Bonferroni adjusted α of 0.0125. When evaluating the results using the adjusted α of 0.0125, the significant result within the *Social Difficulties and Annoyance* conversation activity remains.

Lastly, the results of the statistical analyses performed on the examiner speech are reported in Table 4. The examiners' topic distributions differed significantly between ASD and TD groups within three of the four conversation activities examined: *Emotions*, $F(19, 175) = 2.235$, $p = 0.0035$, with a large effect size, partial $\eta^2 = 0.20$; *So-*

topics	n	df	Pillai	approx. F	df_1	df_2	p	partial η^2
*comp.sys.**	815	1	0.822	414.240	19	1710	<0.001	0.82
*rec.sport.**	915							
comp.sys.ibm.pc.hardware	447	1	0.264	15.008	19	795	<0.001	0.26
comp.sys.mac.hardware	368							
rec.sport.baseball	423	1	0.571	62.722	19	895	<0.001	0.57
rec.sport.hockey	492							
comp.sys.ibm.pc.hardware	219	1	0.020	0.460	19	427	0.976	0.02
"	228							
comp.sys.mac.hardware	198	1	0.044	0.840	19	348	0.659	0.04
"	170							
rec.sport.baseball	206	1	0.041	0.903	19	403	0.579	0.04
"	217							
rec.sport.hockey	247	1	0.029	0.738	19	472	0.780	0.03
"	245							

Table 2: *20Newsgroups*, comparison of LDA topic distribution vectors between and within topics.

		df	Pillai	approx. F	df_1	df_2	p	partial η^2
Emotions	dx	1	0.093	0.941	19	175	0.5334	0.09
Social	dx	1	0.188	2.055	19	169	0.0083	0.19
Friends	dx	1	0.131	1.388	19	175	0.1381	0.13
Loneliness	dx	1	0.135	1.275	19	156	0.207	0.13

Table 3: Child speech, comparison of LDA topic distribution vectors between ASD and TD groups.

cial Difficulties and Annoyance, $F(19, 174) = 3.858$, $p < 0.001$, with a large effect size, partial $\eta^2 = 0.30$; *Friends, Relationships, and Marriage*, $F(19, 176) = 1.833$, $p = 0.0224$, with a large effect size, partial $\eta^2 = 0.17$. There was no significant difference between groups for the *Loneliness* conversation activity. A Bonferroni adjusted α of 0.0125 can be used to address potential Type I error from multiple comparisons. With this adjusted α, a significant group difference within the *Emotions* and *Social Difficulties and Annoyance* activities remains; however, the previous group difference within *Friends, Relationships, and Marriage* is no longer significant.

6 Discussion

The Autistic children and TD children had significantly different topic distributions for one of the four conversation analyzed: *Social Difficulties and Annoyance*. We expected to observe a group difference in all four of the conversation activities instead of only one. Incorporating additional participant-level information such as IQ and age or examining other measures of conversational reciprocity such as the length and complexity of utterances may help shed some light as to why a group difference was only seen in one of the four activities analyzed. In addition, further investigation into sampling context differences between the conversation activities is needed before conclusions can be drawn. This finding illustrates the value of our proposed statistical approach, in that we have numerous ways we could incorporate these additional covariates into our analysis in quantitatively useful ways within the same statistical framework.

The examiners' topic distributions differed significantly between the ASD and TD groups for two of the four activities: *Emotions* and *Social Difficulties and Annoyance*. This is surprising as our initial hypothesis was there would not be any sig-

		df	Pillai	approx. F	df_1	df_2	p	partial η^2
Emotions	dx	1	0.195	2.235	19	175	0.0035	0.20
Social	dx	1	0.296	3.858	19	174	<0.001	0.30
Friends	dx	1	0.165	1.833	19	176	0.0224	0.17
Loneliness	dx	1	0.151	1.557	19	167	0.0726	0.15

Table 4: Examiner speech, comparison of LDA topic distribution vectors between ASD and TD groups.

nificant group differences for the examiners' topic distributions. ADOS examiners are instructed to cover the same questions for each child, regardless of diagnosis, and are trained to a high standard of consistency and repeatability, as the assessment is meant for clinical use. Since one goal of the conversation activities is to foster a dialogue, the examiner would likely avoid actions that could discourage the child from conversing and sharing their interests. It may be the case that the examiners are mirroring the topics introduced by the children during the activities and those topics are being picked up by the topic distributions created by LDA.[6] This could be explored in the future by investigating pairwise group differences.

7 Conclusion

In this paper we presented a novel application of existing statistical methods to evaluate the document-topic distribution vectors created by topic models in order to investigate group differences. By treating the document-topic distribution vectors as compositional data (Aitchison, 1982), we are able to use the ILR transformation (Egozcue et al., 2003) to map the vectors from their original sample sample, the D-part simplex, into the $D-1$ Euclidean space (ILR: $S^D \rightarrow \mathbb{R}^{D-1}$). Once in $\mathbb{R}^{D-1}$, we are able to use classical multivariate analysis tools such as MANOVA (Egozcue et al., 2003).

When applied to an LDA model fitted to the *20Newsgroups* corpus, our method successfully identified that the topic distributions for documents from different categories (computer hardware vs. sports) and also documents from related subcategories (PC hardware vs. Macintosh hardware; baseball vs. hockey) were significantly different. The effect size, measured with partial η^2, also varied across these comparisons, with the effect size being the largest when comparing computer hardware vs. sports and smallest when comparing Macintosh vs. PC hardware. Furthermore, our method did not find that topic distributions are significantly different when comparing groups of documents from the same category.

We also demonstrate the application of this method using LDA and a corpus of child-examiner dialogues of Autistic and TD children, where prior clinical research gave us reason to expect to find group differences. We found that the topic distributions of Autistic and TD children were significantly different during one of the four ADOS conversation activities examined. This result aligns with prior clinical research that Autistic children often have difficulties with topic maintenance in a conversational context. Interestingly, we also found that examiners' topic distributions were significantly different whether they were conversing with an Autistic child or a TD child for two of the four ADOS conversation activities examined. This may indicate that although the examiners are trained to ask the same set of questions irrespective of diagnosis status, tangential topics introduced by the child during the conversation may be mirrored by the examiner and thus are reflected in the associated topic distributions.

There are a few points about the statistical approach outlined in this paper that should be highlighted. Although we demonstrate this method using the document-topic distribution matrix created by LDA, this method can be extended to any topic modeling algorithm that outputs a topic distribution that can be treated as a composition. We decided to use LDA here as it is a well-established technique that has been extended and built upon many times over since it was first introduced in 2003. Another important point to highlight is that, although not shown in here, this analysis has the potential to be extended further with a post-hoc Roy-

[6] An anonymous reviewer brought to our attention that interviewers have been found to adjust their conversational patterns when speaking to patients with other cognitive conditions, such as Alzheimer's disease (Nasreen et al., 2021).

Bargmann step down procedure to explore how much each topic (or combination of topics) contributes to the significant effect of the independent variable (Tabachnick and Fidell, 2013). However, as previously mentioned, the loss of a dimension during the ILR transformation would need to be addressed first. Overall, the statistical approach presented in this paper represents a very promising direction for methods of making topic models more interpretable in a quantitative way, beyond human inspection of topics. In the future we would like to extend this specific analysis to include additional participant-level, independent variables (e.g., age, sex, IQ) by using multivariate analysis of covariance (MANCOVA). Since social communication skill level can vary throughout the ASD spectrum (Tager-Flusberg and Kasari, 2013), we would also like to look at differences within the ASD group by exploring within group variance metrics. We would also like to explore the use of other methods of topic modeling, beyond LDA, for this application.

As the application of topic modeling methods continues to grow into areas such as clinical and behavioral research, so does the need for statistically based methods for evaluation and comparison. Our hope is that the statistical approach described in this paper contributes to bridging that gap by focusing on improving evaluation metrics for existing topic modeling methods.

Limitations

There are several limitations of this analysis that should be mentioned. First, the decision to set k to 20 was specific to the particular clinical discourse corpus used. Our decision was informed by of the type and quantity of questions the examiners are instructed to ask during the ADOS conversation activities; however, it may not always be possible to choose a value for k using existing knowledge of the corpus. Second, as mentioned in section 2, after performing the ILR transformation we lose one dimension from our original topic model's output and go from k to $k-1$ elements in each vector. A consequence of this is that there is no direct mapping between dimensions of the ILR-transformed $\mathbb{R}^{k-1}$ vector and the original k topics after the transformation, though the new dimensions retain the information contained in the original data (as shown by their ability to be used via MANOVA). Depending on the nature of the analysis that one is conducting, this may or may not be an issue; it was not during the present analysis, since we were interested in the overall topic distributions of each document (rather than in specific document-topic associations) but this may not always be the case. A possible direction for future work would be to draw further upon statistical methods from compositional spaces to assist with this issue.

Ethics Statement

This study was approved by the Oregon Health & Science University IRB (Protocol #531) and all research was performed in accordance with their relevant guidelines and regulations.

Acknowledgements

This work was supported in part by the National Institute on Deafness and Other Communication Disorders of the NIH under Awards R01DC012033 (PI: Dr. E. Fombonne) and R01DC015999 (PIs: Dr. S. Bedrick & G. Fergadiotis).

References

John Aitchison. 1982. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–177.

American Psychiatric Association. 2000. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed., text rev.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed.

Christiane A. M. Baltaxe and Nora D'Angiola. 1992. Cohesion in the discourse interaction of autistic, specifically language-impaired, and normal children. *Journal of Autism and Developmental Disorders*, 22(1):1–21.

Mattan S. Ben-Shachar, Daniel Lüdecke, and Dominique Makowski. 2020. effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, 5(56):2815.

Dorothy V. M. Bishop. 2003. *The Children's Communication Checklist, version 2 (CCC-2)*. Pearson, London.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

George E. P. Box. 1949. A general distribution theory for a class of likelihood criteria. *Biometrika*, 36(3/4):317–346.

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. Applications of Topic Models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Lydia Brown. n.d. Identity-First Language. Autistic Self Advocacy Network (ASAN). https://autisticadvocacy.org/about-asan/identity-first-language/ Last accessed on 2023-05-10.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. 2003. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35:279–300.

Michael Friendly, John Fox, and Georges Monette. 2022. *heplots: Visualizing Tests in Multivariate Linear Models*. R package version 1.4-2.

Katherine Gotham, Andrew Pickles, and Catherine Lord. 2009. Standardizing ADOS Scores for a Measure of Severity in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, 39(5):693–705.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 1:5228–35.

Bettina Grün and Kurt Hornik. 2011. topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30.

Lauryn J. Hagg, Stephanie S. Merkouris, Gypsy A. O'Dea, Lauren M. Francis, Christopher J. Greenwood, Matthew Fuller-Tyszkiewicz, Elizabeth M. Westrupp, Jacqui A. Macdonald, and George J. Youssef. 2022. Examining Analytic Practices in Latent Dirichlet Allocation Within Psychological Science: Scoping Review. *Journal of Medical Internet Research*, 24(11):e33166.

Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken?: The incoherence of coherence.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Grace O. Lawley, Steven Bedrick, Heather MacFarlane, Jill K. Dolata, Alexandra C. Salem, and Eric Fombonne. 2023. "Um" and "Uh" usage patterns in children with autism: Associations with measures of structural and pragmatic language ability. *Journal of Autism and Developmental Disorders*, 53:2986–2997.

Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H. Cook, Bennett L. Leventhal, Pamela C. DiLavore, Andrew Pickles, and Michael Rutter. 2000. The Autism Diagnostic Observation Schedule, Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. *Journal of Autism and Developmental Disorders*, 30(3):205–223.

Heather MacFarlane, Alexandra C. Salem, Steven Bedrick, Jill K. Dolata, Jack Wiedrick, Grace O. Lawley, Lizbeth H. Finestack, Sara T. Kover, Angela John Thurman, Leonard Abbeduto, and Eric Fombonne. 2023. Consistency and reliability of automated language measures across expressive language samples in autism. *Autism Research*, 16(4):802–816.

J. Miller and A. Iglesias. 2012. SALT: Systematic analysis of language transcripts [Research version]. *Middleton, WI: SALT Software*.

Tom Mitchell. 1997. *Machine Learning*. McGraw Hill.

Shamila Nasreen, Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Alzheimer's Dementia Recognition From Spontaneous Speech Using Disfluency and Interactional Features. *Frontiers in Computer Science*, 3.

Rhea Paul, Stephanie Miles Orlovski, Hillary Chuba Marcinko, and Fred Volkmar. 2009. Conversational Behaviors in Youth with High-functioning ASD and Asperger Syndrome. *Journal of Autism and Developmental Disorders*, 39(1):115–125.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

K. C. S. Pillai. 1955. Some New Test Criteria in Multivariate Analysis. *The Annals of Mathematical Statistics*, 26(1):117–121.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alexandra C. Salem, Heather MacFarlane, Joel R. Adams, Grace O. Lawley, Jill K. Dolata, Steven Bedrick, and Eric Fombonne. 2021. Evaluating atypical language in Autism using automated language measures. *Scientific Reports*, 11(1):10968.

Julia Silge and David Robinson. 2016. tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS*, 1(3).

Barbara G. Tabachnick and Linda S. Fidell. 2013. *Using Multivariate Statistics*, 6th edition. Pearson.

Helen Tager-Flusberg and Connie Kasari. 2013. Minimally Verbal School-Aged Children with Autism Spectrum Disorder: The Neglected End of the Spectrum. *Autism Research*, 6(6):468–478.

K. Gerald van den Boogaart, Raimon Tolosana-Delgado, and Matevz Bren. 2023. *compositions: Compositional Data Analysis*. R package version 2.0-6.

David Wechsler. 2002. WPPSI-III: Wechsler Preschool and Primary Scale of Intelligence - 3rd ed. *San Antonio, TX: Psychological Corporation*.

David Wechsler. 2003. WISC-IV: Wechsler Intelligence Scale for Children. *San Antonio, TX: Psychological Corporation*.

OpinionConv: Conversational Product Search with Grounded Opinions

Vahid Sadiri Javadi
Conversational AI and Social Analytics (CAISA) Lab
University of Bonn

Martin Potthast
Text Mining and Retrieval (TEMIR) Group
Leipzig University and ScaDS.AI

Lucie Flek
Conversational AI and Social Analytics (CAISA) Lab
University of Bonn

Abstract

When searching for products, the opinions of others play an important role in making informed decisions. Subjective experiences about a product can be a valuable source of information. This is also true in sales conversations, where a customer and a sales assistant exchange facts and opinions about products. However, training an AI for such conversations is complicated by the fact that language models do not possess authentic opinions for their lack of real-world experience. We address this problem by leveraging product reviews as a rich source of product opinions to ground conversational AI in true subjective narratives. With OpinionConv, we develop the first conversational AI for simulating sales conversations. To validate the generated conversations, we conduct several user studies showing that the generated opinions are perceived as realistic. Our assessors also confirm the importance of opinions as an informative basis for decision making.

1 Introduction

In order to elucidate the mechanics of conversational product search, Kotler and Keller (2015) delineated a five-stage process that encapsulates customer decision making (see Figure 1, left). This process suggests that the customer: (1) recognizes a problem or need; (2) searches for information about potential products or services that could resolve the problem or fulfill the need, filtering them until a manageable set of alternatives remains; (3) evaluates and compares these alternatives against each other with regard to personal preferences and third party opinions to inform their decision making; (4) proceeds to make a purchase decision predicated upon this informed evaluation; and finally, (5) exhibits post-decision behaviors that reflect their satisfaction, which completes the process.

Typically, in-store shopping predominantly engages with the second and third stages of this customer decision process. Both the activities of reducing the number of alternatives and evaluating their merits and demerits are conducted in conversations between customers and sales assistants. The absence of such interactions in online environments is perceived as a deficiency in customer service especially with respect to the third stage (Exalto et al., 2018). Customers derive post-purchase satisfaction from personal exchanges, relating to others experience, and having the opportunity to ask questions (Papenmeier et al., 2022). The considerable number of online product reviews available are not a substitute for everyone, since many customers lack the patience to examine many of them, leading to post-purchase dissatisfaction and product returns. Conversational AI has been suggested as a solution (Gnewuch et al., 2017), with the goal of mimicking the conversational strategies of sales assistants (Papenmeier et al., 2022). But despite its importance, previous research on conversational product search almost entirely neglects the third stage, or rather its opinionated aspects (Section 2).

Recent advances in large-scale conversational language models, spearheaded by OpenAI's ChatGPT, are driving a paradigm shift in the development of conversational technologies. Nonetheless, when it comes to expressing opinions pertaining to real-world events or entities, these language models lack the necessary grounding in tangible reality. For an individual to formulate an opinion on a particular subject matter, they require exposure to the subject to relate the new experience to past ones, and importantly, an emotional perception. A language model is only capable of generating what might be termed as a "statistical average" of third-party opinions, if they have been part of its training data. In the context of product search, such opinions would be deemed unauthentic as they are not based on real-world experiences or substantiated knowledge. This lack of authenticity poses challenges to the effective utilization of these models when (personal) opinions play an important role.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 66–76
September 11–15, 2023. ©2023 Association for Computational Linguistics

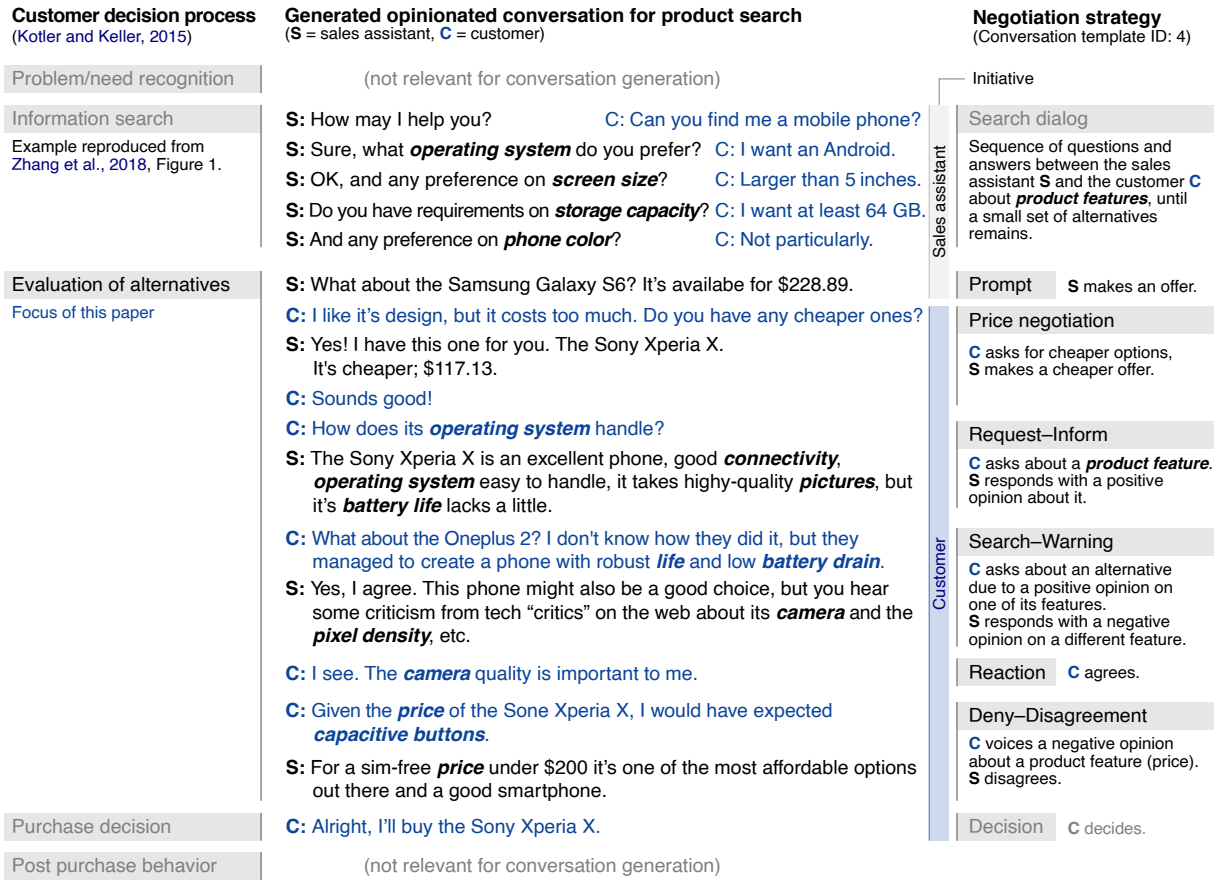


Figure 1: A grounded opinionated conversation generated by OpinionConv based on Conversation Template 4.

In this paper, we focus on the third stage of the customer decision process, for which we contribute the first approach to generate grounded opinionated statements (Section 3). We conceive and operationalize the generation of grounded opinions by positing that a grounded opinion about a product is an opinion which has been verifiably expressed by a minimum of one individual in a product review that specifically discusses the product under scrutiny. Our approach, OpinionConv, combines a product-specific index of reviews for a cohort of products of the same kind with a mechanism to generate realistic opinionated conversational exchanges. While carefully tuned, our approach must still be considered an early prototype. Consequently, before asking real customers to use it, its fundamental capabilities must first be established. We therefore simulate in-store dialogues between a customer and a sales assistant, where both parties incorporate grounded opinions. These conversations are then systematically evaluated in an experimental setup that ascertains the perception of human readers regarding the realism of these dialogs (Section 4).[1]

[1]Code and data: https://github.com/caisa-lab/OpinionConv

2 Related Work

Three lines of research are related to ours: opinionated question answering, conversational product search, and review-based conversation generation.

2.1 Opinionated Question Answering

While factoid Question Answering (QA) systems have a long tradition and some even outperform humans, non-factoid questions, such as opinions, explanations, or descriptions, are still an open problem (Cortes et al., 2021). Cardie et al. (2003) employed opinion summarization to help multi-perspective QA systems identify the opinionated answer to a given question. Yu and Hatzivassiloglou (2003) separated opinions from facts and summarized them as answers. The linguistic features of opinion questions have also been studied (Pustejovsky and Wiebe, 2005; Stoyanov et al., 2005). Kim and Hovy (2005) identified opinion leaders, which are a key component in retrieving the correct answers to opinion questions. Ashok et al. (2020) introduced a clustering approach to answer questions about products by accessing product reviews. Rozen et al. (2021) examined the task of

answering subjective and opinion questions when no (or few) reviews exist. Jiang et al. (2010) proposed an opinion-based QA framework that uses manual question–answer opinion patterns.

Closer to our work, Moghaddam and Ester (2011) address the task of answering opinion questions about products by retrieving authors' sentiment-based opinions about a given target from online reviews. McAuley and Yang (2016) address subjective queries using relevance ranking, and Wan and McAuley (2016) extends this work by considering questions that have multiple divergent answers, incorporating aspects of personalization and ambiguity. AmazonQA (Gupta et al., 2019) is one of the largest review-based QA datasets. Its authors show that it can be used to learn relevance in the sense that relevant opinions are those for which an accurate predictor can be trained to select the correct answer to a question as a function of opinion. SubjQA (Bjerva et al., 2020) includes subjective comments on product reviews.

2.2 Conversational Product Search

Information is often gathered through conversations with a series of questions and answers. Conversational Question Answering (CQA) systems engage in such multi-turn conversations to satisfy a user's information need (Zaib et al., 2021). Despite the attention this task has received in e-commerce (Ricci et al., 2011; Bi et al., 2019; Zhang et al., 2018), building a successful conversational product search system for online shopping still suffers from the lack of realistic dialog datasets for model training (Xiao et al., 2021).

2.3 Review-based conversation generation

Recently, multi-turn QA has grown more prominent (Cambazoglu et al., 2020). Product reviews are one of the sources of information that are being used for conversational product search. Penha et al. (2022) generate review-based explanations for voice-driven product search. Zhang et al. (2018) builds a dataset to answer conversational questions, as illustrated in Figure 1 (Stage 2 "Information Search"). They extract feature–value pairs from reviews and convert each review into a conversation based on the mentioned pairs, but omit opinionated statements. Xu et al. (2019) explores the possibility of turning reviews into a knowledge source to answer questions. Feature-related, non-opinionated statements in reviews are flagged and appropriate questions are formulated.

3 Grounded Product Opinion Generation

This section introduces the OpinionConv construction pipeline to generate grounded opinionated conversations for product search based on product reviews. Figure 2 gives an overview of the pipeline's individual steps, grouped into preprocessing, information search dialog generation (Stage 2 of the customer decision process, which we reproduce from Zhang et al. (2018)), and evaluation dialog generation (Stage 3, our focus).

3.1 Data Source and Preprocessing

As a basis for grounded opinions, we utilize a crawl of Amazon product data including their reviews created by Ni et al. (2019).[2] The metadata enclosed includes product descriptions, multi-level product categories, and product information. For our proof-of-concept, we focus on one of its 24 product categories, *Cell Phones and Accessories*. As a first cleansing step, we reviewed the product data and added missing product details. We found the reviews to be of varying writing quality, especially with respect to basic syntax conventions, like the use of punctuation. We employed the model of Alam et al. (2020) to restore the punctuation, which enabled a more reliable sentence extraction and thus benefited the subsequently applied models, which were largely trained on "cleaner" data.

To extract the product features discussed in the reviews, we use the extraction model of Karimi et al. (2021). It is based on a hierarchical aggregation approach and was trained on the laptop review dataset of SemEval 2014 (Pontiki et al., 2014), performing best at that time. Given a review sentence, the model extracts feature terms on which an opinion has been expressed. On sentences containing such opinion statements, we then applied the sentiment analysis model of Zeng et al. (2019), which is based on self-attention to capture local context and global context features to determine the polarity score of the opinion.

3.2 Information Search Dialog Generation

To generate the information search dialog of the customer decision process (see Figure 1), we reproduce the approach of Zhang et al. (2018). Reproducing the original dialogs turned out to be straightforward, and we verified our success by direct comparison to the data supplied with the original paper. The dialogs are structured as follows: The sales

[2] https://jmcauley.ucsd.edu/data/amazon/

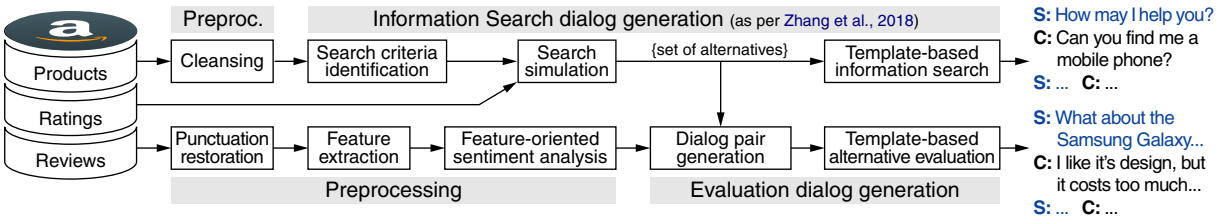


Figure 2: High-level overview of our approach in OpinionConv for generating opinionated multi-turn conversations.

assistant asks for preferences on product features, and the customer answers, narrowing down the set of alternatives. The resulting set of alternatives is fed to the next stage of the customer decision process, the evaluation of alternatives.

3.3 Evaluation Dialog Generation

The generation of an opinion-based evaluation of alternatives dialog is divided into two steps, the generation of pairs of talking points based on reviews of the alternative products, and their combination into a multi-turn conversation as exemplified in Figure 1. For lack of public corpora of in-store conversations, we resort to a template-based approach. The templates are derived from common conversational negotiation strategies from the literature.

Dialog Turn Pair Generation In each turn of an evaluation dialog the customer and the sales assistant discuss the relative value of product features, as well as their benefits and shortcomings compared to alternative products. Sales assistants, by training and/or experience, are usually well-equipped to provide customers with satisfying answers to their questions as well as to respond to opinions that customers express throughout the conversation. The most salient open question in this respect is: How should a sales assistant react to a customer's opinion in the context of a negotiation?

Negotiations combine features of claiming and creating value. Each requires unique strategies and tactics for a negotiator to effectively achieve their objectives while creating the greatest value possible for all parties (Thompson and Hastie, 1990). We take inspiration from three negotiation tactics (Dwyer et al., 1987; Săvescu, 2019): (1) *Distributive negotiation*: This is a competitive win–lose situation. Any value claimed by one party is at the expense of the other. (2) *Integrative negotiation*: The parties create or generate value during the negotiation, and both parties may achieve mutual gains beyond what they would achieve independently, a win–win scenario. (3) *Compatible negotiation*:

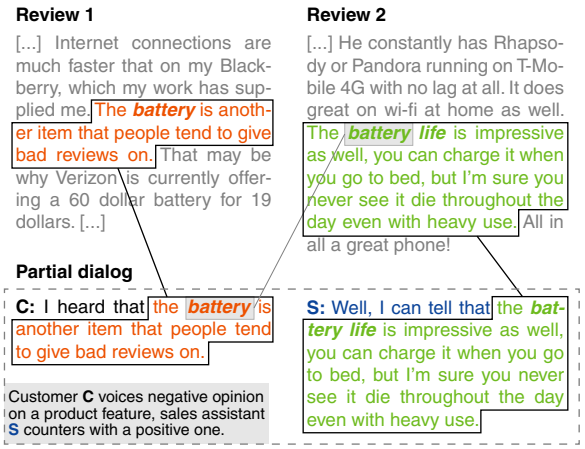


Figure 3: Example of a basic opinionated dialog pair generation step: Given a product feature like "battery", opinionated statements are extracted from reviews of a given product to form part of a dialog between Customer C and Sales Assistant S.

The parties desire the exact same outcome, so that there is no need for any trade-off.

For instance, as illustrated in Figure 3, where Customer C's remark about a product feature (left) is countered by an opinionated counterargument from Sales assistant S (right), we use the Deny–Disagreement tactic, where the customer expresses a negative opinion on a feature of a product in question, whereas the sales assistant disagrees and counters with a positive opinion on the same feature. This tactic may either correspond to a win–lose or a win–win situation, dependent on whose opinion applies more to the customer: If the customer is correct, they lose against the sales assistant, since the product is not switched. If the sales assistant is correct, they both win, since the customer still gets what they wanted, and the sales assistant may still get to sell the product in question.

A key constraint that we enforce by generating grounded opinions (i.e., opinions rooted in a real product reviews as exemplified in Figure 3) is that neither the customer nor the sales assistant can "lie" to each other, as their opinions are backed by a real person's opinion about the product and its

Table 1: Negotiation tactics used in dialog pair templates (P=product, F=feature).

Dialog pair template	Description of negotiation tactic
Request–Inform Question: P-1, F-A, neutral Answer: P-1, F-A, positive	Customer asks about the sales assistant's view on a feature of a product. Sales assistant expresses positive view on it.
Deny–Disagreement Opinion: P-1, F-A, negative Opinion: P-1, F-A, positive	Customer expresses negative opinion on a feature of a product. Sales assistant disagrees and expresses positive opinion on it.
Deny–Switch Product Opinion: P-1, F-A, negative Opinion: P-2, F-A, positive	Customer expresses negative opinion on a feature of a product. Sales assistant switches the product and expresses positive opinion on the same feature wrt. new product.
Deny–Switch Feature Opinion: P-1, F-A, negative Opinion: P-1, F-B, positive	Customer expresses negative opinion on a feature of a product. Sales assistant disagrees and expresses positive opinion on a different feature of the same product.
Search–Agreement Opinion: P-1, F-A, positive Opinion: P-1, F-A, positive	Customer expresses positive opinion on a feature of a product. Sales assistant agrees and expresses another positive opinion it.
Search–Switch Feature Opinion: P-1, F-A, positive Opinion: P-1, F-B, positive	Customer expresses positive opinion on a feature of a product. Sales assistant agrees and expresses positive opinion on different features of the same product.
Search–Warning Opinion: P-1, F-A, positive Opinion: P-1, F-B, negative	Customer expresses positive opinion on a feature of a product. Sales assistant warns the user and expresses negative opinion on different features of the same product.

feature. Thereby the dialog turns are more realistic, despite both parties being simulated. Moreover, our dialog turns enforce a conversational concept flow between (Li et al., 2023), as the product features and their attributes as key concepts are connected.

In Table 1, we list the templates for dialog pairs according to different negotiation tactics derived from the literature; seven patterns are devised, one question–answer pair, and six opinion–opinion pairs. The customer's utterance consists of a feature-specific opinion with either positive or negative opinion for a certain product and one of its features, extracted from one of its reviews. The sales assistant's utterance is a response that expresses either a positive or a negative opinion, not necessarily to the same product or feature.

As can be seen, depending on the type of dialog pair, different negotiation tactis may apply. The sales assistant is allowed to switch the polarity, feature under discussion, and product under negotiation. The mapping of a dialog pair to the negotiation tactics thus depends on the factuality of either opinion expressed by the customer or sales assistant, in case a product switch on the price changes (the price of the new product may be lower, similar, or higher), but also on whether the customer gets what they want. For instance, a switch to a pricier product is certainly worthwhile for the sales assistant, as long as the customer ends up with a desired feature. However, given the necessity of interpreting each generated dialog with respect to its factuality, a mapping between dialog pair and negotiation tactic must be decided on a case-by-case basis, which is beyond the scope of this paper.

A constraint in cases where the product is being changed includes that the sales assistant is allowed to use only two types of products: (1) Retrieved products: The products from the set of alternatives retrieved in the information search dialog at the outset of a conversation based on customer preferences; (2) also viewed products: The products listed in the metadata that have been viewed by other customers. Customers thus can go beyond their original preferences and the set of alternatives.

Template-based Alternative Evaluation The last step of our pipeline generates conversations composed of multiple dialog pairs, based on a negotiation strategy. Considering the diversity of real dialogs and the fact that a coherent conversation should have a smooth transition between turns (Li et al., 2023), we define a diverse set of conversation templates inspired by past studies on negotiation in behavioral economics (Pruitt, 1981; Fisher and Ury, 1981; Thompson et al., 2010), including both high-level (e.g., insisting on your position: Disagreement) and low-level (e.g., focus on interests: Reaction) dialog acts.

Table 2: Example of the combination of dialog pairs in a conversation template.

Pair	Principle	Action	Example
Deny–Switch Product	Insist on position	Express negative sentiment	**B:** What I know about its battery is that the battery keeps draining because the phone is constantly looking for network signal.
	Invent options for mutual gain	Recommend a new product	**S:** If the battery is important for you, we can offer this product: Axon 7 is the same price as OnePlus 3, but it has slightly bigger battery.
Request–Inform	Focus on interests	Look for more information	**B:** What do you think about its speakers?
	Build trust	Express positive sentiment	**S:** It has dual front-facing speakers with good quality.
Search–Agreement	Focus on interests	Search for alternatives	**B:** I heard about this phone: Galaxy S4 that has a super-fast processor and a good battery life.
	Build trust	Confirm consumer's preference	**S:** Yes, that's true. This phone is also a good choice with the one premium hardware, great software and a reasonable price.

Table 3: Demographics of study participants.

Measure	Characteristics	Study 1 (N=100)	Study 2 (N=420)
Gender	Males	41.0%	31.0%
	Females	58.0%	69.0%
	Non-binary	1.0%	0.0%
Age	25 to 34 years	35.0%	38.0%
	35 to 44 years	28.0%	30.1%
	18 to 24 years	21.0%	15.7%
	55 to 64 years	6.0%	13.3%
	45 to 54 years	5.0%	1.8%
	65 years or older	5.0%	1.2%

We follow Zhou et al. (2019) and devise 14 conversation templates with different combinations of the generated question–answer and opinion–opinion pairs. Table 2 exemplifies one of them. We adapt the "CraigslistBargain" setting of He et al. (2018), where a buyer and a seller negotiate the price of a product. But unlike in their work, the sales assistant and the customer negotiate not only the price but primarily the relative merits of product features, whereas price may only be one of them.

4 Evaluation

A volunteer who is asked to pose as a customer in a laboratory user study, and who has no real intention of investing a fairly large amount of his or her own money in the purchase of a product, does not usually have the same information needs as a real customer. At the same time, we consider it unethical to confront real customers with an early prototype of a conversational sales assistant. Before a practical assistant can be developed, the basic means of generating informed opinions must first be established.

In our evaluation, we therefore decided to simulate full conversations between a hypothetical customer and a hypothetical sales assistant as described in the previous section. We then designed two user studies in which we specifically investigated whether human subjects consider these conversations realistic.

Study 1 investigates whether subjective narratives in conversational product search are considered important compared to purely factual exchanges. Study 2 investiages individuals' perceptions of the quality and realism of the conversations generated by OpinionConv. We conducted the two studies by recruiting volunteers living in the US or UK on Prolific.[3] Table 3 shows the total sample size and key demographic information for both studies. As can be seen, we had fewer male participants than females and more than 60% of the participants are between 25 and 45 years old. Key to both our study design is that participants initially believed that the conversations are genuine transcripts of real sales negotiations recorded in a store, instead of generated ones. At the end of the questionnaire, it was revealed that they are not.

4.1 Study 1: Importance of Product Opinions

The first study started with the following instruction: "Below is an automatically generated transcript of a sales conversation. We show two variants: Variant 1 is focused on the customer's preferences and requirements. Variant 2 starts similarly, but then continues with an opinionated discussion." After reading both variants of the same dialog participants were asked "Which of the two variants would you as a customer hold with the sales as-

[3] https://www.prolific.co

sistant while searching for a smartphone?" The survey concluded by asking participants an open-ended question to explain their judgment for the previous question, which also allowed to ascertain that they had actually read the conversations.

As a result, we find that 83% of the 100 participants of Study 1 prefer Variant 2 over the Variant 1, which confirms that they tend to prefer opinionated conversations when searching and evaluating a product rather than exclusively factual ones.

4.2 Study 2: Perceptions of Dialog Realism

For this study, the questionnaire consisted of two separate parts. In the first part, we again let participants believe they are reading a transcript of a real conversation by instructing them as follows: "Suppose you are in an electronics store. While browsing, you happen to overhear part of a conversation between a customer and a sales assistant. Both exchange opinions about the features of one or more products." They are then asked to answer three questions using the 4-point Likert scale: (1) *Definitely yes*, (2) *Rather yes*, (3) *Rather not* and (4) *Definitely not*. As depicted in Figure 4, bottom, we ask questions for the following goals: *Customer understanding*, *Sales assistant answer sufficiency* and *Reasonableness of exchange*. To investigate whether participants' perceptions change significantly after they learned that the conversation was generated, at the beginning of the second part, we reveal the truth and declare that the conversation they just read, was not a real but an automatically generated one. After the disclosure, they were asked answer Questions 4 to 6 using the same Likert-scaled responses as shown in Figure 4 in order to observe any changes of opinion. For each of our 14 conversation templates, we generated ten examples, and for each example, three participants were asked to answer the questions, a total of 140 questionnaires answered by 420 participants.

Figure 4 shows the distribution of participants' responses to each question, outlining the alterations in perception after revealing the automated generation of conversations. Both user (Q1 & Q4) and agent (Q2 & Q5) utterances, as well as the overarching dialog (Q3 & Q6), were subjected to quality evaluation. The data reveals that prior to revealing the truth, over 66% of evaluators deemed the conversation reasonable (both "yes" answers combined), marginally reducing to 64% post-revelation (Q3 & Q6). Regarding participants' assessment of the customer's understanding of the sales assistant (Q1 & Q4), over 78% affirmed it, which reduces to 77% after the disclosure, albeit almost half of participants switched form *definitely yes* to *rather yes*. Regarding the evaluation of the sales assistant's response quality to customer inquiries (Q2 & Q5), over 60% of participants agreed that the responses were sufficient, while the disclosure incited a reduction in both *definitely yes* and *rather yes* responses to over 54%. Altogether, the responses indicate a generally positive reception of conversations generated by OpinionConv, with variation in assessment among different conversation templates. While a rough two thirds of participants agree with this outcome, more than half of participants improve their rating from *definitely not* to *rather not* considering overall reasonableness.

4.3 Participants' Comments

In both studies, participants were asked to explain their judgment in about two sentences, in case of Study 2 once in each part of the questionnaire. With respect to assessing the conversation realism, we mostly observe positive comments. For instance, before disclosing the nature of the conversation one participants commented "The customer was recommended the phone by a friend, and the sales assistant was able to give further information on the phone. Likewise the sales assistant was able to inform the customer on a drawback associated with the phone.", and after "The conversation appeared real as there was flow - i.e. the sales assistant was able to connect with what the customer said and elaborate upon it. Likewise the sales assistant was able to pick up on key details associated with the phone like the camera and OS."

However, we also observe three key concerns raised: (1) Some features are of no interest to be discussed, e.g., "Why would the person asks the sales assistant about colors? That seems out of the ordinary." (2) Some participants judge the conversations based on their personal experience with real sales assistants, e.g., "As always in marketing strategies, he [the sales assistant] was just trying to sell a phone not what he [the customer] wanted." (3) A stronger argumentation is expected by some, e.g., "While the sales assistant did respond in a way that does answer the customer's questions, their responses are not so direct and detailed as to be helpful towards the customer. For example, for the question about the screen, stating that it's 'bright

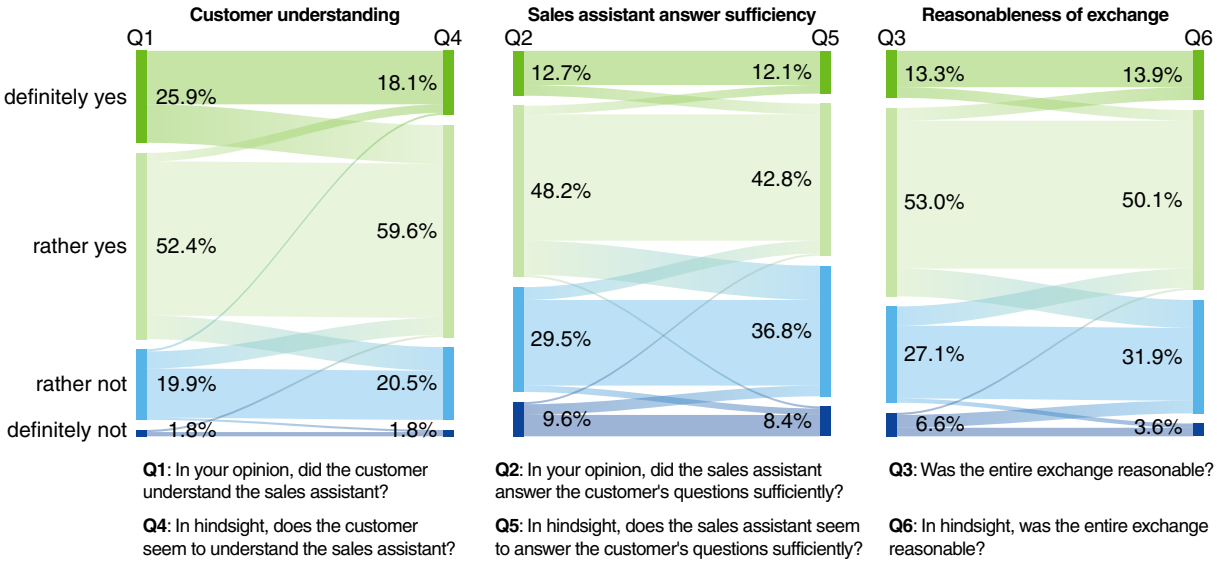


Figure 4: Evaluation results for Study 2; Questions Q1, Q2, and Q3 are asked in the first part of the questionnaire (before disclosing the conversations are generated), Q4, Q5, and Q6 are asked in the second part (after disclosure).

and good quality' would not be convincing enough for me to want to buy the product."

Reading the participants' comments and observing the results of crowd-sourced qualitative evaluations have suggested several new research directions for future work relating to common sense product knowledge and argument generation.

5 Conclusion

We introduce a OpinionConv, a new conversation generation pipeline that generates opinionated multi-turn conversations for product search. OpinionConv was mainly designed to incorporate subjective narratives into conversational product search. The pipeline presented in this work can be easily extended to different domains. Recent progress in conversational systems, such as ChatGPT and YouChat, have shown tremendous improvements in natural language dialog between humans and conversational agents. However, when it comes to holding an opinionated conversation, specifically in product search, they are still limited for lack of grounding in real-world experience about products. This motivated the design of a pipeline to control both the dialog coherence and the information to be mentioned in the utterances. However, it should be mentioned that the trade-off between a coherent conversation and a more diverse conversation needs to be further studied. In order to validate the quality of the conversations generated by OpinionConv, we conduct two extensive human evaluations. Our results confirm the conversational plausibility of the generated dialogs and reveal that people tend to exchange their personal opinions while searching for a product.

In future work, we envision customer-oriented assistant for buying products that assist customers in discussing the merits of products with a sales assistant, grounded in real-world reviews.

6 Limitations

As mentioned in Section 3.1, we focused on the *Cell Phones and Accessories* category of products. However, there is no inherent limitation of our design that prevents future work from including conversations related to other product categories.

Furthermore, an opinion is an observation or a belief that does not need to have evidence to support itself, whereas an argument requires premises. As we discussed in the Section 4.3, study participants expected to have stronger arguments in the generated conversation, rather than only expressing opinions. Therefore, future work should address this aspect utilizing argument mining techniques for generating argumentative dialogues.

Acknowledgements

This work has been partially supported by the German Federal Ministry of Education and Research (BMBF) within the Junior AI Scientists program under the reference 01-S20060. We would like to thank the anonymous reviewers for their valuable feedback.

Ethics Statement

Systems designed to influence humans via communication constitute a highly sensitive topic due to their intrinsically social nature (Stock et al., 2016). Any automated sales assistant comes along with the ethical risk of not only influencing customer opinion but doing so in ways undesired by customers, e.g., to their financial or otherwise personal disadvantage. Naturally, it is the company that deploys a manipulative sales assistance technology that is at fault, but the question of why research that may be misused in this direction has been undertaken in the first place is still pertinent.

Negotiation differs from persuasion in its goal. Negotiation strives to reach an agreement from both sides, while persuasion merely aims to change one specific person's attitude and decision (Wang et al., 2019). Most human sales assistants have no interest in deceiving customers, since that very customer may come back to complain, or not come back to buy further products. Modern marketing strategies typically involve building a trustworthy customer relationship which includes the post-purchase stage of the aforementioned customer decision process, where customer satisfaction is to be maximized. We intend our research to serve as a step towards studying the capabilities of automated sales assistance with the goal of mutually beneficial negotiation. Nevertheless, if it turns out that it is easier for technology to manipulate its users with respect to a purchase decision than to consult them for mutual benefit, this must be found out, and publicly, or else no policies against such exploits can be enforced.

Moreover, an automatic sales assistant deployed by a marketplace must be considered separately from, e.g., an automatic sales assistant deployed by an independent third party (including open source variants). We imagine that not only the former will become available in the future, but also the latter, which will be more trustworthy overall.

References

Tanvirul Alam, Akib Khan, and Firoj Alam. 2020. Punctuation restoration using transformer models for high- and low-resource languages. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 132–142.

Aishwarya Ashok, Ganapathy Natarajan, Ramez Elmasri, and Laurel Smith-Stvan. 2020. SimsterQ: A similarity based clustering approach to opinion question answering. In *Proceedings of the 3rd Workshop on e-Commerce and NLP*, pages 69–76, Seattle, WA, USA. Association for Computational Linguistics.

Keping Bi, Qingyao Ai, Yongfeng Zhang, and W Bruce Croft. 2019. Conversational product search based on negative feedback. In *Proceedings of the 28th acm international conference on information and knowledge management*, pages 359–368.

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. Subjqa: a dataset for subjectivity and review comprehension. *arXiv preprint arXiv:2004.14283*.

B. Barla Cambazoglu, Mark Sanderson, Falk Scholer, and W. Bruce Croft. 2020. A review of public datasets in question answering research. *SIGIR Forum*, 54(2):5:1–5:23.

Claire Cardie, Janyce Wiebe, Theresa Wilson, and Diane J Litman. 2003. Combining low-level and summary representations of opinions for multi-perspective question answering. In *New directions in question answering*, pages 20–27.

Eduardo Gabriel Cortes, Vinicius Woloszyn, Dante Barone, Sebastian Möller, and Renata Vieira. 2021. A systematic review of question answering systems for non-factoid questions. *Journal of Intelligent Information Systems*, pages 1–28.

F Robert Dwyer, Paul H Schurr, and Sejo Oh. 1987. Developing buyer-seller relationships. *Journal of marketing*, 51(2):11–27.

Marja Exalto, Maarten De Jong, Tim De Koning, Axel Groothuis, and Pascal Ravesteijn. 2018. Conversational commerce, the conversation of tomorrow. In *Proceedings of the 14th European Conference on Management, Leadership and Governance, ECMLG*, pages 76–83.

R. Fisher and W. Ury. 1981. *Getting to Yes: Negotiating Agreement Without Giving in*. Houghton Mifflin.

Ulrich Gnewuch, Stefan Morana, and Alexander Mädche. 2017. Towards designing cooperative and social conversational agents for customer service. In *Proceedings of the International Conference on Information Systems - Transforming Society with Digital Innovation, ICIS 2017, Seoul, South Korea, December 10-13, 2017*. Association for Information Systems.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C Lipton. 2019. Amazonqa: A review-based question answering task. *arXiv preprint arXiv:1908.04364*.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2333–2343. Association for Computational Linguistics.

Peng Jiang, Hongping Fu, Chunxia Zhang, and Zhendong Niu. 2010. A framework for opinion question answering. In *2010 6th International Conference on Advanced Information Management and Service (IMS)*, pages 424–427. IEEE.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. Improving BERT performance for aspect-based sentiment analysis. In *4th International Conference on Natural Language and Speech Processing, Trento, Italy, November 12-13, 2021*, pages 196–203. Association for Computational Linguistics.

Soo-Min Kim and Eduard Hovy. 2005. Identifying opinion holders for question answering in opinion texts. In *Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains*, pages 1367–1373.

Philip Kotler and Kevin Lane Keller. 2015. Marketing management. *New Jersey*, 143.

Siheng Li, Wangjie Jiang, Pengda Si, Cheng Yang, Yao Qiu, Jinchao Zhang, Jie Zhou, and Yujiu Yang. 2023. Enhancing dialogue generation with conversational concept flows. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1484–1495. Association for Computational Linguistics.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pages 625–635.

Samaneh Moghaddam and Martin Ester. 2011. Aqa: aspect-based opinion question answering. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 89–96. IEEE.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197.

Andrea Papenmeier, Alexander Frummet, and Dagmar Kern. 2022. "mhm..." - conversational strategies for product search assistants. In *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, March 14 - 18, 2022*, pages 36–46. ACM.

Gustavo Penha, Eyal Krikon, and Vanessa Murdock. 2022. Pairwise review-based explanations for voice product search. In *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, March 14 - 18, 2022*, pages 300–304. ACM.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics.

D.G. Pruitt. 1981. *Negotiation Behavior*. Library and Information Science. Academic Press.

James Pustejovsky and Janyce Wiebe. 2005. Introduction to special issue on advances in question answering. *Language Resources and Evaluation*, 39(2/3):119–122.

Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer.

Ohad Rozen, David Carmel, Avihai Mejer, Vitaly Mirkis, and Yftah Ziser. 2021. Answering product-questions by utilizing questions from other contextually similar products. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 242–253. Association for Computational Linguistics.

D Săvescu. 2019. Some aspects regarding negotiation in business. In *IOP Conference Series: Materials Science and Engineering*, volume 514, page 012042. IOP Publishing.

Oliviero Stock, Marco Guerini, and Fabio Pianesi. 2016. Ethical dilemmas for adaptive persuasion systems. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Veselin Stoyanov, Claire Cardie, and Janyce Wiebe. 2005. Multi-perspective question answering using the opqa corpus. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 923–930.

Leigh Thompson and Reid Hastie. 1990. Social perception in negotiation. *Organizational Behavior and Human Decision Processes*, 47(1):98–123.

Leigh L Thompson, Jiunwen Wang, and Brian C Gunia. 2010. Negotiation. *Annual review of psychology*, 61:491–515.

Mengting Wan and Julian McAuley. 2016. Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 489–498. IEEE.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

Liqiang Xiao, Jun Ma, Xin Luna Dong, Pascual Martinez-Gomez, Nasser Zalmout, Wei Chen, Tong Zhao, Hao He, and Yaohui Jin. 2021. End-to-end conversational search for online shopping with utterance transfer. *arXiv preprint arXiv:2109.05460*.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Review conversational reading comprehension. *arXiv preprint arXiv:1902.00821*.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136.

Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2021. Conversational question answering: A survey. *arXiv preprint arXiv:2106.00874*.

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16):3389.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 177–186.

Yiheng Zhou, He He, Alan W. Black, and Yulia Tsvetkov. 2019. A dynamic strategy coach for effective negotiation. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 367–378. Association for Computational Linguistics.

Dial-M: A Masking-based Framework for Dialogue Evaluation

Suvodip Dey and **Maunendra Sankar Desarkar**
Indian Institute of Technology Hyderabad, India
suvodip15@gmail.com, maunendra@cse.iith.ac.in

Abstract

In dialogue systems, automatically evaluating machine-generated responses is critical and challenging. Despite the tremendous progress in dialogue generation research, its evaluation heavily depends on human judgments. The standard word-overlapping based evaluation metrics are ineffective for dialogues. As a result, most of the recently proposed metrics are model-based and reference-free, which learn to score different aspects of a conversation. However, understanding each aspect requires a separate model, which makes them computationally expensive. To this end, we propose Dial-M, a Masking-based reference-free framework for Dialogue evaluation. The main idea is to mask the keywords of the current utterance and predict them, given the dialogue history and various conditions (like knowledge, persona, etc.), thereby making the evaluation framework simple and easily extensible for multiple datasets. Regardless of its simplicity, Dial-M achieves comparable performance to state-of-the-art metrics on several dialogue evaluation datasets. We also discuss the interpretability of our proposed metric along with error analysis.

1 Introduction

Dialogue systems research has seen massive advancements in recent years. It is not surprising to see models generating high-quality human-like meaningful responses nowadays. Despite this enormous progress, the evaluation of machine-generated dialogues remains a concern. Although many automatic metrics have been proposed, we still have to rely on human evaluation, which is tedious and costly. Thus, improving the quality of automatic dialogue evaluation is essential for the overall development of this evolving area.

The evaluation metrics for dialogue generation can be broadly divided into two classes: reference-based and reference-free. In reference-based metrics, the generated dialogue is evaluated with respect to one more reference utterance(s). The most popular reference-based metrics used in dialogue systems are standard word-overlapping based metrics like BLEU (Papineni et al., 2002), NIST (Lin and Och, 2004), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), Diversity (Li et al., 2016), and Entropy (Zhang et al., 2018b). However, these metrics have been shown to be ineffective because of the one-to-many nature of dialogues (Liu et al., 2016; Yeh et al., 2021). As a result, people started adopting learning-based referenced metrics like ADEM (Lowe et al., 2017), RUBER (Tao et al., 2017), BERT-RUBER (Ghazarian et al., 2019), PONE (Lan et al., 2020), BERTScore (Zhang* et al., 2020), BLEURT (Sellam et al., 2020), FBD (Xiang et al., 2021), Deep AM-FM (Zhang et al., 2021b), etc. However, reference-based metrics are not feasible for evaluation in an online setting where the reference response is unavailable. Also, collecting good-quality candidate responses is costly and requires human annotation. Hence, most of the recent efforts are being made in the direction of reference-free metrics.

In reference-free metrics, the generated dialogue is evaluated without any references. Here, most of the methods formulate the dialogue evaluation problem as one or more classification tasks and use the classification scores as the metric or sub-metrics. Metrics like Maude (Sinha et al., 2020) and DEB (Sai et al., 2020) learn to differentiate between correct and incorrect responses given the context. GRADE (Huang et al., 2020) and DynaEval (Zhang et al., 2021a) leverage graph-based methods, while DEAM (Ghazarian et al., 2022) relies on Abstract Meaning Representation (AMR) to evaluate dialogue coherence. MDD-Eval (Zhang et al., 2022) addresses the issue of multi-domain evaluation by introducing a teacher evaluator. The quality of a generated dialogue depends on multiple factors such as understandability, informativeness, coherence, etc. Metrics like USR (Mehri and Eskenazi, 2020b), USL-H (Phy et al., 2020), FED

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 77–84
September 11–15, 2023. ©2023 Association for Computational Linguistics

(Mehri and Eskenazi, 2020a), HolisticEval (Pang et al., 2020), D-score (Zhang et al., 2021c), QualityAdapt (Mendonca et al., 2022) learn to compute various sub-metrics and then combine them to give a final score. For further improvement, IM^2 (Jiang et al., 2022) combines multiple metrics that are good at measuring different dialog qualities to generate an aggregate score. However, modeling different sub-metric requires a separate model or adapter, increasing the computational cost. Moreover, the decision boundary of the classification-based metrics depends on the quality of negative sampling (Lan et al., 2020), inducing training data bias.

In this work, we aim to address these issues by proposing **Dial-M** [1] , a **M**asking-based reference-free framework for **Dial**ogue evaluation. The central idea of Dial-M is to mask the keywords of the current utterance and use the cross-entropy loss while predicting the masked keywords as the evaluation metric. Doing so avoids the requirement for multiple models and negative sampling, making the framework simple and easily extensible to multiple datasets. The keywords in the current utterance are obtained in an unsupervised manner. We show that Dial-M achieves comparable performance to various state-of-the-art metrics on several evaluation datasets, especially knowledge-grounded datasets like Topical-Chat. We observe that Dial-M can capture different aspects of a conversation. We also show that the Dial-M score can be interpreted by inspecting the masked words, which enables the scope for error analysis.

2 Dial-M Framework

Let $D = \{u_1, u_2, ...\}$ be a multi-turn conversation where u_i represents the utterance at turn i. Let $C = \{c_1, c_2, ...\}$ be the set of conditions where c_i denotes the condition that is used to generate the u_i. The condition can be knowledge, fact, persona, or other relevant information based on the task/dataset. The condition can be absent as well for conversations like chit-chat. For a given turn t, the objective of dialogue generation is to generate u_t given $D_{<t}$ i.e. $\{u_1, ..., u_{t-1}\}$ and C_t i.e. $\{c_1, ..., c_t\}$. The goal of the Dial-M framework is to learn a scoring function $f : (D_{<t}, u_t, c_t) \rightarrow s$ where $s \in \mathbb{R}$ denotes the quality of the generated response (u_t) given $D_{<t}$, u_t and c_t (if available). The details of our proposed framework are described as follows.

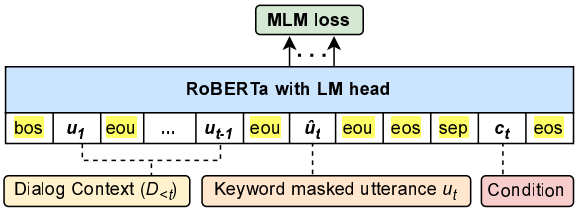


Figure 1: Dial-M Finetuning task.

2.1 Pre-Training

We pre-train the RoBERTa (Liu et al., 2020) model with Masked Language Modeling (MLM) task on various conversational datasets. For a given conversation, the utterances are concatenated with a special token (eou). We consider only dialogue history for this MLM task, i.e., fact, persona, or any other conditions are ignored. We use RoBERTa-base[2] with Language Model (LM) head as our base model. The masking probability is set to 0.15.

2.2 Finetuning

As discussed earlier, state-of-the-art evaluation metrics depend on multiple models to compute the final evaluation score. The main motivation for this work is to develop a lightweight alternative that can be trained using a single model and avoids negative sampling. To achieve this goal, we use a keyword masking task to finetune the pre-trained RoBERTa model (as shown in Fig. 1). For a given turn t, we construct the RoBERTa input as text pair (D_t, c_t) or simply D_t if the condition is absent. The utterances of D_t are concatenated with the special token eou. Let K_t be the set of keywords in the current utterance u_t. Let $\hat{u}_t$ be the representation of u_t after masking the tokens associated with K_t. Then we formulate our denoising task as predicting the masked tokens of u_t given $D_{<t}$, $\hat{u}_t$, and c_t (if available). We use YAKE! (Campos et al., 2018, 2020), an unsupervised feature-based keyword extraction algorithm, to find the keywords. Further detail regarding YAKE! is provided in Appendix A.1. While finetuning, we ignore the utterances with no keywords.

In previous works, the standard MLM task has been used as a proxy for fluency or likability (Mehri and Eskenazi, 2020b; Pang et al., 2020). In contrast, focusing on the keywords helps to capture other important aspects like understandability, naturalness, and informativeness, which we later justify using the results of Table 2. Moreover, formulating the problem as an MLM task and the

[1] Code is available at github.com/SuvodipDey/Dial-M

[2] huggingface.co/roberta-base

Row	Metric	USR-Topical		USR-Persona		PredictiveEngage		HolisticEval	
		P	S	P	S	P	S	P	S
1	BLEU-4 (Papineni et al., 2002)	0.216	0.296	0.135	0.090*	-	-	-	-
2	METEOR (Banerjee and Lavie, 2005)	0.336	0.391	0.253	0.271	-	-	-	-
3	BERTScore (Zhang* et al., 2020)	0.298	0.325	0.152	0.122*	-	-	-	-
4	BERT-RUBER (Ghazarian et al., 2019)	0.342	0.348	0.266	0.248	-	-	-	-
5	MAUDE (Sinha et al., 2020)	0.044*	0.083*	0.345	0.298	0.104	0.060*	0.275	0.364
6	DEB (Sai et al., 2020)	0.180	0.116	0.291	0.373	0.516	0.580	0.584	**0.663**
7	GRADE (Huang et al., 2020)	0.200	0.217	0.358	0.352	**0.600**	**0.622**	**0.678**	**0.697**
8	HolisticEval (Pang et al., 2020)	-0.147	-0.123	0.087*	0.113*	0.368	0.365	**0.670**	**0.764**
9	USR (Mehri and Eskenazi, 2020b)	**0.412**	**0.423**	**0.440**	0.418	**0.582**	**0.640**	0.589	0.645
10	USL-H (Phy et al., 2020)	0.322	0.340	**0.495**	**0.523**	**0.688**	**0.699**	0.486	0.537
11	IM^2-overall (Jiang et al., 2022)	**0.462**	**0.461**	0.438	**0.431**	-	-	-	-
12	Dial-M (ours)	**-0.432**	**-0.463**	**-0.464**	**-0.486**	-0.570	-0.592	**-0.590**	-0.598
	Ablation Study								
13	with Random Masking	-0.320	-0.316	-0.359	-0.345	-0.549	-0.547	-0.607	-0.630
14	w/o Pre-training	-0.391	-0.429	-0.443	-0.489	-0.556	-0.586	-0.567	-0.583
15	w/o Finetuning	-0.290	-0.282	-0.288	-0.258	-0.550	-0.549	-0.592	-0.613
16	w/o Pre-training and Finetuning	-0.248	-0.248	-0.154	-0.144	-0.508	-0.535	-0.540	-0.552

Table 1: Result comparison on various datasets with top-3 scores highlighted in bold. P and S indicate Pearson and Spearman's coefficients, respectively. All values are statistically significant to $p < 0.05$, unless marked by *.

inclusion of dialogue conditions provide the flexibility to extend the framework to different kinds of conversational datasets without any additional annotation. For example, if the output of database queries (like system-act annotation in MultiWOZ (Budzianowski et al., 2018)) is converted into a natural sentence and used as the condition, Dial-M can be utilized for task-oriented conversation.

2.3 Dial-M Metric

To evaluate a generated response u_t, we first extract the set of keywords (K_t) from u_t. For each keyword in K_t, we mask the associated tokens and compute the cross-entropy loss to predict them using the finetuned RoBERTa model. We use the mean of these cross-entropy losses as our evaluation score. Let $k_{t,j}$ be the j^{th} keyword in K_t. Let $T_{t,j}$ be the set of tokens associated with the word $k_{t,j}$. Let $\hat{u}_{t,j}$ be the representation of u_t after masking the tokens $T_{t,j}$. Then the evaluation score (s) of the Dial-M metric is defined as:

$$s = \frac{1}{|K_t|} \sum_{j=1}^{|K_t|} \left(\frac{1}{|T_{t,j}|} \sum_{y \in T_{t,j}} -\log p(y | D_{<t}, \hat{u}_{t,j}, c_t) \right) \quad (1)$$

We use YAKE! to extract the keywords. Since YAKE! is unsupervised and feature-based, it may not find all the relevant keywords. Thus, we also consider the words tagged with specific parts-of-speech (POS) as keywords to increase coverage. If no keyword is found in u_t, we consider all words as keywords. We observed that the utterances with no keywords are generally short and generic responses. As we are using cross-entropy loss, a lower score denotes a better response quality and vice-versa.

3 Experimental Setup

We use DailyDialog (Li et al., 2017), Persona-Chat (Zhang et al., 2018a), Wizard-of-Wikipedia (Dinan et al., 2019), and Topical-Chat (Gopalakrishnan et al., 2019)) for both pre-training and finetuning Dial-M. We show our results on USR (Mehri and Eskenazi, 2020b), PredictiveEngage (Ghazarian et al., 2020), and HolisticEval (Pang et al., 2020) datasets for dialogue evaluation. USR is based on Topical-Chat and Persona-Chat, while PredictiveEngage and HolisticEval are based on DailyDialog. We call the Topical-Chat and Persona-Chat datasets of USR as USR-Topical and USR-Persona, respectively. We use spaCy (Honnibal and Montani, 2017) POS tagger along with YAKE! to find the keywords during evaluation. We analyzed the POS tags of co-occurring words in response (u_t) knowledge (c_t) pair in Topical-Chat train data and selected the most frequent POS tags (*NN*, *NNP*, *NNS*, *JJ*, *CD*, *VB*, *VBN*, *VBD*, *VBG*, *RB*, *VBP*, *VBZ*, *NNPS*, and *JJS*) for our purpose. The rest of the details are provided in Appendix A.2.

4 Result and Analysis

Table 1 compares Dial-M with different metrics on four dialogue evaluation datasets. In Dial-M, a lower score is better, resulting in a negative correlation with the human scores. In Table 1,

Sub-Metric	Metric	USR-Topical		USR-Persona	
		P	S	P	S
Understandable	USR	0.29	0.32	0.12	0.13
	Dial-M	**-0.35**	**-0.40**	**-0.18**	**-0.14**
Natural	USR	0.28	0.30	0.19	0.24
	Dial-M	**-0.37**	**-0.40**	**-0.28**	**-0.28**
Maintains Context	USR	**0.42**	0.38	**0.61**	**0.53**
	Dial-M	-0.37	**-0.40**	-0.40	-0.39
Engaging	USR	**0.46**	**0.46**	0.03	0.02
	Dial-M	-0.43	-0.45	**-0.33**	**-0.34**
Uses Knowledge	USR	0.32	0.34	**0.40**	0.32
	Dial-M	**-0.35**	**-0.37**	-0.34	**-0.37**

Table 2: Correlation with different sub-metrics.

we can first observe that Dial-M outperforms the reference-based metrics (Rows 1-4). Secondly, it achieves comparable performance to state-of-the-art reference-free metrics. Thirdly, Dial-M performs relatively better for knowledge-grounded dialogues (USR-Topical and USR-Persona) than chit-chat (PredictiveEngage and HolisticEval). This is because the keywords of the current utterance generally align with context and the selected knowledge, which may not be the case for chit-chat. Nevertheless, the correlation values of Dial-M are close to the top-3 metrics for the chit-chat datasets. Table 2 shows the correlation of Dial-M with different sub-metrics on the USR dataset. Dial-M maintains a moderate correlation with all the sub-metrics, which justifies the utility of keyword masking in capturing different aspects of a conversation.

Rows 13-16 of Table 1 shows the result of our ablation study. In Row 13, we randomly mask 15% words of u_t instead of having a principled approach of identifying keywords and masking them while finetuning. We can observe that random masking degrades the performance except for HolisticEval. A similar observation can be seen in Row 15, where we do not use any finetuning i.e. the evaluation score is computed using the pre-trained model (described in Section 2.1). This conflicting behavior on HolisticEval can be due to the random chit-chat conversations in the dataset. In Row 14, we do not pre-train RoBERTa on dialogue datasets, which reduces the performance and shows the importance of pre-training. Row 16 displays the result with no training i.e. the scores are computed using the base RoBERTa model, resulting in poor performance.

5 Discussion

In this section, we discuss the interpretability and error analysis of Dial-M scores. Table 3 shows an illustrative example of Dial-M evaluation on a USR-Persona conversation. Let us first analyze the good cases (Responses 1-3). We can observe that Dial-M has given a low score to Response 1 in comparison to Responses 2 and 3, which correlates with the human scores. The reason for this low score can be deduced by looking at the masked words of Response 1, which are connected to both context and condition (persona). In Response 2, masked words like *red* and *blue* are out of context, resulting in a higher Dial-M score. The masked words of Response 3 are slightly out of context in comparison to Response 1, resulting in an average score that is reflected in the human scores as well. Let us now analyze Response 4, which can be treated as a bad case because Dial-M finds it superior even though it is not the best response. The possible reason for the lower human score of Response 4 than Response 1 is the usage of "*i get up early everyday*", which is not mentioned in the persona. However, the phrase "*i get up early*" is very common. Since Dial-M is pre-trained on MLM task, the prediction of "*early*" given "*i get up*" becomes easy, resulting in the lowest score. This is how we can interpret and perform error analysis of the Dial-M scores by inspecting the masked words. We observed that Dial-M generally assigns a low score to short, generic, and frequently used sentences where the masked word can be easily predicted from its neighbors. We aim to address this issue in our future work.

Context ($D_{<t}$)	"hey . where are you from ? i'm from a farm in Wisconsin", "i love ice cream what is your favorite ? mine is chocolate", "mine is mint chocolate chip"
Condition (c_t) (Persona)	my wife and kids are the best. my favorite ice cream flavor is chocolate. i've three children. i'm a plumber. i love going to the park with my three children and my wife.
Response 1	my three ***kids love mint chocolate chip*** !
Human Score	Overall score: [5, 5, 5], Average: 5.0
Dial-M Score	0.1399
Response 2	i ***like*** the ***color red*** . i ***like*** the ***color blue*** .
Human Score	Overall score: [1, 2, 2], Average: 1.67
Dial-M Score	4.3131
Response 3	i ***like chocolate chip cookies***
Human Score	Overall score: [3, 4, 4], Average: 3.67
Dial-M Score	2.4582
Response 4	i get up ***early everyday*** and ***eat ice cream***
Human Score	Overall score: [3, 4, 5], Average: 4.0
Dial-M Score	0.1034

Table 3: Illustrative example of Dial-M evaluation on USR-Persona. Masked words are shown in bold italics.

6 Conclusion

In conclusion, we propose Dial-M, a masking-based reference-free framework for dialogue eval-

uation. We mask the keywords of the current utterance and use the cross-entropy loss while predicting the masked keywords as the evaluation metric. Formulating the problem as a keyword masking task avoids the requirement for multiple models and negative sampling, making the framework simple and easily extensible to multiple datasets. Dial-M achieves comparable performance to state-of-the-art metrics on several dialogue evaluation datasets. We also show the utility of keyword masking in capturing various aspects of a conversation and discuss the interpretability and error analysis of Dial-M scores. We want to explore better keyword extraction strategies in future work. We also want to investigate better techniques to handle the cases where no keywords are detected in the current utterance.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. YAKE! Collection-Independent Automatic Keyword Extractor. In *Advances in Information Retrieval*, pages 806–810, Cham. Springer International Publishing.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7789–7796.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Zhihua Jiang, Guanghui Ye, Dongning Rao, Di Wang, and Xin Miao. 2022. IM^2: an interpretable and multi-category integrated metric framework for automatic dialogue evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11091–11103, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Trans. Inf. Syst.*, 39(1).

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

John Mendonca, Alon Lavie, and Isabel Trancoso. 2022. QualityAdapt: an automatic dialogue quality estimation framework. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 83–90, Edinburgh, UK. Association for Computational Linguistics.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, Siddharth Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI Conference on Artificial Intelligence*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Jiannan Xiang, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. 2021. Assessing dialogue systems with distribution distances. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2192–2198, Online. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021a. DynaEval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5676–5689, Online. Association for Computational Linguistics.

Chen Zhang, Luis D'Haro, Thomas Friedrichs, and Haizhou Li. 2022. Mdd-eval: Self-training on augmented data for multi-domain dialogue evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:11657–11666.

Chen Zhang, Luis Fernando D'Haro, Rafael E. Banchs, Thomas Friedrichs, and Haizhou Li. 2021b. *Deep AM-FM: Toolkit for Automatic Dialogue Evaluation*, pages 53–69. Springer Singapore, Singapore.

Chen Zhang, Grandee Lee, Luis Fernando D'Haro, and Haizhou Li. 2021c. D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2502–2516.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

A Appendix

A.1 YAKE!

YAKE! (Campos et al., 2018, 2020) is a lightweight unsupervised method for automatic keyword extraction. It is a feature-based system for extracting keywords from single documents, which supports texts of different sizes, domains, or languages. YAKE! builds upon unsupervised textual features (like casing, word frequency, word position, etc.) to find the most important keywords of a text, making it applicable to documents written in many different languages without the need for external knowledge. Thus, YAKE! does not rely on dictionaries/thesauri and requires no training against any corpora. However, it performs well and significantly outperforms other unsupervised methods on texts of different sizes, languages, and domains.

A.2 Implementation Details

We implemented Dial-M using PyTorch and Huggingface (Wolf et al., 2020) libraries in Python 3.10. All the experiments are performed on two devices of Nvidia DGX server with 32GB of memory each. The number of parameters in our pre-trained and finetuned model is 125M, the same as the RoBERTa-base model. The whole vocabulary is considered while predicting the tokens for the MLM tasks (both pre-training and keyword masking). The pre-training MLM task is trained for 30 epochs with a batch size 64 on a single GPU. The finetuning task is trained for 10 epochs with a batch size of 96 on two GPUs. We used AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate 1e-5 for both the training. The best model is selected based on minimum validation loss. The results of the other evaluation metrics in Table 1 and Table 2 are taken from the following references - Yeh et al. (2021); Mehri and Eskenazi (2020b); Jiang et al. (2022).

Fig. 2 shows the parts of speech (POS) of the co-occurring words in the response and corresponding knowledge in Topical-Chat (Gopalakrishnan et al., 2019) training data. We use the most frequent POS tags (*NN*, *NNP*, *NNS*, *JJ*, *CD*, *VB*, *VBN*, *VBD*, *VBG*, *RB*, *VBP*, *VBZ*, *NNPS*, and *JJS*) to mask the keywords during evaluation.

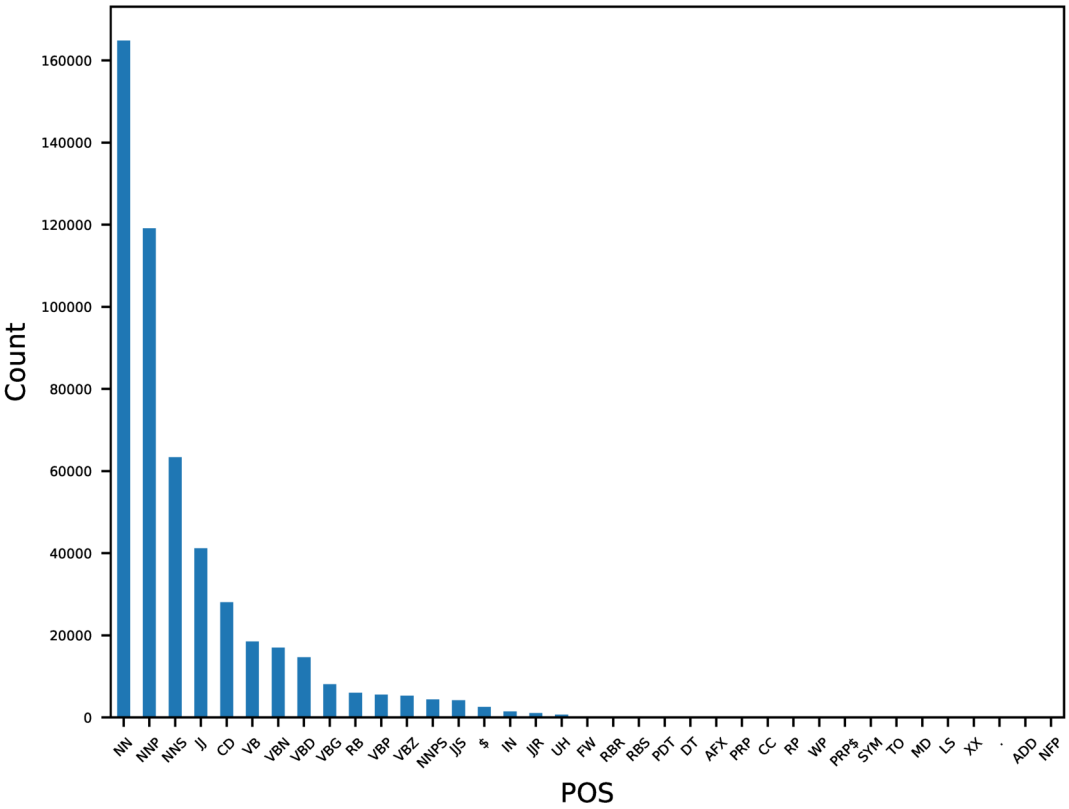


Figure 2: POS analysis on Topical-Chat train data.

From Chatter to Matter: Addressing Critical Steps of Emotion Recognition Learning in Task-oriented Dialogue

Shutong Feng, Nurul Lubis, Benjamin Ruppik, Christian Geishauser, Michael Heck, Hsien-chin Lin, Carel van Niekerk, Renato Vukovic, and Milica Gašić
Heinrich Heine University Düsseldorf, Germany
{fengs,lubis,ruppik,geishaus,heckmi,linh,niekerk,revuk100,gasic}@hhu.de

Abstract

Emotion recognition in conversations (ERC) is a crucial task for building human-like conversational agents. While substantial efforts have been devoted to ERC for chit-chat dialogues, the task-oriented counterpart is largely left unattended. Directly applying chit-chat ERC models to task-oriented dialogues (ToDs) results in suboptimal performance as these models overlook key features such as the correlation between emotions and task completion in ToDs. In this paper, we propose a framework that turns a chit-chat ERC model into a task-oriented one, addressing three critical aspects: data, features and objective. First, we devise two ways of augmenting rare emotions to improve ERC performance. Second, we use dialogue states as auxiliary features to incorporate key information from the goal of the user. Lastly, we leverage a multi-aspect emotion definition in ToDs to devise a multi-task learning objective and a novel emotion-distance weighted loss function. Our framework yields significant improvements for a range of chit-chat ERC models on EmoWOZ, a large-scale dataset for user emotion in ToDs. We further investigate the generalisability of the best resulting model to predict user satisfaction in different ToD datasets. A comparison with supervised baselines shows a strong zero-shot capability, highlighting the potential usage of our framework in wider scenarios.

1 Introduction

Emotion recognition in conversations (ERC) is a crucial task in conversational artificial intelligence research because it lays the foundation for affective abilities in computers such as empathetic response generation (Picard, 1997). Over years, it has shown values in downstream applications such as opinion mining (Colneric and Demšar, 2020) and human-like dialogue modelling (Zhou et al., 2018).

Dialogue systems can be broadly categorised into two categories: (1) chit-chat or open-domain systems and (2) task-oriented dialogue (ToD) systems. Chit-chat systems are set up to mimic human behaviours in a conversation (Jurafsky and Martin, 2009). There are no particular goals associated with the dialogue and the system aims to keep the user engaged with natural and coherent responses. On the other hand, ToD systems are concerned with fulfilling user goals, such as information retrieval for hotel booking (Young, 2002).

```
A: We have a holiday next week, don't we ?
B: Yes, on Monday .
A  What're you going to do ?
B: I'm probably going to spend the day looking at cars.
   ...
```

(a) Chit-chat dialogue from Li et al. (2017)

```
U: I am planning a vacation and really could use some
   help finding a good place to stay in town. I've
   never been to Cambridge before.
S: which side of town do you prefer and what is the
   price range?
U: It doesn't matter. What do you recommend?
S: alexander bed and breakfast is a guest house in the
   centre area. Would you like to book a room?
   ...
```

(b) Task-oriented dialogue from Budzianowski et al. (2018)

Figure 1: Comparison of dialogues about holiday in chit-chat dialogues and task-oriented dialogues.

Recently, the difference between chit-chat and ToD systems have been blurred by the utilisation of pre-trained language models as back-bone to both types of systems. However, emotions in ToDs and chit-chat dialogues play different roles and are therefore expressed differently (Feng et al., 2022). This highlights the need for dedicated emotion modelling methods for each system.

As illustrated in Figure 1, in chit-chat dialogues, speakers make use of emotions to facilitate communication by, for example, raising empathy as a result of emotion-eliciting situations or topics. On the other hand, emotions in ToDs are centred around the user's goal, and therefore emotion cues lie in both the user's wording and the task performance.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 85–103
September 11–15, 2023. ©2023 Association for Computational Linguistics

While many large-scale corpora for emotions in chit-chat dialogues exist (Busso et al., 2008; McKeown et al., 2012; Lubis et al., 2015; Li et al., 2017; Zahiri and Choi, 2018), there are considerably fewer resources for emotions in ToDs. EmoWOZ, which evolved from MultiWOZ, a widely used ToD dataset, is one notable exception (Feng et al., 2022). It contains a novel emotion description that is designed for ToDs and inspired by the Ortony-Clore-Collins (OCC) model (Ortony et al., 1988). Emotion is described in terms of three aspects: **valenced** (positive or negative) reactions towards **elicitors** (operator, user, or event) in a certain **conduct** (polite or impolite). However, due to the nature of ToDs, the occurrence of some emotions (e.g. users expressing feelings about their situations) are very rare, leading to a class imbalance in the corpus.

Similarly, advancements on the ERC task are mainly focused on chit-chat dialogues, involving an array of diverse factors from speaker personality (Majumder et al., 2019) to commonsense knowledge (Ghosal et al., 2020). Nevertheless, since these models are designed for chit-chat dialogues, they overlook how emotions are triggered and expressed with respect to goal completion in task-oriented context. The work of Devillers et al. (2003) is among one of the earliest and very few to address emotion detection in ToDs but uses generic unigram models instead of dedicated approaches.

In this work, we tackle critical steps of ERC in ToDs from three angles: the data, the features, and the learning objective. In particular,

Data: we address the poor ERC performance of particularly rare emotions in ToDs via two strategies of data augmentation (DA),

Features: we leverage dialogue state information and sentiment-aware textual features,

Objective: we exploit the three aspects of emotions, namely valence, elicitor, and conduct, in two ways: as a multi-task learning (MTL) objective and to define a novel emotion-distance-weighted loss (*EmoDistLoss*).

To the best of our knowledge, our work is the first to provide dedicated methods for emotion recognition in ToDs. Our experiments and analyses show that our framework leads to significant improvements for a range of chit-chat ERC models when evaluated on EmoWOZ.

We further investigate the generalisability of the best resulting model to predict user satisfaction in various ToD datasets under zero-shot transfer. Our model achieves comparable results as supervised baselines, demonstrating strong zero-shot capability and potential to be applied in wider scenarios.

2 Related Work

2.1 ERC Datasets

Early work on ERC relied on small scale datasets (Busso et al., 2008; McKeown et al., 2012; Lubis et al., 2015). More recently, a few large-scale datasets have been made available to the research community. They contain dialogues from emotion-rich and spontaneous scenarios such as daily communications (Li et al., 2017) and situation comedies (Zahiri and Choi, 2018).

For ToDs, the majority of available datasets address only one particular aspect of emotions such as sentiment polarity (Saha et al., 2020; Shi and Yu, 2018), user satisfaction (Schmitt et al., 2012; Sun et al., 2021), and politeness (Hu et al., 2022; Mishra et al., 2023). For more fine-grained emotions, Singh et al. (2022) constructed EmoInHindi for emotion category and intensity recognition in mental health and legal counselling dialogues in Hindi, and Feng et al. (2022) released EmoWOZ, which concerns user emotions in human-human and human-machine in information-seeking dialogues. Among these datasets, EmoWOZ has the largest scale, accompanied with a label set tailored to the task-oriented scenario.

2.2 Data Augmentation (DA)

DA is an effective approach to improve model performance by improving data diversity without explicitly collecting more data. While textual DA can be performed in the feature space via interpolation and sampling (Kumar et al., 2019), it is commonly performed in the data space for controllability. Rule-based methods involve operations such as insertion and substitution (Wei and Zou, 2019). While they are easy to implement, the diversity in augmented samples depends on the complexity of the rules. On the contrary, model-based methods are more scalable. These typically include the use of language models (Jiao et al., 2020), translation models (Xie et al., 2020a), and paraphrasing methods (Hou et al., 2018).

Additional training samples can also be obtained from unlabelled data via weak supervision (Ratner et al., 2017). To generate the automatic labels, a single model or an ensemble of models may

be used. This method can be interpreted as self-augmentation (Xu et al., 2022), self-training (Xie et al., 2020b), or distillation (Radosavovic et al., 2017).

DA has also been also deployed in ToD modelling. Hou et al. (2018) generated samples by paraphrasing delexicalised utterances. Gritta et al. (2021) conceptualised ToDs into transitional graphs and generate new dialogue paths by sampling. Heck et al. (2022) proposed a weak supervision framework to address the lack of fine-grained span labels for dialogue state tracking. DA for emotions in ToDs requires careful considerations to avoid emotion mismatch and is not yet explored.

2.3 ERC Models and Features

Text-based ERC is in essence a text classification problem with an emphasis on contextual modelling. Poria et al. (2017) proposed a recurrent neural network (RNN) for multimodal ERC. The follow-up work of Majumder et al. (2019) considered speaker-specific context. ERC performance has been continuously improved by techniques such as incorporating external knowledge (Ghosal et al., 2020) and contrastive learning (Song et al., 2022).

Sentiment-aware Embeddings Word-vector embeddings tailored for a particular natural language processing task can effectively improve the performance for that task (Naseem et al., 2021). In a similar vein, Tang et al. (2014) incorporated sentiment classification objectives in the training of the word embedding model of Collobert and Weston (2008) specifically for sentiment analysis. Yu et al. (2017) refined static word embeddings with the aid of a sentiment lexicon. Later, many sentiment-aware variants of pre-trained language models were obtained by incorporating sentiment-related objectives in training (Xu et al., 2019; Yin et al., 2020; Zhou et al., 2020). They successively achieved state-of-the-art performance in sentiment analysis tasks among language representation models.

2.4 Learning Objectives for ERC Models

ERC is often considered a single-label sequential classification problem. Using softmax cross-entropy loss has been the norm in the training of deep learning ERC models for categorical emotions (Poria et al., 2017; Zhong et al., 2019; Ghosal et al., 2020; Kim and Vossen, 2021) or quantised emotion dimensions (Cerisara et al., 2018; Wang et al., 2020). However, this simplistic cross-entropy loss ignores the inter-class relations and output probabilities on incorrect classes.

Chen et al. (2019) proposed to suppress the output probabilities of incorrect classes equally while minimising the standard cross-entropy loss. Hou et al. (2016) proposed squared earth mover's distance to penalise the misclassifications according to a ground distance matrix that quantifies the dissimilarities between classes for image age estimation and aesthetics estimation.

Although highly suitable for emotions, learning from misclassifications is rarely considered because the distance between emotion classes is hard to quantify. Therefore, we propose to leverage the structured label definition of EmoWOZ to model inter-class similarity.

Multi-task Learning (MTL) is a technique for learning tasks in parallel using a shared representation. It aims to improve generalisation by using the information in training signals of related tasks as an inductive bias (Caruana, 1997). In emotion recognition, auxiliary tasks include topic classification (Wang et al., 2020) and personality traits (Li et al., 2021). When co-labels are not available, it is also possible to leverage aspects of emotion for additional labels such as valence-arousal (Kim et al., 2017). In this work, we exploit the valence-elicitor-conduct labels in EmoWOZ for MTL.

3 Background

3.1 User Emotion Recognition

We formulate the task as recognising one emotion class e_t from a set of n discrete emotions $E = \{e^1, e^2, ..., e^n\}$ in the user turn u_t, given a dialogue history $H_t = [u_t, s_{t-1}, u_{t-1}, ..., s_1, u_1]$, where s denotes system turns and u denotes user turns. Unlike existing chit-chat ERC models, which are often built for static analysis on the dialogue as a whole, real-time ERC in ToDs does not consider future utterances in dialogue.

3.2 User Satisfaction Prediction

User satisfaction prediction aims to predict one satisfaction level c_t from a set of m discrete levels $C = \{c^1, c^2, ..., c^m\}$ in the user turn u_t, given all previous turns $P_t = [s_{t-1}, H_{t-1}]$. This task differs from ERC in that the user turn u_t is not available as a part of model input. Since user satisfaction is highly correlated with the valence aspect in user emotion, this task can also be viewed as

user emotion prediction. This is an important task in building ToD systems and has been used for user simulation and system evaluation (Sun et al., 2021).

4 Emotion Recogniser for Task-oriented Dialogues (ERToD)

In this section, we propose our ERToD framework that adapt chit-chat ERC models to the task-oriented domain, as illustrated in Figure 2.

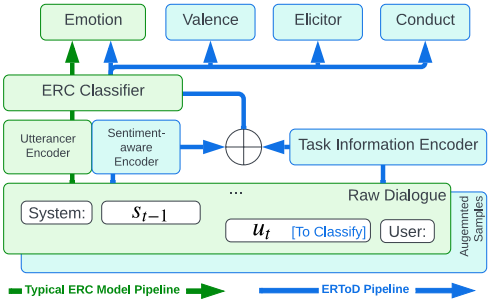


Figure 2: Our proposed ERToD Framework.

4.1 Data Augmentation

Unlike emotions in chit-chat dialogues, resources for emotions in ToDs are very limited. In addition, the data scarcity not only lies in the lack of linguistic diversity but also in the limited domains and actions in which emotions are expressed.

In ToDs, user's emotional expressions have different degrees of connection to the dialogue task. For example, a user can express dissatisfaction towards the system by pointing out the system's mistake. In such a case, simply replacing or paraphrasing the user's utterance based on emotion can potentially break the consistency of the task flow in the context. Such emotions are *context-dependent*.

On the other hand, *context-independent emotions* are expressed without any connection to the user goal, such is the case with abusive utterances. Due to the lack of connection, a simple replacement with a different abusive sentence can fit into the context well without impairing the consistency of task flow in the dialogue.

To obtain augmented samples with meaningful and coherent context, we adopt two different strategies of DA according to the degree of context dependency of emotional expressions.

Context-independent Emotions To augment samples for a target emotion e, we select a user utterance u' with the equivalent label from other dialogue datasets. We then use it to replace the user utterance u_t having label e in the training data while keeping the original context $[s_{t-1}, u_{t-1}, ..., s_1, u_1]$. The new sample is obtained as $H'_t = [u', s_{t-1}, u_{t-1}, ..., s_1, u_1]$.

Context-dependent Emotions We first sample a pool of unlabelled candidate dialogues $H'_t = [u'_t, s'_{t-1}, u'_{t-1}, ..., s'_1, u'_1]$ from other ToD datasets. We train a classifier with an uncertainty estimator to identify the emotion label e_t of the user utterance u_t and its confidence in each candidate:

$$p(e_t), \text{conf}(e_t) = \text{UncertaintyClassifier}(H'_t) \quad (1)$$

The candidate is selected for emotion e_t only if $\text{conf}_t(e)$ is above a confidence threshold θ.

4.2 Task Information Encoder

We use a dialogue state tracker (DST) to determine the status of goal completion at each turn. In ToDs, the dialogue state describes the system's understanding of the the user's goal up to that point in the dialogue (Young et al., 2010). It encodes dialogue progress in an abstractive manner.

Here as a proof of concept, we use an ontology-dependent DST, which means the concepts that the system can talk about are pre-determined. While we can eliminate the ontology dependency by, for example, using an ontology-independent DST and extracting task features from dialogue state description in natural language, this goes beyond the scope of this work. The DST takes the dialogue history to determine SemDS_t, the current dialogue state in semantic form. It is stored as a dictionary that records slots and filled values. SemDS_t is then converted into a vector of 0/1's, indicating whether a particular slot has been filled.

$$V_t = \text{Vectoriser}(\text{SemDS}_t) \quad (2)$$

To account for the change of dialogue state, which depicts how the system performs locally, we concatenate dialogue states of three consecutive turns to obtain a contextual dialogue state vector.

$$\widetilde{V}_t = V_t \oplus V_{t-1} \oplus V_{t-2} \quad (3)$$

$\text{V}_{t\leq 0}$ are zero vectors, representing the state before the dialogue starts. $\widetilde{V}_t$ is then fed into a trainable fully connected (FC) layer.

$$S_t = \text{FC}(\widetilde{V}_t) \quad (4)$$

Feature Fusion for Emotion Classification For a chit-chat ERC model with an arbitrary utterance encoder, $R_t = \text{Encoder}(H_t)$, i.e. R_t is the encoded representation of the dialogue history H_t. The utterance encoder is replaced with a sentiment-aware

encoder in our framework (see Figure 2).

The utterance and the task information encodings are fused via concatenation and fed into the emotion classifier. The output probability of all emotion classes in utterance u_t is given by:

$$p_t = \text{Softmax}(\text{Classifier}(R_t \oplus S_t)) \quad (5)$$

4.3 Learning Objectives

4.3.1 Emotion-Distance Weighted Loss

Emotion classification is a very challenging task due to the subjectivity in the perception of emotion. Since some emotions are more similar to each other than others, it may be advantageous to distinguish marginally wrong recognitions (satisfied vs excited) from extremely wrong ones (satisfied vs dissatisfied). Furthermore, different misclassifications can elicit different user reactions to the dialogue agent. For example, perceiving satisfaction when the user is neutral may or may not annoy the user, but accusing the user of abusive behavior by mistake is a serious offense to the user. Therefore, it is intuitive to penalise misclassifications according to (1) the distance from the label and (2) output probabilities on incorrect labels.

Defining the Emotion Distance Since emotion labels in EmoWOZ are defined in three aspects, we can define the distance between emotion labels in terms of their distance on each aspect. A matrix D is defined where each element $D(i, j)$ is a vector containing the distance between emotion label i and j in each of three aspects (valence, elicitor, and conduct). The matrix D is symmetric with vector-valued entries.

$$D(i, j) = [d_{val}(i, j), d_{eli}(i, j), d_{con}(i, j)] \quad (6)$$

The final distance is obtained by the sum of the distance in each aspect, followed by an addition of 1 and smoothing with the log operator. The addition of 1 ensures that the log distance is still 0 for identical labels.

$$\widetilde{D}(i, j) = \log(\text{sum}(D(i, j)) + 1) \quad (7)$$

Considering Misclassification Probabilities For each sample including the dialogue history H_t, we look at the softmax output from the model.

$$p_t = \text{Classifier}(H_t) \quad (8)$$

We aim to minimise the probability of each misclassification $p_t(e = e_i)$ where $e_i \neq \text{label}_t$. This is done by maximising $1 - p_t(e = e_i)$, the probability of the utterance *not* being wrongly recognised as e_i. We then calculate the log of this probability so that in the case of a perfectly correct recognition, the penalty from misclassification will be 0.

$$f(p_t) = \log(1 - p_t) \quad (9)$$

Obtaining Weights for Misclassifications We obtain the relevant row in matrix D that contains the distance between each emotion and the ground-truth label j of utterance u_t, followed by a normalisation to obtain a vector $w_{t,j}$ of normalised emotion-distance weights for all emotions.

$$o_{t,j} = \text{onehot}(\text{label}_t = j) \quad (10)$$

$$\widetilde{D}(:, j) = \widetilde{D} \times o_{t,j} \quad (11)$$

$$w_{t,j} = \widetilde{D}(:, j)/\text{sum}(\widetilde{D}(:, j)) \quad (12)$$

EmoDistLoss The final loss, which we name *EmoDistLoss*, is calculated from the negative weighted sum of log terms from Equation 9. Since the distance, hence the weight, between identical labels is 0, this calculation does not involve the output probability of the correct label.

$$\textit{EmoDistLoss}_t = -w_{t,j} \cdot f(p_t) \quad (13)$$

4.3.2 MTL via Emotional Aspects

In addition to the emotion classification head, we have a classification head for each emotion aspect from the label definition, namely the valence, the elicitor, and the conduct.

The overall classification loss L is a weighted sum of the loss from softmax outputs of four classification heads $L_{emo}, L_{val}, L_{eli}, L_{con}$ with a hyperparameter α.

$$L = \alpha L_{emo} + \frac{1}{3}(1-\alpha)(L_{val} + L_{eli} + L_{con}) \quad (14)$$

5 User Emotion Recognition in ToDs with ERToD

5.1 Experimental Set-up

5.1.1 Dataset

We train and test our models on EmoWOZ. It contains user emotion annotations for all dialogues from MultiWOZ (Budzianowski et al., 2018) and additional 1000 human-machine dialogues. It contains 7 emotion groups (see Table 1 and Appendix A for details). Four emotion classes are considerably rare: *fearful*, *apologetic*, *abusive*, and *excited*. DA examples can be found in Appendix B. Our primary aim of DA is to address the poor ERC performance on rare emotions rather than building a balanced dataset. While the later aim can be achieved with the aid of large language models for example, this is out of the scope of our work.

Class Name	Valence	Elicitor	Conduct	Count (%)
Neutral	Neutral	Don't Care	Polite	58,656 (70.1%)
Satisfied	Positive	Operator	Polite	17,532 (21.0%)
Dissatisfied	Negative	Operator	Polite	5,117 (6.1%)
Excited	Positive	Event/Fact	Polite	971 (1.2%)
Apologetic	Negative	User	Polite	840 (1.0%)
Fearful	Negative	Event/Fact	Polite	396 (0.5%)
Abusive	Negative	Operator	Impolite	105 (0.2%)

Table 1: EmoWOZ Emotion definition and distribution.

Augmenting Abusive Utterances The user sometimes becomes abusive towards the system. While this correlates with failure to satisfy the user goal, exact abusive expressions uttered by the user are usually independent of the context. Therefore, we apply our DA method for context-independent emotions for *Abusive*. We utilise ConvAbuse, a dataset for nuanced abusive behaviours in chit-chat conversations (Cercas Curry et al., 2021), for more diverse abusive expressions. In ConvAbuse, user utterances are labelled with type, target, strength, and directiveness. We filter for abuses on the system's intellectuality (labelled as `type=intellectual` and `target=system`) to better suit ToD context. We combine each selected utterance with the context of a random abusive utterance in EmoWOZ, resulting in 273 augmented samples.

Augmenting Fearful, Apologetic, and Excited Utterances Expressions of these emotions usually contain task information. *Fearful* and *Excited* usually co-occur with a description of the situation that prompts the user to interact with the system. *Apologetic* is frequently associated with a correction of search criteria. There is a strong connection between these emotion expressions and the progression of the task in the dialogue history. Therefore, we apply our DA method for these context-dependent emotions. We look for samples with desired emotions from other ToD datasets using automatic labels. We train a ContextBERT on EmoWOZ (see Section 5.1.2) with a 30% dropout on the BERT output. We train the model with 10 different seeds and run inferences on the training set of existing ToD datasets: Schema-Guided Dialogue (SGD, Rastogi et al. 2019), Taskmaster-1 (TM-1), and Taskmaster-2 (TM-2) (Byrne et al., 2019). In addition, we filter for common domains of EmoWOZ: *Hotels, RideSharing, Travel, Restaurants* in SGD, *RestaurantTable, PizzaOrdering, CoffeeOrdering, UberLyft* in TM-1, and *HotelSearch, Restaurants, FoodOrdering* in TM-2. The classification confidence is measured by votes from 10 models. We use a confidence threshold of 0.7 and cap the number of augmented samples at 1000 for each emotion, resulting in 268 *fearful*, 872 *apologetic*, and 1000 *excited* samples.

5.1.2 Baselines

We implement ERToD to a range of ERC models that have been used to benchmark EmoWOZ, as listed in Table 2. ContextBERT (Feng et al., 2022) and EmoBERTa (Kim and Vossen, 2021) are simple yet robust transformer-based ERC models, and they have similar spirits except that they respectively use BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as utterance encoder. They are both built on top of BERT by additionally considering dialogue context and speaker roles in the input. DialogueRNN (Majumder et al., 2019) and COSMIC (Ghosal et al., 2020) are RNN-based models. Following (Feng et al., 2022), we use DialogueRNN with either {GloVe(Pennington et al., 2014)+Convolutional Neural Network} or BERT as the utterance encoder. COSMIC additionally extracts features with a pre-trained commonsense model (Bosselut et al., 2019)[1]. It is important to note that after replacing the original utterance encoder with the sentiment-aware encoder (as described in Section 5.1.3), two variants of DialogueRNN essentially become the same model, and so do EmoBERTa and ContextBERT.

5.1.3 Training

In our task information encoder, we use SetSUMBT DST (van Niekerk et al., 2021) from ConvLab-3 toolkit (Zhu et al., 2022). SetSUMBT is a strong DST considering uncertainty with a joint goal accuracy of 52.26% on MultiWOZ 2.1 (Eric et al., 2020). The FC layer in Equation 4 has input/output dimensions of 1083 and 256 respectively and hyperbolic tangent activation (TanH, LeCun et al. 2015). We further replace the utterance encoders of chit-chat ERC models with SentiX, a sentiment-adapted BERT (Zhou et al., 2020).

We use our proposed *EmoDistLoss* for the emotion classification head and cross-entropy loss for MTL heads (valence, elicitor, and conduct). Since the elicitor of *Neutral* emotion is not distinguishable and therefore not explicitly defined in EmoWOZ, we mark the elicitor of *Neutral* samples

[1]COSMIC requires future utterances in recognising the current emotion whereas other models can be configured as either bidirectional or unidirectional. While we use unidirectional set-ups where possible to comply with our task formulation in Section 3.1, we are also interested in how ERToD improves COSMIC for static dialogue analysis in ToDs.

as *don't care*, and their loss in from elicitor classification is ignored. α in Equation 14 is set to 0.4 based on several rounds of hyperparameter tuning.

To calculate the *EmoDistLoss*, we use 1 as the unit distance and define the distance for each emotional aspect as illustrated in Appendix C. For valence, it is commonly adopted to consider negative and positive as two polarities and neutral in the middle (Socher et al., 2013). Therefore, the distance is 2 between positive and negative, and 1 between non-neutral and neutral. For emotion elicitors, we set the distance between *don't care* to any specific elicitor as 0.5 to penalise a "lazy" classifier that wrongly recognises the emotion as neutral. Doing so also results in a consistent shortest distance of 1 between any pair of specific elicitors.

We follow the default training set-up of each model except for ContextBERT. We reduce the context size of ContextBERT from 512 to 128, resulting in stronger performance and faster training.

5.1.4 Evaluation

We report F1 for each emotion. For overall performance, we report both macro F1 and weighted F1. Macro F1 considers each emotion equally and reflects the model's ability to recognise rare emotions. Weighted F1 is the weighted sum of F1 scores of each label. Weights are determined by the proportion of each emotion in the dataset. We exclude *Neutral* from calculating the averages as it makes up more than 70% of labels.

In addition, we also calculate the average emotion distance (AED) between the recognised emotion and the label to quantify how wrong the model is when it misclassifies. The AED of an emotion e is calculated from the average of $\widetilde{D}(\text{label}{=}e, \text{recognised\_emotion})$ of samples whose label is e (see Equation 7). Lower AED means less severe consequences from mistakes, and is therefore more desirable. All experiments are repeated with 10 different seeds.

5.2 ERC Results

Table 2 shows the change in the emotion recognition performance of the selected chit-chat ERC models after incorporating our ERToD framework. ERToD achieves significant improvement in average F1 scores of all models (see Appendix D for examples of model outputs, Appendix E for F1 of individual emotions).

	Base Model		+ ERToD		Difference	
	MF1	WF1	MF1	WF1	MF1	WF1
BERT	50.1	73.5	61.4	77.3	+11.3	+3.8
DialogueRNN+GloVe	40.1	74.6	56.5	78.5	+16.4	+3.9
DialogueRNN+BERT	52.1	75.5	56.5	78.5	+4.4	+3.0
COSMIC	56.3	77.1	57.4	79.6	+1.1	+2.5
EmoBERTa	57.9	**83.0**	**65.9**	**83.9**	+9.0	+0.9
ContextBERT	**59.1**	81.9	**65.9**	**83.9**	+6.8	+2.0

Table 2: Macro- and weighted-average F1 (MF1, WF1) of ERC models before and after incorporating ERToD. Best average F1s are marked in **bold**. All differences are significant with $p < 0.05$.

	Model	Neu.	Sat.	Dis.	Exc.	Apo.	Fea.	Abu.
F1 Score (↑)	ContextBERT	93.5	89.1	69.7	45.6	69.6	33.3	47.0
	+ DA	†**94.2**	†90.5	†71.0	45.3	†72.1	‡38.3	†67.4
	+ DS	†**94.2**	†90.5	†71.3	45.7	†72.7	35.3	†69.4
	+ SentiX	†**94.2**	†**90.6**	†72.2	‡47.1	†73.2	†39.0	†66.1
	+ MTL	†**94.2**	†90.4	†**72.3**	‡47.2	†**73.4**	†41.0	†67.9
	+ ERToD	†94.1	†**90.6**	†**72.3**	†**47.6**	†72.0	†**42.4**	†**69.8**
AED Score (↓)	ContextBERT	0.058	0.094	0.304	0.497	0.269	0.605	0.554
	+ DA	†**0.049**	†0.080	0.312	0.493	‡0.292	0.593	†0.339
	+ DS	†0.053	†0.075	0.296	0.481	0.277	0.582	†0.300
	+ SentiX	†0.052	†0.077	‡0.286	†0.454	0.287	0.596	†0.283
	+ MTL	†0.054	†0.075	‡**0.284**	‡0.456	0.277	0.585	†**0.258**
	+ ERToD	0.056	†**0.070**	0.296	†**0.435**	**0.244**	**0.571**	†0.277

Table 3: F1 (↑) and AED (↓) scores of **Neu**tral, **Sat**isfied, **Dis**satisfied, **Exc**ited, **Apo**logetic, **Fea**rful, and **Abu**sive. † indicates statistically significant difference with $p < 0.05$ and ‡ indicates $p < 0.1$ when comparing with ContextBERT. Best scores are marked in **bold**.

5.3 Ablation Study on ERToD

We perform an ablation study on the best resulting model, ContextBERT-ERToD (Table 3). We add each technique in the order of data-related, feature-related, and loss-related approaches. Averaged scores can be found in Appendix F.

Impact of DA DA helps improve almost all F1 scores even with a relatively small number of additional samples. There is a small and insignificant drop in the F1 of *Excited*, which is also frequently confused among human annotators. Further work to resolve the ambiguities would be beneficial.

Impact of Dialogue State (+DS) Adding dialogue state features further improves most other non-neutral emotions. Although it does not bring advantages for the F1 of *Fearful*, the AED of it continues to improve, showing that the system is making less severe mistakes.

Impact of SentiX Initialising BERT with SentiX parameters further improves the recognition of all other non-neutral emotions except for *Abusive*. This suggests that the sentiment information encoded in SentiX is useful for resolving ambiguity. We suspect that, while SentiX is good at distinguishing the valence of emotion, its effect is

limited for user conduct, the hallmark of *Abusive*.

Impact of MTL MTL improves F1 for all non-neutral emotions except for *Satisfied*. It also achieves the best AED for *Abusive*. This suggests that MTL heads, especially the conduct classification head, help identify emotions in the simpler valence-elicitor-conduct space. There is a slight drop in the F1 score of *Satisfied*, but it is compensated by the improvement in its AED.

Impact of *EmoDistLoss* (+ERToD) The final version of the model achieves the best F1 score in {*Satisfied, Dissatisfied, Excited, Fearful, Abusive*} and the best AED score in {*Satisfied, Excited, Apologetic, Fearful*}, leading to best averaged scores (Table F8). This shows penalising misclassifications according to emotion distance, which is only possible thanks to the emotion model, further helps recognise ambiguous emotions.

For the degradation of both scores in *Neutral*, we hypothesise that the model recognises non-neutral emotions more boldly than annotators, who are more cautious about subtle emotional cues.

6 Zero-shot User Satisfaction Prediction

6.1 Experimental Set-up

6.1.1 Dataset

We evaluate our model with **User Satisfaction Simulation** (USS) dataset where user utterances are annotated with 5-level satisfaction ratings (Sun et al., 2021). Dialogues in USS come from 5 different ToD datasets:

Jing Dong Dialogue Corpus (JDDC, Chen et al., 2020) is a multi-turn Chinese dialogue dataset for E-commerce customer service. USS contains 54.5k user satisfaction annotations for 3300 dialouges sampled from JDDC. Since JDDC is in Chinese, we translated it into English with Google Translate API first.

Schema-guided Dialogues (SGD, Rastogi et al., 2020) is a multi-domain, task-oriented conversations between a human and a virtual assistant. These conversations involve interactions with services and APIs spanning 20 domains, such as banks, events, media, calendar, travel, and weather. USS contains 13.8k user satisfaction annotations for 1000 dialogues sampled from SGD. Although we use SGD for DA, our DA samples do not overlap with SGD dialogues in USS.

Recommendation Dialogue (ReDial, Li et al., 2018) is an annotated dataset of dialogues, where users recommend movies to each other. USS contains 11.8k user satisfaction annotations for 1000 dialogues sampled from ReDial.

Coached Conversational Preference Elicitation (CCPE, Radlinski et al., 2019) is a dialogue dataset where the " assistant" is tasked with eliciting the "user" preferences about movies collected in the Wizard-of-Oz framework. USS contains 6.8k user satisfaction annotations for 500 dialogues sampled from CCPE.

MultiWOZ (Budzianowski et al., 2018) is a multi-domain task-oriented dialogue dataset collected in the Wizard-of-Oz framework spanning 7 domains such as restaurant, hotel, and attraction. USS contains 12.5k user satisfaction annotations for 1000 dialogues sampled from MultiWOZ. Since we trained our ERC model on EmoWOZ, which was based on MultiWOZ, we excluded it in our evaluation.

6.1.2 Baselines

We compare our zero-shot results with supervised models of Sun et al. (2021) and Kim and Lipani (2022). HiGRU (Yang et al., 2016) and BERT (Devlin et al., 2019) were the best two models trained by Sun et al. (2021) to benchmark USS dataset when it was first released. SatAct and SatActUtt are T5-based models (Raffel et al., 2020). SatAct is trained to predict user satisfaction and user action in a MTL set-up, whereas SatActUtt additionally incorporates user utterance generation. For satisfaction prediction, these models were set up to predict a 5-level rating during training.

These baseline models were trained on each one of the five ToD subsets in USS with a 10-fold cross-validation. Although non-3 ratings were upsampled by 10 times in their training, the training data size is still smaller than that of ContextBERT-ERToD (68.9k emotion annotations, EmoWOZ and DA samples altogether).

6.1.3 Zero-shot Inference

We experimented with ContextBERT-ERToD, the best resulting model from ERC training. After training the model for ERC, we fixed its parameters and ran inference with USS dataset for zero-shot user satisfaction prediction. To adapt to user satisfaction prediction set-up, we excluded information about the user turn at t from the model

input as well as the dialogue state. Specifically, for utterance encoding, we excluded u_t from the dialogue history to have $H_t = [s_{t-1}, u_{t-1}, ..., s_1, u_1]$. For task information encoding, we shifted the context window in Equation 3 by one and have $\widetilde{V}_t = V_{t-1} \oplus V_{t-2} \oplus V_{t-3}$ as the new contextual dialogue state vector.

6.1.4 Evaluation

In the works of baseline models, satisfaction ratings {1,2} were considered the negative class and {3,4,5} as the positive. To map the emotion prediction from our ERC model to binary satisfaction ratings, it is intuitive to leverage the valence aspect of emotions. Emotion classes with a negative valence were considered *Not Satisfied* and those with a positive valence as *Satisfied*. The emotion *Apologetic* is an exception among emotions with a negative valence. Since its elicitor is the user him/herself, it should not be considered as a sign of user dissatisfaction. Regarding the emotion class *Neutral*, we mapped it to *Satisfied* because the original evaluation set-up of baseline models considered the medium satisfaction rating, 3, as the positive class.

Overall, we considered {*Neutral, Apologetic, Excited, Satisfied*} as the positive class and {*Fearful, Dissatisfied, Abusive*} as negative.

6.2 Results

	JDDC	SGD	ReDial	CCPE
HiGRU (Sun et al., 2021)	17.1	8.6	8.3	27.4
BERT (Sun et al., 2021)	18.5	4.8	12.5	24.5
SatAct (Kim and Lipani, 2022)	-	71.3	-	16.5
SatActUtt (Kim and Lipani, 2022)	-	**84.7**	-	73.4
ContextBERT-ERToD (0-shot)	**50.8**	78.8	**78.1**	**77.6**

Table 4: Binary F1 scores on different USS subsets. Best scores are marked in **bold**.

Following existing work, we first report binary F1 for direct comparison. In Table 4, ContextBERT-ERToD performs comparably with SatActUtt and significantly outperforms other models. This shows that our ERToD framework in combination with the ERC model generalises well to user satisfaction prediction.

7 Conclusion

In this work, we propose ERToD, a framework to address three critical steps in learning and effectively adapt chit-chat ERC models to recognise emotions in ToDs. We propose two strategies of DA for different emotions to improve ERC performance in ToDs on rare emotions. We further leverage dialogue state and sentiment-aware embeddings for a richer feature representation. In addition, we apply MTL and devise a novel loss function, *EmoDistLoss*, which take the similarities between emotions into account. Our framework significantly improves existing chit-chat ERC models' performance in recognising user emotions in ToDs. By further applying our best resulting model to perform the task of user satisfaction prediction, we show that our method generalises well on other similar valence-related classification tasks in ToDs.

As more sophisticated and powerful dialogue systems such as ChatGPT arise, there is an urge to recognise, understand and handle the emotion of the user, especially in the age where online abuse is omnipresent. The long-term aim of this work is to obtain valuable insight for downstream ToD modelling tasks. This allows further investigation of emotion regulation strategies on the system side to improve task performance and user satisfaction, and to prevent undesirable user behaviours.

8 Acknowledgements

S. Feng, N. Lubis, M. Heck, and C. van Niekerk are supported by funding provided by the Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award endowed by the Federal Ministry of Education and Research, while C. Geishauser, H-C. Lin, B. Ruppik, and R. Vukovic are supported by funds from the European Research Council (ERC) provided under the Horizon 2020 research and innovation programme (Grant agreement No. STG2018804636). Computing resources were provided by Google Cloud.

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural*

Language Processing, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 745–754, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hao-Yun Chen, Pei-Hsin Wang, Chun-Hao Liu, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. 2019. Complement objective training. In *International Conference on Learning Representations*.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The JDDC corpus: A large-scale multi-turn Chinese dialogue dataset for E-commerce customer service. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 459–466, Marseille, France. European Language Resources Association.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *International Conference on Machine Learning*.

Niko Colneric and Janez Demšar. 2020. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing*, 11:433–446.

L. Devillers, L. Lamel, and I. Vasilescu. 2003. Emotion detection in task-oriented spoken dialogues. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 3, pages III–549.

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

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4096–4113, Marseille, France. European Language Resources Association.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Milan Gritta, Gerasimos Lampouras, and Ignacio Iacobacci. 2021. Conversation graph: Data augmentation, training, and evaluation for non-deterministic dialogue management. *Transactions of the Association for Computational Linguistics*, 9:36–52.

Michael Heck, Nurul Lubis, Carel van Niekerk, Shutong Feng, Christian Geishauser, Hsien-Chin Lin, and Milica Gašić. 2022. Robust Dialogue State Tracking with Weak Supervision and Sparse Data. *Transactions of the Association for Computational Linguistics*, 10:1175–1192.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. 2016. Squared earth mover's distance-based loss for training deep neural networks. *ArXiv*, abs/1611.05916.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245, Santa Fe, New

Mexico, USA. Association for Computational Linguistics.

Zhiqiang Hu, Roy Kaa-Wei Lee, and Nancy F. Chen. 2022. Are current task-oriented dialogue systems able to satisfy impolite users? *ArXiv*, abs/2210.12942.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Nam Kyun Kim, Jiwon Lee, Hun Kyu Ha, Geon Woo Lee, Jung Hyuk Lee, and Hong Kook Kim. 2017. Speech emotion recognition based on multi-task learning using a convolutional neural network. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 704–707.

Taewoon Kim and Piek Vossen. 2021. EmoBERTa: Speaker-aware emotion recognition in conversation with RoBERTa. *ArXiv*, abs/2108.12009.

To Eun Kim and Aldo Lipani. 2022. A multi-task based neural model to simulate users in goal oriented dialogue systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2115–2119, New York, NY, USA. Association for Computing Machinery.

Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and William Campbell. 2019. A closer look at feature space data augmentation for few-shot intent classification.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*.

Yang Li, Amirmohammad Kazameini, Yash Mehta, and Erik Cambria. 2021. Multitask learning for emotion and personality detection. *ArXiv*, abs/2101.02346.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2015. Construction and analysis of social-affective interaction corpus in english and indonesian. In *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 202–206. IEEE.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2023. Genpads: Reinforcing politeness in an end-to-end dialogue system. *PLOS ONE*, 18(1):1–20.

Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. 2021. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(5).

Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding

movie preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 353–360, Stockholm, Sweden. Association for Computational Linguistics.

Ilija Radosavovic, Piotr Dollár, Ross B. Girshick, Georgia Gkioxari, and Kaiming He. 2017. Data distillation: Towards omni-supervised learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4119–4128.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *AAAI Conference on Artificial Intelligence*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré . 2017. Snorkel. *Proceedings of the VLDB Endowment*, 11(3):269–282.

Tulika Saha, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning. *PLOS ONE*, 15(7):1–28.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated spoken dialog corpus of the CMU let's go bus information system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3369–3373, Istanbul, Turkey. European Language Resources Association (ELRA).

Weiyan Shi and Zhou Yu. 2018. Sentiment adaptive end-to-end dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1509–1519, Melbourne, Australia. Association for Computational Linguistics.

Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. EmoInHindi: A multi-label emotion and intensity annotated dataset in Hindi for emotion recognition in dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5829–5837, Marseille, France. European Language Resources Association.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2499–2506, New York, NY, USA. Association for Computing Machinery.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland. Association for Computational Linguistics.

Carel van Niekerk, Andrey Malinin, Christian Geishauser, Michael Heck, Hsien-chin Lin, Nurul Lubis, Shutong Feng, and Milica Gasic. 2021. Uncertainty measures in neural belief tracking and the effects on dialogue policy performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7914, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiancheng Wang, Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2020. Sentiment classification in customer service dialogue with topic-aware multitask learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9177–9184.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020a. Unsupervised data augmentation for consistency training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020b. Self-training with noisy student improves imagenet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Yifei Xu, Jingqiao Zhang, Ru He, Liangzhu Ge, Chao Yang, Cheng Yang, and Ying Nian Wu. 2022. Sas: Self-augmentation strategy for language model pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11586–11594.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706, Online. Association for Computational Linguistics.

Steve Young. 2002. Talking to machines (statistically speaking). In *Seventh International Conference on Spoken Language Processing*.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, Copenhagen, Denmark. Association for Computational Linguistics.

Sayyed Zahiri and Jinho D. Choi. 2018. Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. In *Proceedings of the AAAI Workshop on Affective Content Analysis*, AFFCON'18, pages 44–51, New Orleans, LA.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Jie Zhou, Junfeng Tian, Rui Wang, Yuanbin Wu, Wenming Xiao, and Liang He. 2020. SentiX: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 568–579, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qi Zhu, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gašić, and Minlie Huang. 2022. Convlab-3: A flexible dialogue system toolkit based on a unified data format.

A Emotion Definitions in EmoWOZ

Elicitor	Valence	Conduct	OCC Emotion Tokens	EmoWOZ Emotion	Implication of User
Operator	Positive	Polite	Admiration, gratitude, love	Satisfied, liking, appreciative	Satisfied with the operator because the goal is fulfilled.
		Impolite		Not applicable to EmoWOZ	
	Negative	Polite	Reproach, anger, hate	Dissatisfied, disliking	Dissatisfied with the operator's suggestion or mistake.
		Impolite		Abusive	Insulting the operator when the goal is not fulfilled.
User	Positive	Polite	Pride, gratification	Not applicable to EmoWOZ	
		Impolite			
	Negative	Polite	Shame, remorse, hate	Apologetic	Apologising for causing confusion to the operator.
		Impolite		Not modelled in EmoWOZ	Insulting the operator for no reason.
Events, facts	Positive	Polite	Happy-for, gloating, love, satisfaction, relief, joy	Excited, happy, anticipating	Looking forward to a good event (e.g. birthday party).
		Impolite		Not applicable to EmoWOZ	
	Negative	Polite	Distress, resentment, hate, fears-confirmed, pity, disappointment	Fearful, sad, disappointed	Encountered a bad event (e.g. robbery and option not available).
		Impolite		Not applicable to EmoWOZ	
-	Neutral	Polite	-	Neutral	Describing situations and needs.
		Impolite		Not modelled in EmoWOZ	No emotion but rude (e.g. using imperative sentences).

Table A1: EmoWOZ labels and similar emotions tokens from the OCC emotion model. For simplicity, emotion words in blue are used to represent each emotion category.

B Examples of Augmented Samples

B.1 Augmentation with Automatic Label

Source: Taskmaster-1
Dialogue ID: dlg-02edb443-9d6f-4553-af6e-f69778eb0fc5

...
S: Any other restaurant you were thinking about?
U: Yes, what about Char's at Tracy Mansion?
S: You mean the one on 1829 N Front St, Harrisburg, PA 17102-2213
U: **Yes, I've heard great things about that [DA Candidate]**

ContextBERT Ensemble Prediction: (Emotion = Excited, Confidence = 100%)

Figure B.1: DA sample for emotion *Excited*.

Source: SGD
Dialogue ID: 93_00124

...
U: I plan to check in on the 7th of this month.
S: Okay. You want to reserve 1 room at 1 Hotel Brooklyn Bridge in New York, for a 1 day stay on March 7th?
U: **Sorry, I misremembered the date. I actually want to check into one room on Saturday this week. [DA Candidate]**

ContextBERT Ensemble Prediction: (Emotion = Apologetic, Confidence = 100%)

Figure B.2: DA sample for emotion *Apologetic*.

Source: SGD Dialogue ID: 1_00028
U: **Hey! Today feeling so tired... Wish to have some food outside in a good restaurant... Can you find me one?** **[DA Candidate]** ...
ContextBERT Ensemble Prediction: (Emotion = Fearful, Confidence = 90%), (Emotion = Excited, Confidence = 10%)

Figure B.3: DA sample for emotion *Fearful*. Please note that although this class is called "fearful" for simplicity, user's negative emotion due to any undesirable events that is out of the control of the operator also belongs to this category in EmoWOZ according to Table A1.

B.2 Augmentation with Existing Dataset and Utterance Replacement

Source: ConvAbuse
DA Candidates:
1. Are you an idiot 2. You are stupid privitive word-repeater 3. Why do you not answer my question 4. Foolish ..
EmoWOZ sample for *Abusive* Dialogue ID: DMAGE2157.json
... U: need a good resturant near macy's in nyc S: You are more than welcome! U: can you help S: The address is Parkside, Cambridge . The phone is 01223358966 . The postcode is none . Thank you for using our services . U: **your horrible and not helpful at all** [Label = Abusive]

Figure B.4: DA sample for emotion *Abusive*. Candidate DA samples from ConvAbuse can be used to replace the abusive user utterance in EmoWOZ without causing any context inconsistency.

C Emotional Aspect Distance Definition

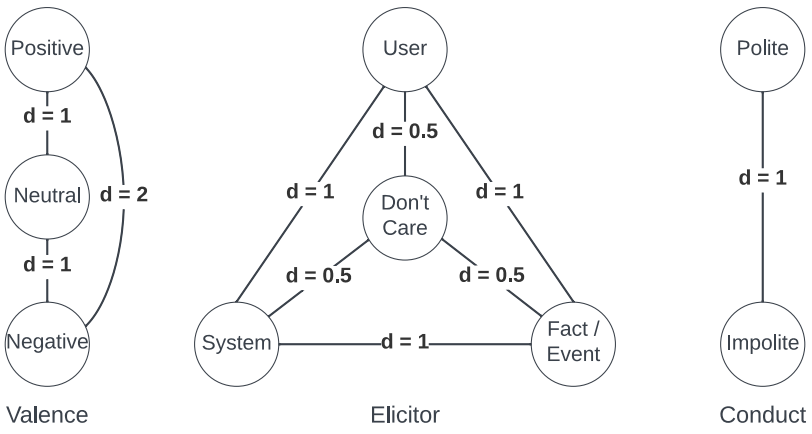


Figure C.1: Distance definition for Equation 6.

D Examples of Model Recognitions

Dialogue ID: DMAGE3777.json

...
U: Can you help me?
S: The phone is 01223358966. The postcode is none.
U: Ok
S: The address is Parkside, Cambridge.
U: **Nice** **[To classify, label = Dissatisfied]**

	BERT	DialRNN-G	DialRNN-B	COSMIC	EmoBERTa	ContextBERT
Base	Satisfied	Dissatisfied	Satisfied	Satisfied	Satisfied	Neutral
+ ERToD	Neutral	Dissatisfied	Dissatisfied	Dissatisfied	Dissatisfied	Dissatisfied

Figure D.1: Model Recognitions on dialogue DMAGE3777 in EmoWOZ.

Dialogue ID: PMUL2437.json

...
S: There are 21 restaurants available in the centre of town. How about a specific type of cuisine?
U: I need to know the food type and postcode and it should also have multiple sports
S: I am sorry I do not understand what you just said. Please repeat in a way that makes sense.
U: **Get me the food type and the post code** **[To classify, label=Dissatisfied]**

	BERT	DialRNN-G	DialRNN-B	COSMIC	EmoBERTa	ContextBERT
Base	Neutral	Dissatisfied	Neutral	Neutral	Dissatisfied	Neutral
+ ERToD	Neutral	Dissatisfied	Dissatisfied	Dissatisfied	Dissatisfied	Dissatisfied

Figure D.2: Model Recognitions on dialogue PMUL2437 in EmoWOZ

E Detailed ERC Performance on Each Emotion

Model	Neutral	Satisfied	Dissatisfied	Excited	Apologetic	Fearful	Abusive
BERT	89.8	88.8	35.1	42.9	70.4	36.2	27.5
DialogueRNN+GloVe	83.5	86.4	51.4	32.7	57.7	12.7	0.0
DialogueRNN+BERT	86.9	87.6	47.5	39.4	**71.5**	41.3	25.6
COSMIC	89.8	88.4	50.7	44.4	70.9	**52.0**	31.6
EmoBERTa	**94.0**	**90.3**	**71.0**	44.9	70.6	31.3	39.3
ContextBERT	93.5	89.1	69.7	**45.6**	69.6	33.3	**47.0**

Table E2: F1 scores of selected chit-chat ERC models BEFORE incorporating ERToD framework. The best score for each emotion is marked in **bold**.

	Neu.		Sat.		Dis.		Exc.		Apo.		Fea.		Abu.		M-Avg		W-Avg	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
BERT	90.3	89.3	88.4	89.2	38.9	38.6	**47.7**	39.1	69.7	71.5	47.7	30.0	42.1	22.4	55.7	48.5	74.5	74.5
DialRNN-GloVe	**97.6**	73.0	78.5	**95.9**	36.5	**87.6**	22.2	**65.7**	44.7	**82.5**	11.2	18.9	0	0	32.2	**58.4**	65.0	**91.4**
DialRNN-BERT	94.0	80.7	84.7	90.7	34.8	75.3	36.5	42.9	68.3	75.0	46.7	37.5	28.6	23.5	49.9	57.5	70.4	84.2
COSMIC	93.1	86.8	86.2	90.7	42.3	64.4	43.7	45.3	**71.9**	70.1	**65.0**	**43.3**	**77.3**	20.0	**64.4**	55.6	74.0	81.7
EmoBERTa	94.2	**94.0**	**88.7**	92.2	**74.6**	69.5	45.6	42.6	73.0	70.3	37.9	27.2	54.0	24.7	62.3	54.4	**82.9**	83.8
ContextBERT	93.4	93.7	88.5	89.8	72.6	67.2	46.4	45.4	68.3	71.6	37.9	30.0	64.5	**37.6**	63.0	57.0	82.3	81.8

Table E3: **P**recision and **R**ecall scores of selected chit-chat ERC models BEFORE incorporating ERToD framework. We report scores of each emotion: **Neu**tral, **Sat**isfied, **Dis**satisfied, **Exc**ited, **Apo**logetic, **Fea**rful, **Abu**sive, as well as **M**acro- and **W**eighted **Av**era**g**ed scores. The best score for each emotion is marked in **bold**. Neutral is excluded when calculating the averaged scores. For better presentation, DialogueRNN is shortened to DialRNN.

Model	Neutral	Satisfied	Dissatisfied	Excited	Apologetic	Fearful	Abusive
BERT	92.4	90.4	43.7	**49.7**	75.4	39.5	**69.7**
DialogueRNN+GloVe	92.6	90.1	51.4	43.9	**77.6**	42.4	33.8
DialogueRNN+BERT	92.6	90.1	51.4	43.9	**77.6**	42.4	33.8
COSMIC	91.1	89.5	58.1	45.6	73.3	36.3	41.6
EmoBERTa	**94.0**	**90.5**	**72.3**	47.9	71.9	**43.4**	**69.7**
ContextBERT	**94.0**	**90.5**	**72.3**	47.9	71.9	**43.4**	**69.7**

Table E4: F1 scores of selected chit-chat ERC models AFTER incorporating ERToD framework. The best score for each emotion is marked in **bold**.

	Neu.		Sat.		Dis.		Exc.		Apo.		Fea.		Abu.		M-Avg		W-Avg	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
BERT	91.0	93.8	88.9	92.0	57.5	35.5	**51.2**	48.9	**81.6**	70.3	48.1	33.9	**74.8**	65.9	67.0	57.7	79.8	76.3
DialRNN-GloVe	91.3	**94.0**	**89.7**	90.5	60.9	41.5	44.4	45.6	76.5	77.3	42.6	**38.3**	54.3	30.0	61.4	53.9	80.6	76.5
DialRNN-BERT	91.3	**94.0**	**89.7**	90.5	60.9	41.5	44.4	45.6	76.5	77.3	42.6	**38.3**	54.3	30.0	61.4	53.9	80.6	76.5
COSMIC	**94.4**	88.3	86.9	92.3	51.6	**68.9**	38.7	**57.4**	68.2	**79.3**	36.2	**38.3**	44.7	38.8	54.4	62.5	75.9	**84.6**
EmoBERTa	94.3	93.9	88.9	**92.4**	**75.6**	68.0	45.7	50.7	70.8	74.4	**54.6**	35.6	72.4	**68.2**	**68.0**	**64.9**	**83.5**	84.3
ContextBERT	94.3	93.9	88.9	**92.4**	**75.6**	68.0	45.7	50.7	70.8	74.4	**54.6**	35.6	72.4	**68.2**	**68.0**	**64.9**	**83.5**	84.3

Table E5: **P**recision and **R**ecall scores of selected chit-chat ERC models AFTER incorporating ERToD framework. We report scores of each emotion: **Neu**tral, **Sat**isfied, **Dis**satisfied, **Exc**ited, **Apo**logetic, **Fea**rful, **Abu**sive, as well as **M**acro- and **W**eighted **Av**era**g**ed scores. The best score for each emotion is marked in **bold**. Neutral is excluded when calculating the averaged scores. For better presentation, DialogueRNN is shortened to DialRNN.

Model	Neutral	Satisfied	Dissatisfied	Excited	Apologetic	Fearful	Abusive
BERT	+2.6	+1.6	+8.6	+6.8	+5.0	+3.3	+42.2
DialogueRNN+GloVe	+9.1	+3.7	+0.0	+11.2	+19.9	+29.7	+33.8
DialogueRNN+BERT	+5.7	+2.5	+3.9	+4.5	+6.1	+1.1	+8.2
COSMIC	+1.3	+1.1	+7.4	+1.2	+2.4	**-15.7**	+10.0
EmoBERTa	0.0	+0.2	+1.3	+3.0	+1.3	+12.1	+30.4
ContextBERT	+0.5	+1.4	+2.6	+2.3	+2.3	+10.1	+22.7

Table E6: Change of F1 scores of selected chit-chat ERC models after incorporating ERToD framework. The only degradation in performance is marked in **bold**.

In terms of F1 scores, ERToD results in improvement in all emotions except for *fearful* in COSMIC (Table E6). We further investigate this exception. While most of fearful utterances are located at the beginning

of the dialogue in the training and development set in EmoWOZ, the position of such utterances are more evenly distributed in the test set as well as the augmented samples. Upon toggling the development set and the test set for evaluation, we observe that the F1 of fearful by COSMIC drops significantly (52.0% $\rightarrow$ 28.8%) while that of COSMIC-ERToD remains roughly unchanged (35.5% $\rightarrow$ 37.6%). The trend in all other results remains unchanged.

The drastically different performance of COSMIC on the development and the test set suggests that COSMIC develops a positional bias from the training set of EmoWOZ. At the same time, COSMIC-ERToD performs similarly on both non-training sets, likely relying more on textual and task information. The limited performance of COSMIC-ERToD is likely due to the extra false-positives at the later stage of dialogues.

	Neu.		Sat.		Dis.		Exc.		Apo.		Fea.		Abu.		M-Avg		W-Avg	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
BERT	+0.7	+4.5	+0.5	+2.8	+18.6	-3.1	+3.5	+9.8	+11.9	-1.2	+0.4	+3.9	+32.7	+43.5	+11.3	+9.2	+5.3	+1.8
DialRNN-GloVe	-6.3	+21.0	+11.2	-5.4	+24.4	-46.1	+22.2	-20.1	+31.8	-5.2	+31.4	+19.4	+54.3	+30.0	+29.2	-4.5	+15.6	-14.9
DialRNN-BERT	-2.7	+13.3	+5.0	-0.2	+26.1	-33.8	+7.9	+2.7	+8.2	+2.3	-4.1	+0.8	+25.7	+6.5	+11.5	-3.6	+10.2	-7.7
COSMIC	+1.3	+1.5	+0.7	+1.6	+9.3	+4.5	-5.0	+12.1	-3.7	+9.2	-28.8	-5.0	-32.6	+18.8	-10.0	+6.9	+1.9	+2.9
EmoBERTa	+0.1	-0.1	+0.2	+0.2	+1.0	-1.5	+0.1	+8.1	-2.2	+4.1	+16.7	+8.4	+18.4	+43.5	+5.7	+10.5	+0.6	+0.5
ContextBERT	+0.9	+0.2	+0.4	+2.6	+3.0	+0.8	-0.7	+5.3	+2.5	+2.8	+16.7	+5.6	+7.9	+30.6	+5.0	+7.9	+1.2	+2.5

Table E7: The difference in **P**recision and **R**ecall scores of selected chit-chat ERC models before and after incorporating ERToD framework. We report scores of each emotion: **Neu**tral, **Sat**isfied, **Dis**satisfied, **Exc**ited, **Apo**logetic, **Fea**rful, **Abu**sive, as well as **M**acro- and **W**eighted **Av**era**g**ed scores. The best score for each emotion is marked in **bold**. Neutral is excluded when calculating the averaged scores. For better presentation, DialogueRNN is shortened to DialRNN.

F Averaged Scores for the Ablation Study

	Model	Macro Avg	Weighted Avg
F1 Score (↑)	ContextBERT	59.1	81.9
	+ DA	†64.1	†83.4
	+ DS	†64.1	†83.5
	+ SentiX	†64.8	†83.7
	+ MTL	†65.3	†83.7
	+ ERToD	†**65.7**	†**83.9**
AED Score (↓)	ContextBERT	0.387	0.168
	+ DA	†0.351	†0.159
	+ DS	†0.335	†0.151
	+ SentiX	†0.331	†0.149
	+ MTL	†0.322	†0.147
	+ ERToD	†**0.316**	†**0.145**

Table F8: Ablation Study of ERToD. † indicates statistically significant difference with $p < 0.05$ when comparing with ContextBERT. The best score in each category is in **bold**. For each of the additional methods: DA = Data Augmentation, DS = Dialogue State Features, SentiX = Sentiment-aware Text Embedding, MTL = Multi-task Learning. Neutral is excluded when calculating the averaged scores.

	Neu.		Sat.		Dis.		Exc.		Apo.		Fea.		Abu.		M-Avg		W-Avg	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
ContextBERT	93.4	93.9	88.5	**92.4**	72.6	68.0	46.4	50.7	68.3	74.4	37.9	**35.6**	64.5	68.2	63.0	64.9	82.3	84.3
+ DA	93.9	94.4	89.4	91.6	75.6	67.2	47.2	44.6	75.0	70.0	53.1	30.6	70.1	65.9	**68.4**	61.6	84.0	83.1
+ DS	93.8	**94.6**	**90.1**	90.9	74.5	68.4	**47.9**	44.6	75.8	69.0	50.7	27.8	69.9	69.4	68.1	61.7	**84.2**	82.9
+ SentiX	94.1	94.3	89.5	91.7	76.0	69.1	47.5	49.3	**76.7**	70.3	50.9	32.2	66.0	66.5	67.8	63.2	84.1	83.9
+ MTL	94.2	94.0	88.9	91.5	**76.4**	**70.6**	45.7	**49.8**	76.6	**71.6**	51.2	35.0	67.0	**72.4**	67.6	**65.1**	83.8	**84.2**
+ ERToD	**94.3**	94.1	88.9	91.9	75.6	69.3	45.7	48.8	70.8	70.8	**54.6**	34.4	**72.4**	70.0	68.0	64.2	83.5	84.1

Table F9: Ablation study on **P**recision and **R**ecall scores of ERToD. We report scores of each emotion: **Neu**tral, **Sat**isfied, **Dis**satisfied, **Exc**ited, **Apo**logetic, **Fea**rful, **Abu**sive, as well as **M**acro- and **W**eighted **Av**era**g**ed scores. The best score for each emotion is marked in **bold**. For each of the additional methods: DA = Data Augmentation, DS = DialogueState Features, SentiX = Sentiment-aware Text Embedding, MTL = Multi-task Learning.. Neutral is excluded when calculating averaged scores.

Analyzing Differences in Subjective Annotations by Participants and Third-party Annotators in Multimodal Dialogue Corpus

Kazunori Komatani **Ryu Takeda**
SANKEN, Osaka University
Ibaraki, Osaka 567-0047, Japan
{komatani,rtakeda}@sanken.osaka-u.ac.jp

Shogo Okada
JAIST
Nomi, Ishikawa 923-1292, Japan
okada-s@jaist.ac.jp

Abstract

Estimating the subjective impressions of human users during a dialogue is necessary when constructing a dialogue system that can respond adaptively to their emotional states. However, such subjective impressions (e.g., how much the user enjoys the dialogue) are inherently ambiguous, and the annotation results provided by multiple annotators do not always agree because they depend on the subjectivity of the annotators. In this paper, we analyzed the annotation results using 13,226 exchanges from 155 participants in a multimodal dialogue corpus called Hazumi that we had constructed, where each exchange was annotated by five third-party annotators. We investigated the agreement between the subjective annotations given by the third-party annotators and the participants themselves, on both per-exchange annotations (i.e., participant's sentiments) and per-dialogue (-participant) annotations (i.e., questionnaires on rapport and personality traits). We also investigated the conditions under which the annotation results are reliable. Our findings demonstrate that the dispersion of third-party sentiment annotations correlates with agreeableness of the participants, one of the Big Five personality traits.

1 Introduction

To achieve adaptive human-machine (or human-robot) dialogue, it is necessary to estimate the human user's subjective impressions and emotions during the dialogue. The user's satisfaction with the dialogue can be increased by appropriately changing the dialogue content in accordance with the user's emotions. Estimated subjective impressions and emotions can also be utilized to evaluate the dialogue.

The difficulty here is that such impressions and feelings are inherently subjective, and it is impossible to objectively determine unique references for subjective content. References are necessary for training and evaluating machine learning models. Even when manual annotations are performed, the results among annotators do not always agree, which is a common problem in annotations of subjective labels.

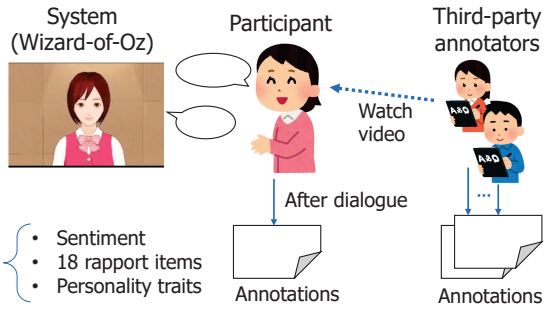


Figure 1: Subjective annotations given by participants themselves and by third-party annotators.

In this paper, we analyze the disagreement among human annotation results, especially the differences between annotations given by participants themselves and by third-party annotators (Fig. 1). Specifically, we conducted investigations on per-exchange annotations, i.e., sentiments, and per-dialogue (per-participant) annotations, i.e., questionnaires measuring the participant's rapport and personality traits. We used the Osaka University Multimodal Dialogue Corpus Hazumi, which we had previously constructed (Komatani and Okada, 2021), for the analysis. Our findings show that third-party annotators tend to give subjective annotations on the basis of their rather simple impressions compared to the participants themselves, who may not always fully express their inner states during the dialogue. We also clarify why automatic estimation performances of per-exchange sentiments from multimodal features differ between cases in which the reference sentiments were given by the participants themselves and by the third-party annotators, where the latter usually obtains better performances.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 104–113
September 11–15, 2023. ©2023 Association for Computational Linguistics

We then investigate the conditions under which estimated sentiments given by third-party annotators would be reliable by examining the dispersion of the annotation results. The estimation of users' sentiments based on multimodal data with machine learning will never be perfect, so it would be helpful to know whether the estimation results can be reliable for each user on the basis of other information sources. In this paper, after showing how the dispersion of sentiments given by third-party annotators correlates with machine learning performance, we demonstrate that this dispersion is negatively correlated with one of the personality traits, namely, agreeableness. This finding indicates that a personality trait can be a useful clue for determining the reliability of the sentiment estimation results.

2 Related Work

We here describe related studies on adaptive dialogue systems, emotion recognition, datasets of multimodal dialogues, reference labels for subjective annotations, and personality traits, in that order.

It is essential that dialogue system responses be adaptive to user states. In task-oriented dialogues, task success rates can be improved and the number of turns to task completion can be reduced by adapting system responses in accordance with several user types (Komatani et al., 2005). As for non-task-oriented dialogues, personalization based on the user's domain expertise has been attempted (Ferrod et al., 2021). System responses are preferably based on various modalities such as vision and prosody in addition to textual input. A variety of studies have examined text-based chatbots based on large pre-trained language models (e.g., (Adiwardana et al., 2020; Roller et al., 2021)). Currently, studies on dialogue systems have been actively expanded from the text-based perspective to a multimodal one, as evidenced by a recent dialogue competition using a humanoid robot (Minato et al., 2022).

User impressions (such as emotions) can be an important clue for adaptive dialogue systems. In particular, adapting to the user's emotions is essential for social interaction (Barros et al., 2021). Moreover, different types of information, including multimodal information (e.g., vision and prosody), can be utilized to recognize the user's emotions, as can physiological signals (Katada et al., 2022; Wu et al., 2022). In this paper, emotion is treated as sentiments per exchange.

A famous multimodal dialogue corpus with emotion labels is the IEMOCAP dataset (Busso et al., 2008), which contains dialogues between actors in role-playing scenarios. The Emotional Dyadic Motion CAPture (IEMOCAP) dataset is a well-known dataset used to recognize emotion during dialogues (Busso et al., 2008). It is a well-controlled dataset in the sense that data were collected by asking actors to speak with designated emotions. Therefore, this dataset contains objective reference labels for each emotion, i.e., the designated emotions. In contrast, our Hazumi dataset (Komatani and Okada, 2021) utilized in this paper consists of natural and spontaneous dialogues. Thus, there are no objective reference labels. We opted to use this dataset because our objective is to analyze the differences between several manual annotation results and discuss reference labels for subjective annotations.

Prior studies in the fields of social signal processing and affective computing have examined how to determine the ground truth of subjectively assigned labels (Spodenkiewicz et al., 2018; Bourvis et al., 2021; Maman et al., 2022). Maman et al. (2022) proposed three strategies for utilizing self-assessment labels and external assessment labels in training data for two dimensions of a group engagement state (called cohesion) and compared their prediction performances. Wang et al. (2023) recently proposed a method to train a classifier that fits better with the annotation results in medical binary classification tasks. In this paper, we do not train a classifier but analyze what happened in a multimodal dialogue data. We also extend analysis from single to several subjective annotations, i.e., per-exchange annotation and per-dialogue annotations.

Emotion depends on individual users, e.g., their personality traits (such as the Big Five (Goldberg, 1990)). Personality traits also play an important role in a variety of user-adapted interactions (Mairesse and Walker, 2010; Mota et al., 2018; Fernau et al., 2022; Yamamoto et al., 2023). The personality traits of a robot and human interlocutors are known to be effective for engagement estimation in human-robot interactions (Salam et al., 2017), and correlation between the engagement and the personality traits given per dialogue has been investigated in human-robot and human-human interactions (Celiktutan et al., 2019). In this work, we

Table 1: Hazumi versions and corresponding annotations.

Version	Recording environment	No. of participants (dialogues)	No. of exchanges	Self-sentiment	Third-party sentiment	18 rapport items	Personality traits
Hazumi1712	in-person	29	2,422		√		
Hazumi1902		30	2,514	√	√	√	
Hazumi1911		30	2,859	√	√	√	√
Hazumi2010	online	33	2,798		√	√	√
Hazumi2012		63	5,334		√	√	√
Hazumi2105		29	2,235		√	√	√
Total		214	18,162				

comprehensively analyzed the relationship between the user's personality traits on the basis of per-dialogue questionnaire results and per-exchange sentiments.

3 Target Corpus

We utilized the multimodal dialogue corpus Hazumi, which we had previously constructed (Komatani and Okada, 2021). It is a dataset that can be used extensively for research and development purposes[1]. Table 1 lists the various versions of the Hazumi corpus along with their recording environments, numbers of participants and exchanges, and annotations. It has six versions: 1712, 1902, 1911, 2010, 2012, and 2105, where the numbers correspond to the year and month the data collection started; for example, the collection of Hazumi1911 data began in November 2019. The first three versions were collected in-person and the following three were collected online due to the COVID-19 pandemic. Each dialogue lasted approximately 15 to 20 minutes.

The annotation unit at the utterance level is the exchange. An exchange is defined from the beginning of a system utterance to the beginning of the next system utterance. The data contain 18,162 exchanges in total; the mean duration was 13.10 seconds and its standard deviation was 7.80.

3.1 Dialogue data details

In Hazumi, the system used by the participants for talking was MMDAgent (Lee et al., 2013), which was operated by the Wizard-of-Oz (WoZ) method in which the virtual agent was controlled by a human operator (Wizard) located in another room. The Wizard controlled a graphical user interface built for this task while remotely observing the participants. Since the operators were trained to select the next utterance while the participant was still speaking (approximately ten seconds), there was a short wait time before the agent started responding.

The dialogue was chit-chat, meaning there was no specific task to be completed. The conversations were in Japanese and spanned several topics such as travel and movies. The Wizard attempted to select utterances that would engage the participants for a longer time. Specifically, the Wizard changed topics when the participants seemed uninterested, and listened when the participants seemed interested and were actively talking.

The participants were recruited from the general public through a recruiting agency for the in-person collection and through crowdsourcing for the online collection. A total of 214 participants (99 men, 115 women) were included, ranging in age from their 20s to 70s. They were given no special instructions, such as requests to act out their emotions strongly. Data were collected only from participants who signed a consent form that stated the data could be distributed to researchers for research and development purposes.

3.2 Subjective annotations

Manual annotations were given at the utterance and dialogue levels. The right half of Table 1 shows the types of subjective annotations and the Hazumi versions to which they were annotated.

3.2.1 Per-exchange annotations

Sentiment is scored on a 7-point scale representing how much the participant enjoyed the dialogue. Annotators gave it once per exchange, while watching the recorded videos of the dialogues. The sentiment annotation given by the third-party annotators is called *third-party sentiment*. For Hazumi1902 and Hazumi1911, the sentiment was also given by the participants themselves, which is called *self-sentiment*. They watched the recorded video and provided annotations immediately after their dialogue.

[1]The corpus has been distributed by the Informatics Research Data Repository at the National Institute of Informatics (NII-IDR). https://www.nii.ac.jp/dsc/idr/en/rdata/Hazumi/

Table 2: Cronbach's alpha values among five third-party annotators for per-dialogue annotations.

	Personality traits (Big Five)					Average of 18 rapport items
	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness	
Hazumi1911	0.835	0.761	0.622	0.620	0.696	0.876
Hazumi2010	0.911	0.791	0.560	0.697	0.883	0.883
Hazumi2012	0.867	0.827	0.598	0.599	0.747	0.856
Hazumi2105	0.903	0.843	0.663	0.645	0.786	0.813

3.2.2 Per-dialogue annotations

As per-dialogue annotations, participants answered two questionnaires after completing their dialogue: *18 rapport items*, which measured their rapport in the dialogue, and *Personality traits*, which examined their personality traits. Five third-party annotators also answered the same questionnaires about the participants from a third-party perspective after watching the recorded videos of the dialogues (i.e., they did not just read the transcribed texts).

The *18 rapport items* questionnaire was developed by social psychologists and originally consisted of 18 English adjectives[2] (Bernieri et al., 1996). It aims to examine the interlocutor's rapport and the results indicate how the dialogue was perceived. We utilized 18 questionnaire items with the 18 adjectives translated and converted into Japanese sentences (Kimura et al., 2005), such as "1. The dialogue was well-coordinated," "2. The dialogue was boring," and "18. The dialogue was slow." Each item is scored on an 8-point scale.

The second questionnaire asked about the participants' *personality traits* modeled on the Big Five, that is, extraversion, agreeableness, conscientiousness, neuroticism, and openness (Goldberg, 1990; Vinciarelli and Mohammadi, 2014). We used the 10-item personality inventory translated into Japanese (TIPI-J) (Oshio et al., 2012), which measures the Big Five with ten items. The items are scored on a 7-point scale, with two questions for each of the traits, one of which is an inverted item. Each of the Big Five scores is the sum of the two question items, one of which corresponds to the inverted item subtracted from 8 (i.e., the minimum is 2 and the maximum is 14).

As a preliminary analysis, Table 2 shows the Cronbach's alpha values among the five third-party annotators for the two kinds of per-dialogue annotations. An annotation result is considered consistent if the Cronbach's alpha is greater than 0.8. As we can see, extraversion, agreeableness, and openness tended to be around 0.8 or above, while conscientiousness and neuroticism tended to be below 0.8. This is consistent with the results of personality trait annotation agreement rates in other studies (Aran and Gatica-Perez, 2013). The values of the Cronbach's alpha for the average of the 18 rapport items also tended to be consistent.

[2]All adjectives appear in Table 4.

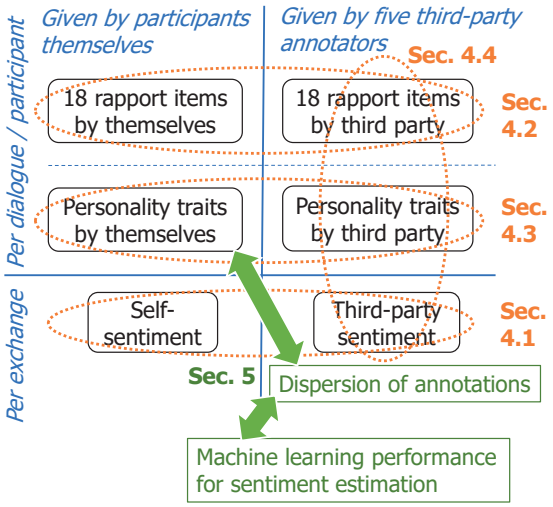


Figure 2: Positioning of the analyses.

4 Analyses on Relationship Between Annotations Given by Participants and Third-Party Annotators

We analyzed the correlations between the manual annotations given by the participants themselves and by five third-party annotators. Sentiments were analyzed using Hazumi1902 and Hazumi1911 due to their availability (see Table 1). As for the two annotations given per dialogue, we used the data of the four versions after Hazumi1911, which consist of 13,226 exchanges from 155 participants. Figure 2 depicts the positioning of the analyses we conducted.

If any correlation is found between two metrics corresponding to the annotations, it will provide useful insights for the machine learning design. For example, it would be effective to use one of the metrics as input when estimating the other by machine learning. The correlation would also be

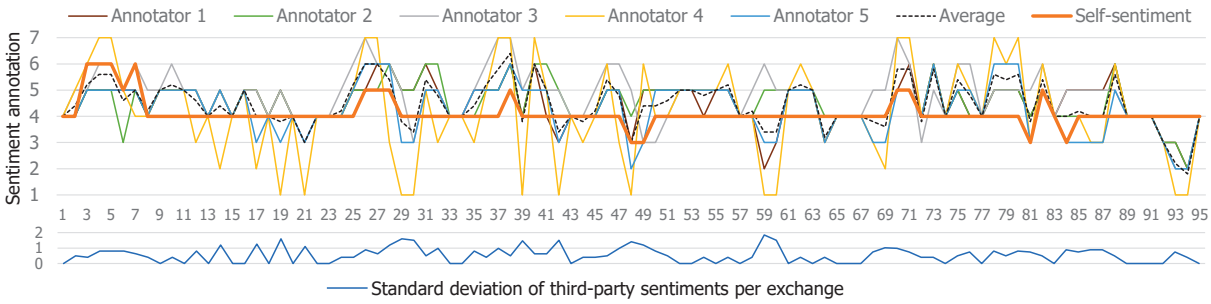


Figure 3: Example of sentiment annotation results and standard deviations (participant ID: 1911M4001).

helpful in designing multi-task learning with deep neural networks in which some layers are shared (Hirano et al., 2019). The two metrics can be utilized together to improve the machine learning performance.

4.1 Sentiment

Figure 3 shows an example of the sentiment annotation results for a male participant in his 40s (participant ID: 1911M4001). Horizontal and vertical axes indicate time in units of exchange and the annotation results on the 7-point scale, respectively. The solid lines in different colors represent the third-party sentiments by the five annotators (1 male, 4 female; Annotator 5 was male). The thick orange line in the center is self-sentiment, which does not agree with the third-party sentiments. The third-party sentiments by the five annotators share certain trends but do not completely agree. The correlation coefficient between the self-sentiment and the average of the third-party sentiments was 0.45. The figure also shows standard deviations of the third-party sentiments per exchange at the bottom, which will be used in Section 5.

Table 3 shows the correlation coefficients between self-sentiments and third-party sentiments, which were calculated per participant. The macro average of all correlation coefficients was 0.43. The maximum was 0.79 and the minimum was 0.01, indicating large individual differences. These results clarify that the self-sentiments and third-party sentiments are not necessarily correlated, as reported in (Truong et al., 2012).

This is why automatic estimation performances from multimodal features differ between cases in which self-sentiments and third-party sentiments are used as the references (Katada et al., 2022), where the latter obtained better performances. Third-party sentiments can be perceived from outside the participants, which suggests that computers attempting to estimate the sentiments can utilize the same information that the third-party annotators use. Self-sentiment is more difficult to estimate because it is not necessarily perceivable from the outside, even by human third-party annotators. Additional use of physiological signals has thus improved the estimation performance of self-sentiment (Katada et al., 2022). The signals can be regarded as extra information that third-party annotators can perceive.

Table 3: Correlation between self-sentiments and third-party sentiments.

	No. of participants	Macro average	(max., min.)
Hazumi1902	30	0.45	(0.69, 0.11)
Hazumi1911	30	0.41	(0.79, 0.01)
Total	60	0.43	(0.79, 0.01)

We also confirmed here that the correlation coefficients differ among participants and that the sentiment annotations results differ among the third-party annotators. We therefore attempted to use the deviation of the third-party sentiments in Section 5.

4.2 18 rapport items

We investigated the correlation between the answers by participants themselves and the averages of third-party annotators for each of the 18 rapport items. Table 4 lists the correlation coefficients in descending order. Excluding the three below the solid line, all correlations were statistically significant ($p < 0.05$). The correlation between the averages of the correlation coefficients was 0.34 (bottom line), and it was also statistically significant ($p = 0.023$).

Thus, the averaged answers to the 18 rapport items, which correspond to the posterior evaluation of the dialogue, showed a correlation between the participants themselves and the averages of the third-party annotations. The results in Table 4

Table 4: Correlation coefficients of all 18 rapport items between self- and third-party annotations.

5*	unsatisfying	0.38
9	engrossing	0.35
2*	boring	0.32
17	worthwhile	0.29
8*	awkward	0.27
16*	dull	0.25
10*	unfocused	0.23
6*	uncomfortably paced	0.23
1	well-coordinated	0.22
12*	intense	0.21
11	involving	0.21
14	active	0.20
4	harmonious	0.20
7*	cold	0.19
18*	slow	0.17
13	friendly	0.13
15	positive	0.09
3	cooperative	0.07
Average of 18 items		0.34

'*' denotes inverted items.

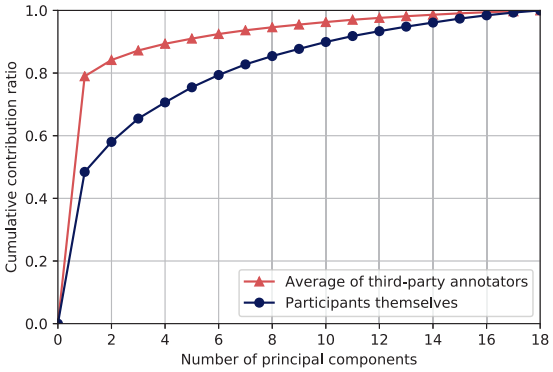


Figure 4: Cumulative contribution ratios by PCA for 18 rapport items.

also suggest that the upper-level items are mostly related to the content of the conversation (e.g., unsatisfying, engrossing, and boring). In contrast, the lower-level items are related to the feeling and atmosphere of the dialogue (e.g., friendly, positive, and cooperative).

We also applied principal component analysis (PCA) to the results of the answers to the 18 rapport items for each of those by participants themselves and the averages by the third-party annotators. Table 4 lists the cumulative contribution ratio of the PCA. The contribution ratios of the first principal components were 0.790 and 0.484 for the answers by the third-party annotators and participants themselves, respectively. These results indicate that one dimension could explain about 80% of the answers by the third-party annotators; in other words, the third-party annotators tended to answer the 18 items on the basis of rather simple impressions of positive or negative. In contrast, the participants presumably answered after considering more complicated inner impressions of the dialogue that they were actually participating in.

4.3 Personality traits

Table 5 shows the correlations between the personality traits reported by the participants themselves and the averages given by the five third-party annotators. The correlation coefficients for extraversion were consistently large and statistically significant among the versions, but the overall tendency appears to be that the other personality traits by the participants themselves do not necessarily correlate with the averages by the third-party annotators. Openness had the next largest correlation coefficient, followed by conscientiousness. The reason extraversion had high correlation coefficients is that it (by definition) tends to be more easily expressed during dialogue. This result is consistent with an experiment in the psychology field (Borkenau et al., 2009) in which extraversion was reported to be highly consistent between self-rating and rating by others.

It makes sense that the annotation results do not necessarily correlate if the personality traits of the participants are not sufficiently expressed in the dialogue, e.g., for neuroticism and agreeableness. This is because third-party annotators do not know the participants and score personality traits based only on their impression during the dialogue.

4.4 Relation among annotation results by third-party annotators

We investigated the correlations among the above annotation results given by the third-party annotators for sentiments, 18 rapport items, and personality traits. Table 6 lists the correlation of each of the five personality traits with the averages of the 18 rapport items and sentiments. As we can see, the average of the 18 rapport items correlated with all of the five personality traits with statistical significance, especially for agreeableness, extraversion, and openness, whose correlation coefficients were 0.68, 0.53, and 0.52, respectively. Similarly, the average of sentiments correlated with three personality traits (openness, agreeableness, and extraversion) with statistical significance; their correlation coefficients were 0.36, 0.30, and 0.21, respectively. In addition, the average of the 18 rapport items

Table 5: Correlation between personality traits given by participants themselves and averages given by third-party annotators.

	No. of participants	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Hazumi1911	30	0.53	0.08	0.43	0.25	0.29
Hazumi2010	33	0.58	−0.44	0.17	0.10	0.34
Hazumi2012	63	0.39	0.19	0.11	0.14	0.19
Hazumi2105	29	0.57	0.37	0.06	0.21	0.17
Total	155	0.49	0.06	0.16	0.15	0.21

Underlined values indicate statistical significance ($p < 0.05$).

Table 6: Correlation of personality traits with 18 rapport items and sentiments. All are averages given by five third-party annotators.

	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Average of 18 rapport items	0.53	0.68	0.21	−0.22	0.52
Average of sentiments	0.21	0.30	0.12	0.00	0.36

Underlined values indicate statistical significance ($p < 0.05$).

also correlated with the average of sentiments with statistical significance; its correlation coefficient was 0.55.

These results confirm that there were correlations between the three annotation results given by the five third-party annotators, thereby demonstrating that these three metrics can help each other in their estimation using machine learning. For example, in a dialogue where the participant seemed to enjoy talking, the average of the sentiments was high, the average of the 18 rapport items was also high, and the participant's extraversion, cooperativeness, and openness also seemed high. This simple tendency is echoed our discussion about the results of the PCA analysis in Section 4.2: that is, the third-party annotators tended to annotate on the basis of rather simple impressions of positive or negative.

5 Analyses on Dispersion of Third-Party Sentiments

We here focus on the dispersion of sentiments given by the five third-party annotators (third-party sentiments). We discuss the conditions under which the third-party sentiments would be reliable.

5.1 Formulating dispersion of third-party sentiments

The bottom line in Fig. 3 shows the standard deviations of the third-party sentiments for each exchange. Using this as a basis, we formulate the dispersion of third-party sentiments as the averages of the standard deviations, as follows.

Let $dispersion(i)$ denote the dispersion of third-party sentiments for a participant i (i.e., dialogue). Values a_{ijk} denote third-party sentiments for the j-th exchange ($j = 1, ..., J_i$) in the dialogue with participant i by the k-th third-party annotator ($k = 1, ..., K$). The values of sentiments are annotated on a 7-point scale, i.e., $a_{ijk} \in \{1, ..., 7\}$. J_i denotes the total number of exchanges in the dialogue with participant i, which is 95 in the example in Fig. 3. K denotes the number of third-party annotators, i.e., $K = 5$. Standard deviations of the annotated sentiments

$$stdev(i, j) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(a_{ijk} - \overline{a_{ij}})^2} \quad (1)$$

can be calculated per exchange. Here, $\overline{a_{ij}}$ denotes the averages of the third-party sentiments given by K annotators for the j-th exchange. We define $dispersion(i)$ of third-party sentiments for a participant i (i.e., dialogue) as the average of the standard deviations $stdev(i, j)$, i.e.,

$$dispersion(i) = \frac{1}{J_i}\sum_{j=1}^{J_i} stdev(i, j). \quad (2)$$

5.2 Relationship between dispersion and machine learning performance

Here, we discuss the relationship between the dispersion of third-party sentiments and the performance of machine learning. This explains why we focused on the dispersion.

It is known empirically that machine learning performs better when the manual annotations agree more. For example, in an emotion recognition task for spoken utterances, it was reported that the recognition performance based on machine learning was

Table 7: Correlation between the dispersion of third-party sentiments and personality traits.

	No. of participants	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Hazumi1911	30	0.24	−0.44	0.16	0.12	0.27
Hazumi2010	33	0.38	−0.38	−0.15	0.11	0.04
Hazumi2012	63	−0.13	−0.20	0.00	0.08	−0.05
Hazumi2105	29	−0.20	−0.13	0.29	−0.04	0.03
Total	155	−0.05	−0.26	−0.05	0.03	−0.04

Underlined values indicate statistical significance ($p < 0.05$).

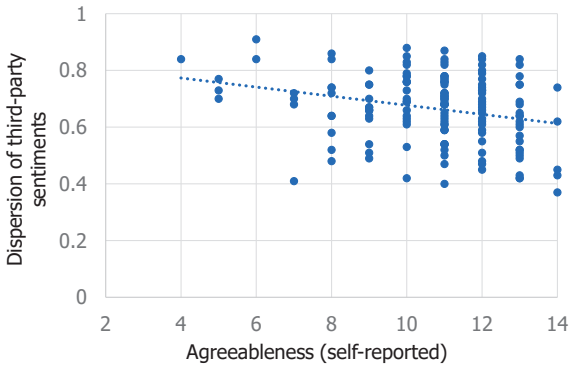


Figure 5: Correlation between the dispersions and the scores of agreeableness for each participant.

better on units to which multiple annotators gave the same labels (Seppi et al., 2008). As a preliminary investigation, we also calculated the mean square error of the sentiment estimation results as the regression from multimodal features (Katada et al., 2020) using 2,468 exchanges in Hazumi1911, where the references were the average of third-party sentiments. The correlation coefficient between the mean square errors and the standard deviations of the third-party sentiments per exchange was 0.342, which is statistically significant ($p = 1.57 \times 10^{-68}$). In other words, the error in machine learning results tends to be larger for exchanges with large standard deviations of third-party sentiments.

These results suggest that the sentiment estimation performance based on machine learning tends to be lower for parts with large deviations in human judgment.

5.3 Correlation between dispersion and personality traits

We investigated the correlation between the dispersion of third-party sentiments and the personality traits of each participant. The personality traits utilized here are those reported by the participants themselves. Table 7 lists the correlation coefficients between each of the five personality traits and the dispersions of third-party sentiments per participant (calculated by Eq. (2)), for each of the four versions and in total. We can see here that agreeableness negatively correlates with the dispersion of third-party sentiments. Specifically, the correlation coefficient with agreeableness was -0.26 for the total, which is statistically significant ($p = 9.1 \times 10^{-4}$).

Figure 5 shows the dispersions of the third-party sentiments and the scores of agreeableness. Each point denotes 155 participants from Hazumi1911 to Hazumi2105. Horizontal and vertical axes denote the score of agreeableness and the dispersions of the third-party sentiments for each participant. We can see here that there is a negative correlation between these two metrics. In other words, there were fewer dispersions of the third-party sentiments for the participants who recognized themselves as more agreeable. This result can be interpreted as a phenomenon that the more agreeable the participant is, the more he/she tries to express his/her sentiments in a way that the interlocutor (and thus the third-party annotators) can recognize.

The results of the sentiment estimation for highly agreeable users thus tend to be reliable, given the low dispersion of the third-party sentiments, which tend to correlate with machine learning performance, as discussed in Section 5.2.

6 Conclusion

In this paper, we investigated the correlation of subjective annotation results between the participants themselves and five third-party annotators. We found that some are correlated, which will potentially be useful in machine learning to estimate one of the annotation targets, such as the participants' sentiments, their evaluation of dialogues (18 rapport items), or their personality traits.

We also investigated the dispersion of the sentiments given by the five third-party annotators. We showed that a difference in annotation results correlates with the estimation error of machine learning and found that the dispersion was negatively cor-

related with agreeableness, one of the Big Five personality traits.

These results can provide insights into the development of adaptive dialogue systems: specifically, a personality trait can be used as a clue to determine whether or not to rely on the sentiment recognition results. One of our future works is to estimate the user's personality traits before and during the dialogue. The system can then utilize the personality trait to decide how actively to adapt to the user on the basis of the discussion in this paper. Personality traits such as neuroticism are not expressed by users during dialogues such as chit-chat and thus are difficult for the system and third-party annotators to observe. The analyses in this paper considered all of the Big Five traits, but it will be necessary to select personality traits observable in the dialogue accordingly, e.g., extraversion.

The results presented in this paper are based on our Japanese dataset Hazumi. Various factors such as the behavior of the participants and annotators, for example, can be involved. Further investigation is needed to confirm the generalizability of the obtained results to other languages and cultures, as well as to different experimental settings including dialogue tasks and instructions to the participants.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers JP22H00536 and JP19H05692, and JST Moonshot R&D Grant Number JPMJPS2011.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Oya Aran and Daniel Gatica-Perez. 2013. One of a kind: Inferring personality impressions in meetings. In *Proc. International Conference on Multimodal Interaction (ICMI)*, pages 11–18.

Pablo Alves De Barros, Ana Tanevska, and Alessandra Sciutti. 2021. Affect-aware learning for social robots. In *Adjunct Proc. Conference on User Modeling, Adaptation and Personalization*, pages 130–132.

Frank J. Bernieri, John S. Gillis, Janet M. Davis, and Jon E. Grahe. 1996. Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology*, 71(1):110–129.

Peter Borkenau, Steffi Brecke, Christine Mottig, and Marko Paelecke. 2009. Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, 43(4):703–706.

Nadege Bourvis, Aveline Aouidad, Michel Spodenkiewicz, Giuseppe Palestra, Jonathan Aigrain, Axel Baptista, Jean-Jacques Benoliel, Mohamed Chetouani, and David Cohen. 2021. Adolescents with borderline personality disorder show a higher response to stress but a lack of self-perception: Evidence through affective computing. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 111:110095.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Na rayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2019. Multimodal human-human-robot interactions (MHHRI) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, 10(4):484–497.

Daniel Fernau, Stefan Hillmann, Nils Feldhus, Tim Polzehl, and Sebastian Möller. 2022. Towards personality-aware chatbots. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 135–145.

Roger Ferrod, Federica Cena, Luigi Di Caro, Dario Mana, and Rossana Grazia Simeoni. 2021. Identifying users' domain expertise from dialogues. In *Adjunct Proc. Conference on User Modeling, Adaptation and Personalization*, pages 29–34.

R. Lewis Goldberg. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, pages 1216–1229.

Yuki Hirano, Shogo Okada, Haruto Nishimoto, and Kazunori Komatani. 2019. Multitask prediction of exchange-level annotations for multimodal dialogue systems. In *Proc. International Conference on Multimodal Interaction (ICMI)*, page 85–94.

Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. 2020. Is she truly enjoying the conversation? Analysis of physiological signals toward adaptive dialogue systems. In *Proc. International Conference on Multimodal Interaction (ICMI)*, pages 315–323.

Shun Katada, Shogo Okada, and Kazunori Komatani. 2022. Effects of physiological signals in different types of multimodal sentiment estimation. *IEEE Transactions on Affective Computing*.

Masanori Kimura, Masao Yogo, and Ikuo Daibo. 2005. Expressivity halo effect in the conversation about emotional episodes (in Japanese). *The Japanese Journal of Research on Emotions*, 12(1):12–23.

Kazunori Komatani and Shogo Okada. 2021. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G. Okuno. 2005. User modeling in spoken dialogue systems to generate flexible guidance. *User Modeling and User-Adapted Interaction*, 15(1):169–183.

Akinobu Lee, Keiichiro Oura, and Keiichi Tokuda. 2013. MMDAgent –a fully open-source toolkit for voice interaction systems–. In *Proc. IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, pages 8382–8385.

François Mairesse and Marilyn A. Walker. 2010. Towards personality-based user adaptation: Psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20(3):227–278.

Lucien Maman, Gualtiero Volpe, and Giovanna Varni. 2022. Training computational models of group processes without groundtruth: The self- vs external assessment's dilemma. In *Companion Publication of the 2022 International Conference on Multimodal Interaction (ICMI)*, pages 18–23.

Takashi Minato, Ryuichiro Higashinaka, Kurima Sakai, Tomo Funayama, Hiromitsu Nishizaki, and Takayuki Nagai. 2022. Overview of dialogue robot competition 2022. *arXiv preprint arXiv:2210.12863*.

Pedro Mota, Maike Paetzel, Andrea Fox, Aida Amini, Siddarth Srinivasan, and James Kennedy. 2018. Expressing coherent personality with incremental acquisition of multimodal behaviors. In *Proc. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 396–403.

Atsushi Oshio, Shingo Abe, and Pino Cutrone. 2012. Development, reliability, and validity of the Japanese version of ten item personality inventory (TIPI-J) (in Japanese). *The Japanese Journal of Personality*, 21(1):40–52.

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. 2021. Recipes for building an open-domain chatbot. In *Proc. European Chapter of the Association for Computational Linguistics (EACL)*, pages 300–325, Online.

Hanan Salam, Oya Celiktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Chetouani. 2017. Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access*, 5:705–721.

Dino Seppi, Anton Batliner, Björn Schuller, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, and Vered Aharonson. 2008. Patterns, prototypes, performance: classifying emotional user states. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 601–604.

Michel Spodenkiewicz, Jonathan Aigrain, Nadège Bourvis, Séverine Dubuisson, Mohamed Chetouani, and David Cohen. 2018. Distinguish self- and hetero-perceived stress through behavioral imaging and physiological features. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 82:107–114.

Khiet P. Truong, David A. van Leeuwen, and Franciska M.G. de Jong. 2012. Speech-based recognition of self-reported and observed emotion in a dimensional space. *Speech Communication*, 54(9):1049–1063.

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Chongyang Wang, Yuan Gao, Chenyou Fan, Junjie Hu, Tin Lun Lam, Nicholas Donald Lane, and Nadia Berthouze. 2023. Learn2agree: Fitting with multiple annotators without objective ground truth. In *ICLR 2023 Workshop on Trustworthy Machine Learning for Healthcare*.

Yuyan Wu, Miguel Arevalillo Herráez, Stamos Katsigiannis, and Naeem Ramzan. 2022. On the benefits of using hidden markov models to predict emotions. In *Proc. Conference on User Modeling, Adaptation and Personalization*, pages 164–169.

Kenta Yamamoto, Koji Inoue, and Tatsuya Kawahara. 2023. Character adaptation of spoken dialogue systems based on user personalities. In *Proc. International Workshop on Spoken Dialogue System Technology (IWSDS)*.

Frame-oriented Summarization of Argumentative Discussions

Shahbaz Syed [†] **Timon Ziegenbein** [‡] **Philipp Heinisch** [♠]
Henning Wachsmuth [‡] **Martin Potthast** [†◇]
[†]Leipzig University [‡]Leibniz University Hannover [♠]Bielefeld University [◇]ScaDS AI
<shahbaz.syed@uni-leipzig.de>

Abstract

Online discussions on controversial topics with many participants frequently include hundreds of arguments that cover different framings of the topic. But these arguments and frames are often spread across the various branches of the discussion tree structure. This makes it difficult for interested participants to follow the discussion in its entirety as well as to introduce new arguments. In this paper, we present a new rank-based approach to extractive summarization of online discussions focusing on argumentation frames that capture the different aspects of a discussion. Our approach includes three retrieval tasks to find arguments in a discussion that are (1) relevant to a frame of interest, (2) relevant to the topic under discussion, and (3) informative to the reader. Based on a joint ranking by these three criteria for a set of user-selected frames, our approach allows readers to quickly access an ongoing discussion. We evaluate our approach using a test set of 100 controversial Reddit ChangeMyView discussions, for which the relevance of a total of 1871 arguments was manually annotated.

1 Introduction

Web-based forums like Reddit facilitate discussions on all kinds of topics. Given the size and scope of some communities (known as "Subreddits"), multiple individuals regularly participate in the discussions of timely controversial topics, such as on ChangeMyView.[1] Notably, the volume of arguments tends to grow substantially in a tree-like response structure wherein each branch forms a concurrent discussion thread. These threads develop in parallel as different perspectives are introduced by the participants. After a discussion subsides, the resulting collection of threads and their arguments often represents a comprehensive overview of the most pertinent perspectives (henceforth, referred to as *frames*) put forth by the participants.

Frames help shape one's understanding of the topic and deliberating one's own stance (Entman, 1993; Chong and Druckman, 2007). However, in large discussions, prominent arguments as well as the various frames covered may be distributed in arbitrary (and often implicit) ways across the various threads. This makes it challenging for participants to easily identify and contribute arguments to the discussion. Large online forums like Reddit typically provide features that enable the reorganization of posts, for example, based on their popularity, time of creation, or in a question–answer format. A popularity-based ranking may seem beneficial, but Kano et al. (2018) discovered that an argument's popularity is not well correlated with its informativeness. Furthermore, a popularity-based ranking does not cover the breadth of frames of a discussion, as we will show in this paper (Section 4.1).

In this paper, we cast discussion summarization as a ranking task with an emphasis on frame diversity, thereby introducing a new paradigm to discussion summarization in the form of *multiple* summaries per discussion (one per frame). Previous research has focused on creating a single summary per discussion instead (Section 2). As illustrated in Figure 1, we first assign arguments to one or more frames. Next, we re-rank arguments in a frame according to their topic relevance. Additionally, we also rank them based on their informativeness via post-processing. Finally, we fuse these rankings to create the final ranking from which the top-k candidates can be used as an *extractive* summary of the discussion centered around a specific frame.

In our experiments, we explore various state-of-the-art methods to realize the three steps of our approach. Our results suggest that: (1) Utilizing retrieval models together with query variants is an effective method for frame assignment, reducing the reliance on large labeled datasets. Here, our approach outperforms a state-of-the-art supervised baseline. (2) Re-ranking arguments of a frame

[1](CMV) https://www.reddit.com/r/changemyview/

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 114–129
September 11–15, 2023. ©2023 Association for Computational Linguistics

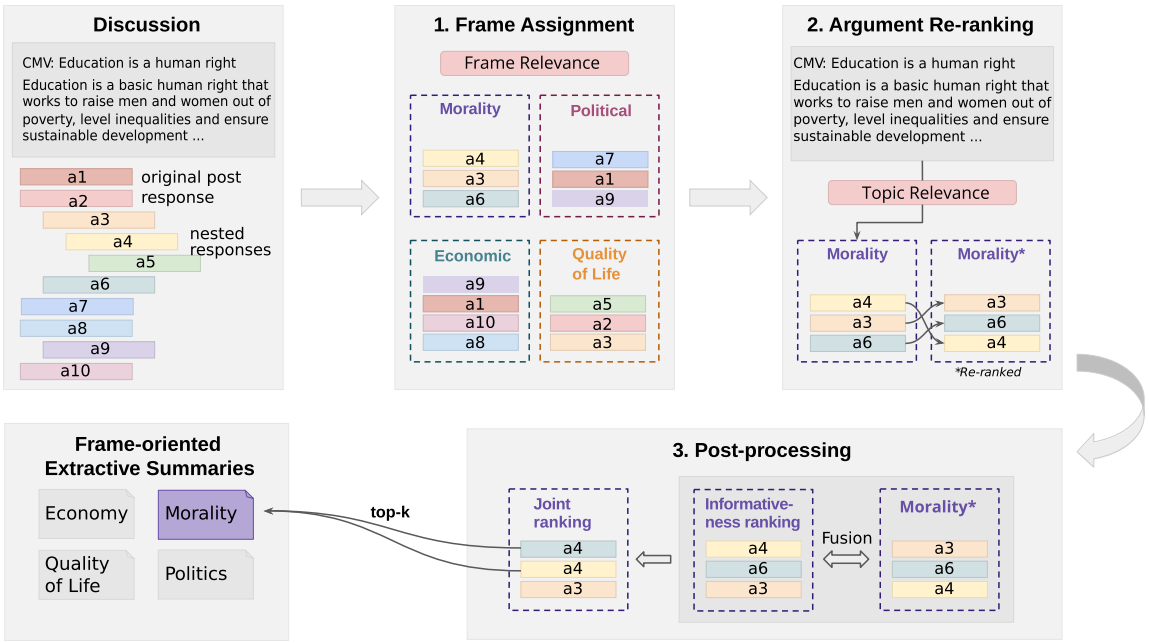


Figure 1: The proposed modular approach to frame-oriented discussion summarization: *1. Frame assignment* assigns arguments to frames ensuring frame relevance. *2. Argument re-ranking* ensures topic relevance of a frame's arguments (here, the *morality* frame is exemplified). *3. Post-processing* fuses the re-ranked arguments with an informativeness ranking. The top-k arguments are then taken as an extractive summary of the discussion.

based on content overlap with the discussion topic is more effective than retrieval-based approaches for ensuring the relevance of the frame's arguments to the topic. (3) Post-processing the argument rankings based solely on content features is insufficient to signal informativeness.

In summary, our contributions include: (1) A fully unsupervised frame assignment approach that assigns one or more frame labels to every argument within a discussion (Section 3.1). (2) An argument retrieval approach that ranks frame-specific arguments based on their topic relevance and informativeness (Section 3.2). (3) A dataset consisting of 1871 arguments sourced from 100 ChangeMyView discussions, where each argument has been judged in terms of frame relevance, topic relevance, and informativeness (Section 4) which forms the basis for an extensive comparative evaluation (Section 5).[2]

2 Related Work

Previous approaches to summarizing discussions can be broadly classified into two categories: *discussion unit extraction* and *discussion unit grouping*. We survey the literature on discussion summarization according to these two categories, followed by the literature on *argument framing*.

[2]Code and data: https://github.com/webis-de/SIGDIAL-23

2.1 Discussion Unit Extraction

Extraction-based approaches use either heuristics or supervised learning to identify important units, such as key phrases, sentences, or arguments within a discussion, then presented as the summary.

Tigelaar et al. (2010) identified several features for identifying key sentences from the discussion, such as the use of explicit author names to detect the response-tree structure, quoted sentences from the preceding arguments, and author-specific features such as participation and talkativity. They found that, while these features can be helpful, summarizing discussions primarily involves balancing coherence and coverage in the summaries. Ren et al. (2011) developed a hierarchical Bayesian model trained on labeled data to track the various topics within a discussion and a random walk algorithm to greedily select the most representative sentences for the summary. Ranade et al. (2013) extracted relevant and sentiment-rich sentences from debates, using lexical features to create indicative summaries. Bhatia et al. (2014) leveraged manually annotated dialogue acts to extract key posts as a concise summary of discussions on question-answering forums (Ubuntu, TripAdvisor). This dataset was further extended with more annotations by Tarnpradab et al. (2017) who proposed a

hierarchical attention network for extractive summarization of forum discussions. Egan et al. (2016) extracted key content from discussions via "point" extraction derived from a dependency parse graph structure, where a point is a verb together with its syntactic arguments.

Closely related to the domain we consider, Kano et al. (2018, 2020) studied the summarization of non-argumentative discussions on Reddit. They found that using the karma scores of posts was not correlated with their informativeness and that combining both local and global context features for comments was the most effective way to identify informative ones. Therefore, we do not rely on karma scores in our post-processing module (Section 4.2) and instead extract several content features for computing informativeness.

The outlined approaches all create a single summary for the entire discussion via end-to-end models. In contrast, we model the extraction of informative arguments organized by frames, thus enabling diverse summaries for a discussion. Furthermore, our experiments with unsupervised retrieval models for frame assignment (Section 4.2) enable us to assess the need to create labeled datasets beforehand to develop strong frame-oriented summarization models tailored to discussions.

2.2 Discussion Unit Grouping

Grouping-based approaches first categorize a discussion's units into explicit (or implicit) classes, such as queries, aspects, topics, dialogue acts, argument facets, or expert-labeled keypoints, and then generate individual summaries for each class. They rely on specific reference points to organize a discussion's units, providing flexibility to the readers by allowing them to choose from diverse summaries that best fit their information needs.

Qiu and Jiang (2013) modeled the discovery of latent viewpoints to group arguments based on two user characteristics: *user identity*, as arguments from the same user are likely to contain the same viewpoint; and *user interaction*, as users with different viewpoints may express disagreement or attack each other, while those with similar viewpoints may support each other. Misra et al. (2015) used summarization to discover repeating arguments and grouped them into facets. Reimers et al. (2019) proposed agglomerative clustering via contextual embeddings to identify similar arguments on a sentence level based on their aspects.

Nguyen et al. (2021) proposed an unsupervised approach to class-specific abstractive summarization of customer reviews with the goal of reducing generic and uninformative content in summaries. They model reviews in the context of topical classes of interest, which are treated as latent variables. These classes represent their reference points as latent variables to be discovered through supervised or reinforcement learning. In contrast, our frame inventory provides a more controlled—and thus more interpretable—set of reference points for discussion summarization. More recently Shapira et al. (2022) proposed a query-assisted, sentence-level interactive summarization approach for news reports using reinforcement learning. Their approach consists of two subtasks of query-based sentence selection and generating query suggestions to enable an interactive setting. In our scenario, we enable this interaction via the predefined set of frames.

Summarizing public debates, Bar-Haim et al. (2020a,b) investigated mapping similar arguments to expert-written key points. Bražinskas et al. (2021) summarized product reviews by selecting subsets of informative reviews, treating the choice of review subset as a latent variable that is learned by a model trained on a dataset compiled from professional product review forums. Amplayo et al. (2021) proposed aspect-controlled opinion summarization via employing multi-instance learning on a labeled dataset to identify aspects in reviews for grouping followed by summarization. The reference points of these approaches are defined either through manual annotations or distant supervision. Some of these reference points are highly topic-specific, requiring them to be created manually for each topic, for instance, the key points from Bar-Haim et al. (2020a). In contrast, we use a fixed and topic-independent set of reference points, namely media frames (Boydstun et al., 2014), grounded in framing theory (Chong and Druckman, 2007).

2.3 Argument Framing

Framing theory was initially utilized to categorize (political) newspaper articles in order to manifest the specifically reported perspective (Neuman et al., 1992; Semetko and Valkenburg, 2006; Boydstun et al., 2014). It was first introduced to the field of argumentation by Naderi and Hirst (2017). Later, Ajjour et al. (2019) modeled framing in argumentation more systematically, introducing automatically extracted, fine-grained, issue-specific frame labels.

Heinisch and Cimiano (2021) successfully combined computational argumentation with framing theory by showing a latent connection between the different frame granularities for the media frames defined by Boydstun et al. (2014). Hartmann et al. (2019) also used frame-labeled data from newswire corpus to successfully train frame classifiers for political discussions via multi-task and adversarial learning. Following the literature, we use the media frames due to their wide adoption in categorizing arguments (Card et al., 2015; Chen et al., 2021).

3 Ranking-based Summarization

This section describes our ranking-based approach to the extractive summarization of online discussions, centered around argumentation frames (Figure 1). First we describe our novel unsupervised approach for frame assignment, followed by methods for re-ranking arguments of a frame based on their relevance to the discussion topic and informativeness. The top-k arguments from the joint ranking are taken as the frame's summary.

3.1 Frame Assignment

Our approach to frame assignment IRFRAME is completely unsupervised in that it employs information retrieval models to rank arguments in a discussion by their *frame relevance*. Here, we consider arguments as documents and frames as queries. This offers a basic and interpretable alternative to frame assignment that does not require labeled data to train supervised models. We investigated both lexical and dense retrieval models.

We used an existing inventory of media frames to organize the arguments in a discussion. This originates from Boydstun et al. (2014) and consists of the 15 frames listed in Table 1. This inventory aims to support an issue-generic frame categorization of political communication. In the context of discussions on Reddit CMV, these issue-generic frames ideally cover a wide variety of controversial topics. The *other* frame is a catch-all category for frames that do not fit into any of the others. We excluded it from our experiments as it is not well-defined, and thus difficult to evaluate. For full frame descriptions see Table 4 in the appendix.

Employing query variants—semantically related queries derived from the primary query—has been shown to improve the retrieval performance (Benham et al., 2019). Thus, we manually created ten query variants for each frame to retrieve and rank all arguments in the discussion based on their frame relevance. Each variant is a high-quality sentence describing the various *aspects* of a frame. We manually curated these sentences from the Wikipedia pages of the frame labels as well as those of the various aspects mentioned in their descriptions (in Table 4). For example, a query variant for the frame *cultural identity* is: "Cultural identity is defined as the identity of a group or culture or of an individual as far as one is influenced by one's belonging to a group or culture and is similar to, and overlaps, with identity politics". The complete list of query variants for all frames is provided in the supplementary material. The output of this module is a ranked list of arguments for each frame, which is then used for extractive summarization (Section 3.2).

Frame Inventory	
Capacity & Resources	Health & Safety
Constitutionality & Jurisprudence	Morality
Crime & Punishment	Policy Prescription & Evaluation
Cultural Identity	Political
Economic	Public Opinion
External Regulation & Reputation	Quality of Life
Fairness & Equality	Security & Defense
	Other

Table 1: Inventory of frames proposed by Boydstun et al. (2014) to track the media framing on policy issues.

We first obtained ten rankings of the arguments (one for each query variant) and then combined these via reciprocal rank fusion (Cormack et al., 2009) to obtain the final list of ranked arguments for a frame. We also compare our approach with a supervised baseline, SUPERFRAME, a classifier finetuned on a set of labeled arguments (details in Section 4.2).

3.2 Extractive Summarization

Building upon the frame assignment component described above that ensures frame relevance, we now perform an *extractive* summarization of the discussion by re-ranking the frame-relevant arguments based on their relevance to the discussion topic and informativeness. This modular approach to summarizing discussions does not require expensive ground-truth summaries, and is thus more scalable than supervised approaches. We first describe the argument re-ranking module followed by the post-processing module.

Argument Re-ranking Besides being relevant to a frame, arguments in the summary must also be relevant to the discussion topic. Thus, we re-rank the frame's arguments according to their *topic relevance*. In our scenario, a "topic" is the combination of the title and the reasoning of the original post on CMV. We propose two approaches for computing topic relevance. The first approach computes content overlap (lexical and semantic) between each argument and the topic. We used Jaccard similarity for lexical overlap, and for semantic overlap, we used the cosine similarity between the contextual sentence embeddings of an argument and the topic. Arguments within a frame are then re-ranked by their overlap scores. The second approach employs retrieval models and (re-)ranks the frame's arguments using the entire topic as the query (details in Section 4).

Post-processing Parallel to the aforementioned re-ranking by topic relevance, we derive a separate re-ranking of the frame's arguments based on their *informativeness*. Our goal is to prioritize content-rich and argumentative texts in the top-k arguments of our approach. We operationalize this through *content scoring* and *argumentativeness scoring*. For content scoring we employed a set of content-specific features such as named entities, noun phrases, the number of discourse markers, and the number of children an argument has in the discussion. Next, for argumentativeness scoring, we trained a topic-based argumentativeness scoring model (details in Section 4). The informativeness score of an argument is the sum of its content score and the argumentativeness score. We then re-rank the frame's arguments by this score.

Frame-oriented Extractive Summaries Given the list of arguments first ranked by frame relevance, then re-ranked by topic relevance, we fuse this ranking with the standalone informativeness ranking from the post-processing module (via reciprocal rank fusion) to derive the final ranking. The top-k arguments from this ranking are taken as the *extractive* summary of the discussion. A key benefit of our ranking-based extractive summarization approach is the flexibility to determine the summary length (i.e., k) by the user according to the discussion's length and their information need. Thus we refrain from setting a specific length budget for the summary.

4 Data and Experiments

This section describes the dataset on which our approach was evaluated, the various retrieval models with their respective parameters, and the content features that we used in our experiments. Also described is the supervised baseline for frame classification SUPERFRAME that we implemented to assign multiple frames to each argument.

4.1 Data

We constructed a dataset of 100 long discussions from CMV, dated January 2020, using the Pushshift Reddit dataset (Baumgartner et al., 2020). For the purpose of this study, we defined a long discussion as a post with at least 100 comments. As preprocessing, we filtered out comments that were deleted by their authors, removed by moderators due to violating community rules, or posted by bots (e.g., DeltaBot, RemindMeBot). The average length of the posts in our dataset is 304 words, with a minimum of 83 words and a maximum of 1611 words. These posts have a total of 25,385 comments, with an average of 253 comments per discussion. The shortest discussion has 105 comments, while the longest has 1066 comments. The average length of a comment is 90 words, with a minimum of 2 words and a maximum of 1589 words excluding the quoted text from either the post or the parent comments they responded to.[3]

Popularity Ranking We investigated to what extent does ranking the arguments only by their popularity (via karma scores on Reddit) cover all the top-k arguments of the frames in the discussion (as assigned by our approach). To quantify this, we computed the mean coverage of the top 10 arguments across all frames and models by their popularity ranking. We considered discussions with at least 500 arguments and ranked them by their popularity scores provided by the Reddit API. Then, at each rank, we computed the percentage of top 10 arguments from all frames that have been covered by the popular arguments. Figure 2 shows that in order to completely cover the top 10 arguments from all frames, a user must read through hundreds of arguments. This encourages us to investigate novel approaches to group arguments in a discussion via

[3]The strict community guidelines of CMV (https://www.reddit.com/r/changemyview/wiki/rules) ensure that comments are primarily argumentative. Therefore, in this paper, we consider each comment to be an argument and do not perform any argument mining.

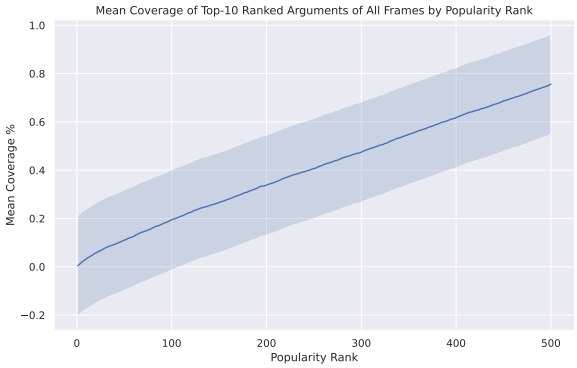


Figure 2: Mean coverage percentage by popularity rank of the top 10 (unique) frame arguments as assigned by our approaches.

frames instead of solely relying on their popularity. A similar conclusion was drawn by Kano et al. (2018) who investigated the effectiveness of popularity scores as a feature for summarizing Reddit discussions.

4.2 Experiments

We first describe the models and parameters for our approaches to frame assignment and extractive summarization. We then describe the supervised baseline for frame assignment.

Frame Assignment We experimented with three retrieval models for IRFRAME to retrieve frame-relevant arguments: BM25 (Robertson et al., 1994), SBERT (Reimers and Gurevych, 2019), and ColBERT (Khattab and Zaharia, 2020). The latter two are dense retrieval models based on contextual embeddings to match arguments to frames, addressing the limitation of BM25 not finding arguments with exact lexical matches to our query variants. We used the Okapi BM25 model with default settings (k=1.5, b=0.75),[4] initialized SBERT with the `all-mpnet-base-v2` model, and used ColBERT-v2 (Santhanam et al., 2022).[5]

Argument Re-ranking We experimented with two approaches to re-rank the arguments retrieved by IRFRAME: content overlap and retrieval-based re-ranking. Content overlap considers both lexical and semantic overlap between the topic and the argument. For lexical overlap, we used Jaccard similarity and for semantic overlap, we used SBERT (`all-mpnet-base-v2` model). For the retrieval-based re-ranking, we experimented with BM25 and ColBERT, with the topic as the query, to (re-)rank the frame's arguments. We excluded SBERT as an additional retrieval model since it is already integrated in the content overlap approach.

Post-processing Informativeness is computed based on the content richness and the argumentativeness of the arguments. Content is scored as the sum of the ratios of named entities, discourse markers, and noun phrases found in the argument and the number of children for an argument in the discussion. We used spaCy (Honnibal et al., 2020) for text tokenization and extraction of the named entities and noun phrases.[6] For discourse markers, we used a lexicon of claim-related words constructed by Levy et al. (2017) for identifying claim-containing sentences. The ratios of named entities and noun phrases were on the token level, while the ratio of discourse markers was on the word level, all normalized by the arguments' lengths. For argumentativeness, we developed *ArgDetector*,[7] a RoBERTa model (Liu et al., 2019) fine-tuned on the dataset by Schiller et al. (2022), containing 150 controversial topics with 144 sentences labeled for their argumentativeness, given the topic. Implementation details are described in Appendix A.

SUPERFRAME This is the supervised baseline for frame assignment. Extending the state-of-the-art frame classification model of Heinisch and Cimiano (2021), we developed a new classifier trained on an external frame-labeled dataset. The existing classifier of Heinisch and Cimiano (2021), utilizes a recurrent neural network to assign a *single* frame to an argument, and combines it with a model that predicts a cluster of frame labels from the inventory of Ajjour et al. (2019) in a multi-task setting. Particularly longer arguments, however, often contain multiple frames. Thus, assigning a single frame to an argument may not be sufficient (Reimers et al., 2019). We therefore extend the model to predict *multiple* frames for an argument. Given the probability distribution of the classification model $P = (p_{f_1}, \cdots, p_{f_k})$ over a set of frames $\mathcal{F} = \{f_1, \ldots, f_k\}$, $k \geq 2$, we apply nucleus sampling (Holtzman et al., 2020) to predict multiple frames for an argument. Specifically, given a cumulative probability mass threshold τ, we assign

[4]We used the Rank BM25 toolkit (Brown, 2020)

[5]We used PyTerrier (Macdonald and Tonellotto, 2020) for the ColBERT pipeline

[6]We used the `en_core_web_md` model.

[7]https://huggingface.co/pheinisch/roberta-base-150T-argumentative-sentence-detector

the minimal subset of frames $F \subseteq \mathcal{F}$ such that:

$$\sum_{f \in F} p_f \geq \tau$$

When the model is very confident in predicting one frame, it is hence likely that an argument is classified to that frame. In cases where the model has lower confidence in its prediction, the argument may consist of multiple frames. This overcomes the limitation of clustering-based approaches and classifiers which strictly assign a single frame to arguments that may contain multiple ones (Reimers et al., 2019; Heinisch and Cimiano, 2021).

To train SUPERFRAME, we used the Media Frames Corpus by Card et al. (2015) consisting of 14,515 news articles with text spans manually annotated for the frame classes in Table 1. Following Heinisch and Cimiano (2021), we trained two variants of the classifier, a *single-task* and a *multi-task* classifier which additionally used the framing dataset by Ajjour et al. (2019) with 12,326 labeled arguments. Both models were based on BiLSTMs, used GloVe embeddings,[8] and trained up to 12 epochs using early stopping. We truncated the input to 75 words with a batch size of 64. To choose between the *single-task* and *multi-task* variants, three of the authors first manually assigned frame(s) for 150 arguments. We then predicted the frames for these arguments using both variants.[9] We opted for higher precision as our goal is to minimize mislabeling arguments with an unrelated frame that can negatively impact the resulting frame-oriented summaries. Since frame assignment is a subjective task (Card et al., 2015) and the boundaries of the frame classes are fuzzy (Reimers et al., 2019; Budzynska et al., 2022), we observed some diversity in our manual annotations. Specifically, we observed that 92% of all the annotated arguments have at least one frame, which was assigned by only a single annotator (minority), indicating different perceptions of observing specific frames in texts. On average, an argument was assigned 3.8 frames (or 1.3 and 0.4 considering the majority and full agreements, respectively).

Table 2 presents the precision scores of both variants with cumulative probability threshold $\tau = 0.9$. Assigning only the most probable frame as predicted by the *single-task* model results in a precision of 59.6% (micro-average) and 49.6% (macro-average), respectively. The *multi-task* model is slightly better at predicting rare frame classes (+0.5% macro-average) but worse at predicting the frequent ones (-7.2% micro-average). Assigning multiple frames per argument increases the effectiveness of the *single-task* model by +0.9% (micro-average), and especially the prediction of rare frame classes, increasing the macro-average prevision by +5.8% (at $\tau = 0.9$).

Model	Minority	Majority	Full
single-task	59.6 / 49.6	**41.7 / 34.1**	**38.8 / 28.8**
single-$\tau = .8$	55.0 / 45.5	32.6 / 27.6	34.8 / 27.6
single-$\tau = .9$	**60.5 / 55.4**	27.8 / 24.5	30.4 / 23.7
multi-task	52.4 / 50.1	27.9 / 22.7	38.4 / 29.5
multi-$\tau = .8$	56.4 / 55.0	33.0 / 26.6	27.4 / 20.1
multi-$\tau = .9$	51.0 / 46.9	26.7 / 21.7	25.4 / 17.9

Table 2: Precision scores (micro / macro %) of the SUPERFRAME model variants at different annotator agreements and thresholds τ for multi-frame prediction.

Considering only the majority-labeled frame classes as ground truth restricts the set of manually assigned frame classes, and hence, reduces the precision scores. On this restricted subset of frame labels, the *single-task* model performs best in nearly all cases, by predicting only the most probable frame class due to the sparsity of the manually assigned frame classes. This variant of the *single-task* model which predicts only a single frame for an argument has a micro-averaged precision of 41.7% and 38.8% in the majority and full agreement scenarios, respectively. Despite this, we extended the *single-task* variant to predict multiple frames per argument, resulting in a high overlap with ground truth frame labels from at least one annotator as well as benefiting from a higher recall. This also avoids having sparse sets of arguments assigned under rare frames.

In conclusion, our internal evaluation supports using the *single-task* model, as opposed to the findings of Heinisch and Cimiano (2021) due to our emphasis on precision while the *multi-task* variant primarily encourages the model in its recall-generalization ability. On average, SUPERFRAME (*single-task* variant) assigned 2.6 frames per argument, with a minimum of 1 and a maximum of 8.2. The frequency counts of all frames in both posts and arguments are shown in Appendix Table 5.

[8] https://nlp.stanford.edu/data/glove.840B.300d.zip

[9] We also experimented with multiple preprocessing methods (e.g. generating a conclusion or ranking the sentences) before automatically predicting the frames. However, these methods negatively impacted the frame prediction.

5 Evaluation

Given that our entire approach is based on retrieval models, we evaluated it manually via relevance judgments. We followed the evaluation style of TREC (Harman, 1993) as best practice. Our evaluation was comprised of judging the *frame relevance*, the *topic relevance*, and the *importance* (in the discussion's context) of arguments retrieved by our models. Following the TREC protocol, we first created 50 evaluation topics, each comprising a post's title, the post itself, and a frame of interest (see supplementary material). To obtain a sufficiently large set of arguments to pool from, we then selected only those discussions for which all models assigned at least 20 arguments to each of the five most frequent frames identified in the comments: *cultural identity*, *economic*, *quality of life*, *public opinion*, and *political* (see Table 5 in the Appendix for the full list). We retrieved arguments for each evaluation topic and performed pooling at depth 5 using TrecTools (Palotti et al., 2019), resulting in 1871 unique arguments to be judged.

5.1 Pilot Study

Multi-annotator relevance judgments can often result in low agreement due to the subjective nature of defining *relevance* and the varying perspectives of annotators (Voorhees, 1998; Bailey et al., 2008; McDonnell et al., 2016; Thomas et al., 2022). Additionally, judges may experience inconsistencies in their decisions as the task progresses (Scholer et al., 2011). To mitigate these issues, we conducted a pilot study with 100 arguments (not included in the main evaluation) to train three annotators and gather feedback for improving the main evaluation interface. The annotators were Computer Science graduates with backgrounds in NLP and IR.

Task Design Following McDonnell et al. (2016), we used a four-point scale for assessing the frame and topic relevance, and the importance of an argument with these options: *definitely not*, *probably not*, *probably*, and *definitely* relevant/important.[10] In assessing importance, we asked annotators to indicate the relevance of an argument to a discussion by answering this question: "How important is the argument to be included in a *summary* of the discussion?". We also experimented with an automatic summary (Nathan, 2016) for long arguments to reduce the cognitive load of the annotators. They were instructed to use the summary if they found it helpful, otherwise to read the entire argument (for details, see Appendix B, Figure 3).

Pilot Agreement and Feedback We measured the inter-annotator agreement (IAA) for the three evaluated criteria using Krippendorff's α, similar to Card et al. (2015). The resulting α values were 0.22 for frame relevance, 0.33 for topic relevance, and 0.22 for importance, respectively. While the agreement is thus limited, the values are consistent with the findings of Card et al. (2015) in their annotation of frame-relevant text spans for the Media Frames Corpus, particularly the frame relevance α value. From feedback, we improved the task design for the main evaluation. Firstly, we removed the automatic summary for each argument since it did not provide significant help. Secondly, we rephrased the importance question to "How important is the argument to be included in the *discussion* of the given topic?" to make it more straightforward, since we did not have ground-truth summaries of the discussions at hand. Annotators also reported that assessing the relevance of an argument for a *single* frame was too restrictive, since an argument may belong to multiple frames, which aligns with the observations of Card et al. (2015). Therefore, we allowed them to assign multiple frames to an argument if the currently-assigned one was not relevant. Accordingly, we proceeded with the main evaluation by assigning each annotator an independent set of arguments to judge. This allowed us to collect more relevance judgments while ensuring a certain level of *shared* understanding of the task.

5.2 Main Evaluation Results

The evaluated models are shown in Table 3.[11] We obtained relevance judgments for a total of 1871 arguments and calculated nDCG@5 (Järvelin and Kekäläinen, 2002) as the effectiveness measure (mean over all topics). Described below are the key findings for each module of our ranking-based extractive summarization framework.

Frame Relevance Our frame assignment approach (*IRFr* with BM25) outperforms other models for identifying frame-relevant arguments in a

[10] We mapped these labels to numerical values ranging from 0 (*definitely not* relevant/important) to 3 (*definitely* relevant/important) for computing nDCG scores.

[11] Model names in Table 3 shortened for brevity. SUPERFRAME → *SupFr* denotes the baseline, IRFRAME → *IRFr* denotes our frame assignment approach, Argument Reranking → *_rr* (via overlap and retrieval models), and Postprocessing → *_post*

discussion with an nDCG@5 of 0.573. Among the retrieval models, BM25 performs better than SBERT and ColBERT, also for re-ranking by topic relevance. Upon further inspection, we found that BM25 often retrieves longer arguments compared to the embedding-based SBERT and ColBERT models. This may provide annotators with more context for informed judgments compared to the shorter arguments. Given the computational costs of running dense retrieval models in real-time, it is promising that a relatively simple and explainable model performs well on our query variants. For the baseline (*SupFr*), combinations with argument re-ranking (via BM25 and topic overlap) also perform reasonably well. However, as various query variants can be easily designed, our *IRFr* approach is more flexible and can be adapted to other domains and topics without the need for labeled data.

Topic Relevance Argument re-ranking by overlap (**_rr_overlap*) outperforms retrieval models for ensuring topic relevance of a frame's arguments. This benefits both *IRFr* and *SupFr* frame assignment approaches with an nDCG@5 scores of 0.847 and 0.785 for the top two models, respectively. Among the retrieval models, BM25 slightly outperforms ColBERT. Given the intuitive nature of content overlap, we conclude that it is favorable to use for re-ranking arguments in a frame.

Importance None of the post-processed models (using informativeness) appear in the top-5 for ranking arguments by importance in the context of the discussion. Instead, argument re-ranking by topic relevance performs best, with nDCG@5 of 0.381 combined with *SupFr* for frame assignment. This contradicts our intuition of post-processing to promote important arguments in the final ranking. As future work, we plan to investigate using context features of the arguments (Kano et al., 2018), as well as pairwise judgments for importance (Zopf, 2018; Luo et al., 2022).

Model	**nDCG@5**		
	Frame	Topic	Imp.
Our Approach			
IRFr_BM25	**0.573**[1]	0.708	0.375[2]
IRFr_SBERT	0.480	0.525	0.303
IRFr_ColBERT	0.522	0.659	0.361[3]
IRFr_BM25_rr_BM25	0.516	0.781[3]	0.349
IRFr_BM25_rr_overlap	0.560[2]	**0.847**[1]	0.350[5]
IRFr_BM25_rr_ColBERT	0.540[4]	0.761	0.358[4]
IRFr_BM25_rr_BM25_post	0.489	0.735	0.297
IRFr_BM25_rr_overlap_post	0.522	0.755	0.339
IRFr_BM25_rr_ColBERT_post	0.526	0.719	0.325
Supervised Baseline			
SupFr_rr_BM25	0.545[3]	0.765[4]	**0.381**[1]
SupFr_rr_overlap	0.536[5]	0.785[2]	0.334
SupFr_rr_ColBERT	0.529	0.764[5]	0.348
SupFr_rr_BM25_post	0.493	0.714	0.322
SupFr_rr_overlap_post	0.493	0.734	0.348
SupFr_rr_ColBERT_post	0.487	0.709	0.329

Table 3: nDCG@5 for the manual relevance judgments for frame relevance, topic relevance, and importance. The best results for each evaluated criterion are highlighted in bold, alongside the rankings for the five best models. We evaluated our frame assignment approach (*IRFr*) against the supervised baseline (*SupFr*), combined with our argument re-ranking (*_rr*) and post-processing components (*_post*). We see that our approach to frame assignment results in the best models for frame and topic relevance and is also competitive for argument importance.

6 Conclusion and Future Work

We introduced a novel ranking-based approach to frame-oriented (extractive) discussion summarization in web-based forums, aiming to enhance the accessibility and comprehension of large-scale online discussions for participants. Our approach involves three key steps: frame assignment, argument re-ranking, and post-processing. Specifically, we developed unsupervised methods for both frame and topic assignment leveraging standard retrieval models. Extensive experiments on a dataset of 1871 arguments from 100 ChangeMyView discussions demonstrate the effectiveness of our approach in ensuring frame and topic relevance in the summary, outperforming a state-of-the-art supervised baseline for frame assignment. Nevertheless, further exploration is needed to enhance summary informativeness through post-processing.

In the future, we plan to develop practical applications that leverage our approach for scalable exploration of online discussions guided by argumentation frames. Moreover, we will explore the application of our approach to summarize discussions in various Subreddits beyond ChangeMyView and across different debate portals.

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2915–2925, Hong Kong, China. Association for Computational Linguistics.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, pages 667–674. ACM.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4029–4039. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 39–49. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 830–839. AAAI Press.

Rodger Benham, Joel M. Mackenzie, Alistair Moffat, and J. Shane Culpepper. 2019. Boosting search performance using query variations. *ACM Trans. Inf. Syst.*, 37(4):41:1–41:25.

Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. 2014. Summarizing online forum discussions – can dialog acts of individual messages help? In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2127–2131, Doha, Qatar. Association for Computational Linguistics.

Amber E. Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9424–9442, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dorian Brown. 2020. Rank-BM25: A Collection of BM25 Algorithms in Python.

Katarzyna Budzynska, Chris Reed, Manfred Stede, Benno Stein, and Zhang He. 2022. Framing in communication: From theories to computation (dagstuhl seminar 22131). *Dagstuhl Reports*, 12(3):117–140.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2021. Controlled neural sentence-level reframing of news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2683–2693, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dennis Chong and James N. Druckman. 2007. *Framing theory*, Annual Review of Political Science, pages 103–126.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 758–759. ACM.

Charlie Egan, Advaith Siddharthan, and Adam Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 134–143, Berlin, Germany. Association for Computational Linguistics.

Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, 390:397.

Donna Harman. 1993. Overview of the second text retrieval conference (TREC-2). In *Proceedings of The Second Text REtrieval Conference, TREC 1993,*

Gaithersburg, Maryland, USA, August 31 - September 2, 1993, volume 500-215 of *NIST Special Publication*, pages 1–20. National Institute of Standards and Technology (NIST).

Mareike Hartmann, Tallulah Jansen, Isabelle Augenstein, and Anders Søgaard. 2019. Issue framing in online discussion fora. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1401–1407, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Heinisch and Philipp Cimiano. 2021. A multi-task approach to argument frame classification at variable granularity levels. *it - Information Technology*, 63(1):59–72.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Ryuji Kano, Yasuhide Miura, Motoki Taniguchi, Yan-Ying Chen, Francine Chen, and Tomoko Ohkuma. 2018. Harnessing popularity in social media for extractive summarization of online conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1145, Brussels, Belgium. Association for Computational Linguistics.

Ryuji Kano, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. 2020. Identifying implicit quotes for unsupervised extractive summarization of conversations. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 291–302, Suzhou, China. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, Copenhagen, Denmark. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ge Luo, Hebi Li, Youbiao He, and Forrest Sheng Bao. 2022. Prefscore: Pairwise preference learning for reference-free summarization quality assessment. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5896–5903. International Committee on Computational Linguistics.

Craig Macdonald and Nicola Tonellotto. 2020. Declarative experimentation in information retrieval using pyterrier. In *ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020*, pages 161–168. ACM.

Tyler McDonnell, Matthew Lease, Mücahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? collecting annotator rationales for relevance judgments. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, 30 October - 3 November, 2016, Austin, Texas, USA*, pages 139–148. AAAI Press.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440, Denver, Colorado. Association for Computational Linguistics.

Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 536–542, Varna, Bulgaria. INCOMA Ltd.

Paco Nathan. 2016. PyTextRank, a Python implementation of TextRank for phrase extraction and summarization of text documents.

W Russell Neuman, Russell W Neuman, Marion R Just, and Ann N Crigler. 1992. *Common knowledge: News and the construction of political meaning*. University of Chicago Press.

Thi Nhat Anh Nguyen, Mingwei Shen, and Karen Hovsepian. 2021. Unsupervised class-specific abstractive summarization of customer reviews. In *Proceedings of The 4th Workshop on e-Commerce and NLP*, pages 88–100, Online. Association for Computational Linguistics.

Joao Palotti, Harrisen Scells, and Guido Zuccon. 2019. Trectools: an open-source python library for information retrieval practitioners involved in trec-like campaigns. SIGIR'19. ACM.

Minghui Qiu and Jing Jiang. 2013. A latent variable model for viewpoint discovery from threaded forum posts. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1031–1040, Atlanta, Georgia. Association for Computational Linguistics.

Sarvesh Ranade, Jayant Gupta, Vasudeva Varma, and Radhika Mamidi. 2013. Online debate summarization using topic directed sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM 2013, Chicago, IL, USA, August 11, 2013*, pages 7:1–7:6. ACM.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Zhaochun Ren, Jun Ma, Shuaiqiang Wang, and Yang Liu. 2011. Summarizing web forum threads based on a latent topic propagation process. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 879–884, New York, NY, USA. Association for Computing Machinery.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3715–3734. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2022. On the effect of sample and topic sizes for argument mining datasets.

Falk Scholer, Andrew Turpin, and Mark Sanderson. 2011. Quantifying test collection quality based on the consistency of relevance judgements. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 1063–1072. ACM.

Holli A. Semetko and Patti M. Valkenburg Valkenburg. 2006. Framing European politics: A Content Analysis of Press and Television News. *Journal of Communication*, 50(2):93–109.

Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Ido Dagan, and Yael Amsterdamer. 2022. Interactive query-assisted summarization via deep reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2551–2568, Seattle, United States. Association for Computational Linguistics.

Sansiri Tarnpradab, Fei Liu, and Kien A Hua. 2017. Toward extractive summarization of online forum discussions via hierarchical attention networks. In *The Thirtieth International Flairs Conference*.

Paul Thomas, Gabriella Kazai, Ryen White, and Nick Craswell. 2022. The crowd is made of people: Observations from large-scale crowd labelling. In *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, March 14 - 18, 2022*, pages 25–35. ACM.

Almer S. Tigelaar, Rieks op den Akker, and Djoerd Hiemstra. 2010. Automatic summarisation of discussion fora. *Nat. Lang. Eng.*, 16(2):161–192.

Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 315–323. ACM.

Markus Zopf. 2018. Estimating summary quality with pairwise preferences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1687–1696. Association for Computational Linguistics.

A Argumentativeness Scoring

The dataset from Schiller et al. (2022) consists of topics formulated as phrases as opposed to the topic titles in CMV, which are often formulated as claims. To unify this, we manually transformed their topics by appending them with stance-indicative phrases (e.g., "Abortion" → "Abortion should be banned"). We trained the RoBERTa model for the binary classification task with default training parameters: a learning rate of 5e-5, 5% of the training data for

Frame	Description
Capacity & Resources	The lack of or availability of physical, geographical, spatial, human, and financial resources, or the capacity of existing systems and resources to implement or carry out policy goals.
Constitutionality & Jurisprudence	The constraints imposed on or freedoms granted to individuals, government, and corporations via the Constitution, Bill of Rights and other amendments, or judicial interpretation. This deals specifically with the authority of government to regulate, and the authority of individuals/corporations to act independently of government.
Crime & Punishment	Specific policies in practice and their enforcement, incentives, and implications. Includes stories about enforcement and interpretation of laws by individuals and law enforcement, breaking laws, loopholes, fines, sentencing and punishment. Increases or reductions in crime.
Cultural Identity	The social norms, trends, values and customs constituting culture(s), as they relate to a specific policy issue.
Economic	The costs, benefits, or monetary/financial implications of the issue (to an individual, family, community or to the economy as a whole).
External Regulation & Reputation	A country's external relations with another nation; the external relations of one state with another; or relations between groups. This includes trade agreements and outcomes, comparisons of policy outcomes or desired policy outcomes.
Fairness & Equality	Equality or inequality with which laws, punishment, rewards, and resources are applied or distributed among individuals or groups. Also the balance between the rights or interests of one individual or group compared to another individual or group.
Health & Safety	Healthcare access and effectiveness, illness, disease, sanitation, obesity, mental health effects, prevention of or perpetuation of gun violence, infrastructure and building safety.
Morality	Any perspective—or policy objective or action (including proposed action)— that is compelled by religious doctrine or interpretation, duty, honor, righteousness or any other sense of ethics or social responsibility.
Policy Prescription & Evaluation	Particular policies proposed for addressing an identified problem, and figuring out if certain policies will work, or if existing policies are effective.
Political	Any political considerations surrounding an issue. Issue actions or efforts or stances that are political, such as partisan filibusters, lobbyist involvement, bipartisan efforts, deal-making and vote trading, appealing to one's base, mentions of political maneuvering. Explicit statements that a policy issue is good or bad for a particular political party.
Public Opinion	References to general social attitudes, polling and demographic information, as well as implied or actual consequences of diverging from or getting ahead of public opinion or polls.
Quality of Life	The effects of a policy, an individual's actions or decisions, on individuals' wealth, mobility, access to resources, happiness, social structures, ease of day-to-day routines, quality of community life, etc.
Security & Defense	Security, threats to security, and protection of one's person, family, in-group, nation, etc. Generally an action or a call to action that can be taken to protect the welfare of a person, group, nation sometimes from a not yet manifested threat.
Other	Any frames that do not fit into the above categories.

Table 4: Descriptions of frames as per Boydstun et al. (2014). We substituted the term "policy" with the phrase "actions/decisions" to align the frame definitions with the individualistic style of arguments in CMV. Similarly, in *External Regulation & Reputation*, we substituted "United States" with "country" to generalize it.

warmup, early stopping, and a batch size of 32. On the test split provided by Schiller et al. (2022), our fine-tuned model performs with a macro-F1 of 67%, which is comparable with the results from the best model reported in Schiller et al. (2022). A text is labeled as argumentative if the output probability from the finetuned classifier is higher than 50%. Given an input text and the discussion topic we take the mean scores of its constituent sentences as the text's argumentativeness score.

Posts		Comments	
Frame	**Count**	**Frame**	**Count**
Cultural Identity	53	Cultural Identity	13,540
Quality of Life	37	Economic	8931
Economic	33	Quality of Life	8559
Public Opinion	26	Public Opinion	7257
Health & Safety	22	Political	5177
Political	19	Health & Safety	4927
Morality	12	Morality	4237
Policy Prescription & Evaluation	10	Policy Prescription & Evaluation	4108
Fairness And Equality	10	Constitutionality & Jurisprudence	3226
Constitutionality & Jurisprudence	9	Fairness & Equality	2457
Security & Defense	1	Crime & Punishment	898
Crime & Punishment	1	Security & Defense	515
		External Regulation & Reputation	216
		Capacity & Resources	169

Table 5: Counts of frames in posts and comments in our dataset of 100 discussions as predicted by `SuperFrame`. Since each text can be assigned multiple frames, the counts include duplicates. Here, we observe that there are two additional frames found in the comments: *External Reputation & Regulation*, *Capacity & Resources* that are not found in the posts.

B Annotation Interface

Annotation interfaces for the pilot study and the main evaluation are shown in Figures 3 and 4, respectively. We improved the interface for our main evaluation based on annotator feedback from the pilot study with the following changes: (1) We substituted "probably" with "rather" in our scales to indicate a clearer relevance judgment. (2) For non-argumentative texts or meta-arguments (e.g. "I agree.", "I don't understand what you mean." etc.), we allowed annotators to mark the text as *noisy* and skip it. (3) We asked annotators to select at least one relevant frame if the current frame was (definitely/rather) not relevant, with the possibility of selecting multiple frames if required.

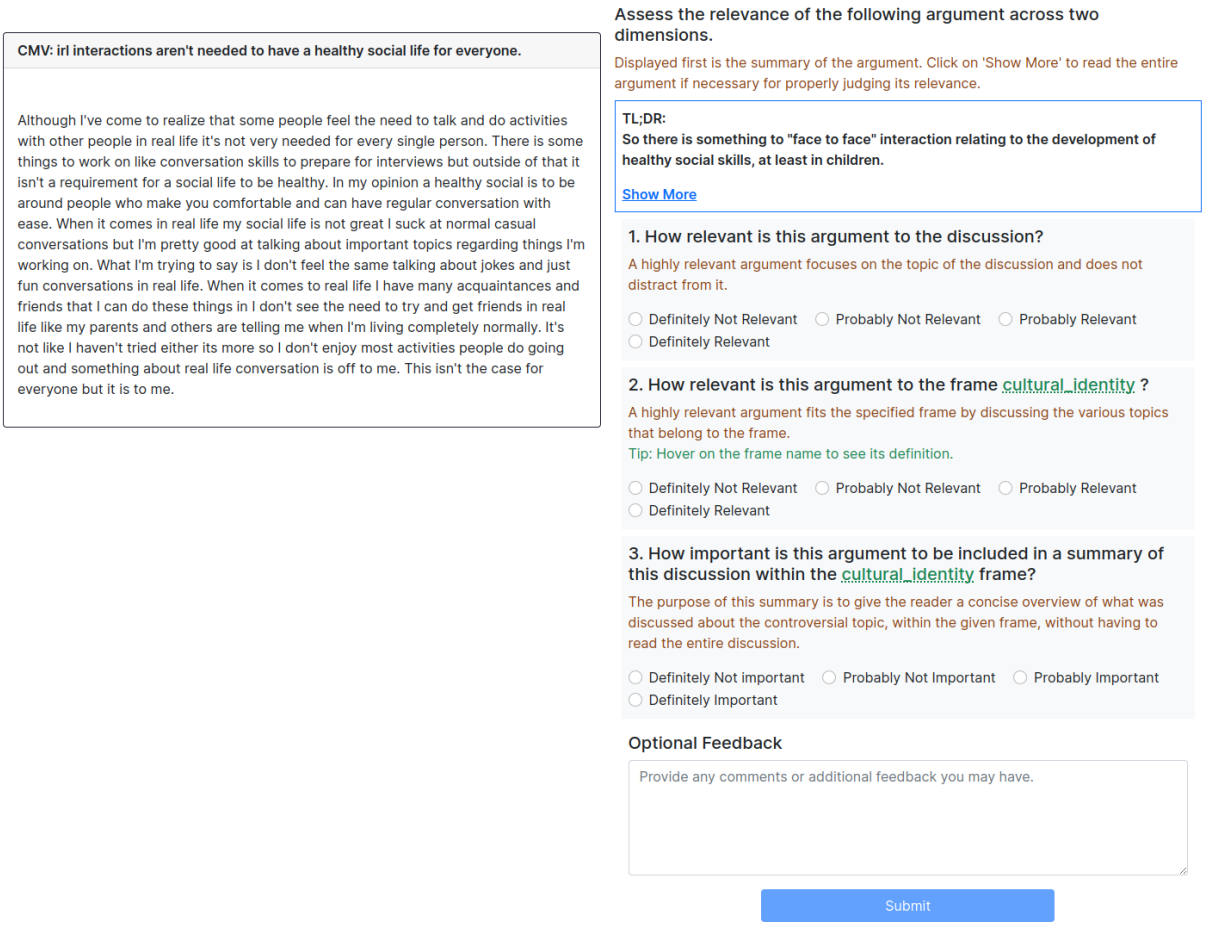


Figure 3: Annotation interface for the **pilot study**. Annotators were provided a summary of the argument alongside the entire argument. There was no option to mark a text as noisy/non-argumentative. Furthermore, the importance of an argument was assessed based on how likely it was to be included in a frame-oriented *summary* of the discussion.

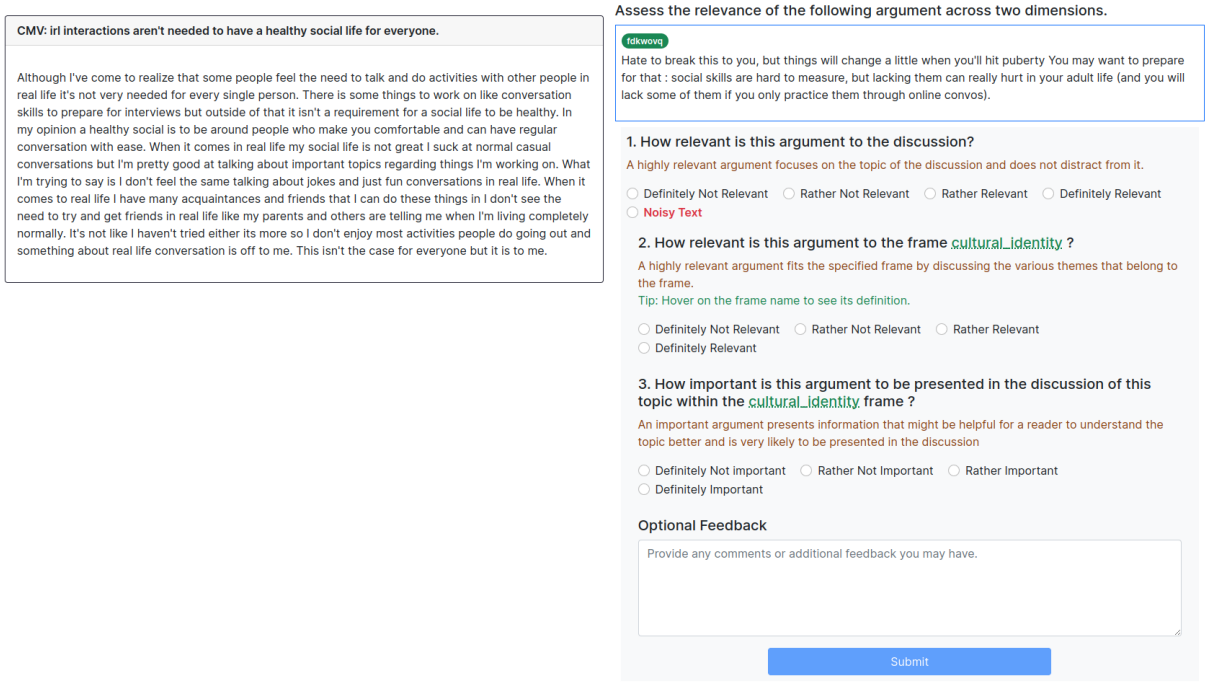


Figure 4: Annotation interface for the **main evaluation**. First, we removed the summary of the argument and always showed the complete argument. Next, we allowed marking a text as "noisy" and skip answering the remaining questions. Finally, as it was difficult to decide if an argument was important enough to be included in a summary of the discussion before reading the entire discussion, we rephrased the important question as the likelihood of including an argument in the *discussion* of the topic.

Towards Multilingual Automatic Open-Domain Dialogue Evaluation

John Mendonça[1,2,*], **Alon Lavie**[3,4] and **Isabel Trancoso**[1,2]
[1] INESC-ID, Lisbon
[2] Instituto Superior Técnico, University of Lisbon
[3] Carnegie Mellon University, Pittsburgh
[4] Phrase, Pittsburgh
{john.mendonca, isabel.trancoso}@inesc-id.pt
alavie@cs.cmu.edu

Abstract

The main limiting factor in the development of robust multilingual open-domain dialogue evaluation metrics is the lack of multilingual data and the limited availability of open-sourced multilingual dialogue systems. In this work, we propose a workaround for this lack of data by leveraging a strong multilingual pretrained encoder-based Language Model and augmenting existing English dialogue data using Machine Translation. We empirically show that the naive approach of finetuning a pretrained multilingual encoder model with translated data is insufficient to outperform the strong baseline of finetuning a multilingual model with only source data. Instead, the best approach consists in the careful curation of translated data using MT Quality Estimation metrics, excluding low quality translations that hinder its performance.

1 Introduction

Open-domain dialogue systems have gained substantial attention in the NLP (Natural Language Processing) and ML (Machine Learning) fields, thanks to their increasingly human-like behaviour (Thoppilan et al., 2022; Shuster et al., 2022). Their impressive generation capabilities can be attributed to new milestones in model development and scaling (Adiwardana et al., 2020), and the amount of data used during training. Despite this research and development effort, advertised generation capabilities were only attainable in a select few languages (typically English or Chinese) due to low resources in dialogue for other languages (Zhang et al., 2022b). More recently, however, the advent of LLMs (Large Language Models) finetuned with Reinforcement Learning from Human Feedback such as ChatGPT (Ouyang et al., 2022) has opened the path for high-quality and easily accessible multilingual dialogue generation.

Similarly, automated open-domain dialogue evaluation has also been largely limited to evaluating a select few languages. Word-overlap based metrics from NLG (Natural Language Generation) such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) are agnostic to language, only requiring a reference response. However, these metrics are known to correlate poorly with human judgments due to the multifaceted nature of dialogue (Liu et al., 2016). Reference-free metrics such as USR (Mehri and Eskenazi, 2020) and USL-H (Phy et al., 2020), however, require dialogue data for training. Considering most open-source dialogue data is in English, these models are expected to underperform significantly in other languages. Additionally, most open sourced dialogue systems are also limited to English, further disincentivising multilingual research.

One solution to the issues previously mentioned is to leverage MT (Machine Translation). With MT services becoming more affordable and consistent, some authors resort to translation when developing their multilingual dialogue systems (Schuster et al.,

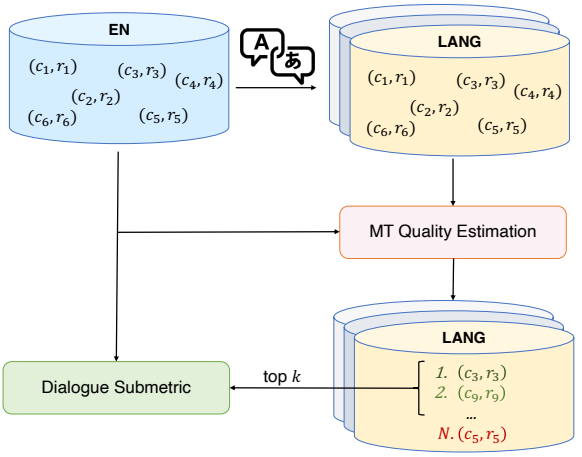


Figure 1: Proposed architecture. The original dialogue dataset is transformed into context-response pairs (c_n, r_n) and translated using MT. The final dialogue submetric is trained using a combination of the original English data and the top k sentences or (c_n, r_n) from each language, depending on the submetric.

[*] Work conducted as a visiting scholar at CMU.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 130–141
September 11–15, 2023. ©2023 Association for Computational Linguistics

2019; Anastasiou et al., 2022). This can either be included as a module in the system's pipeline – allowing the use of proven English generation models for other languages; or as a cross-lingual transfer method – by translating training data.

In this paper, we extend the approach of training using data generated by MT for the development of multilingual models for evaluation of open-domain dialogue responses. We experiment with and evaluate several different possible workarounds for this problem. Namely, we leverage the availability of strong pretrained multilingual encoders as a foundation for training multilingual dialogue evaluation models. As a first step, we translate existing publicly-available English dialogue data into the target languages. We then explore multiple alternative ways to leverage this translated data in order to finetune and train monolingual and multilingual dialogue evaluation models for two specific dialogue submetrics. To address the impact of low quality translations, we propose using an MT Quality Estimation (QE) model to rank the translations and investigate the impact of finetuning models with varying amounts of quality-ranked data. Figure 1 illustrates the proposed approach.

The performance of these alternative models is evaluated on a curated test set of dialogues which were human-annotated with dialogue quality scores for two subqualities. The original English test set was translated using MT and then post-edited by editors into six different target languages (PT-Portuguese, DE-German, FR-French, ZH-Chinese, ES-Spanish and JA-Japanese). The quality scores from the human annotations of the original English dialogues were then carried over to the target-language dialogues. Our finetuned multilingual dialogue evaluation models exhibit strong correlations with human judgements, comparable to LLMs, indicating it is possible to leverage multilingual dialogue evaluation metrics without the constraints LLMs currently possess (costs, latency, etc.). We hope this will encourage other researchers to update existing metrics using our proposed multilingual finetuning approach.

In summary, the primary contributions of this work are as follow:

- We evaluate cross-lingual transfer and translation augmented training approaches using MT for the task of training multilingual dialogue evaluation models, showing that, on average, the best performance is achieved by finetuning with subsets consisting of only the best translations. We found that, depending on the subquality and target language, the optimal amount of translated data can be as low as 5% and as high as 75%.
- We translate and release DailyDialog and a corresponding test set of human quality annotations in 6 languages to facilitate future benchmarking of multilingual dialogue evaluation metrics[1].

2 Background

2.1 Open-Domain Dialogue Evaluation Metrics

The recent trend in open-domain dialogue evaluation is to train dialogue submetrics using well-defined self-supervised tasks which correlate well with their corresponding subqualities. The most used self-supervised task is Next Sentence Prediction (NSP), as it is known to correlate well with subqualities that evaluate *"Context Awareness"*. Examples of this include: *Uses Context* (Mehri and Eskenazi, 2020), *Sensibleness* (Phy et al., 2020; Mendonca et al., 2022) and *Relevance* (Zhao et al., 2020; Zhang et al., 2022a). Other subqualities include: *Fluency, Grammatically Correct* or *Understandability*, which use word-level noising techniques to generate negative samples (Phy et al., 2020; Mendonca et al., 2022; Zhang et al., 2022a); and *Specificity*, which uses an MLM (Masked Language Modelling) score (Mehri and Eskenazi, 2020; Phy et al., 2020; Zhang et al., 2022a). For overall quality, these submetrics are typically combined using different methods (e.g. empirical observation, trained Linear Regression or multilayer perceptrons).

To the best of our knowledge, there has not been any published research on cross-lingual transfer and/or development of trained multilingual metrics for open-domain dialogue evaluation.

2.2 Multilingual Text Classification

Despite the lack of research on multilingual dialogue evaluation, extending text classification to other languages is a well established subfield of research in NLP. The main constraint for multilingual performance parity is the lack of task-specific resources in the vast majority of written languages. Given the creation of these resources is

[1] github.com/johndmendonca/DialEvalML

both time consuming and expensive, most research effort has been geared towards general-purpose cross-lingual representations that are learned in an unsupervised way, therefore leveraging the unstructured data available in the wild. Large multilingual Transformer-based models (e.g mBERT, XLM-RoBERTa, and mT5) have been successfully used in a variety of classification tasks (Conneau et al., 2020; Pires et al., 2019; Xue et al., 2021). The standard approach for cross-lingual transfer is to finetune on existing domain data in a source language and perform inference in a target language. However, this approach typically lags behind models specifically trained with in-domain (both task and language) data.

As a solution to this problem, Pfeiffer et al. (2020) propose learning language-specific adapter modules via MLM on unlabelled target-language data followed by task-specific adapter modules by optimising a target task on labelled data in the source language. Task and language adapters are stacked, allowing cross-lingual transfer to the target language by substituting the target-language adapter at inference.

Bornea et al. (2021) propose an augmentation strategy where a corpus of multilingual silver-labelled QA pairs is generated by combining the original English training data with MT-generated data. A language adversarial training and arbitration framework bring the embeddings closer to each other, making the model language invariant.

To the best of our knowledge, there has not been any research on the utilization of MT Quality Estimation (QE) scoring as a means for identifying and demoting or excluding poorly translated data in such cross-language training scenarios.

3 Problem Formulation

The goal of reference-free turn-level dialogue evaluation is, given a dialogue history (frequently denoted as context) c of varying amount of turns, and a response r, to learn a scoring function that assigns a score $f(c, r) \rightarrow s$. This scoring function is compared against human judgements, which annotate the same context-response pairs. These responses are evaluated using a scaling method, for instance, a binary $(0, 1)$ judgement or a $[1, 5]$ scale, where the lowest value means lowest quality and highest value maximum quality. The notion of quality varies wildly depending on the annotation. In this work, we evaluate dialogue in two dimensions:

- **Understandability** An understandable response is one that can be understood without context. Such responses may contain minor typos that do not hinder the comprehension of the response.

- **Sensibleness** A sensible response is one that takes into account its preceding context.

Most automatic evaluation metrics reformulate the problem as regression. Performance is then evaluated using Pearson and Spearman correlations with human annotations.

3.1 Automatic Dialogue Evaluation Metrics

The majority of competitive metrics for dialogue evaluation include models trained in a self-supervised way for Valid Sentence Prediction (VSP) and Next Sentence Prediction (NSP) (Yeh et al., 2021; Zhang et al., 2021). As such, the focus of this work was to evaluate multilingual dynamics for these models, which can then be employed on existing metrics.

VSP: Valid Sentence Prediction In this paper, we followed the approach used by Phy et al. (2020) and initially proposed by Sinha et al. (2020). A regression model was trained to differentiate between positive samples and synthetic negative samples. **Positive** samples are perturbed by randomly applying one of the following: (1) no perturbation, (2) punctuation removal, (3) stop-word removal. **Negative** samples are generated by randomly applying one of the following rules: (1) word reorder (shuffling the ordering of the words); (2) word-drop; and (3) word-repeat (randomly repeating words).

NSP: Next Sentence Prediction The task of predicting sensibleness can be considered a binary (NSP) task, distinguishing a positive example from a semantically negative one, given a context. A discriminative regression model was trained using the following sampling strategy: **positive** responses are drawn directly from the dialog; **negative** responses are randomly selected and a token coverage test discards semantically similar sentences. All responses are processed using the positive-sample heuristic used by VSP.

4 Cross-lingual Transfer Learning

The goal of the experiments described in this section was to evaluate different basic approaches of

cross-lingual transfer for the task of automatic dialogue evaluation. For encoder model training, we leveraged Machine Translation (MT) by fully translating an English source dialogue dataset and then finetuning monolingual and multilingual models using these translations.

4.1 Experimental Setup

4.1.1 Dataset

All experiments in this paper were based on the **DailyDialog (Li et al., 2017)** dataset, a high-quality human-human open-domain dialogue dataset focused on day-to-day conversations. After processing, we obtained train/dev splits of 58,515/25,078 and 89,707/38,449 per language for the VSP and NSP models, respectively. For training and evaluation, the post-processed dataset was translated into the target languages using MBART50 (Liu et al., 2020). We opted for using MBART50 as it is a relatively lightweight open sourced model with a large language coverage.

For the test set, we leveraged the annotations from Phy et al. (2020). These human annotations evaluate five responses from two retrieval methods, two generative methods, and one human-generated response for 50 contexts. These responses were annotated in terms of *Understandability* and *Sensibleness* [2]. We translated this set using Unbabel's[3] translation service. A total of 300 sentences were translated, corresponding to the 50 shared contexts and 250 responses. The translations were then split into smaller tasks and were corrected by editors from a commercial provider. Editors were specifically asked to retain any source disfluencies or hallucinations stemming from low quality response generation (e.g. *"I'm afraid you can't. I'm afraid you can't."*; *"Au contraire, you need to be a bahn."*). This ensured the original human quality annotations remained valid for the translation. A secondary senior editor reviewed the edited content as a whole.

4.1.2 Finetuned Encoders

We used XLM-RoBERTa (Conneau et al., 2020) as the encoder model for the experiments. This model is the multilingual version of RoBERTa, pretrained on CommonCrawl data containing 100 languages. For both the VSP and NSP models, we added a regression head on top of the encoder model.

EN – Zero-shot inference As a baseline for our results, we conducted zero-shot inference on the target languages using a model finetuned only on the original English data.

LANG – Target-Language Finetuning We finetuned the encoder with target-language translated dialogue data only. The downside of this approach is that a unique model needs to be trained for each target language. However, this method can be scaled to every language, including new ones, and is optimised to perform best in that language.

ML – Multilingual Finetuning Instead of finetuning a new model for each target language, one can finetune a single multilingual model by combining all of the translated data. In this case, the resulting single trained model is then used to evaluate responses in all languages. However, its performance may suffer in languages it has not seen during finetuning, even if they are supported by the encoder model. Furthermore, unlike target-language finetuned, the multilingual model is optimised jointly for all included languages.

MAD-X In this approach, we trained a VSP and NSP task adapter using the original English data by stacking the task adapter with a pretrained English language adapter (kept frozen during training). For zero-shot inference, the English language adapter was replaced by the target-language counterpart, while keeping the trained task adapter in place.

4.1.3 Large Language Model

As an additional strong baseline, we leveraged `gpt-3.5-turbo` (colloquially known as ChatGPT) as an evaluator of Understandability and Sensibleness. The context (exclusively for Sensibleness) and response was provided as input, together with the prompt *"{Given the context,} evaluate from 1-5 the response in terms of {dimension}. Provide the score and nothing else."*. This prompt, paired with a temperature setting of 0.0 attempted to minimises the variability of the output. Nevertheless, we report a standard deviation of (.003, .003) and (.001, .001) for Understandability and Sensibleness correlations, respectively, across 3 runs.

4.2 Results

The correlation results for all subqualities and the overall quality are presented in Table 1.

[2] Annotations for *Specificity* and *Overall Quality* were also conducted, but were excluded since they do not map to the learned metrics under study.

[3] unbabel.com

	EN		PT		DE		FR		ZH		ES		JA		AVG	
	Pr.	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**
Understandability																
EN	.376	.187	**.366**	.167	.328	.172	.351	.120	**.318**	**.202**	.342	.204	**.204**	.176	.327	.194
LANG	-	-	.176	.164	.214	.138	*.052*	.034	.274	.156	.219	.144	.185	.132	.214	.146
ML	.336	.117	.176	.167	.262	.150	*.012*	.015	.225	.138	.117	.158	*.091*	*.092*	.174	.126
MAD-X	.363	.166	.189	.103	.237	.122	.168	*.078*	.305	.168	.217	.119	.119	.129	.228	.126
ChatGPT	**.397**	**.334**	.365	**.230**	**.332**	**.263**	**.369**	**.273**	.276	.182	**.394**	**.263**	**.228**	**.223**	**.337**	**.263**
Sensibleness																
EN	.658	.676	.636	.651	.657	.655	.646	.656	.640	.656	.646	.657	.590	.599	.639	.649
LANG	-	-	**.649**	.661	.669	**.699**	.635	.655	.634	**.671**	.629	.669	.617	**.640**	.642	**.664**
ML	.651	.691	.606	**.675**	.634	.680	.605	**.669**	.642	.667	.596	.676	.599	.637	.619	**.664**
MAD-X	.660	.681	.614	.604	.664	.652	.624	.624	.608	.647	**.688**	.661	.558	.595	.631	.638
ChatGPT	**.746**	**.724**	.636	.626	**.683**	.675	**.695**	.666	**.655**	.645	.680	**.677**	**.625**	.610	**.674**	.662

Table 1: Average correlation results across 3 runs with different seeds. **Pr.** denotes Pearson and **Sp.** denotes Spearman. **Bold** denotes best performance, *Italic* $p < 0.05$.

Understandability The results show that, on average, the best performing encoder approach is the zero-shot inference using the English model (**EN**). Both the target-language finetuning (**LANG**) and multilingual finetuning approaches (**ML**) have much lower performances, indicating that translation augmentation is detrimental for this task. We also note that the **MAD-X** approach, although performing slightly better than ML and LANG, still lags behind EN considerably. In any case, ChatGPT largely outperforms other models on both metrics.

Sensibleness The best performing encoder approach for this subquality is LANG. Intuitively this makes sense, given that during finetuning the model is exposed to target-language data for the language it is being evaluated on. Furthermore, the performance difference between the different approaches is relatively much smaller, which indicates the Sensibleness subquality is less sensitive to MT quality. When comparing these results with ChatGPT, we observe a much smaller performance gap, with the best encoder models slightly outperforming on Spearman.

5 MT Quality-aware finetuning

The effects of noise introduced to the training data is a subject of intense research in the literature (Zhang et al., 2017; Hu et al., 2020; Swayamdipta et al., 2020). It is expected that, for this task, noise is introduced by low quality translations, reducing the performance of trained models. This issue was identified in Section 4, where for the VSP model in particular, the models trained using translations performed much worse than the baseline approach. Our hypothesis is that some translations heavily disrupt morphosyntactic cues used to infer response fluency, as shown in Table 2. We acknowledge that these low quality translations may also reduce

EN: Yes, I'd like to see the receipt.
Oh ! I see you bought the watch last week.
PT: Sim, gostava de ver o receio.
Oh! Vejo-te a fazer o relógio na semana passada.
QE score: -0.670

EN: Just look around ? Ah, that's boring.
ES: ¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡¡
QE score: -1.481

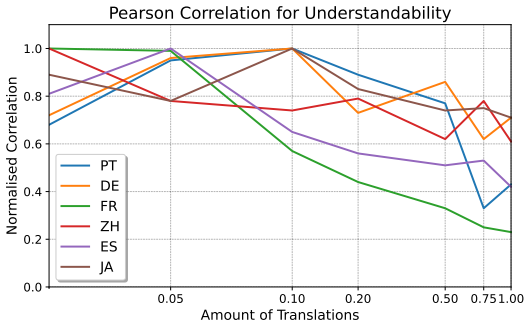
EN: Eight tens, six ones and large silver for others.
ZH: 八个十个,六个十个,其他十个十个十个十个十...
QE Score: -1.312

Table 2: Examples of low quality translations with corresponding QE score. Red denotes MT error, with underline in the source sentence indicating the closest alignment of the error. Blue denotes keywords that refer to prior context.

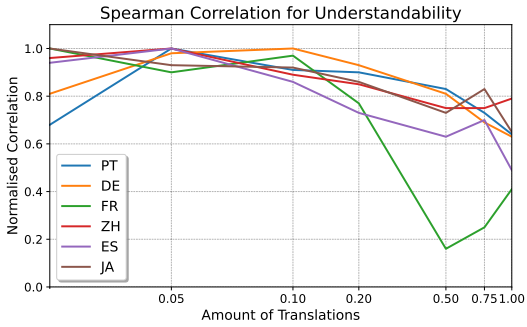
the quality of the response by disrupting keywords that point to the context (which is important for Sensibleness), or even more subtle quality cues (e.g. loss of empathy, inconsistency with named entities). However, the NSP model is trained to discriminate between the original response and randomly selected response from the corpus. As such, the model's prediction will remain invariant to most translation errors.

These observations, paired with the fact encoder models only slightly underperform ChatGPT (a much larger and expensive model), motivate the work described in this section. We hypothesise that, by ameliorating the MT noise via identifying and filtering low quality translations, the encoder model performance can outperform LLMs such as ChatGPT, at a fraction of the cost.

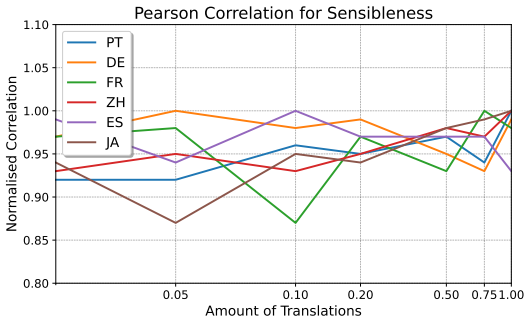
Since there are no available references, an MT QE (Specia et al., 2018) automatic metric is used for this purpose. Formally, an MT QE model is a scoring function that assigns a score given a source sentence and hypothesis translation. The unboundedness and uncalibrated nature of this score across languages results in the need for a cumbersome



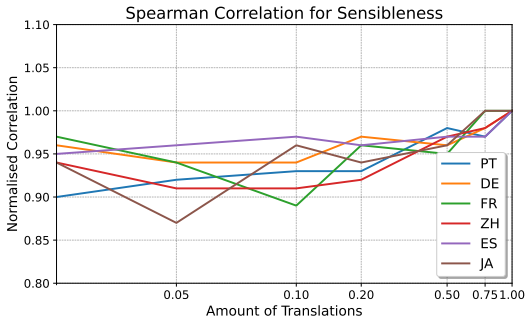
(a) Pearson Correlation, Understandability.



(b) Spearman Correlation, Understandability.



(c) Pearson Correlation, Sensibleness.



(d) Spearman Correlation, Sensibleness.

Figure 2: Normalised Pearson and Spearman correlation for the Understandability and Sensibleness submetric with varying amount of translated training data. Numeric results available in Appendix B.

analysis for each individual language in order to determine a threshold for filtering. Instead, we propose to use QE scores for response ranking, for each target language. This ensures a standardised method for filtering, improving the scalability of this method to new languages.

5.1 Experimental setup

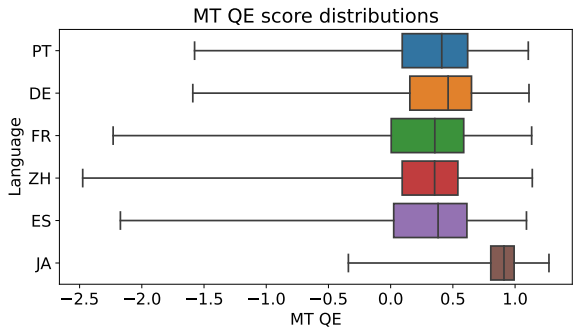


Figure 3: MT QE unnormalised score boxplot per language.

In order to confirm our hypothesis, we retrained all models using different amounts of translated data (100, 75, 50, 20, 10 and 5%). The ranking of the translations was conducted by scoring them using the WMT20 COMET-QE-DA model (Rei et al., 2020). For the VSP model, we ranked the individual sentences, and then applied negative sampling. For the NSP model, we ranked the positive and negative samples separately and then merged them together. Figure 3 presents the unnormalised score boxplot per language for all sentences (context and responses) for DailyDialog.

One of the things we noticed when finetuning the monolingual models was that the VSP models had large variations in performance. This can be attributed to (1) the low amount of training data, especially when using very few examples (5%, 10%), and (2) low quality translations, which is the research question this experiment attempts to answer. Since the true impact of low quality translations is obfuscated by other factors, we decided to finetune the LANG models starting from the EN checkpoint instead of the pretrained XLM-RoBERTa, and include the zero-shot results as 0%.

5.2 Results

LANG For the monolingual models, we plot normalised correlation results with the amount of MT data used during finetuning in Figure 2. The ***Understandability*** correlation results show that the optimal amount of translated data is language dependent, but with a clear indication that the inclu-

	EN		PT		DE		FR		ZH		ES		JA		AVG	
	Pr.	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**
Understandability																
0 (EN)	.376	.187	.366	.167	.328	.172	.351	.120	.318	.202	.342	.204	.204	.176	.327	.194
5	.403	.182	.490	.219	.344	.172	**.385**	*.091*	.320	**.235**	**.429**	.236	**.230**	.179	**.372**	.211
10	.377	.180	**.514**	.227	**.381**	.193	.294	*.091*	**.338**	.214	.385	.212	.216	.175	.358	.206
20	.384	.177	.478	.236	.333	.203	.153	*.087*	.318	.219	.315	.214	.174	.168	.308	.202
50	**.413**	.201	.481	**.242**	**.381**	.213	*.103*	*.053*	.310	.200	.315	.221	.219	.149	.317	.200
75	.311	.145	.247	.211	.320	.195	*.047*	*.048*	.163	.149	.111	.198	*.108*	.127	.187	.158
100	.336	.117	.176	.167	.262	.150	*.012*	*.015*	.225	.138	.117	.158	*.091*	*.092*	.174	.126
ChatGPT	.397	**.334**	.365	.230	.332	**.263**	.369	**.273**	.276	.182	.394	**.263**	.228	**.223**	.337	**.263**
Sensibleness																
0 (EN)	.658	.676	.636	.651	.657	.655	.646	.656	.640	.656	**.646**	.657	.590	.599	.639	.649
5	.637	.674	.629	.632	.627	.648	.637	.656	.629	.646	.626	.647	.567	.596	.621	.640
10	.642	.675	.639	.664	.661	.669	.636	.661	.637	.656	.635	.668	.575	.604	.632	.654
20	.650	.689	.627	.670	.649	.681	.627	.666	.621	.661	.637	.673	.568	.614	.626	.660
50	.667	.691	**.642**	.687	.650	.672	.621	.662	.652	.664	.629	.673	.600	**.642**	.637	.666
75	.677	.712	.629	**.694**	.679	**.702**	.633	**.679**	**.661**	**.673**	.643	**.695**	.593	.635	.645	**.679**
100	.651	.691	.606	.675	.634	.680	.605	.669	.642	.667	.596	.676	.599	.637	.619	.664
ChatGPT	**.746**	**.724**	.636	.626	**.683**	.675	**.695**	.666	.655	.645	.680	.677	**.625**	.610	**.674**	.662

Table 3: Average correlation results across 3 runs with different seeds for multilingual models when varying the amount of translated data.

sion of more translations decreases performance significantly. Instead, a lower amount of translations (5-10%) yields optimal performance. This shows that this small finetuning step is essentially adapting a model that was already finetuned for the downstream task to the target-language domain. For ***Sensibleness***, we see that the inclusion of more translations yields the best results. As such, we can conclude that low-quality MT does not adversely affect performance. We hypothesise this is due to MT being able to correctly translate keywords that indicate context awareness. Since we are only concerned about relevance, the overall sentence may still contain MT errors and be scored highly.

ML The correlation results for the multilingual models are presented in Table 3. For ***Understandability***, we note that, on average, and similar to LANG, the best performance is attained with the minimum amount of translated data (ML-5), with the performance decreasing when more translations are added. Comparing these results with ChatGPT, we observe an improvement in performance, but our encoder models are still generally weaker when using Spearman as a metric. For ***Sensibleness***, decreasing the amount of data reduces the performance of the model. However, we note a decrease in performance when including the full amount of translated data (ML-100). This may be due to the inclusion of the worst translations – typically hallucinations – which is compounded by training on all languages. Unlike in Understandability, here we see that ChatGPT still outperforms the best encoder model in terms of Pearson correlation.

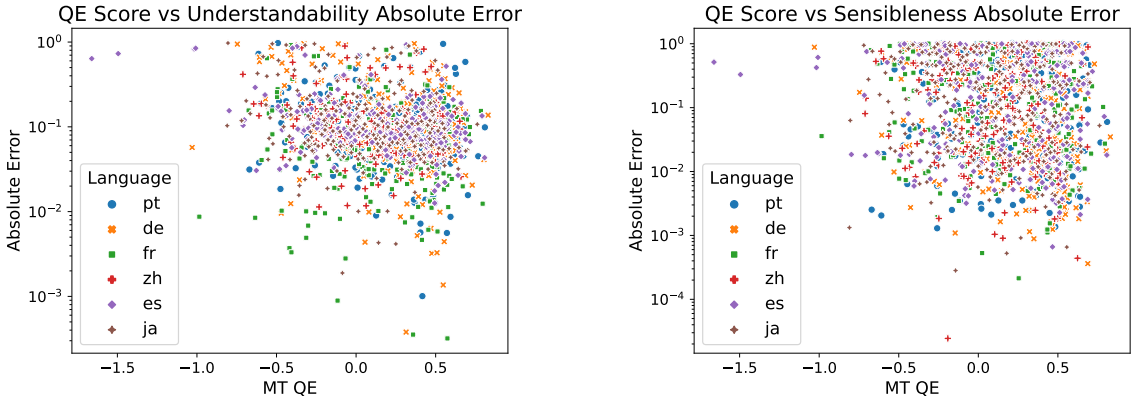
5.3 Effect of low-quality translation during prediction

One might ask if a low-quality translation can induce the submetrics to output a different score. Intuitively, we hypothesise each model will attribute different scores in the face of low quality translations. More specifically, given the results presented in previous sections, we expect the test prediction error to be:

- **Negatively correlated with the MT QE scores for VSP.** We know this model is highly sensitive to low quality translations, since MT errors frequently affect the fluency of the response (as identified in previous sections);
- **Weakly correlated for the NSP model.** The model showed robustness when including more translations during training, with performance decreasing only when we included all translations (ML-100) during training.

In order to evaluate these assumptions, the correlation plots of the MT QE z-scores (obtained independently for each language) against the submetric absolute error using the best ML models (ML-5 for VSP and ML-75 for NSP) for the test set are presented in Figure 4.

For the ***Understandability*** subquality, we note that there is a slight negative correlation between the absolute error and the MT QE score. This is also confirmed by a calculated Pearson Correlation value of -0.245. For the ***Sensibleness*** subquality, the relationship between these two measures is less obvious. For instance, we note that, unlike for Understandability, maximum deviations



(a) Understandability scatter plot.

(b) Sensibleness scatter plot.

Figure 4: Scatter plot comparing the test set MBART50 per-language QE z-scores (x-axis) versus the per sample Absolute Prediction Error (y-axis in log scale) for Understandability and Sensibleness subqualities.

are spread evenly across the QE scale, which points to the model erroneously predicting Sensibleness irrespective of the translation quality. Conversely, we also note a higher density of accurate predictions with lower QE scores. These results, paired with the calculated Pearson Correlation value of -0.129, confirm our hypothesis that the NSP model is more agnostic of MT quality than VSP.

CTX: Também me apercebi desta questão. E a automatização dos processos do escritório é essencial. **RES:** Sim, fazer tudo manualmente demora demasiado.	
EN-VSP: .394	**EN-NSP:** .824
ML-VSP: 1.00	**ML-NSP:** 1.00
Unders.: 1.00	**Sensibl:** 0.00
CTX: Ja, ich leite die Jungs am Kai. **RES:** Wow, das klingt nach einem fantastischen Job, de du da bekommen hast.	
EN-VSP: .963	**EN-NSP:** .315
ML-VSP: .941	**ML-NSP:** .981
Unders.: 1.00	**Sensibl:** 1.00

Table 4: Examples of subquality predictions from the test set.

5.4 Example test predictions

We present representative examples of our best ML models' prediction (ML 5/75) in Table 4. In the first example, the baseline English model fails to appropriately identify the understandability of the response. In the second example, we see that the multilingual model is able to correctly identify that the response takes into account the job presented in the context (manager) by complimenting it ("fantastic job"), which the EN model failed to identify.

6 Conclusions

This paper explored the use of cross-lingual knowledge transfer for the novel task of automatic multilingual dialogue evaluation. We evaluated different strategies for this task, including zero-shot inference, MAD-X and Machine Translation augmentation. Empirically we showed that the naive approach of leveraging MT for augmentation is insufficient to outperform the baseline of English finetuning with a multilingual encoder-based LM, let alone a strong LLM. Instead, by filtering out low quality translations, we were able to reduce the gap of performance on ChatGPT, outperforming it on select correlation metrics. Experimental results showed that we obtain the best performance when training encoder models with the following proportions of MT-QE: 5% for Understandability and 75% for Sensibleness.

One could argue the notion of quality is intrinsically related to cultural norms. For instance, Japanese speakers may prefer a polite conversation, whereas German speakers might prefer a more direct interaction. A future research direction is to evaluate generative model responses in different languages using annotators exposed to the culture associated with a given language. In addition to ensuring the evaluation of the response meets the criteria of "quality" in different cultures, it would also allow for a qualitative analysis of the differences in the notion of quality between languages.

Limitations

Perhaps the main limitation of this work is the restricted amount of languages studied. Ideally, we would have used a more comprehensible set of languages, including low-resource ones, to evaluate the consistency of the conclusions drawn from the experiments.

Another limitation is the focus on a single open-domain dialogue dataset. Dialogue evaluation metrics are known to correlate poorly when evaluated on unseen datasets (Yeh et al., 2021). As such, it is not certain that the observations presented in this work would hold for other datasets, or even different annotations (Mehri et al., 2022).

Finally, the pretrained encoder, MT and QE models used in this work are not fully representative of all available models. We acknowledge that the optimal amount of filtering is likely to be different, depending on the combination of models used.

Ethics Statement

This work leverages dialogues and annotations developed exclusively by English-speakers. This introduces an English-centric bias with respect to the notion of quality (and subqualities) in dialogues. Although not evaluated in depth in this work, there could be a chance that the models erroneously yield lower scores to responses not conforming to English notions of quality responses.

The original dialogue dataset and generated responses were checked for personally identifiable information or offensive content by the original authors. Although highly unlikely, we acknowledge the translations may contain offensive content resulting from decoding.

The post-editing conducted in this work used a crowdsourcing platform that awarded users a fair wage according to their location.

Acknowledgements

This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Responsible.AI), and by national funds through *Fundação para a Ciência e a Tecnologia* (FCT) with references PRT/BD/152198/2021 and UIDB/50021/2020, and by the P2020 program MAIA (LISBOA-01-0247-FEDER-045909).

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Dimitra Anastasiou, Anders Ruge, Radu Ion, Svetlana Segărceanu, George Suciu, Olivier Pedretti, Patrick Gratz, and Hoorieh Afkari. 2022. A machine translation-powered chatbot for public administration. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 329–330, Ghent, Belgium. European Association for Machine Translation.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. 2021. Multilingual transfer learning for qa using translation as data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12583–12591.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Wei Hu, Zhiyuan Li, and Dingli Yu. 2020. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An

empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

John Mendonca, Alon Lavie, and Isabel Trancoso. 2022. QualityAdapt: an automatic dialogue quality estimation framework. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 83–90, Edinburgh, UK. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, W.K.F. Ngan, Spencer Poff, Naman Goyal, Arthur D. Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *ArXiv*, abs/2208.03188.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and

Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Chen Zhang, João Sedoc, L. F. D'Haro, Rafael E. Banchs, and Alexander I. Rudnicky. 2021. Automatic evaluation and moderation of open-domain dialogue systems. *ArXiv*, abs/2111.02110.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Pengfei Zhang, Xiaohui Hu, Kaidong Yu, Jian Wang, Song Han, Cao Liu, and Chunyang Yuan. 2022a. MME-CRS: Multi-Metric Evaluation Based on Correlation Re-Scaling for Evaluating Open-Domain Dialogue. *arXiv preprint arXiv:2206.09403*.

Qingyu Zhang, Xiaoyu Shen, Ernie Chang, Jidong Ge, and Pengke Chen. 2022b. Mdia: A benchmark for multilingual dialogue generation in 46 languages.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Online. Association for Computational Linguistics.

A Training setup and Hyperparamters

We used the XLM-R Large encoder model downloaded from HuggingFace [4] for all experiments. A token representing the speaker was added for each turn, and a history length of 3 turns was used. We applied a regression head consisting of a 2-layer MLP with a hidden size of 1024 and a hyperbolic tangent function as activation for prediction. All parameters were trained/finetuned using Adam optimizer (Kingma and Ba, 2015).

The task adapters were trained using the recipe from Mendonca et al. (2022), using a learning rate of 1e-4 and training for 10 epochs, with a batch size of 32. We used the existing language adapters from AdapterHub whenever possible (EN, ZH, JA) and trained the remaining using the AdapterHub's MLM recipe [5] on Wikipedia data[6]. The fully finetuned models used a learning rate of 3e-6 and were trained for 3 epochs using a batch size of 16. Evaluation was conducted every 1,000 steps for the smaller training sets and 10,000 steps for the larger ones (75% and 100 %). The best performing model on the evaluation set was selected for testing.

For the dialogue data preprocessing we used spaCy [7] and the corresponding core language models. For the translations we used `facebook/mbart-large-50-one-to-many-mmt` from HuggingFace. Batch size was set to 16 and decoding was conducted using beam search, with the number of beams set to 4.

We used a single Quadro RTX 6000 24GB GPU for all experiments.

B Additional Results

Table 5 presents the monolingual model results for the experiments of Section 5. Due to time and computational constraints, we only conduct these experiments using a single seed.

[4] huggingface.co/xlm-roberta-large

[5] github.com/adapter-hub

[6] dumps.wikimedia.org

[7] spacy.io

	EN		PT		DE		FR		ZH		ES		JA		AVG	
	Pr.	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**	**Pr.**	**Sp.**
Understandability																
0	.347	.192	.381	.176	.353	.184	**.349**	**.106**	**.406**	.251	.372	.210	.268	**.223**	.354	.212
5			**.534**	**.259**	.469	.223	.347	*.095*	.318	**.263**	**.459**	**.223**	.236	.208	**.387**	**.231**
10			.563	.236	**.489**	**.227**	.199	*.102*	.300	.233	.300	.191	**.303**	.206	.357	.218
20			.499	.233	.356	.211	.153	*.082*	.323	.223	.257	.163	.251	.191	.312	.201
50			.433	.214	.418	.185	*.117*	*.017*	.250	.198	.233	.140	.225	.163	.289	.175
75			.186	.189	.306	.158	*.089*	*.026*	.319	.198	.243	.156	*.226*	.185	.245	.169
100			.240	.165	.347	.144	*.082*	*.043*	.248	.206	.191	.109	.216	.146	.239	.155
Sensibleness																
0	.621	.654	.618	.627	.667	.668	.621	.644	.605	.647	.628	.628	.577	.592	.620	.635
5			.615	.636	.687	.657	.632	.628	.618	.629	.599	.631	.538	.553	.616	.626
10			.647	.646	.672	.655	.562	.596	.607	.626	**.635**	.637	.587	.606	.619	.630
20			.639	.644	.680	.679	.627	.640	.620	.633	.615	.634	.582	.595	.626	.638
50			.651	.680	.654	.671	.601	.631	.637	.665	.613	.639	.603	.609	.626	.647
75			.634	.670	.640	.681	**.643**	.664	.629	.673	.615	.639	.608	.635	.627	.656
100			**.671**	**.693**	**.681**	**.698**	.631	**.666**	**.650**	**.688**	.589	**.659**	**.617**	**.633**	**.637**	**.666**

Table 5: Average correlation results for the monolingual models when varying the amount of translated data.

Dialog Action-Aware Transformer for Dialog Policy Learning

Huimin Wang[1*], **Wai-Chung Kwan**[2,3*], **Kam-Fai Wong**[2,3]
[1]Jarvis Lab, Tencent, Shenzhen, China
[2]The Chinese University of Hong Kong, Hong Kong, China
[3]MoE Key Laboratory of High Confidence Software Technologies, China
{hmmmwang}@tencent.com
{wckwan,kfwong}@se.cuhk.edu.hk

Abstract

Recent works usually address Dialog policy learning DPL by training a reinforcement learning (RL) agent to determine the best dialog action. However, existing works on deep RL require a large volume of agent-user interactions to achieve acceptable performance. In this paper, we propose to make full use of the plain text knowledge from the pre-trained language model to accelerate the RL agent's learning speed. Specifically, we design a dialog action-aware transformer encoder (DaTrans), which integrates a new fine-tuning procedure named masked last action task to encourage DaTrans to be dialog-aware and distils action-specific features. Then, DaTrans is further optimized in an RL setting with ongoing interactions and evolves through exploration in the dialog action space toward maximizing long-term accumulated rewards. The effectiveness and efficiency of the proposed model are demonstrated with both simulator evaluation and human evaluation.

1 Introduction

A task-oriented dialog system that can serve users on certain tasks has increasingly attracted research efforts. Dialog policy learning (DPL) aiming to determine the next abstracted system output plays a key role in pipeline task-oriented dialog systems (Kwan et al., 2023). Recently, it has shown great potential for using reinforcement learning (RL) based methods to formulate DPL (Young et al., 2013; Su et al., 2016; Peng et al., 2017). A lot of progress is being made in demonstration-based efficient learning methods (Brys et al., 2015; Cederborg et al., 2015; Wang et al., 2020; Li et al., 2020; Jhunjhunwala et al., 2020; Geishauser et al., 2022). Among these methods, dialog state tracking (DST), comprising all information required to determine the response, is an indispensable module. However, DST inevitably accumulates errors from each module of the system.

Recent pre-trained language models (PLMs) gathering knowledge from the massive plain text show great potential for formulating DPL without DST. Recently, the studies on PLMs for dialog, including BERT-based dialog state tracking (Gulyaev et al., 2020) and GPT-2 based dialog generation (Peng et al., 2020; Yang et al., 2021) are not centred on DPL. To this end, we proposed the **D**ialog **A**ction-oriented transformer encoder termed as **DaTrans**, for efficient dialog policy training. DaTrans is achieved by a dialog act-aware fine-tuning task, which encourages the model to distil the dialog policy logic. Specifically, rather than commonly used tasks, like predicting randomly masked words in the input (MLM task) and classifying whether the sentences are continuous or not (NSP task) (Devlin et al., 2019), DaTrans is fine-tuned by predicting the masked last acts in the input action sequences (termed as MLA task). After that, DaTrans works as an RL agent which evolves toward maximizing long-term accumulated rewards through interacting with a user simulator. Following the traditional RL-based dialog policy learning framework, the main novelty of DaTrans is that it integrates a proposed dialog action-aware fine-tuning task (MLA), which helps to extract action-specific features from historical dialog action sequences to improve dialog policy learning. The empirical results prove the excellent performance of DaTrans. Our main contributions include 1) We propose the DaTrans that integrates the dialog act-aware fine-tuning task to extract the dialog policy logic from the plain text; 2) We validate the efficiency and effectiveness of the proposed model on a multi-domain benchmark with both simulator and human evaluation.

[*] Equal Contribution

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 142–148
September 11–15, 2023. ©2023 Association for Computational Linguistics

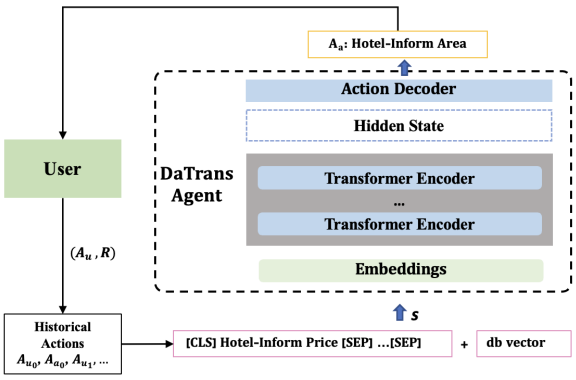


Figure 1: The Illustration of **D**ialog **A**ction-oriented Transformer Encoder (**DaTrans**). In this example, **DaTrans** generates the dialog action A_a based on historical actions.

2 Approach

We cast the dialog policy learning problem as a Markov Decision Process and optimize the policy with deep reinforcement learning approaches. RL usually involves an interactive process (as shown in Figure 1), during which the dialog agent's behavior should choose actions that tend to increase the long-turn sum of rewards given by the user. It can learn to do this over time, by systematic trials and errors until reaches the optimal. In our setting, the dialog agent is encoded with the proposed DaTrans, which perceives the state s and determines the next action A_a. We consider a transformer decoder-based policy model, which takes text concatenating of tuples containing a domain name, an intent type, and slot names as input and determines the next action.

2.1 DaTrans

We apply Deep Q-learning to optimize dialog policy. $Q_\theta(s, a)$, approximating the state-action value function parameterized θ, is implemented based on DaTrans as illustrated in Figure 1. In each turn, perceiving the state s that consists of historical action sequences and a database vector denoting the matches of the current constraints, DaTrans determines the dialog action a with the generated value function $Q_\theta(\cdot|s)$. Historical action sequences are tokenized started from $[CLS]$, followed by the tokenized actions separated and ended with $[SEP]$. Then the transformer encoder gets the final hidden states denoted $[t_0..t_n] = encoder([e_0..e_n])$ (n is the current sequence length, e_i is the embedding of the input token). The contextualized sentence-level representation t_0, is passed to a linear layer named action decoder $\boldsymbol{T}$ to generate:

$$Q_\theta(s, a) = \boldsymbol{T}_a(encoder(Embed(s))) \quad (1)$$

where $Embed$ is the embedding modules of transformer encoder, $\boldsymbol{T}_a$ denoted the a_{th} output unit of $\boldsymbol{T}$. Based on DaTrans, the dialog policy is trained with ϵ-greedy exploration that selects a random action with probability ϵ, or adopts a greedy policy $a = argmax_{a'} Q_\theta(s, a')$. In each iteration, $Q_\theta(s, a)$ is updated by minimizing the following square loss with stochastic gradient descent:

$$\begin{aligned} \mathcal{L}_\theta &= \mathbb{E}_{(s,a,r,s')\sim D}[(y_i - Q_\theta(s, a))^2] \\ y_i &= r + \gamma \max_{a'} Q'_\theta(s', a') \end{aligned} \quad (2)$$

where $\gamma \in [0, 1]$ is a discount factor, D is the experience replay buffer with collected transition tuples (s, a, r, s'), s is the current state, r refers to the reward, and $Q'(\cdot)$ is the target value function, which is only periodically updated, and s' is the next state. By differentiating the loss function with regard to θ, we derive the following gradient:

$$\begin{aligned} \nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')\sim D}[(r + \\ \gamma max_{a'} Q'_{\theta'}(s', a') - Q_\theta(s, a))\nabla_\theta Q_\theta(s, a)] \end{aligned} \quad (3)$$

In each iteration, we update $Q(.)$ using mini-batch Deep Q-learning.

2.2 Dialog Action-aware Fine-tuning

A vanilla transformer decoder without pre-training can encumber the learning of dialog policy since it is totally unaware of the text and dialog logic. Meanwhile, well-pre-trained models like BERT, due to the generality of pre-training tasks and corpus, are still difficult with competent in dialog modeling. The NSP task encourages BERT to model the relationship between sentences, which may benefit natural language inference, however, biased dialog policy learning due to the inconsistency between success and continuity of sentences, e.g. discontinuous sentences can form a successful dialog. Also, the MLM task allows the word representation to fuse the left and right context, while the dialog agent is only allowed to access the

left one. Considering that the ability to reason the next dialog action plays a key role in dialog policy, we replace the MLM and NSP task with a novel fine-tuning task: predicting masked last dialog action (MLA). MLA is based on a dialog action-aware fine-tuning corpus, each piece of which is a dialog session composed of the annotated historical action sequences, for example, "*[CLS] Police-Inform Name [SEP] Police-Inform Phone Addr Post [SEP] general-thank none [SEP]*", (denoted as **sentence A**). Then we randomly cut between two consecutive actions of a session, and select the first half with the masked last act as input. For example, we cut **sentence A** between the 2_{nd} and the 3_{rd} action, and mask the last act to get the input: "*[CLS] Police-Inform Name [SEP] [MASK]..[MASK]*". The label for the masked tokens is "*Police - Inform Phone Addr Post*". Significantly, the proposed MLA task for BERT is actually different from auto-regression. The way auto-regression works is after each token is produced, that token is added to the sequence of inputs and this new sequence becomes the input to the model in its next step. However, in DaTrans, the MLA task works as predicting the last dialog action word by word without adding a new predicted word.

The goal of MLA is to minimize the cross-entropy loss with input tokens $w_0, w_1, .., w_n$:

$$\mathcal{L}^{mla} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=n-k+1}^{n}\log \boldsymbol{p}(w_j^i | w_{0:j-1,j+1:n}^i) \tag{4}$$

where $w_{0:j-1,j+1:n}^i = w_0^i \cdots w_{j-1}^i, w_{j+1}^i..w_n^i$, $\boldsymbol{p}$ is the action decoder head for predicting masked tokens. $w_j^i \in \{0 \cdots v-1\}$ is the label for the masked token, v is the required vocabulary size, and m is the number of dialog sessions. Besides, n and k are the length of the input and masked action sequences, respectively.

3 Experiments and Results

We first conduct the simulator evaluation to assess the DaTrans' performance of learning efficiency, the robustness of fine-tuning Corpus, and domain adaptation. Besides, the case study and human evaluation are conducted and the results are presented in Section D & E in Appendix. In our experiment, NLU and NLG modules are ignored since the interactions are made with dialog actions. Notably, DaTrans can be equipped with any NLU and NLG models. Two datasets, MultiWoz (Budzianowski et al., 2018) and Schema-Guided dialog (SGD) (Rastogi et al., 2020) are involved. We leverage a public available agenda-based user simulator (Zhu et al., 2020) setup on MultiWoz. The details of the dataset, implementation, and the user simulator are illustrated in the Appendix.

3.1 Baseline Agents

We compare the performance of the proposed DaTrans with the state-of-art model JOIE (Wang and Wong, 2021), vanilla BERT, and its variants of different optimization and fine-tuning settings. [1] **DQN** agent is trained with a deep Q-Network. **BERT** agent is equipped with BERT as the encoder that replaces the fully connected layer in DQN. **BERT**$_{\texttt{MWoz}}$ agent is with BERT pre-trained with MLM and NSP tasks on MultiWoz. **JOIE** agent (Wang and Wong, 2021) is a collaborative multi-agent framework factoring the joint action space and learning each part by a different agent. **DaTrans**$_{\texttt{MWoz}}$ is our proposed agent that is pre-trained with MLA task as described in Section 3.1 on MultiWoz dataset.

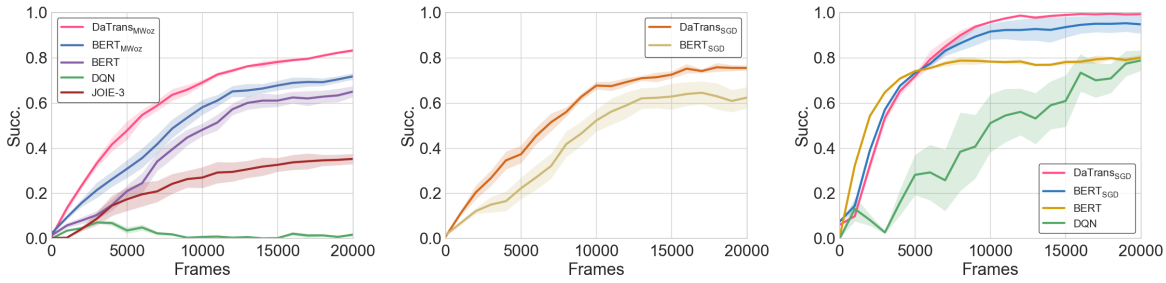
Table 1: The simulation performance of different agents. Succ. denotes the final success rate, Turn and Reward are the average turn and the average reward of the whole training process, respectively.

Model	Succ.↑	Turn↓	Reward↑
DaTrans$_{\texttt{MWoz}}$	**0.84**	**10.21**	**27.35**
BERT$_{\texttt{MWoz}}$	0.72	12.14	14.21
BERT	0.64	14.75	-15.47
DQN	0.01	19.51	-53.66
JOIE-3	0.38	15.98	-21.42

3.2 Simulator Evaluation

All agents are evaluated with the success rate (Succ.) at the end of the training, average turn (Turn), average reward (Reward). The main simulation results are shown in Table. 1 and Figure 2(a). The results indicate that the proposed DaTrans$_{\texttt{MWoz}}$ learns faster and achieves

[1] "optimization" refers to the interactive training process with Reinforcement Learning. "pre-train" means the process of PLMs trained with massive plain text. Besides, we use both "pre-train" and "fine-tuning" to refer to the self-supervised training process of BERT with annotated historical action sequences.



(a) Main results. (b) Domain adaptation. (c) Altering fine-tuning corpus.

Figure 2: Comparison of the success rate evolving during the training process.

a better convergence in in-domain evaluation. DaTrans$_{\texttt{MWoz}}$, pre-trained with the mask last act task (MLA) on the MultiWoz corpus achieves the best Succ. (on average 0.84) with the highest learning efficiency in BERT-based models. The performance of DaTrans$_{\texttt{MWoz}}$ reveals that our MLA pre-training task can not only encode the characteristics of dialog policy for efficiency improvement but also show better transfer abilities because dropping it BERT$_{\texttt{MWoz}}$ degrades the performance of DaTrans$_{\texttt{MWoz}}$. Additionally, BERT is consistently the worst in BERT-based models, which is not surprising since it is only initialized with official BERT's pre-trained weights without in-domain fine-tuning. The generality of fine-tuning corpus and task, domain awareness, and knowledge transferability of BERT are poor. Furthermore, without any fine-tuning, JOIE and DQN are worse than BERT-based agents. Finally, the comparison results of Turn and Reward are illustrated in Table. 1. It depicts that DaTrans$_{\texttt{MWoz}}$ achieves the shortest average turn and highest average reward, which is consistent with the learning curves in Figure 2(a).

Effect of fine-tuning Corpus. We further test the effect of different fine-tuning corpus on the performance. The models are pre-trained on SGD and optimized on MultiWoz to investigate the influence of fine-tuning corpus. We denote DaTrans$_{\texttt{SGD}}$ as a variant of DaTrans which is pre-trained on SGD and optimized on MultiWoz. We only compared the results of fine-tuning on SGD, because the agents who have fine-tuned on MultiWoz have seen the dialogue logic of MultiWoz, so it is of little significance to optimize the comparison on MultiWoz. Besides, we don't optimize the models with RL on SGD because we didn't find an open-source simulator for SGD. Thus, we only take SGD to explore the effect of corpus and domain adaptation. The core conclusion indicated from Figure 2(b) is that DaTrans is robust to the different fine-tuning corpus. Firstly, the proposed MLA pre-training task does better in extracting the knowledge of dialog action sequence, especially the structure information that is invariant over domains. As a consequence, DaTrans$_{\texttt{SGD}}$ outperforming BERT$_{\texttt{SGD}}$.

Domain Adaptation. To assess the ability for new task adaptation, we compare the agents that continually learn a new domain Restaurant, starting from being well-trained on the other six domains (i.e. Train, Hotel, Hospital, Taxi, Police, Attraction). Figure 2(c) shows the performances of new task adaptation for dialog policy learning. The results confirm that DaTrans pre-trained with masked last action task is capable of quickly adapting to the new environment compared to DaTrans$_{\texttt{SGD}}$ and BERT$_{\texttt{SGD}}$. Besides, pre-training counts because removing it (BERT) damages the results.

4 Conclusion and Future Work

In this paper, we investigate the pre-trained language model enhancing the reinforcement learning agent for dialog policy learning. We propose DaTrans, which is equipped with a new fine-tuning task that masks the last dialog action to extract the dialog logic for efficient dialog policy learning. The evaluation results show the effectiveness of the proposed DaTrans in terms of learning efficiency and domain adaptation ability.

Limitations

Due to the high cost of interactions with human users, the dialog policy model was trained in a simulated environment rather than real-world scenarios. Our approach is able to construct a highly responsive dialog system because it shortens the required interaction turns, and reduces labour costs associated with interactive training with human users. However, it is worth noting that the model optimized in our experiments may not be suitable for dealing with real-world users, thus simulation evaluation results alone are not sufficient to prove DaTrans's superiority. Despite this limitation, as there are few studies dedicated to investigating PLMs advanced dialog policy learning, We hope that DaTrans will inspire further research in this field in the future.

Acknowledgements

We appreciate the constructive and insightful comments provided by the anonymous reviewers. This research work is partially supported by CUHK under Project No. 3230377.

References

Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E Taylor, and Ann Nowé. 2015. Reinforcement learning from demonstration through shaping. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Thomas Cederborg, Ishaan Grover, Charles L Isbell, and Andrea L Thomaz. 2015. Policy shaping with human teachers. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

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

Christian Geishauser, Carel van Niekerk, Hsien-Chin Lin, Nurul Lubis, Michael Heck, Shutong Feng, and Milica Gasic. 2022. Dynamic dialogue policy for continual reinforcement learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 266–284.

Pavel Gulyaev, Eugenia Elistratova, Vasily Konovalov, Yuri Kuratov, Leonid Pugachev, and Mikhail Burtsev. 2020. Goal-oriented multi-task bert-based dialogue state tracker. *arXiv preprint arXiv:2002.02450*.

Megha Jhunjhunwala, Caleb Bryant, and Pararth Shah. 2020. Multi-action dialog policy learning with interactive human teaching. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 290–296.

Wai-Chung Kwan, Hong-Ru Wang, Hui-Min Wang, and Kam-Fai Wong. 2023. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intelligence Research*, 20(3):318–334.

Sungjin Lee, Qi Zhu, Ryuichi Takanobu, Zheng Zhang, Yaoqin Zhang, Xiang Li, Jinchao Li, Baolin Peng, Xiujun Li, Minlie Huang, and Jianfeng Gao. 2019. ConvLab: Multi-domain end-to-end dialog system platform. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 64–69, Florence, Italy. Association for Computational Linguistics.

Ziming Li, Julia Kiseleva, and Maarten de Rijke. 2020. Rethinking supervised learning and reinforcement learning in task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3537–3546, Online. Association for Computational Linguistics.

Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2231–2240, Copenhagen, Denmark. Association for Computational Linguistics.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 172–182, Online. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696. Number: 05.

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Continuously learning neural dialogue management. *arXiv preprint arXiv:1606.02689*.

Huimin Wang, Baolin Peng, and Kam-Fai Wong. 2020. Learning efficient dialogue policy from demonstrations through shaping. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6355–6365.

Huimin Wang and Kam-Fai Wong. 2021. A collaborative multi-agent reinforcement learning framework for dialog action decomposition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7882–7889.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14230–14238.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang. 2020. ConvLab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 142–149, Online. Association for Computational Linguistics.

A Dataset

Two datasets are involved: 1) MultiWoz (Budzianowski et al., 2018), a large-scale fully annotated corpus of human-human conversations; 2) Schema-Guided dialog (SGD) (Rastogi et al., 2020), multi-domain, task-oriented conversations between a human and a virtual assistant. MultiWOz contains 8,434 pieces of corpus covering 9 domains, while SGD consists of 16,142 pieces of dialog sessions involving 16 domains.

B Implementation Details.

We adopt BERT_{base} (uncased) with default hyperparameters in Huggingface Transformers (Wolf et al., 2020) as the backbone transformer encoder model. We pre-train and optimize BERT-based models on one RTX 2080Ti GPU and GTX TITAN X. The pre-training batch size is 8. The learning rate for the BERT-based model is 0.00003. The action decoder of DaTrans is a linear layer with 400 output units corresponding to the 400 action candidates. Meanwhile, we set the discount factor γ as 0.9. Besides, we apply the rule-based agent from ConvLab (Lee et al., 2019) to warm start the policy with 1000 dialog epochs.

C User Simulator

We leverage a public available agenda-based user simulator (Zhu et al., 2020) setup on MultiWoz. During training, the simulator initializes with a user goal and takes a system action as input and outputs the user action with a reward. The reward is set as -1 for each turn to encourage short turns and a positive reward $(2 \cdot T)$ for successful dialog or a negative reward of $-T$ for failed one, where T (set as 40) is the maximum number of turns in each dialog. A dialog is considered successful only if the agent helps the user simulator accomplish the goal and satisfies all the user's search constraints.

D Human Evaluation

We further conduct a human evaluation to validate the simulation results. We choose the agents trained with 10000 epochs. Before the test, all evaluators are instructed to interact with the agents to achieve their goals. In each session, a randomly selected goal and a random agent are assigned to a user. They can terminate the dialog if they think the session is doomed to fail. At the end of each session, the user is required to judge if the dialog is a success or a failure. We collect 50 conversations for each agent. The results are illustrated in Table. 2. We see that the human evaluation results further convince the simulator evaluation.

Table 2: The Human performance of different agents. The evaluation is conducted at 10000 epochs in Figure 2(a) for all agents. Succ. denotes success rate.

Model	Succ.↑
$\text{DaTrans}_{\texttt{MWoz}}$	0.68
$\text{BERT}_{\texttt{MWoz}}$	0.58
BERT	0.46
DQN	0.00
JOIE-3	0.24

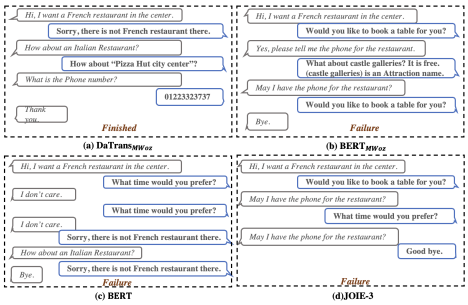


Figure 3: Sampled dialogue examples generated by $\text{DaTrans}_{\texttt{MWoz}}$, $\text{BERT}_{\texttt{MWoz}}$, BERT, DQN, JOIE3. The grey boxes convey the queries from the users while the blue boxes are the responses from the agents. At the bottom of the boxes, we marked whether the session is successful or not.

E Case Study

To further explore the performance of the agents after training, we randomly sampled some real examples generated for a shared restaurant goal. From the samples placed in Fig. 3, some explicable clues are found. In this example, $\text{BERT}_{\texttt{MWoz}}$ fails because it makes mistakes in the restaurant's dialogue logic though it recognizes the right domain. Besides, the response involving "castle galleries" indicates $\text{BERT}_{\texttt{MWoz}}$ suffers from disturbance from other task Attraction. As for BERT and JOIE3, it seems that the knowledge regarding restaurant has not been mastered. Only $\text{DaTrans}_{\texttt{MWoz}}$ systematically handles the issues by taking reasonable actions.

The Wizard of Curiosities: Enriching Dialogues with Fun Facts

Frederico Vicente, Rafael Ferreira, David Semedo, João Magalhães
Universidade NOVA de Lisboa
NOVA LINCS
Lisbon, Portugal
fm.vicente@campus.fct.unl.pt, rah.ferreira@campus.fct.unl.pt,
df.semedo@campus.fct.unl.pt, jm.magalhaes@fct.unl.pt

Abstract

Introducing curiosities in a conversation is a way to teach something new to the person in a pleasant and enjoyable way. Enriching dialogues with contextualized curiosities can improve the users' perception of a dialog system and their overall user experience. In this paper, we introduce a set of curated curiosities, targeting dialogues in the cooking and DIY domains. In particular, we use real human-agent conversations collected in the context of the Amazon Alexa TaskBot challenge, a multimodal and multi-turn conversational setting. According to an A/B test with over 1000 conversations, curiosities not only increase user engagement, but provide an average relative rating improvement of 9.7%.

1 Introduction

The concept of curiosity has for decades been debated by neuroscientists and psychologists. According to Kidd and Hayden (2015), it can be framed into two research views: (1) curiosity as a natural impulse for seeking extended cognition; and (2) a phenomenon related to exploring, playing, learning, and the desire for information. Berlyne (1966) went even further, meditating about how humans had inherently a special type of curiosity, an epistemic curiosity, meaning that above the exploration and information-seeking need, humans also strive for knowledge.

Multimodal conversational task assistants (Gottardi et al., 2022) seek to guide users in accomplishing complex tasks (*e.g.* "Cooking a Strawberry Pie" or "Fixing a broken chair"), in an objective, concise, and engaging manner. Naturally, conversations are rich in knowledge and senses, that are transmitted to users in a dosed manner, towards a successful completion of the task, such that at all phases, knowledge complexity is managed. From the user's perspective, executing a task can be cognitively demanding, potentially involving learning new procedures, using new tools, and following complex task instructions. Thus, conversational assistants should not only ensure a smooth completion of the tasks but also seek to make the task execution a pleasant and entertaining experience that appeals to human senses and curiosity (Dean et al., 2020). To that end, we propose to enrich conversational task assistants with contextualized fun facts, exploiting humans' curiosity-driven information-seeking traits (Kidd and Hayden, 2015). As seen in the work of Konrád et al. (2021), trivia facts have a positive impact on conversations with virtual agents, if used correctly. Hence, in this paper, when dialoguing about a complex task, the user is guided through a sequence of steps as shown in the example in appendix A. Any attempt to fruitfully extend a conversation flow must be done with care. Thus, dialog curiosities should be used as a dialog-enriching element that seeks to maximize user satisfaction. User's psychological factors aligned with the agent efficacy and correctness will be determining aspects. Inspired by Berlyne (1966)'s work, and by the computational model of curiosity of Wu et al. (2012), we propose the introduction of dialog curiosities closely contextualized with certain flows of a conversation, to improve user satisfaction/engagement.

In this context, our contributions are twofold: first, we propose a manually curated dataset of curiosities for the recipes and DIY domains; second, we propose a robust method to naturally insert curiosities in dialogues[1]. An A/B test with over 1000 conversations, conducted with real Alexa users, showed that the proposed approach achieves a relative rating improvement of 9.7%.

2 Curiosities Dataset

In this section, we explain the curiosities dataset creation process, that seeks to fill the existing gap

[1] https://github.com/Mr-Vicente/Curiosity-Dataset

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 149–155
September 11–15, 2023. ©2023 Association for Computational Linguistics

with regards to dialog curiosities for task assistants in the recipes and DIY domains. In particular, we considered the following principles: i) the curiosities' length matters significantly; ii) curiosities should be simple since dense and complex facts could have a negative impact on user engagement; and iii) the quality of each curiosity is more important than the number of curiosities.

2.1 Dataset Categories and Statistics

The dataset consists of a total of 1351 curiosities, with 754 curiosities for the cooking domain and 597 for the DIY domain, which are the target domains of the Alexa TaskBot challenge (Gottardi et al., 2022). Some examples of the curiosities general classes are listed below.

Sample Recipe concepts. Fruit (*e.g.* Avocado, Vitamin C); Meat (*e.g.* beef); Seafood (*e.g.* shrimp); tools (*e.g.* spatula); cuisine concepts (*e.g.* temperature); Popular countries' food (*e.g.* pizza, sushi); U.S. National food days.

Sample DIY concepts. American DIY statistics; DIY tools (*e.g.* hammer); Gardening (*e.g.* lawn mower); Garage (*e.g.* car, bike). House furniture (*e.g.* bookshelf); DIY tasks U.S. National days.

2.2 Curiosities Dataset Creation

The dataset was created by a manual process of searching and curating information found online. We started by considering a main class of a concept, for example "Fruit", and used Google search to find curiosities. After this first process, we get into more specific concepts, such as "Avocado". We complement our dataset with diverse temporally contextualized curiosities. Specifically, we employed a template-based approach to generate curiosities from national food days.

All the curiosities were manually curated to fit the characteristics and specifications identified, ensuring their quality and appropriateness for dialog and its domain.

2.3 Length per Curiosity

The length of a sentence can significantly affect the user's comprehension, especially in voice-based interactions, such as Alexa. Figure 1 shows the length distribution in words of the dataset for both domains. We deposited careful attention to conforming the curiosities length distribution to an average of 15 words, avoiding long sentences to maximize the readers' comprehension.

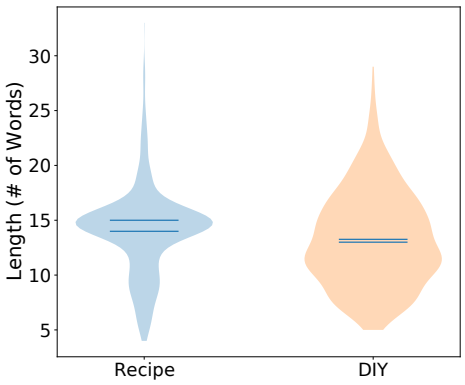


Figure 1: Curiosities length distribution.

3 Dialogues and Contextual Curiosities

One of the objectives of our work is to incorporate curiosities into a conversational assistant's flow while users are being guided through a task in natural language (Colas et al., 2020). However, inserting the right curiosity in the right dialog turn is a non-trivial task. Moreover, matching curiosities to a particular task with human effort would produce a high-quality reward, but, in practice, it is intractable due to the large amount of both tasks ($>$ 100k) and curiosities ($>$ 1k). To this end, we propose two distinct automatic approaches as described in the following sections.

3.1 Extracting Relevant Information

Recipes and DIY articles in their raw form consist of structured text composed of various elements such as the title, steps, and possibly extra information (*e.g.* ingredients, categories, and short descriptions). As a first step, we pre-process the tasks' content to match the curiosities by considering the most relevant content, taking into account the current phase of the conversation. Thus, for the recipes, we extracted the title, the steps, and the ingredients, whereas, for the DIY tasks, we extracted only the title and steps. The main goal is to capture fine-grained and task-specific details from each task, that will allow us to link a task to highly relevant curiosities, as described next.

3.2 Matching Curiosities to Dialogues

Given the information extracted from the tasks, we considered two approaches to match the curiosities to the conversation: (1) a text-based method, and (2) a semantic similarity search using pre-trained LM embeddings (Reimers and Gurevych, 2019).

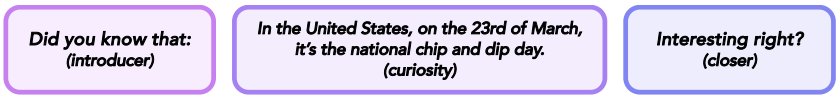


Figure 2: Example of a curiosity enclosed by an *opener* and a *closer*.

Text-based Method. In this approach, we first process the curiosities and the tasks' text by removing punctuation, stopwords, verbs, plurals, and domain-specific common words such as "hours" and "degrees". Then, we calculate two bag-of-words vectors considering the domain-specific words and both the curiosity and the current step of the task. After this, we perform the cosine similarity between both of these vectors for each curiosity available in the dataset and re-rank them according to this score.

Semantic Similarity Search Method. Matching a curiosity to the context of a dialogue requires some level of language understanding that goes beyond keyword matching. Hence, we considered a two-stage process considering a Sentence-BERT (Reimers and Gurevych, 2019) model that first separately encodes the task's content and all curiosities to map them to a common embedding space, allowing the assessment of the similarity between both. This is followed by a re-ranking cross-encoder method to further improve the results. Details of the algorithm are in Appendix B.

3.3 Inserting Curiosities in Dialogue

In a conversational task assistant, the primary objective is to assist the user in accomplishing a task (Gottardi et al., 2022). Therefore, the introduction of curiosities in the conversation should improve the dialog flow and maximize engagement. This requires a careful and contextualized blend of curiosities throughout the conversation.

Curiosity Offer/Backoff. Curiosities should improve the user experience, without negatively affecting the quality of a dialogue (Zheng et al., 2021). To ensure the overall users' satisfaction, and avoid non-intrusive behaviors, we designed a dialogue curiosity offer/backoff strategy (see Appendix C for the full algorithm).

An important aspect of our offer/backoff strategy is that we consider the user's cognitive load, and we never introduce curiosities at the beginning of a dialog, or when the user is listening to long steps ($\geq$200 words). This aims to keep the user focused, to provide short responses that account for the users' attention span. At these points of the dialogue, there are multiple voice instruction commands being explained to the user. Prompting and telling a curiosity would only cause confusion and cognitive overload.

We opted to ask the user at the end of a task step if they want to hear a curiosity (Appendix A, blue text). Given the question, the user can accept, deny, or ignore the request. If the user denies or ignores the curiosity, we opted to not prompt the user again, since the user might not have interest in this feature or may become frustrated. If the user accepts the curiosity, the bot responds with a fun fact following the structure discussed next.

Curiosities Openers and Closers. To smoothly insert individual curiosities in the dialog flow, while keeping the conversational gist, we propose a curiosity-to-dialog scheme, that encompasses curiosity linguistic *openers* and *closers*. To deliver a curiosity with the right tone of voice, we select an *opener* from a pre-defined list, to introduce the curiosity. Similarly, to gracefully end the insertion of a curiosity, we appended a *closer* phrase after the curiosity sentence. Given that the *closer* needs to act as a bridge between the curiosity and the main dialog flow, we formulated a set of ending sentences for the terminator phrase, with the aim of making them sound exciting, while signaling the end of the curiosity sub-flow. An example of a curiosity along with its corresponding *opener* and *closer* phrases, is illustrated in Figure 2.

4 Experimental Results

In this section, we detail the A/B testing setup and discuss the obtained results.

4.1 A/B Testing Setup

To measure the impact of introducing curiosities in a conversation, we performed A/B testing with Alexa device users, in the context of the Alexa Prize TaskBot Challenge 2021 (Gottardi et al., 2022).

The implemented dialogue system interacted with thousands of real users (Ferreira et al., 2022). The dialog state tracking is based on a BERT intent detector (Tavares et al., 2023) and the task retriever

Table 1: A/B testing results: system A engaged users in curiosities and system B had no possibility of curiosities at all. In system A, the user can accept, deny or ignore the curiosity recommendation.

Sys	User action	Conversations	Rating
A	Accepted (≥ 1)	526 (50.8%)	3.94
	Not-accepted	211 (20.4%)	3.55
B	Curios. disabled	299 (28.9%)	3.62

is based on a conversational search method (Ferreira et al., 2021). At the end of a conversation, the user is prompted to give a 1 to 5 rating regarding the quality of the conversation. We use the ratings as the success metric of the proposed work. We performed this study using an A/B testing method, by considering a version of the system with curiosities (A) and without curiosities (B). To ensure that we had high-quality data, we only considered conversations with a minimum of 3 turns, resulting in a total of 1036 conversations.

4.2 Dialogue Curiosities A/B tests

In Table 1, we summarize the A/B testing results that we conducted. We had 71.1% of the conversations in system A and 28.9% in system B. In system A, the user had the option to hear the curiosity and to decline it. Hence, 50.8% of the conversations had curiosities and 49.2% had no curiosities. In all systems, users were anonymous and randomly assigned to our system. Table 1 also relates the users' acceptance of curiosities to average ratings. The results show that users that accept at least one curiosity give on average a higher rating (3.94) compared to users that are not interested or that simply ignored the curiosity (3.55). Overall, this increase in rating shows that users that interact with the curiosities appear to be more engaged in the conversation, which in turn leads to a higher rating.

4.3 Ratings per Number of Curiosities

In this section, we examine system A results in more detail. Overall, we observed a positive result with 70% accepting a curiosity, 18% ignoring (the user does not confirm, *e.g.* "next step"), and 12% denying. Moreover, the relation between the number of curiosities per conversation and the rating is another positive result, Table 2. From these results, we can see that when curiosities are present in a dialogue, the rating is consistently higher than when no curiosities are said. In particular, we see a rating improvement from 3.55 against 3.74 in the worst-case scenario, and 4.13 in the best scenario. The mode is one curiosity per conversation, which corresponds to an average rating of 3.95, i.e. a relative improvement of 9.7%. These are encouraging results, showing that the users are receptive to listening to curiosities in the conversation which in turn leads to increased user satisfaction.

Table 2: System A's results breakdown: the number of provided curiosities and average rating.

Curiosities	None	1	2	≥ 3
Conversations	211	479	32	15
Avg. Rating	3.55	3.95$_{(+9.7\%)}$	3.74	4.13

4.4 Ratings by Curiosities Matching Method

We also examined the impact of the dialogue-curiosity matching methods of Section 3.2. Table 3 shows the results obtained with both methods. The two methods achieve high ratings, with the Semantic Similarity method obtaining slightly higher ratings, thus being more preferable.

Table 3: Rating by curiosity matching method.

Method	**Count**	**Rating**
Semantic Similarity	344 (64.18%)	3.99
Text-based	192 (35.82%)	3.86

5 Critical Discussion and Limitations

Manually Curated vs Hallucinated Curiosities. As an alternative to manually curated curiosities, current LLMs can generate curiosities contextualized to the conversation. We tested this strategy but observed that, often these curiosities are false and incorrect. Hence, this is not a viable solution when the dialog system guides a user through a complex manual task where reliability is key. An example of a false hallucination that we observed is "*Microwaves don't heat the food, they heat the water molecules in it, this causes them to vibrate which is what causes the heat. This vibration is good for your body because it causes your cells to produce more energy.*".

Long-term Effect of Curiosities. We studied the effect of curiosities during a period of 6 months with a controlled A/B testing. However, due to privacy issues, we did not track users, preventing

us from studying the long-term effects of curiosities in recurring users.

Selection bias. Our study is limited to users that own an Alexa device, and to users that participated in the Alexa Prize TaskBot challenge. Moreover, in this setup, we only give fun facts to users who accept the offer of a fun fact. This creates a slight "selection bias" because users who accept a fun fact were probably the ones who were already enjoying the interaction and might have been more likely to give a higher rating. Likewise, users who rejected the curiosity were probably not enjoying the interaction and may have lowered their rating.

To obtain a reference rating (a neutral baseline), we disabled the fun facts functionality and tested the system. Table 1 provides an analysis that sheds some light on this issue. With fun facts disabled, the average rating is 3.62; users who refuse the fun fact, rate the system -0.07 points lower than the neutral baseline; users who accept the fun fact generally rate the system +0.32 points above the neutral baseline. Given the setup, in the future, we will study ways of mitigating possible sources of selection bias.

6 Conclusions

In this paper, we presented a novel approach to introducing curiosities in conversations. Specifically, we curated a dataset of curiosities in the recipes and DIY domains and evaluated the impact of introducing these curiosities in real human-agent conversations in the Alexa TaskBot challenge. We assessed the impact of curiosities in a conversational task assistant setting, and the results allow us to conclude that introducing curiosities in a non-intrusive manner and in the context of the dialog can increase user engagement and improve their appreciation of the dialogue system.

These findings have important implications for the design of conversational systems and can inform future research on incorporating curiosities in conversations to enhance the user experience. As future work, we will (1) investigate the use of generative models for creating factually grounded curiosities (Ouyang et al., 2022; Touvron et al., 2023) and compare them to manually curated curiosities; and (2) investigate methods that can contextualize the curiosities according to a graph of entities (Gonçalves et al., 2023).

Acknowledgments

This work has been partially funded by the FCT project NOVA LINCS Ref. UIDP/04516/2020, by the Amazon Science - TaskBot Prize Challenge and the CMU|Portugal projects iFetch CMUP LISBOA-01-0247-FEDER-045920), and by the FCT Ph.D. scholarship grant UI/BD/151261/2021. Any opinions, findings, and conclusions in this paper are the authors' and do not necessarily reflect those of the sponsors.

References

D. E. Berlyne. 1966. Curiosity and Exploration. *Science* 153, 3731 (1966), 25–33. https://doi.org/10.1126/science.153.3731.25

Cédric Colas, Tristan Karch, Nicolas Lair, Jean-Michel Dussoux, Clément Moulin-Frier, Peter Dominey, and Pierre-Yves Oudeyer. 2020. Language as a cognitive tool to imagine goals in curiosity driven exploration. *Advances in Neural Information Processing Systems* 33 (2020), 3761–3774.

Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. 2020. See, hear, explore: Curiosity via audio-visual association. *Advances in Neural Information Processing Systems* 33 (2020), 14961–14972.

Rafael Ferreira, Mariana Leite, David Semedo, and João Magalhães. 2021. Open-Domain Conversational Search Assistant with Transformers. In *ECIR (1) (Lecture Notes in Computer Science, Vol. 12656)*. Springer, 130–145.

Rafael Ferreira, Diogo Silva, Diogo Tavares, Frederico Vicente, Mariana Bonito, Gustavo Goncalves, Rui Margarido, Paula Figueiredo, Helder Rodrigues, David Semedo, and Joao Magalhaes. 2022. TWIZ: A conversational Task Wizard with multimodal curiosity-exploration. In *Alexa Prize TaskBot Challenge Proceedings*.

Gustavo Gonçalves, Joao Magalhaes, and Jamie Callan. 2023. Conversational search with random walks over entity graphs. In *ACM ICTIR*. ACM.

Anna Gottardi, Osman Ipek, Giuseppe Castellucci, Shui Hu, Lavina Vaz, Yao Lu, Anju Khatri, Anjali Chadha, Desheng Zhang, Sattvik Sahai, Prerna Dwivedi, Hangjie Shi, Lucy Hu, Andy Huang, Luke Dai, Bofei Yang, Varun Somani, Pankaj Rajan, Ron Rezac, Michael Johnston, Savanna Stiff, Leslie Ball, David Carmel, Yang Liu, Dilek Hakkani-Tur, Oleg Rokhlenko, Kate Bland, Eugene Agichtein, Reza Ghanadan, and Yoelle Maarek. 2022. Alexa, let's work together: Introducing the first Alexa Prize TaskBot Challenge on conversational task assistance. *Alexa Prize TaskBot Challenge Proceedings*.

C. Kidd and B. Y. Hayden. 2015. The Psychology and Neuroscience of Curiosity. *Neuron* 88, 3 (Nov 2015), 449–460.

Jakub Konrád, Jan Pichl, Petr Marek, Petr Lorenc, Van Duy Ta, Ondřej Kobza, Lenka Hýlová, and Jan Šedivý. 2021. Alquist 4.0: Towards Social Intelligence Using Generative Models and Dialogue Personalization. arXiv:2109.07968 [cs.CL]

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR* abs/2203.02155 (2022). https://doi.org/10.48550/arXiv.2203.02155 arXiv:2203.02155

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

Diogo Tavares, Pedro Azevedo, David Semedo, Ricardo Sousa, and Joao Magalhaes. 2023. Task Conditioned BERT for Joint Intent Detection and Slot-filling. In *Progress in Artificial Intelligence - 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Faial, Portugal, September 5 - 8, 2023, Proceedings*. Springer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023). https://doi.org/10.48550/arXiv.2302.13971 arXiv:2302.13971

Qiong Wu, Chunyan Miao, and Zhiqi Shen. 2012. A curious learning companion in Virtual Learning Environment. In *FUZZ-IEEE 2012, IEEE International Conference on Fuzzy Systems, Brisbane, Australia, June 10-15, 2012, Proceedings*. IEEE, 1–8. https://doi.org/10.1109/FUZZ-IEEE.2012.6251362

Lulu Zheng, Jiarui Chen, Jianhao Wang, Jiamin He, Yujing Hu, Yingfeng Chen, Changjie Fan, Yang Gao, and Chongjie Zhang. 2021. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *Advances in Neural Information Processing Systems* 34 (2021), 3757–3769.

A Example Dialogues

Table 4: Example of dialogue with curiosities. We highlight the question, opener, and closer.

System	User
Hi, this is an Alexa Prize Taskbot. I'm excited to help you with recipes or DIY tasks! Try asking me: How to cook salmon, or How to Decorate a Pot.	I want to make chicken tacos.
Sure, I found Chicken Tacos recipes in the Whole Foods Market. Would you like to 'start', or 'hear the next recipes'.	Start the recipe.
Chicken Tacos will be yummy. Step 1: Preheat oven to 375°F. To keep going with this recipe, say 'next'.	Next.
Step 2. Brush butter onto both sides of the tortillas. And look, would you like me to astonish you with a fun fact?	Sure.
Did you know that: The word tortilla comes from the Spanish word torta which means round cake. I love sharing these facts!	Next Step.

Table 5: Curiosity in the context of the recipe "Buttermilk Biscuits".

Step Text and Question
Step 2. Fold dry ingredients into buttermilk with your hands until a sticky dough forms. And look, would you like me to astonish you with a fun fact?
Fun Fact
How crazy is it that: In the US, on the 14th of May, it's the national buttermilk biscuit day. I found this really interesting! Anyway, to continue, say 'next'.

Table 6: Curiosity in the context of the DIY task "How to Braid Hair".

Step Text and Question
Step 3. Start with a secure base. If you tie hair into a ponytail with a hair tie, your braid will be easier to handle. And by the way, would you be interested in a fun fact about this?
Fun Fact
Alert! Alert! Fun fact time! The average person has between 100k and 150k strands of hair. This blew my mind! Anyway, to continue, say 'next'.

B Semantic Similarity Curiosity-Matching Algorithm

Algorithm 1: Curiosity Matching

Input : $Tasks$: List of tasks
Input : $n \leftarrow 10$: int (top-n candidate curiosities)
Input : $m \leftarrow 3$: int (top-m candidate curiosities-task matches)
for *each task in Tasks* **do**
  Separate task's content into title, steps (and ingredients) using special tokens;
  Encode the task's content;
  Encode the domain-specific curiosities;
  Calculate the cosine similarity between the task's content and the curiosities;
  Select the top-n curiosities;
  Apply a Cross-Encoder model to all n pairs and select the top-m pairs;
end

C Curiosities Offer/Backoff Algorithm

Algorithm 2: Curiosities Offer/Backoff

Input : T: Task
Input : n_steps: int
Input : $curr\_step$: int
Input : $last\_fact\_step$: int
Input : $questions\_asked$: int
Output : $ask\_curiosity$: bool
$k \leftarrow 6$;
$max\_questions \leftarrow (n\_steps \;//\; k) + 1$;
if $questions\_asked \geq max\_questions$ **then**
  $ask\_curiosity \leftarrow$ **False**;
else if $curr\_step \neq 1$ ***and*** $curr\_step = last\_fact\_step + k$ ***and*** $last\_fact\_step \leq curr\_step$ ***and*** $curr\_step \neq (n\_steps - 1)$ **then**
  $ask\_curiosity \leftarrow$ **True**;
else
  $ask\_curiosity \leftarrow$ **False**;
return $ask\_curiosity$;

The Road to Quality is Paved with Good Revisions: A Detailed Evaluation Methodology for Revision Policies in Incremental Sequence Labelling

Brielen Madureira[1] **Patrick Kahardipraja**[1] **David Schlangen**[1,2]
[1]Computational Linguistics, Department of Linguistics, University of Potsdam, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
{madureiralasota,kahardipraja,david.schlangen}@uni-potsdam.de

Abstract

Incremental dialogue model components produce a sequence of output prefixes based on incoming input. Mistakes can occur due to local ambiguities or to wrong hypotheses, making the ability to revise past outputs a desirable property that can be governed by a policy. In this work, we formalise and characterise edits and revisions in incremental sequence labelling and propose metrics to evaluate revision policies. We then apply our methodology to profile the incremental behaviour of three Transformer-based encoders in various tasks, paving the road for better revision policies.

1 Introduction

Since the dawn of Wikipedia, users have made 1.7×10^9 edits to its pages. Its most revised entry contains 56,713 revisions, all documented in the page history.[1] In such an active community, conflicts inevitably occur. Editors can begin competing to override each other's contributions, causing dysfunctional *edit warrings*.[2] To help regulate the environment, an editing policy is in force, aiming at making edits constructive and improving quality.[3]

Edits, revisions and policies are key concepts in incremental processing, where a model must rely on partial input to generate partial output. Incrementality can help optimise reactivity, naturalness, quality and realism in interactive settings (Schlangen and Skantze, 2011). This is particularly relevant in dialogue models whose NLU components need to operate on incoming input, *e.g.* while performing NER, slot filling or disfluency detection, or doing simultaneous translation.

Local ambiguities in the linguistic input and transient mistakes by the model can result in wrong partial hypotheses, so that the ability to *revise*, by *editing* previous outputs, is desirable (Kahardipraja et al., 2023). Beyond monitoring the occurrence of edits, it is also beneficial to have a *policy* regulating when and which revisions should be made, reducing the occurrence of undesirable edits. Existing literature using consolidated incremental evaluation metrics falls short in capturing relevant nuances of the incremental behaviour in terms of revisions.

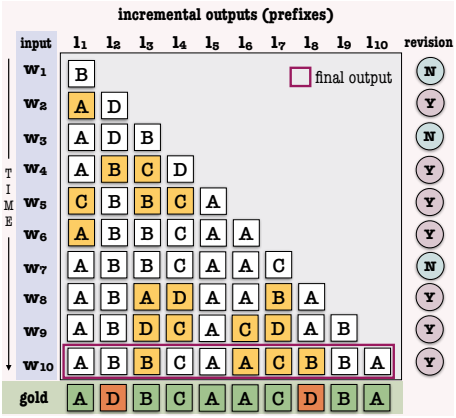


Figure 1: Constructed example of an incremental chart containing output prefixes with marked edits (yellow) and revisions in incremental sequence labelling. Red stands for wrong final predictions wrt. the gold standard.

In this work, we propose an evaluation methodology for revision policies in incremental sequence labelling. A constructed example is shown in Figure 1, with revisions indicated in the right column. Specifically, our contributions to address the identified evaluation gap are: A formalisation of revision policy in incremental sequence labelling, characterising types of edits and of revisions (§4.1-4.2); a proposal of specialised evaluation metrics for revision policies, accompanied by a discussion on the desired behaviour of incremental processors (§4.4-4.5); and a demonstration of our methodology with an analysis of the revision policy in three sequence labelling Transformer-based models (§5).[4]

[1]According to Wikimedia Statistics and wiki Special.
[2]https://en.wikipedia.org/wiki/Wikipedia:Edit_warring
[3]https://en.wikipedia.org/wiki/Wikipedia:Editing_policy

[4]Our implementation is available at https://github.com/briemadu/inc-eval-revisions with accompanying documentation on how to run the evaluation for other models.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 156–167
September 11–15, 2023. ©2023 Association for Computational Linguistics

2 Motivation

Incremental natural language processing[5] has *time* at front line, being pivotal for interactive settings. At each time step, models must operate on partial input to deliver partial output, but sometimes previous decisions have to be revised. For example, at time step 4 in Figure 1, the labels for the input tokens 2 and 3 were edited into new states. With regard to revisions, at least three types of incremental processors exist, as summarised in Table 1:

1. Inherently incremental but monotonic models. They keep an internal state that is updated and used to extend the output at each time step, but cannot revise previous outputs.

2. Non-incremental models used with a *restart-incremental* interface, being forced to perform a full recomputation at each time step. Such models revise the output as a by-product of their recomputations.

3. Incremental models with a dedicated policy to detect the need to perform revisions only when deemed necessary and, more specifically, deciding which parts of the output prefix need to be revised and how.

		non-incremental	incremental
revisions	no	n/a	strictly monotonic outputs
	yes	recomputation policy doing revisions as a by-product	revision policy

Table 1: Types of incremental processors.

Monotonicity avoids instability in the output, allowing subprocesses to start immediately, as it is certain that the outputs will not change. However, they never recover from mistakes, which is one of the drawbacks of employing vanilla RNNs and LSTMs (Hochreiter and Schmidhuber, 1997).

Models that depend on the availability of full sentences at once can be "incrementalised" with the *restart-incremental* paradigm (Schlangen and Skantze, 2011), causing revisions to occur via recomputations.[6]

Cutting-edge NLP models currently rely on Transformers (Vaswani et al., 2017), which are non-incremental. Using them in a *restart-incremental* fashion requires recomputing from scratch at every time step, which we hereby name the *naive recomputation policy*. It is a very expensive policy because, for a sequence of n tokens, the complexity is $\sum_{i=1}^{n} i^2$ (*i.e.* the n-th square pyramidal number). Besides, this naive approach wastes computational budget, because not all recomputations cause revisions. The results reported by Kahardipraja et al. (2023), for example, show that only around 25% of the recomputations actually changed the output prefix. The disadvantages of the naive policy can be alleviated by a smarter policy that cuts down the number of time steps with recomputations.

Still, beyond deciding when to *recompute*, a revision policy par excellence should directly guide the more specific decision of when (and what) to actually *revise*, and must be evaluated accordingly.

3 Related Literature

Revisability is in the nature of incremental processing: Hypothesis revision is a necessary operation to correct mistakes and build up a high-quality final output (Schlangen and Skantze, 2011). Still, there is a trade-off between requiring that later modules handle a processor's revisions and buying stability by reducing some of its incrementality, which makes the concept of *hypothesis stability* very relevant (Baumann et al., 2009). Beuck et al. (2011a) argue that performing revisions should not take as long as the initial processing, so as to retain the advantages of incremental processing. They propose two strategies: Allowing revisions only within a fixed window or limiting their types. Empirically determining how often a model changes the output is an aspect of their analysis we also rely on.

The restart-incremental paradigm was investigated for Transformer-based sequence labelling by Madureira and Schlangen (2020) and Kahardipraja et al. (2021); recently, adaptive policies were proposed to reduce the computational load (Kaushal et al., 2023; Kahardipraja et al., 2023). Rohanian and Hough (2021) and Chen et al. (2022) explored adaptation strategies to use Transformers for incremental disfluency detection. In simultaneous translation, where policies are a central concept (Zheng et al., 2020a; Zhang et al., 2020), the restart-incremental approach is in use and revisions are studied (Arivazhagan et al., 2020; Sen et al., 2023).

[5] For a review, see Köhn (2018). In other contexts, also referred to as real-time processing (Pozzan and Trueswell, 2015) or streaming (Kaushal et al., 2023).

[6] Also called *incremental interface* (Beuck et al., 2011a) or *beat-driven approach* (Baumann et al., 2011).

latency, quality, stability	simultaneous translation	Arivazhagan et al. (2020) Ma et al. (2020)
quality, responsiveness, robustness, stability	speech recognition and diarization	Baumann et al. (2009) Addlesee et al. (2020)
similarity, timing, diachronic	general	Baumann et al. (2011)
fluency, latency, quality, recovery capabilities, timing	simultaneous interpreting (MT and speech synthesis)	Baumann et al. (2014)
decisiveness, monotonicity, stability, timeliness	POS tagging	Beuck et al. (2011a)
amount of predicted information, connectedness, delay, inclusiveness, monotonicity, quality	parsing	Beuck et al. (2011b, 2013) Köhn and Menzel (2014)
cognitive aspects, efficiency	neural coreference resolution	Grenander et al. (2022)
jumpiness, position	reference resolution	Schlangen et al. (2009)
accuracy, integration, representational similarity	sequence-to-sequence	Ulmer et al. (2019)
consistency, diminishing returns, interruptibility, monotonicity, preemptability, (recognisable) quality	anytime algorithms	Zilberstein (1996)

Table 2: Overview of relevant properties for incremental evaluation in various tasks.

Sequence labelling is a staple of various incremental linguistic tasks possibly used in dialogue systems, like SRL (Konstas et al., 2014), POS-tagging (Beuck et al., 2011a), dialogue act segmentation (Manuvinakurike et al., 2016), disfluency detection (Hough and Schlangen, 2015) and dependency parsing (Honnibal and Johnson, 2014).

Revision Categorisation and Prediction Approaches to categorise the properties of revisions or edits exist in various areas. Faigley and Witte (1981) examine the effects and causes of revisions in writing, providing a taxonomy on whether revisions change meaning and bring new information. Afrin and Litman (2018) classify revision quality by whether they improve student essays. Antonio et al. (2020) categorise revisions and edits in WikiHow in terms of what they cause to the text. Wikipedia's edits have also been classified according to factuality and fluency (Bronner and Monz, 2012) and intents (Rajagopal et al., 2022). Other typologies and taxonomies have been proposed for translation revisions (Fujita et al., 2017) and multilingual NLG revision operations (Callaway, 2003).

Vaughan and McDonald (1986) outline three phases of the revision process in NLG: Recognition, editing and re-generation. Revision rules have been applied for incremental summarisation by Robin (1996). Non-incremental revision learning models also exist, relying on revision rules for dependency parsing (Attardi and Ciaramita, 2007) or classification in POS-tagging (Nakagawa et al., 2002). Predicting stability and accuracy of hypotheses is a relevant task (Selfridge et al., 2011), which allows to distinguish hypotheses that will survive and are thus more reliable (Baumann et al., 2009).

Incremental Evaluation Table 2 presents an overview of relevant properties for incremental evaluation. In their seminal work, Baumann et al. (2011) define three general categories of metrics for incremental processing: *similarity*, *timing* and *diachronic*, which can be employed in incremental sequence labelling. They are suitable for capturing *e.g.* instability (edit overhead), quality of prefixes (correctness) and lag (correction time). Kaushal et al. (2023) propose streaming exact match, comparing prefixes with the final gold standard. While these metrics capture instability and correctness of output prefixes, we lack a standard way to evaluate the quality of the performed revisions. We thus complement their work by proposing fine-grained metrics focusing on revisions and recomputations.

4 Evaluation Methodology

In this section, we present our evaluation methodology for incremental sequence labelling with a focus on revisions. After formalising the task, we characterise revisions and edits, define policies and revision-oriented metrics and discuss the ideal behaviour of incremental sequence labelling models.

4.1 Formalisation

We begin by formalising incremental sequence labelling tasks, extending the similar definition of streaming sequence tagging by Kaushal et al.

(2023) with *edits* and *revisions*. Like them, we assume an idealised format where incremental units are well-defined, fixed and complete input tokens, and a model that produces a label for every new input token, so that the output is necessarily extended at every time step. Note, however, that incremental processors may have to operate at sub-token level or with transitional input, which requires the capability of retracting decisions and adjusting to varying length in real-time. In some models, outputs may not have an immediate one-to-one correspondence to the input (*e.g.* due to a delay strategy (Baumann et al., 2011), or to techniques like opportunistic decoding (Zheng et al., 2020b)) and parallel hypotheses can be kept in memory. See Schlangen and Skantze (2011) for details.

Let $L = \{L_1, \ldots, L_M\}$ be a set of labels. In standard sequence labelling, the task is to map an input sequence of n tokens $(w_i)_{i=1}^n$ to an output sequence of n labels $(l_i)_{i=1}^n$, $l_i \in L$. Each output label l_i classifies its corresponding token w_i. The task is more complex than plain token-level classification because the sequential nature of the input and the output need to be taken into account when predicting labels. If available, a gold-standard sequence $(g_i)_{i=1}^n$, with $g_i \in L$, is used to evaluate the correctness of the predicted output sequence.

In an incremental setting, the input is provided in a piecemeal fashion, one token at a time. At each time step $t = 1, 2, \ldots, n$, an increasing input prefix $(w_i)_{i=1}^t$ is available to the model and an output prefix $(l_i)_{i=1}^t$ is predicted. Therefore, an input sequence with n tokens will result in n output prefixes $p_1, p_2, \ldots, p_n$, which we consider to be partial hypotheses for the final output. Each p_i is a sequence of i labels, containing one additional label at the right in relation to p_{i-1}. The last hypothesis p_n is the final decision of the model, having observed the full input. The complete sequence of prefixes can be represented as a lower triangular matrix, whose cells c_i^j contain the label assigned to w_i at time j and each row i contains p_i. We can represent the incremental input and output in an *incremental chart* (IC) as follows:

w_1	$p_1 =$	l_1^1				
w_1, w_2	$p_2 =$	l_1^2	l_2^2			
w_1, w_2, w_3	$p_3 =$	l_1^3	l_2^3	l_3^3		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
$w_1, w_2, \ldots, w_n$	$p_n =$	l_1^n	l_2^n	l_3^n	$\cdots$	l_n^n
	gold $=$	g_1	g_2	g_3	$\cdots$	g_n

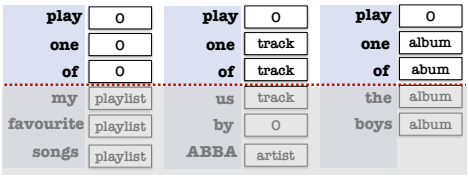


Figure 2: Illustrative example of multiple locally valid hypotheses for the prefix *play one of*. Only after more input is processed definite labels can be assigned.

At each time step t, the observation of the new input token w_t causes the model to i) extend the output sequence with one label for w_t (an addition) and ii) optionally also change its current hypotheses $l_1, \ldots, l_{t-1}$ for previous tokens (substitutions).

An *edit* occurs at time t for label i if $l_i^t \neq l_i^{t-1}$, meaning that the model's prediction for w_i's label changed. A *revision* occurs when, apart from the compulsory addition, a prefix changes at time t in relation to the previous prefix, *i.e.* when at least one label is edited.[7] In Figure 1, revisions occur at time steps $2, 4, 5, 6, 8, 9$ and 10. Highlighted labels in the prefixes are edits.

Gold Standard Evaluation can be done with respect to incremental or non-incremental gold standards (Baumann et al., 2011). Often, only the non-incremental version is available, *i.e.* the labels on the complete sequence, assigned having all left and right context taken into account. A genuinely incremental gold standard contains step-by-step gold prefixes encoding interpretations that are *locally valid* until right context renders it invalid, as illustrated in Figure 2.[8] Since it is usually not available, we can instead "incrementalise" the final gold standard by deriving all its prefixes as hard labels. But this approach somewhat unfairly expects that, even at steps with multiple locally valid interpretations, the model commits to the final decision without observing the input that actually induces that interpretation as correct and the others as wrong. Moreover, using an independent gold standard conflates the external overall performance of the model with the quality of its internal incrementality; an alternative is to consider the final output of the model as a silver standard (Baumann et al., 2011). The correctness of labels and prefixes is then measured with a metric M with respect to the defined target.

[7] The addition is not taken into account here, as it has no precedent label to be compared to at this point. The first time step is by definition not a revision, since there is no prefix yet.

[8] For existing examples, see Hrycyk et al. (2021), Rawat and Barres (2022) and Beuck et al. (2011b).

	Quality	**Edits** (labels)	Example	**Revisions** (prefixes)	Example
Convenience	convenient	change incorrect label	(5,1)	change incorrect prefix	5
	inconvenient	change correct label	(4,2)	change correct prefix	4
Effectiveness	effective	incorrect label → correct	(5,4)	improve prefix correctness	6
	ineffective	incorrect label → incorrect	(9,3)	do not change prefix correctness	9
	defective	correct label → incorrect	(4,3)	worsen prefix correctness	4
Novelty	innovative	label → new state	(9,6)	n/a	n/a
	repetitive	label → previous state	(6,1)	n/a	n/a
(Local) Recurrence	recurrent	subsequence with > 1 edit	(9,3)	subsequence with > 1 revision	8
	steady	subsequence with 1 edit	(4,2)	subsequence with 1 revision	2
Oscillation	oscillating	label with > 1 edit	(6,1)	> 1 revision	all
	stable	label with 1 edit	(4,2)	single revision	-
Company	accompanied	prefix with > 1 edit	(9,6)	prefix with > 1 edit	5
	isolated	prefix with 1 edit	(6,1)	prefix with 1 edit	6
Connectedness	connected	other neighbouring edit	(9,4)	only connected edits	9
	disconnected	no neighbouring edits	(5,1)	only disconnected edits	2
	both	n/a	n/a	both types of edits	5
Distance	short range	near current time step	(5,4)	only short range edits	2
	long range	far from current time step	(9,3)	only long range edits	6
	both	n/a	n/a	both types of edits	5
Definiteness	definite	label → final state	(4,2)	prefix → final state	10
	temporary	label → temporary state	(5,3)	prefix → temporary state	8
Time	intermediate	input still partial	(5,4)	input is still partial	4
	final	at final time step	(10,3)	at the final time step	10

Table 3: Characterisation of edits and revisions. The examples refer to Figure 1, pointing to the (time step, label index) positions for edits and time steps for revisions. Here the gold standard is used to judge prefix correctness.

4.2 Characterisation of Revisions and Edits

In this section, we propose a detailed characterisation for the types of edits and revisions based on ten dimensions, summarised in Table 3, as a means to evaluate revision policies. In the next paragraphs, we assume that either a genuine or a constructed incremental sequence of target prefixes has been selected according to the current needs. We will use Figure 1 and its gold standard as examples.[9]

To characterise edits, we consider the state of an output label in the current prefix in relation to its state in the previous prefix, which are different by definition. They relate to a label's development in time (vertically in their IC column) or to the prefix they belong to (horizontally in their IC row). The dimensions to characterise edits serve the purpose of defining the qualities of the revisions, which operate on prefixes.

4.2.1 Edits

The main aspect to account for is whether labels need to be edited in the first place and, if yes, whether they are edited into the desired state. Edits on correct labels are *inconvenient*, and also *defective*, since the label will fatally change into a wrong label. This happens, for instance, at l_2 in step 4, as the correct label D is edited into a wrong B. Edits on incorrect labels are *convenient* and can be *effective* (if it enters into a correct state, like l_4 at $t = 5$, which changes from an incorrect D to a correct C) or *ineffective* (if it enters into another incorrect state, *e.g.* l_3 at $t = 9$, which changed from an incorrect A to a still incorrect D).

Other dimensions can be used to analyse the behaviour of the processor. *Innovative* edits cause the label to change into a new state. For instance, l_6 becomes a C for the first time at $t = 9$. In the next step, it is edited back into its previous state A, and we consider it to be a *repetitive* edit.

Local *recurrence* refers to whether the edit occurs in isolation in neighbouring time steps (edit subsequences in an IC's column). *Oscillation* refers to how many edits occur in its complete column, just one (*stable*) or more (*oscillating*). For instance, l_3 has two groups of recurrent edits along the time axis, whereas l_2 has one *steady* and stable edit.

Company characterises whether the edit occurs with other edits in a prefix (same IC's row). In Figure 1, l_6 is edited together with other labels at $t = 9$, whereas l_1 is edited in isolation at $t = 6$. *Ac-*

[9]More examples are available in the code repository.

companied edits can be either *connected* (*i.e.* with directly neighbouring edited labels, as in $t = 4$) or *disconnected* to the other edits in its prefix.

Short or *long range* refers to how far the edited label is from the current time step, defined by a distance parameter d. If we set $d = 2$, the edit that changes l_4 into a C at $t = 5$ is short range because it is less than 2 time steps away from the current token being processed. On the other hand, l_3 is edited at $t = 9$, very distant from the right frontier.

Edits can also be *definite* or *temporary*. Definite edits make the label enter into its final state, like l_2 at $t = 4$. Temporary edits are those like the B for l_3 at $t = 5$: It still gets edited further before a final decision is reached (here, also a B). Besides, edits can occur in *intermediate* steps during processing, when the input sequence is incomplete, or at the *final* time step, when the full sequence is available.

4.2.2 Revisions

Similar to edits, revisions are *inconvenient* if they occur on correct prefixes (that should not change), and thus also *defective*, because correctness necessarily decreases. The prefix at $t = 3$ is correct, so the revision at $t = 4$ causes the labels to become wrong. *Convenient* revisions are *effective* if they improve correctness, like at $t = 6$ where the number of correct labels in the prefix increases from 3 to 4, otherwise they can be *ineffective* (edits occur but correctness remains the same, like at $t = 9$) or again *defective*.

Revisions are *locally recurrent* when other revisions occur in neighbouring time steps. We see that from $t = 4$ to $t = 6$. The revision at $t = 2$ is *steady*, as no other revisions occur immediately before or after it. If only one revision occur while a sequence is processed, it is *stable*, otherwise it is *oscillating*. In our example, all revisions are therefore oscillating.

Company, *connectedness* and *distance* refer to what types of edits the revision causes. At the second time step, the prefix contains only a *disconnected* and *short range* edit, whereas at the fifth time step we observe *accompanied* edits, one *connected* and one *disconnected* group and one short and two long range edits.

Definite revisions create prefixes that will not be further edited. In our example, this only happens in the last time step; all others are *temporary*. *Intermediate* revisions happen when the input is not yet completed, otherwise they are *final*.

4.2.3 Recomputations

In models that detach recomputations from revisions, the recomputations should also be evaluated. Recomputations are *active* if they actually result in a revision, otherwise they are *inactive*. The quality of the resulting revisions can then be evaluated with the characteristics above.

4.3 Policies

To perform good revisions, a model must decide *when* to recompute or revise. For that decision, both a *revision policy* and a *recomputation policy* can be generally defined as:

$$\pi : \text{IC} \rightarrow [0, 1] \qquad \pi(\text{IC}_t) = \Pr(r|\text{IC}_t) \qquad (1)$$

It gives the probability of performing a revision or recomputation r, respectively, given the state of the incremental chart at time t.[10] When $\Pr(r|\text{IC}_t) > \tau$, where τ is a threshold hyperparameter, a revision/recomputation is performed. If the revisions are not a mere consequence of full recomputations, the model must then also decide *what* and *how* to edit.

4.4 Metrics

Traditional sequence labelling evaluation metrics like accuracy or F1 can be computed on label, sequence or dataset level. The incremental dimension requires its own metrics, some of which we discussed in §3. Here, we propose specific metrics to evaluate revision and/or recomputation policies. For each time step t in a sequence, either a revision (R) occurred, which is sometimes effective (R_e), or only an addition (A). Assuming we have established a metric for prefix correctness,[11] we know whether the prefix at $t - 1$ was correct (C) or incorrect (I). That results in a distribution of N actions in $\{R, A\} \times \{C, I\}$. From these counts, we derive the metrics in Table 4, computed either per sequence or over the whole dataset. Models that have the option to *recompute* (R') can also be evaluated in $\{R', \neg R'\} \times \{C, I\}$ with two additional metrics.

Since only *effective* revisions are actually desired, the R in the numerators can be replaced by R_e for a more focused evaluation. Revisions can

[10] It is also possible to make the policy dependent only in a portion of the IC, as done *e.g.* by Kahardipraja et al. (2023).

[11] A binary variable or a continuous variable, like accuracy, with a defined threshold for tolerated incorrectness.

		The fraction of...
Rate of Revision	R/N	time steps in which the model revises
Rate of Recomputation	R'/N	time steps in which the model recomputes
Rate of Active Recomputation	$(R' \cap R)/R'$	recomputations that actually causes a revision
R-Pertinence	$(R \cap I)/R$	revisions that edit incorrect prefixes (adapted precision)
R-Appropriateness	$(R \cap I)/I$	incorrect prefixes that are revised (adapted recall)
A-Pertinence	$(A \cap C)/A$	additions upon correct prefixes (adapted precision)
A-Appropriateness	$(A \cap C)/C$	correct prefixes that are not revised (adapted recall)
R_e-Pertinence	$(R_e \cap I)/R$	revisions that effectively edit incorrect prefixes
R_e-Appropriateness	$(R_e \cap I)/I$	incorrect prefixes that are revised effectively

Table 4: Proposed metrics for evaluating recomputation and revision policies. N is the total number of time steps.

be further weighted by how often and how far in the sentence processing they happen. Similarly, edits can be assessed by their correction time and survival time (Baumann, 2013).

4.5 Ideal Processor

Let us now delineate the ideal behaviour of a revision policy for an incremental sequence labelling model. A utopian model would always output the correct label and thus never need to produce edits or revisions (Kahardipraja et al., 2023).[12] But due to the incremental nature of language processing, models should not be penalised for building hypotheses that are *locally valid*, as long as a revision is timely triggered. That is, however, complex to know in raw textual input where local ambiguities are not identified. Instead, we can characterise an outlook according to desirable principles and available resources. In scenarios with an infinite time budget, we can simply wait for the input to be complete. If computation budget can be afforded, restart-incrementality is a good fit. But the constraints are not always so loose.

An ideal revision policy should thus revise as rarely as possible for stability. If a prefix/label is correct, the policy should avoid revising it, whereas an incorrect prefix/label should be revised (maybe not immediately, but eventually). It should always trigger effective, convenient, and definite revisions, preferably in earlier time steps.[13] Recurrent or oscillating revisions cause more instability and should be avoided. Innovative edits are preferable (as long as they are effective), and short range is better to be combined with delay strategies. Connectedness is a relevant dimension for BIO labelling schemes: If, for instance, the beginning label is edited, ideally the middle labels should change simultaneously. Finally, accompanied edits can be further evaluated in their relation to each other and the linguistic input. A good recomputation policy should, additionally, always result in active revisions.

In terms of metrics, R-Pertinence and A-Appropriateness should be exactly 1, *i.e.* all revisions should occur upon incorrect prefixes and all correct prefixes should not be revised. A-Pertinence and R-Appropriateness should be as high as possible, but cannot be expected to be exactly 1 because it may take some time steps until the input that actually resolves the ambiguity or mistake is observed.

5 Architecture Profiling

We now apply our methodology to profile the revision policy behaviour of three models: The reference restart-incremental Transformer and the two TAPIR variations, which have a recomputation policy, proposed by Kahardipraja et al. (2023). We evaluate them on three sequence labelling tasks: Slot filling (Coucke et al., 2018), POS tagging (Silveira et al., 2014) and NER (Tjong Kim Sang and De Meulder, 2003), using the final output as gold standard.[14] Note that the same profiling can be applied to any model with the ability of performing revisions on any sequence labelling task.

Quantitative Assessment Table 5 shows that the recomputation policy implemented in TAPIR reduces the number of restarts to between 10% and 25% in comparison to the restart incremental ap-

[12] That is indeed the case for strictly monotonic models if we use their final output as gold standard.

[13] In the beginning, the absence of both right and left context makes prediction harder. Towards the end, the availability of more left context should lead to less, and better, revisions.

[14] Here we use only the buffer outputs to evaluate the resulting revisions on prefixes that would have been passed on to downstream processors. We do not consider the temporary outputs of the LSTM that the original model had access to when deciding to perform a recomputation. Please refer to the original paper for the details on non-incremental and incremental performance on these tasks.

	% recomputation			% active recomputation			% revision		
	NER	POS	Slot	NER	POS	Slot	NER	POS	Slot
Rest.Incremental-Transformer	100.00	100.00	100.00	7.77	19.29	21.23	7.77	19.29	21.23
TAPIR-LTReviser	13.77	24.52	20.34	20.23	39.55	39.44	2.78	9.69	8.02
TAPIR-TrfReviser	10.36	20.23	21.41	25.36	34.09	33.65	2.62	6.89	7.20

Table 5: Rate of (active) recomputations and of revisions for each model and task.

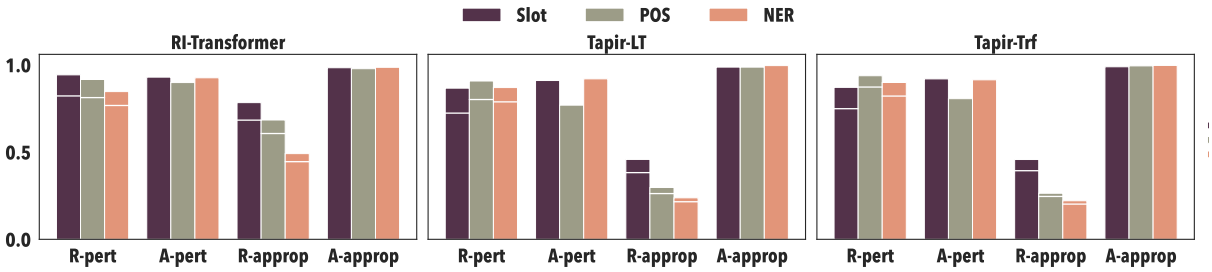


Figure 3: Revision metrics for all models and tasks. The white lines represent only the effective revisions.

proach, considerably alleviating the computation load; the number of revisions is also 2 to 3 times lower. Still, only up to 40% of the remaining recomputations are active, which means that the use of computational budget is still suboptimal. Furthermore, in Figure 3 we see that A-Appropriateness is very close to 1, as it should be. R-Pertinence is slightly below the ideal 1, but still greater than 0.8 in all cases, although it is around 0.1 lower when only effective revisions are considered. A-Pertinence is at similar values, with a lower result for POS-tagging. R-Appropriateness and R_e-appropriateness, however, are low in the restart-incremental Transformer and becomes even lower in the TAPIR models.

This may be evidence that the TAPIR models are waiting for more input before deciding to recompute an incorrect prefix, which is in line with the shifts in the distributions we observe in Figure 4. TAPIR tends to have more revisions towards the end of the sentence than the restart-incremental Transformer. This strategy can indeed help revisions be more effective, given that more left context is available, but it also results in having to wait longer for final decisions, which is not ideal.

The cumulative distributions of the fraction of time steps with revisions per sentence, shown in Figure 5, illustrate that the policy reduces the number of revisions per sentence: 50% or less of the sentences have no revisions in the naive policy, which makes all recomputation effort be used to perform only an addition, while TAPIR's policy caused more sentences to not trigger revisions.

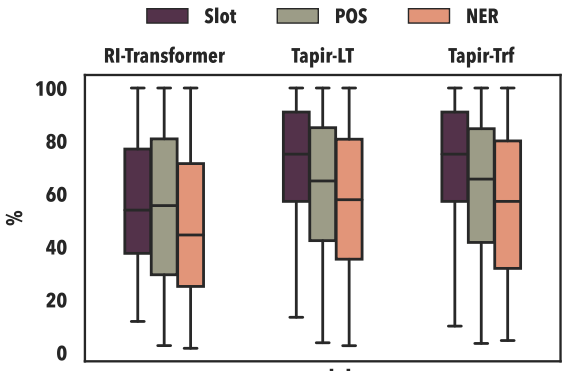


Figure 4: How far in the sentence processing (% of time steps or tokens) revisions occur.

Qualitative Assessment Figures 6 and 7 show the percentages of edits and revisions types to characterise TAPIR-TrfReviser's policy. In terms of edits, most are effective, convenient, innovative and steady. Only around 50% are short range, which means that delay strategies would have limited improvements in reducing edit overhead. For slot filling, around 20% of the edits occur in the last time

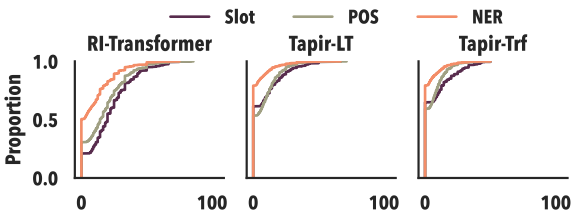


Figure 5: Proportion of time steps with revisions per sentence (cumulative).

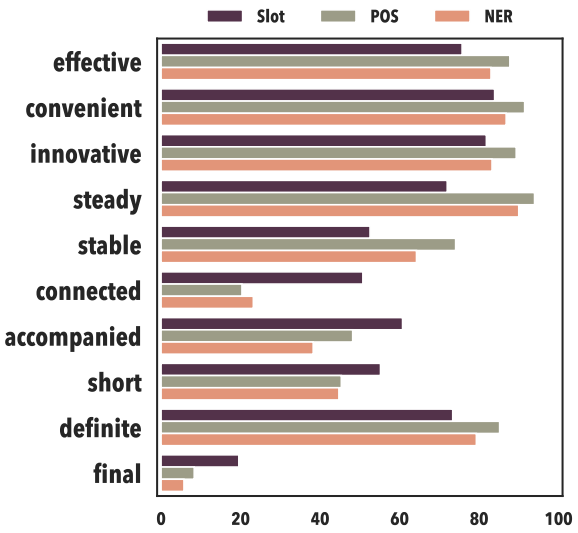


Figure 6: Edits by TAPIR-TrfReviser's policy.

	Slot	POS	NER
effective	74.9	87.3	82.1
defective	16.6	7.8	12.5
ineffective	8.5	4.9	5.3
convenient	87.1	93.8	89.9
inconvenient	12.9	6.2	10.1
steady	59.0	85.0	83.0
recurrent	41.0	15.0	17.0
oscillating	72.9	76.5	63.3
stable	27.1	23.5	36.7
isolated edit	61.8	72.2	78.3
accompanied edits	38.2	27.8	21.7
connected edits	28.8	8.9	11.6
disconnected edits	67.9	87.1	85.9
dis and connected edits	3.3	4.0	2.5
short range	50.7	45.0	43.4
long range	31.9	41.9	48.4
short and long range	17.5	13.0	8.3
temporary	46.3	51.9	41.6
definite	53.7	48.1	58.4
intermediate	80.1	91.5	94.3
final	19.9	8.5	5.7

Figure 7: Revisions by TAPIR-TrfReviser's policy.

step, which is undesired, because it means that the intermediate predictions for these labels are wrong until the model processes the full sentence.

Regarding revisions, TAPIR's policy works best for POS-tagging in terms of effectiveness, convenience, oscillation and recurrence, and worse for slot filling. Most of the edits are isolated, which means that recomputations have been performed for the full partial input to only result in one edit. The proportion of short vs. long range and temporary vs. definite revisions was, in general, balanced. We also see that proportionally fewer revisions occurred in the final step. Although the high percentage of intermediate revisions is high, Figure 4 shows that they are happening towards the end, which prevents incremental subprocessors to reliably count on the intermediate outputs. Slot filling is, here, an example of the occurrence of final revisions being less than ideal.

Based on these results, we conclude that TAPIR's policy is very successful in reducing the number of recomputations and also in revising less, but there is room for improving the quality of the resulting revisions, both in terms of metrics and of characteristics. This speaks for a more dedicated revision policy that could avoid full recomputations and use the state of the incremental chart and internal representations of the model for a more fine-grained prediction of which labels should change.

6 Conclusion

In this work, we have argued that the importance of a solid evaluation framework for revision policies in incremental sequence labelling cannot be overstated. Despite being very useful to capture some incremental aspects like instability or timeliness, existing evaluation metrics set aside other major strands of revisions. To fill that void, we have introduced metrics, characteristics and rationale to support the analysis of revision policies. This methodology serves as a tool to ascertain their quality, to determine their appropriateness in different contexts and to compare different policies.

We identify a few more roads to quality: The creation of incremental gold standards containing locally valid hypothesis, the development of fine-grained revision policies predicting what to revise and a more systematic integration of linguistic aspects of the input into the evaluation procedure. For those willing to drive those routes, we hope our methodology has paved the road well.

Acknowledgements

We thank the anonymous reviewers for their valuable comments and suggestions. We also thank Kaushal et al. (2023) for a conversation on this topic at EACL, in particular about the locally valid hypotheses.

References

Angus Addlesee, Yanchao Yu, and Arash Eshghi. 2020. A comprehensive evaluation of incremental speech recognition and diarization for conversational AI. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3492–3503, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tazin Afrin and Diane Litman. 2018. Annotation and classification of sentence-level revision improvement. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–246, New Orleans, Louisiana. Association for Computational Linguistics.

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5721–5729, Marseille, France. European Language Resources Association.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Giuseppe Attardi and Massimiliano Ciaramita. 2007. Tree revision learning for dependency parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 388–395, Rochester, New York. Association for Computational Linguistics.

Timo Baumann. 2013. *Incremental Spoken Dialogue Processing: Architecture and Lower-level Components*. Ph.D. thesis, Universität Bielefeld, Germany.

Timo Baumann, Michaela Atterer, and David Schlangen. 2009. Assessing and improving the performance of speech recognition for incremental systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 380–388, Boulder, Colorado. Association for Computational Linguistics.

Timo Baumann, Srinivas Bangalore, and Julia Hirschberg. 2014. Towards simultaneous interpreting: the timing of incremental machine translation and speech synthesis. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Papers*, pages 163–168, Lake Tahoe, California.

Timo Baumann, Okko Buß, and David Schlangen. 2011. Evaluation and Optimisation of Incremental Processors. *Dialogue and Discourse*, 2(1):113–141.

Niels Beuck, Arne Köhn, and Wolfgang Menzel. 2011a. Decision strategies for incremental POS tagging. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 26–33, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Niels Beuck, Arne Köhn, and Wolfgang Menzel. 2013. Predictive incremental parsing and its evaluation. In *Computational Dependency Theory*, pages 186–206. IOS Press.

Niels Beuck, Arne Köhn, and Wolfgang Menzel. 2011b. Incremental parsing and the evaluation of partial dependency analyses. In *Proceedings of the 1st International Conference on Dependency Linguistics*, pages 290–299.

Amit Bronner and Christof Monz. 2012. User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 356–366, Avignon, France. Association for Computational Linguistics.

Charles Callaway. 2003. Multilingual revision. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*, Budapest, Hungary. Association for Computational Linguistics.

Angelica Chen, Vicky Zayats, Daniel Walker, and Dirk Padfield. 2022. Teaching BERT to wait: Balancing accuracy and latency for streaming disfluency detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 827–838, Seattle, United States. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint*, arXiv:1805.10190.

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College composition and communication*, 32(4):400–414.

Atsushi Fujita, Kikuko Tanabe, Chiho Toyoshima, Mayuka Yamamoto, Kyo Kageura, and Anthony Hartley. 2017. Consistent classification of translation revisions: A case study of English-Japanese student translations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 57–66, Valencia, Spain. Association for Computational Linguistics.

Matt Grenander, Shay B. Cohen, and Mark Steedman. 2022. Sentence-incremental neural coreference resolution. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 427–443, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Matthew Honnibal and Mark Johnson. 2014. Joint incremental disfluency detection and dependency parsing. *Transactions of the Association for Computational Linguistics*, 2:131–142.

Julian Hough and David Schlangen. 2015. Recurrent Neural Networks for Incremental Disfluency Detection. In *Interspeech 2015*, pages 849–853.

Lianna Hrycyk, Alessandra Zarcone, and Luzian Hahn. 2021. Not so fast, classifier – accuracy and entropy reduction in incremental intent classification. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 52–67, Online. Association for Computational Linguistics.

Patrick Kahardipraja, Brielen Madureira, and David Schlangen. 2021. Towards incremental transformers: An empirical analysis of transformer models for incremental NLU. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1178–1189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Kahardipraja, Brielen Madureira, and David Schlangen. 2023. TAPIR: Learning adaptive revision for incremental natural language understanding with a two-pass model. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4173–4197, Toronto, Canada. Association for Computational Linguistics.

Ayush Kaushal, Aditya Gupta, Shyam Upadhyay, and Manaal Faruqui. 2023. Efficient encoders for streaming sequence tagging. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 418–429, Dubrovnik, Croatia. Association for Computational Linguistics.

Arne Köhn. 2018. Incremental natural language processing: Challenges, strategies, and evaluation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2990–3003, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Arne Köhn and Wolfgang Menzel. 2014. Incremental predictive parsing with TurboParser. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 803–808, Baltimore, Maryland. Association for Computational Linguistics.

Ioannis Konstas, Frank Keller, Vera Demberg, and Mirella Lapata. 2014. Incremental semantic role labeling with Tree Adjoining Grammar. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 301–312, Doha, Qatar. Association for Computational Linguistics.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020. SIMULEVAL: An evaluation toolkit for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online. Association for Computational Linguistics.

Brielen Madureira and David Schlangen. 2020. Incremental processing in the age of non-incremental encoders: An empirical assessment of bidirectional models for incremental NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 357–374, Online. Association for Computational Linguistics.

Ramesh Manuvinakurike, Maike Paetzel, Cheng Qu, David Schlangen, and David DeVault. 2016. Toward incremental dialogue act segmentation in fast-paced interactive dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 252–262, Los Angeles. Association for Computational Linguistics.

Tetsuji Nakagawa, Taku Kudo, and Yuji Matsumoto. 2002. Revision learning and its application to part-of-speech tagging. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 497–504, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Lucia Pozzan and John C Trueswell. 2015. Revise and resubmit: How real-time parsing limitations influence grammar acquisition. *Cognitive Psychology*, 80:73–108.

Dheeraj Rajagopal, Xuchao Zhang, Michael Gamon, Sujay Kumar Jauhar, Diyi Yang, and Eduard Hovy. 2022. One document, many revisions: A dataset for classification and description of edit intents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5517–5524, Marseille, France. European Language Resources Association.

Mrinal Rawat and Victor Barres. 2022. Real-time caller intent detection in human-human customer support spoken conversations. In *Communication in Human-AI Interaction Workshop*.

Jacques Robin. 1996. Evaluating the portability of revision rules for incremental summary generation. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 205–214, Santa Cruz, California, USA. Association for Computational Linguistics.

Morteza Rohanian and Julian Hough. 2021. Best of both worlds: Making high accuracy non-incremental transformer-based disfluency detection incremental. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3693–3703, Online. Association for Computational Linguistics.

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental reference resolution: The task, metrics for evaluation, and a Bayesian filtering model that is sensitive to disfluencies. In *Proceedings of the SIGDIAL 2009 Conference*, pages 30–37, London, UK. Association for Computational Linguistics.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue and Discourse*, 2(1):83–111.

Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams. 2011. Stability and accuracy in incremental speech recognition. In *Proceedings of the SIGDIAL 2011 Conference*, pages 110–119, Portland, Oregon. Association for Computational Linguistics.

Sukanta Sen, Rico Sennrich, Biao Zhang, and Barry Haddow. 2023. Self-training reduces flicker in retranslation-based simultaneous translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3734–3744, Dubrovnik, Croatia. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Dennis Ulmer, Dieuwke Hupkes, and Elia Bruni. 2019. Assessing incrementality in sequence-to-sequence models. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 209–217, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Marie M. Vaughan and David D. McDonald. 1986. A model of revision in natural language generation. In *24th Annual Meeting of the Association for Computational Linguistics*, pages 90–96, New York, New York, USA. Association for Computational Linguistics.

Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2020. Learning adaptive segmentation policy for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2280–2289, Online. Association for Computational Linguistics.

Baigong Zheng, Kaibo Liu, Renjie Zheng, Mingbo Ma, Hairong Liu, and Liang Huang. 2020a. Simultaneous translation policies: From fixed to adaptive. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2847–2853, Online. Association for Computational Linguistics.

Renjie Zheng, Mingbo Ma, Baigong Zheng, Kaibo Liu, and Liang Huang. 2020b. Opportunistic decoding with timely correction for simultaneous translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 437–442, Online. Association for Computational Linguistics.

Shlomo Zilberstein. 1996. Using anytime algorithms in intelligent systems. *AI magazine*, 17(3):73–73.

The effect of conversation type on entrainment: Evidence from laughter

Bogdan Ludusan and **Petra Wagner**
Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC,
Bielefeld University, Germany
{bogdan.ludusan, petra.wagner}@uni-bielefeld.de

Abstract

Entrainment is a phenomenon that occurs across several modalities and at different linguistic levels in conversation. Previous work has shown that its effects may be modulated by conversation extrinsic factors, such as the relation between the interlocutors or the speakers' traits. The current study investigates the role of conversation type on laughter entrainment. Employing dyadic interaction materials in German, containing two conversation types (free dialogues and task-based interactions), we analyzed three measures of entrainment previously proposed in the literature. The results show that the entrainment effects depend on the type of conversation, with two of the investigated measures being affected by this factor. These findings represent further evidence towards the role of situational aspects as a mediating factor in conversation.

1 Introduction

An aspect frequently observed in conversation is the fact that interlocutors become more similar to each other during their interaction, a phenomenon called, among other terms, entrainment. It has been seen to occur for different linguistic levels (e.g., syntactic Branigan et al., 2000, lexical Brennan and Clark, 1996; Nenkova et al., 2008, acoustic Pardo, 2006; Levitan et al., 2015), but also with respect to non-verbal behaviour (Edlund et al., 2009). Moreover, entrainment effects can be seen both on the form level (adopting the same structures), and on the temporal level, through an increase in temporal co-ordination between interlocutors.

Different points of view on the mechanisms behind entrainment exist, with some viewing it as an automatic process (Pickering and Garrod, 2004), while others arguing that the occurrence of entrainment depends on social factors (Pardo, 2012). This latter viewpoint seems to be supported by studies finding that various conversation aspects (e.g., the role of the interlocutors in the conversation Beňuš et al., 2014; Reichel et al., 2018, their relation Menshikova et al., 2021) or individual factors (e.g., speaker traits Lewandowski and Jilka, 2019, native language Kim et al., 2011) may modulate or interact with entrainment.

Laughter is one of the most often encountered non-verbal vocalisations in spoken interaction (Trouvain and Truong, 2012a), having a wide range of roles in communication, including social (Glenn, 2003) and linguistic (Mazzocconi et al., 2020; Ludusan and Schuppler, 2022). Laughter has been found to be subject to entrainment effects. Interlocutors become more similar in their acoustic realization of laughter, as well as in the timing of their laughter productions (Trouvain and Truong, 2012b; Ludusan and Wagner, 2019). Laughter production may be affected by external factors, such as the gender of the speaker or the familiarity of the interlocutors (Smoski and Bachorowski, 2003). However, no evidence exists towards these factors modulating the amount of entrainment in laughter, with previous works investigating these aspects finding no effect of familiarity on entrainment measures (Trouvain and Truong, 2012b; Ludusan and Wagner, 2022).

We investigate here the effect of one conversation factor, namely the conversation type, on entrainment. We define by conversation type the nature of the interaction, considering it to be either task-based, in which the conversation partners have a specific task to solve during their interaction, or free dialogue, in which interlocutors chat freely about topics of their choice. In particular, we evaluate the role of conversation type (free dialogue vs. two different types of task-based dialogues) on three measures of laughter entrainment.

2 Materials

Materials from two corpora, the GRASS corpus (Schuppler et al., 2014) and the DUEL corpus (Hough et al., 2016) were used for the experiments.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 168–174
September 11–15, 2023. ©2023 Association for Computational Linguistics

Type	Class	Corpus	Duration [min]	#Dyads	Gender			Age	#Laughter events
					f-f	f-m	m-m		
free	GR	GRASS	769	13	4	4	5	30.5	2272
task	DA	DUEL	103	7	4	2	1	22.7	442
task	FS	DUEL	104	8	2	5	1	23.1	737

Table 1: Information on the data used in this analysis: conversation type (free dialogue or task-based), conversation class (DA/FS/GR), the source corpus (DUEL/GRASS), total duration, number of dyads included, gender composition of the dyads (f-f, f-m, or m-m), average age of the speakers, and number of produced laughter events.

The GRASS corpus (GR) contains both read materials and conversations between two persons. We employed here the latter subset of the corpus, in which the interlocutors (19 dyads), native speakers of Austrian German, were recorded chatting for one hour straight. The interlocutors knew each other beforehand, being either colleagues, friends, family members or couples. They were asked to chat about whichever subject(s) they desired, with some pairs simply continuing the discussion they had before the recording started. This resulted in spontaneous conversations including a wide variety of topics, such as about vacations, local issues, work, family or relationship problems and public figures. The materials were orthographically transcribed and annotated for conversational phenomena, including laughter (both laughs and speech-laughs).

The second corpus, DUEL, contains dyadic interactions between native speakers of three languages: French, German and Mandarin Chinese. Two different scenarios from the German part of the corpus were employed here: Dream Apartment (DA) and Film Script (FS). For the DA scenario, the interlocutors were told they had a large sum of money to design and furnish an apartment they would have to share. In the FS task, they were supposed to come up with the script for a film, based on an embarrassing moment, which could have been inspired from personal experience. The considered materials were recorded by 10 dyads/scenario (which differed between the two scenarios). The dyads were all students, the majority of them being colleagues/friends, but also some pairs consisting of strangers. The corpus was orthographically transcribed and annotated for laughter and other conversational phenomena.

In order to control for the effect the relation between interlocutors might have on entrainment, we did not consider in our analysis the recordings from the GRASS corpus that involved family members or couples (6 dyads). Similarly, we excluded those between strangers from the DUEL corpus (4 dyads). In this way, the dyads from both corpora were either colleagues or friends. Detailed information on the datasets considered in the analyses and their characteristics can be found in Table 1.

3 Methods

We investigated three measures previously employed in the study of laughter entrainment, all of which were computed at the dyad level. They included both temporal-related entrainment measures such as the amount of overlapping laughter produced by the interlocutors and the synchrony of the produced laughter, and form-related ones, namely the difference in maximum intensity between non-consecutive and consecutive laughter produced by the speakers in the dyad. We examined whether the results of these measures varied with the conversation type (free vs. task-based dialogue), while also considering a second analysis level, the conversation class (examining here three classes: GR, DA, FS).

The first measure, the amount of overlapping laughter, was inspired by the temporal alignment proposed by Trouvain and Truong (2012b) as a measure of laughter entrainment. A higher amount of overlapping laughter implies a higher level of entrainment. The measure was determined by counting all events in which the two interlocutors were laughing at the same time (we took into account any amount of overlap), as well as the total number of laughter events produced during the interaction. We then applied logistic regression models to test the differences between the various conditions (conversation type/class), by considering the odds of overlapping laughter, represented by the pair (overlapping laughter counts, total laughter counts - overlapping laughter counts) as dependent variable of the model and the condition as predictor.

For the synchrony measure, we applied the process described in Ludusan and Wagner (2019).

However, since we had recordings of different lengths within and across datasets, we did not split the recordings into a fixed number of bins. Instead, we used bins of equal duration – 90 seconds (15 minutes / 10 bins, as in Ludusan and Wagner 2019). We then counted the number of laughter events produced by each speaker in each bin and computed the synchrony, defined as the Spearman ρ correlation coefficient between the vectors composed of the binned laughter counts of the interlocutors in a conversation. Positive values of this measure represent entrainment. These first two measures were computed on the data from all 28 dyads included in the study.

The form-related measure characterizes the similarity of consecutive laughter pairs produced by the interlocutors in terms of maximum speech signal intensity (Ludusan and Wagner, 2022). The intensity was computed by means of the Praat software (Boersma and Weenink, 2020), employing a minimum pitch of 75 Hz and subtracting the microphone DC offset. The maximum value over each laughter event was then considered for this entrainment measure. Consecutive laughter pairs are composed of the laughter event of a speaker either overlapping with or followed within one second, by a laughter produced by their interlocutor (similar to the definition of antiphonal laughter in Smoski and Bachorowski 2003). We then compared the difference in intensity between the laughter events of a consecutive pair ($intD_C$) with the same measure computed between the events of non-consecutive laughter pairs ($intD_N$). Non-consecutive pairs were composed of a laughter event from a consecutive laughter pair, and a randomly sampled laughter produced by the interlocutor, except for the one in the same consecutive pair (see Ludusan and Wagner 2022 for more details). The measure was then defined as: $intD_N$ - $intD_C$, with positive values denoting entrainment. This measure was analyzed for 27 dyads, those which produced at least 5 consecutive laughter pairs (one all-male dyad from the GRASS subset was removed).

In addition to comparing these three measures across conversation types, we also determined whether the obtained values represent entrainment or not. For the intensity-based measure, a positive value significantly different from 0 denotes entrainment, and the opposite effect for negative values. For the overlapping laughter and the synchrony measures, we determined whether the dyads achieved entrainment, by comparing their value with those obtained for all pseudo-dyads, similarly to previous work on entrainment (e.g, Ramseyer and Tschacher, 2010). For each dyad in the investigated subset, we created pseudo-dyads, by putting together the speech of each speaker within the dyad with all other speakers in that subset, but the one from the same dyad. For each created pseudo-dyad, the two entrainment measures were computed and the average value across all pseudo-dyads was compared to the entrainment measure of the actual dyad. If the latter was significantly higher than the former, it represented entrainment, while a significantly lower value meant disentrainment.

Finally, there are characteristics which we could not control for in the analyzed data and which may influence laughter production and possibly, indirectly, its entrainment. Therefore, we examined any effect that dyad gender composition (two classifications: f-f/f-m/m-m or same/mixed-gender) or age (two measures: absolute age difference or average age of the dyad) may have on the entrainment measure.

For all analyses except for the ones pertaining to the overlapping laughter measure (which employed logistic regression), linear regression models were fitted with the respective measure values as dependent variable and the various factors investigated as predictors. In case the residuals of the fitted models were found to be not normally distributed (by means of a Shapiro-Wilk test), we applied a corresponding non-parametric method: either a Wilcoxon rank-sum test (for two groups), or a Kruskal-Wallis test (for three groups). To determine whether the studied measures show entrainment on each subset we compared them (either with the 0 level or with the value obtained for the pseudo-dyads) by means of t-tests or Wilcoxon tests (if the samples were not normally distributed). All statistical analyses were run using the appropriate functions of the R software (R Core Team, 2020).

4 Results

The values of the three investigated measures across the considered conversation types and classes are illustrated in Figure 1 and Figure 2, respectively.

In terms of percentage of overlapping laughter between the interlocutors, both conversation types showed entrainment (Figure 1, left panel),

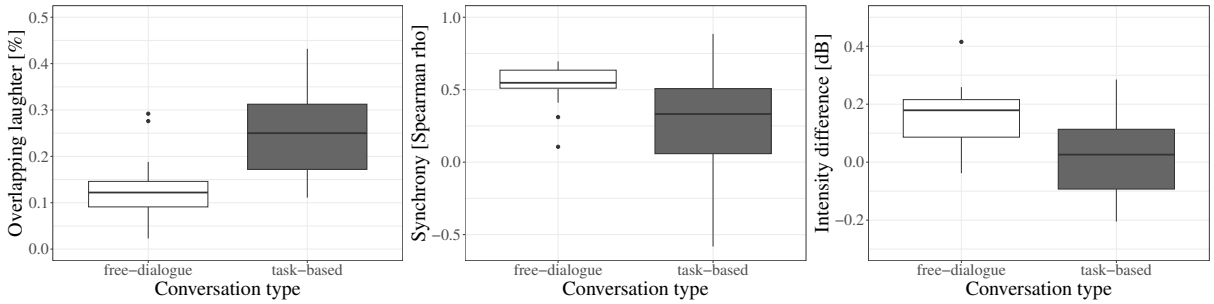


Figure 1: The results of the investigated entrainment measures, with respect to the considered conversation types: overlapping laughter (left panel), synchrony (middle panel) and form-related measure (right panel). The horizontal line represents the median value, the hinges of the boxes the first and third quartiles, and the whiskers going up to 1.5*IQR (inter-quartile range) from the hinges.

as revealed by Wilcoxon tests ($p = 2.4e^{-4}$ for free dialogues and $p = 6.1e^{-5}$ for task-based dialogues). We then investigated the effect of conversation type on entrainment, by using it as predictor in a logistic regression model (AIC = 252.3). The difference between the two types was found to be significant ($\beta = 0.740, z = 8.26, p < 2e^{-16}$). When looking at conversation classes (Figure 2, left panel), entrainment was observed for GR and for both classes included in the task-based data: DA ($t = 8.26, p = 1.7e^{-4}$) and FS ($t = 5.70, p = 7.3e^{-4}$). The ANOVA analysis of the logistic model fitted with the overlapping laughter odds as dependent variable and the class as independent variable (Akaike Information Criterion, AIC = 254.2), revealed a significant effect of class ($\chi^2 = 67.3, p = 2.4e^{-15}$). Moreover, the model showed that the differences between GR and each of the other two classes were significant: DA ($\beta = 0.752, z = 6.07, p = 1.3e^{-9}$) and FS ($\beta = 0.733, z = 7.08, p = 1.5e^{-12}$). No significant difference was found between the DA and FS. Lastly, we verified, by means of logistic regression, whether the age (mean or difference) of the conversation partners or the dyad composition (exact composition or same/mixed) may play a role in the production of overlapping laughter. All but the age difference showed a significant effect, although the fit of these models was worse than that of the models employing the conversation class or type as predictor (the best of these four models had an AIC of 296.5 – lower AIC represents a better model).

For the synchrony measure, we observed entrainment for both free and task-based dialogues (Figure 1, middle panel): $t = 9.32, p = 7.6e^{-7}$ and $t = 3.28, p = 0.005$, respectively. The difference between conversation types was not significant, as given by a Wilcoxon rank sum test ($p = 0.339$). At the level of conversation classes (Figure 2, middle panel), entrainment effects were observed only for FS ($t = 2.99, p = 0.020$), in addition to GR. A Kruskal-Wallis test showed no significant overall difference between conversation classes ($\chi^2 = 2.33, p = 0.312$), but pairwise differences were found between GR and DA, using a Wilcoxon test ($p = 0.024$). Additional Kruskal-Wallis tests revealed no significant effects of age or dyad gender composition.

The last measure, defined as the difference in maximum intensity between non-consecutive and consecutive laughter pairs (Figure 1, right panel), was found to entrain for free dialogues ($t = 4.92, p = 4.6e^{-4}$), but not for the task-based ones ($t = 0.44, p = 0.67$). A significant difference was observed between conversation types, as given by the ANOVA of the fitted linear model ($F = 7.96, p = 0.009$). A similar linear regression model, using the intensity difference as dependent variable and the conversation class as predictor was then fitted (Figure 2, right panel). The ANOVA analysis of this model revealed a significant overall effect of class ($F = 5.50, p = 0.011$), with the difference between GR and DA reaching significance ($\beta = -0.207, z = -3.30, p = 0.003$). None of the subsequent linear models, fitted with the gender make-up of the dyad and the age measures as predictors, showed a significant effect of these factors.

5 Discussion and conclusions

Our findings paint a complex relationship between the investigated entrainment measures and the different conversation types/classes considered here. We found entrainment across the various dialogues types/classes, and differences between types and some classes (overlapping laughter), entrainment

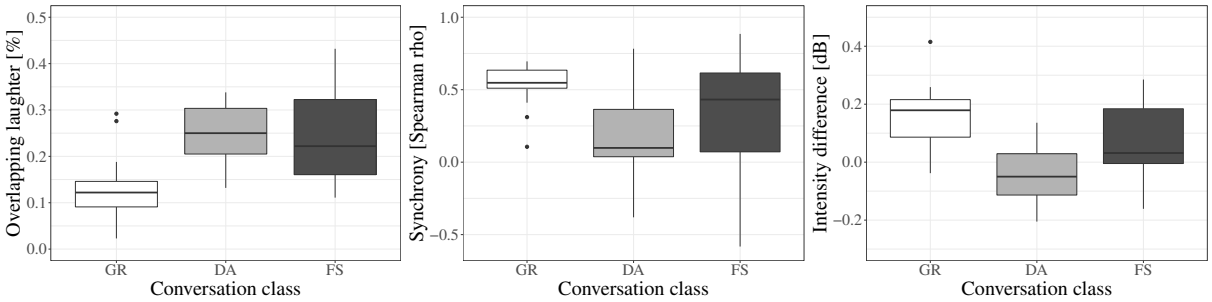


Figure 2: The results of the investigated entrainment measures, with respect to the considered conversation classes: overlapping laughter (left panel), synchrony (middle panel) and form-related measure (right panel). The horizontal line represents the median value, the hinges of the boxes the first and third quartiles, and the whiskers going up to 1.5*IQR (inter-quartile range) from the hinges.

across types, but not for all classes, and some differences between classes (synchrony), entrainment for one type only and differences between some classes (intensity measure). An effect of conversation type/class was observed when controlling for the relation between interlocutors, while other dimensions of variability between the different subsets used (age of interlocutors, gender composition of the dyad) had either no significant effect, or explained the differences in entrainment worse than the conversation type/class.

Another factor of variability may be the fact that the interlocutors in the analyzed corpora spoke different varieties of German and came from slightly different cultures. Yet, evidence from studies that examined laughter entrainment measures cross-linguistically/culturally (Ludusan and Wagner, 2019, 2022), showed no language/culture differences for more distant language pairs (German-Chinese and French-Chinese) than the ones here. One could assume, instead, that the observed differences stem from the fact that task-based interactions require a higher cognitive load, and previous studies have shown that a higher cognitive load may impede entrainment (Abel and Babel, 2017). However, our results did not show an inverse relation between the level of entrainment and the difficulty of the task. Some of the values of the studied measures revealed either the opposite tendency or similar trends between task-based and free dialogue interactions. These findings indicate that what is being captured by our conversation type factor differs from cognitive load.

The results obtained for the overlapping laughter measure, with the free dialogue/GR values being significantly lower than for the other cases, may seem surprising, especially considering that synchrony, another measure of temporal alignment, suggests rather the opposite. It might be that the overlap measure employed here is too strict. Since mirthful laughter, which is predominant in the FS data and partly in the DA recordings, is generally longer than social laughter, it is more likely that, when the conversation partner joins in laughing in response to a mirthful laughter, their laughter will overlap that of their interlocutor. A more appropriate measure could be one that takes into account also the interval immediately following the produced laughter, such as the antiphonal laughter definition of Smoski and Bachorowski (2003).

To conclude, our findings represent further evidence for entrainment not being a fully automatic process (Pardo, 2012), but that different factors (here, the conversation type) may influence it and should be taken into account when investigating this phenomenon. As future work, on the one hand, we would like to tease apart the effect of conversation type on entrainment from that potentially brought by laughter type, since the employed dialogues contain different types of laughter. One the other hand, our results raise further questions about the potential effect of conversation type on the entrainment of other levels. Thus, extending this investigation to conversation elements/linguistic levels previously shown to be subject to entrainment is highly desirable. This will shed further light on the role of entrainment in human communication and will also allow more realistic implementations of this phenomenon in spoken dialogue systems (e.g., Stoyanchev and Stent, 2009; Duplessis et al., 2017).

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 461442180.

References

Jennifer Abel and Molly Babel. 2017. Cognitive load reduces perceived linguistic convergence between dyads. *Language and Speech*, 60(3):479–502.

Štefan Beňuš, Agustín Gravano, Rivka Levitan, Sarah Ita Levitan, Laura Willson, and Julia Hirschberg. 2014. Entrainment, dominance and alliance in supreme court hearings. *Knowledge-Based Systems*, 71:3–14.

Paul Boersma and David Weenink. 2020. Praat: doing phonetics by computer [Computer program]. Version 6.1.35, retrieved 1 December 2020 from http://www.praat.org/.

Holly P Branigan, Martin J Pickering, and Alexandra A Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Guillaume Dubuisson Duplessis, Chloé Clavel, and Frédéric Landragin. 2017. Automatic measures to characterise verbal alignment in human-agent interaction. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 71–81.

Jens Edlund, Julia Hirschberg, and Mattias Heldner. 2009. Pause and gap length in face-to-face interaction. In *Proceedings of INTERSPEECH*, pages 2779–2782.

Phillip Glenn. 2003. Towards a social interactional approach to laughter. In *Laughter in Interaction*, pages 7–34. Cambridge University Press.

Julian Hough, Ye Tian, Laura de Ruiter, Simon Betz, Spyros Kousidis, David Schlangen, and Jonathan Ginzburg. 2016. DUEL: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *Proceedings of LREC*, pages 1784–1788.

Midam Kim, William S. Horton, and Ann R. Bradlow. 2011. Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology*, 2(1):125–156.

Rivka Levitan, Štefan Beňuš, Agustín Gravano, and Julia Hirschberg. 2015. Acoustic-prosodic entrainment in slovak, spanish, english and chinese: A cross-linguistic comparison. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 325–334.

Natalie Lewandowski and Matthias Jilka. 2019. Phonetic convergence, language talent, personality and attention. *Frontiers in Communication*, 4:18.

Bogdan Ludusan and Barbara Schuppler. 2022. To laugh or not to laugh? the use of laughter to mark discourse structure. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 76–82.

Bogdan Ludusan and Petra Wagner. 2019. Laughter dynamics in dyadic conversations. In *Proceedings of INTERSPEECH*, pages 524–528.

Bogdan Ludusan and Petra Wagner. 2022. Laughter entrainment in dyadic interactions: Temporal distribution and form. *Speech Communication*, 136:42–52.

Chiara Mazzocconi, Ye Tian, and Jonathan Ginzburg. 2020. What's your laughter doing there? A taxonomy of the pragmatic functions of laughter. *IEEE Transactions on Affective Computing*, 13(3):1302–1321.

Alla Menshikova, Daniil Kocharov, and Tatiana Kachkovskaia. 2021. Lexical Entrainment and Intra-Speaker Variability in Cooperative Dialogues. In *Proceedings of INTERSPEECH*, pages 1957–1961.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of ACL-08: HLT, Short Papers*, pages 169–172.

Jennifer Pardo. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, 119(4):2382–2393.

Jennifer Pardo. 2012. Reflections on phonetic convergence: Speech perception does not mirror speech production. *Language and Linguistics Compass*, 6(12):753–767.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Fabian Ramseyer and Wolfgang Tschacher. 2010. Nonverbal synchrony or random coincidence? How to tell the difference. In Anna Esposito, Nick Campbell, Carl Vogel, Amir Hussain, and Anton Nijholt, editors, *Development of Multimodal Interfaces: Active Listening and Synchrony. Lecture Notes in Computer Science, vol 5967*, pages 182–196. Springer.

Uwe Reichel, Stefan Beňuš, and Katalin Mády. 2018. Entrainment profiles: Comparison by gender, role, and feature set. *Speech Communication*, 100:46–57.

Barbara Schuppler, Martin Hagmüller, Juan Andres Morales-Cordovilla, and Hannes Pessentheiner. 2014. GRASS: the Graz corpus of Read And Spontaneous Speech. In *Proceedings of LREC*, pages 1465–1470.

Moria Smoski and Jo-Anne Bachorowski. 2003. Antiphonal laughter between friends and strangers. *Cognition and Emotion*, 17(2):327–340.

Svetlana Stoyanchev and Amanda Stent. 2009. Lexical and syntactic adaptation and their impact in deployed spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual*

Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pages 189–192.

Jürgen Trouvain and Khiet Truong. 2012a. Comparing non-verbal vocalisations in conversational speech corpora. In *Proceedings of the LREC Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, pages 36–39.

Jürgen Trouvain and Khiet P Truong. 2012b. Convergence of laughter in conversational speech: effects of quantity, temporal alignment and imitation. In *Proceedings of the International Symposium on Imitation and Convergence in Speech*, pages 37–38.

'What are you referring to?' Evaluating the Ability of Multi-Modal Dialogue Models to Process Clarificational Exchanges

Javier Chiyah-Garcia* **Alessandro Suglia*†** **Arash Eshghi*†** **Helen Hastie***
*Heriot-Watt University, Edinburgh, United Kingdom
†AlanaAI, Edinburgh, United Kingdom
{fjc3,a.suglia,a.eshghi,h.hastie}@hw.ac.uk

Abstract

Referential ambiguities arise in dialogue when a referring expression does not uniquely identify the intended referent for the addressee. Addressees usually detect such ambiguities immediately and work with the speaker to *repair* it using meta-communicative, Clarificational Exchanges (CE[1]): a *Clarification Request* (CR) and a response. Here, we argue that the ability to generate and respond to CRs imposes specific constraints on the architecture and objective functions of multi-modal, visually grounded dialogue models. We use the SIMMC 2.0 dataset to evaluate the ability of different state-of-the-art model architectures to process CEs, with a metric that probes the contextual updates that arise from them in the model. We find that language-based models are able to encode simple multi-modal semantic information and process some CEs, excelling with those related to the dialogue history, whilst multi-modal models can use additional learning objectives to obtain disentangled object representations, which become crucial to handle complex referential ambiguities across modalities overall[2].

1 Introduction

In dialogue, people work together on a moment by moment basis to achieve shared understanding and coordination (Clark, 1996; Clark and Brennan, 1991; Goodwin, 1981; Healey et al., 2018; Mills, 2007). A key mechanism people use to repair misunderstandings when they occur is via meta-communicative, clarificational exchanges (CE): a clarification request (CR) followed by a response (see Fig. 1). CRs are a highly complex phenomenon: they are multi-modal (Benotti and Blackburn, 2021), highly context-dependent with different forms and interpretations (Purver, 2004; Purver and Ginzburg, 2004), and can occur at different levels of communication on Clark's (1996) joint action ladder (Schlangen, 2004; Benotti and Blackburn, 2021). But while the crucial role of generating and responding to CRs in dialogue systems has long been recognised (San-Segundo et al., 2001; Rieser and Moore, 2005; Rodríguez and Schlangen, 2004; Rieser and Lemon, 2006), CRs still remain an understudied phenomenon (Benotti and Blackburn, 2021), especially in the context of recent successes in multi-modal dialogue modelling (Suglia et al., 2021; Wang et al., 2020; Chen et al., 2020; Guo et al., 2022; Das et al., 2017; Chen et al., 2021; Agarwal et al., 2020). There is recent work related to identifying when to pose a CR (Madureira and Schlangen, 2023; Zhu et al., 2021; Shi et al., 2022), but few evaluate the ability of models to process their responses (Gervits et al., 2021; Aliannejadi et al., 2021).

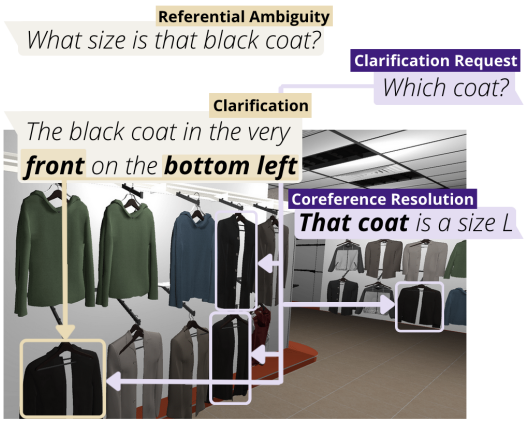


Figure 1: Example referential ambiguity and clarification in SIMMC 2.0 dialogues.

In this paper, we use CRs as a testbed for studying and evaluating different neural dialogue model architectures (see also Madureira and Schlangen (2023)). We focus on *referential CRs* occurring at level three of Clark's (1996) action ladder: that of *understanding*. We provide a framework for

[1]Not to be confused with, but related to Clarification Ellipsis as used in e.g. Fernández and Ginzburg (2002)

[2]The source code and evaluation experiments are available at https://github.com/JChiyah/what-are-you-referring-to

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 175–182
September 11–15, 2023. ©2023 Association for Computational Linguistics

evaluating how well multi-modal dialogue models are able to exploit referential CEs to resolve ambiguous referential descriptions. We use this framework to probe several state-of-the-art models proposed for the SIMMC 2.0 Challenge (Kottur et al., 2021) trained to resolve situated multi-modal coreferences with CEs found in the SIMMC 2.0 dataset itself.

The results indicate that the ability of a model to exploit CRs to resolve referential ambiguities depends on the level of granularity of the model's cross-modal representations, i.e. how well information about different object attributes is represented. In particular, we find that the model that includes a training objective designed for predicting object attributes in a multi-task setup performs significantly better than the rest which was not optimised with this objective. This is in line with findings in Suglia et al. (2020) who show that having disentangled object representations (Bengio et al., 2013) allows models to better partition the search space of potential referents; and thereby better exploit effective object attributes in disambiguation.

2 Dataset

We used the SIMMC 2.0 dataset (Kottur et al., 2021), which is a collection of multi-modal task-oriented dialogues, where both the system and the agent are situated in the same virtual environment. The dataset dialogues have a high degree of ambiguity and use rich referring expressions due to the overlap of many similar-looking objects (e.g., 5 red t-shirts in view); dialogues with references to multiple and previously discussed objects (mean 4.5 unique objects referenced per dialogue, SD: 2.4); and changing points of view throughout dialogues with partially observed objects. Thus, referential ambiguities in both the visual and conversational contexts are common. Furthermore, other common datasets do not contain coordination phenomena exhibited in SIMMC 2.0 (i.e. GuessWhat?! (de Vries et al., 2017)) or have a mixture of CRs which focuses solely on multi-modal referential ambiguities (e.g., Photobook (Haber et al., 2019)).

2.1 Dataset Details

In the SIMMC 2.0 dataset (Kottur et al., 2021), the agent acts as the shopping assistant to a user in a virtual shop. It encompasses the domains of fashion and furniture over 11,244 dialogues and it was collected using a mix of dialogue self-play and crowd-sourcing. The dataset is originally split into `train/dev/devtest/test-std` with 65% / 5% / 15% / 15% of the dialogues respectively.

Each dialogue is complemented by images of the environment scene and rich metadata annotations. Some dialogues have multiple scene images with partially overlapping sets of objects, requiring models to carry over information from previous viewpoints. On average, dialogues have 5.2 utterance pairs (*user-assistant*) and associated scenes have a mean of 27.6 objects, with some of them reaching up to a maximum of 141 items. Table 1 shows a dialogue from the dataset, refer to Appendix B for further samples.

	USR	*Hello, do you have any jackets for me to look at?*
	SYS	*Sure, what do you think of the light grey jacket hanging up high at the back left?*
	USR	*Do you have anything with a similar size range to the black sweater beside the light grey jacket?*
	SYS	*Sorry, I don't have anything similar to that*
Before-CR	USR	*What size is that sweater anyways?* [Referential Ambiguity]
CR	SYS	*The black one?* [Clarification Request]
After-CR	USR	*Yes exactly* [Clarification]

- Tags in CE: Individual Property

Table 1: Sample dialogue with a CE from the SIMMC 2.0 dataset.

Since the gold data from the `test-std` split is not available, we used the `devtest` data for our evaluation. Thus, some of the model object F1 scores may differ from their respective papers by a few decimals.

2.2 CRs in SIMMC 2.0

We focus on the clarificational sub-dialogues from the SIMMC 2.0 dataset. During the challenge, the dataset authors proposed several tasks, two of which are relevant here: Multi-modal Disambiguation (detecting whether the system has enough information to identify a unique object or is ambiguous) and Multi-modal Coreference Resolution (find the objects mentioned by the user). The dataset provides annotations that mark whether a turn is ambiguous or not, and which objects are referred to. Models were implicitly required to handle them as

part of longer conversations, although the challenge did not explore clarifications in-depth. We choose this dataset for studying CRs for two main reasons: 1) it contains complex multi-modal dialogues with gold labels for referential ambiguity; 2) it focuses on tasks such as disambiguation and coreference resolution in multi-modal settings that are directly related with the problem of CR resolution.

2.3 Clarification Taxonomy

To evaluate how models handle CEs, we need to understand their ability to exploit fine-grained contextual information across modalities beyond level three of Clark's (1996) action ladder. Therefore, we derive a taxonomy of different types of clarifications depending on the information or *Disambiguating Property* exploited to resolve them: 1) **Individual Property**, such as object colour or state (i.e., *"The red jacket hanging"*); 2) **Dialogue History**, such as referring to previously mentioned objects (i.e., *"the one you recommended"*); and 3) **Relational**, such as position or their relation to other objects in the scene (i.e., *"the left shirt, next to the central rack"*).

These types are not mutually exclusive, and thus we often find that CRs are resolved with complementary information (i.e., *"The green dress on the right"*). Refer to Appendix B for discourse and taxonomy samples.

3 Experimental Setup

3.1 Clarification Extraction and Tagging

This section gives a summary of how we extracted the clarifications from the SIMMC 2.0 dataset using the gold annotations and tagged them using our taxonomy from Section 2.3.

When a turn is annotated as ambiguous, the system generates a CR (e.g., *"which one do you mean?"*). We label as **Before-CR** the user utterances preceding a CR (the user gave ambiguous information); whereas we label as **After-CR** the following user utterances that resolve the ambiguity. We obtain a subset of CEs (10% of all system turns are CRs) which we use for the analysis. Finally, we use a keyword-based method to tag the disambiguating properties exploited for clarifications (cf. Appendix A).

3.2 Metrics

We follow the SIMMC 2.0 evaluation protocol and measure coreference resolution performance using **Object F1**, derived as the mean of recall and precision for the predicted objects at each turn, as defined in (Kottur et al., 2021).

Along with object F1, we look at the difference in F1 between the turns before and after a clarification. Intuitively, a model that can process clarifications will improve after one, reflecting a higher F1 in the set of turns after a CR. Similarly, the turns before a CR may perform poorly, signalling confusion or uncertainty in general. We take this as the **Relative Delta** Δ to compare it across models.

3.3 Models

For our evaluation, we selected publicly available state-of-the-art models that took part in the SIMMC 2.0 challenge[3]. We give the relevant model details below, but please refer to original papers for additional architectural information.

Language-based We use two GPT-2-based (Radford et al., 2019) models: the Baseline ($Baseline_{GPT-2}$) from Kottur et al. (2021) (36.6% Object F1 ↑); and an improved version from one of the challenge participant teams (Hemanthage and Lemon, 2022), $GroundedLan_{GPT-2}$ (67.8% F1 ↑). Both models are similar and treat the task as a generation task, and are jointly trained with other goals in the challenge (coreference resolution, dialogue state tracking and response generation).

Vision-and-Language We take LXMERT-based (Tan and Bansal, 2019) model (Chiyah-Garcia et al., 2022) ($VisLan_{LXMERT}$, 68.6% F1 ↑) that combines the images from the visual scenes and the dialogue to predict the coreferenced objects at each turn. It extracts object attributes from a Detectron2 model (Wu et al., 2019) to use as textual descriptions along with the visual features. For each object in the scene, it outputs a probability for the object being referenced in that turn and selects those above a threshold. This model is only trained on coreference resolution.

Language-Vision-and-Relational We use the model of the coreference challenge winner team (Lee et al., 2022) ($MultiTask_{BART}$, 74% F1 ↑), a BART-based model (Lewis et al., 2020) trained to handle all challenge tasks. A pretrained ResNet model (He et al., 2016) encodes each object along with its non-visual attributes, a learnable embedding that is later mapped to match the dimension

[3]Not all models were public and some had missing code or weights.

Model	$Baseline_{GPT-2}$			$GroundedLan_{GPT-2}$			$VisLan_{LXMERT}$			$MultiTask_{BART}$		
Split	**Before-CR**	**After-CR**	**Δ**	**Before-CR**	**After-CR**	**Δ**	**Before-CR**	**After-CR**	**Δ**	**Before-CR**	**After-CR**	**Δ**
All Turns	34.3 (.01)			67.8 (.01)			68.6 (.01)			74.0 (.01)		
CR Turns	36.4 (.01)	29.1 (.01)	-20.1%	64.8 (.01)	67.7 (.01)	+4.4%	65.7 (.01)	69.2 (.01)	+5.4%	66.9 (.01)	74.3 (.01)	+11.1%
Disambiguating Property												
Individual Property	35.4 (.02)	27.4 (.01)	-22.7%	65.0 (.02)	68.0 (.02)	+4.6%	65.1 (.02)	69.3 (.01)	+6.4%	68.0 (.02)	**75.7** (.01)	+11.3%
Dialogue History	47.6 (.04)	43.7 (.04)	-8.2%	81.7 (.03)	82.1 (.03)	+0.4%	81.7 (.03)	**84.6** (.03)	+3.5%	67.2 (.04)	75.7 (.04)	+12.6%
Relational Context	32.9 (.02)	25.0 (.02)	-24.1%	62.4 (.02)	63.7 (.02)	+2.1%	62.7 (.02)	65.0 (.02)	+3.7%	66.5 (.02)	**72.6** (.02)	+9.1%

Table 2: Evaluation results for models at handling CEs with different disambiguating properties. Measured in **Object F1** ↑ (SD) and **Relative Delta** Δ.

of BART. The model is jointly optimised on multiple tasks, including several secondary tasks that enable learning disentangled object representations (Bengio et al., 2013) through object attribute slot prediction for each coreferenced object. The object location is also encoded through the bounding box information and a location embedding layer. Finally, the canonical object IDs are used to ground relations between the object locations, the visual and non-visual attributes.

4 Experiments

Referential Ambiguities Firstly, we explore whether referential ambiguities are an issue for models and if clarifications are thus needed. From the initial two rows of Table 2, we observe that, aside from the $Baseline_{GPT-2}$ model, all other models perform worse in turns **Before-CR** than when evaluating **All Turns**. This implies that indeed those utterances lack information to uniquely identify the referent objects, causing referential ambiguities for models and a lower object F1.

We also find that the F1 is higher in turns **After-CR** compared to turns **Before-CR** in all models but $Baseline_{GPT-2}$. This suggests that models can at least process clarifications in some cases. The $VisLan_{LXMERT}$ and $MultiTask_{BART}$ models even benefit with increased performance in **After-CR** turns compared to **All Turns**.

Regarding the surprisingly high scores for the $Baseline_{GPT-2}$ in turns **Before-CR** and low for **After-CR** , we suspect that it is due to the model exploiting linguistic phenomena along with smart use of previously mentioned objects and their canonical IDs, as explained in (Chiyah-Garcia et al., 2022). The model's performance drops dramatically when it is crucial to carry over cross-turn information and ground it in dialogue which is required **After-CR** .

Disambiguating Properties Using the CR taxonomy (cf. Section 2.3), we probe how models perform at exploiting different information with subsets of clarifications (bottom of Table 2).

All models but the baseline show a similar performance in **Before-CR** turns that exploit an Individual Property. $GroundedLan_{GPT-2}$ and $VisLan_{LXMERT}$ show a moderate F1 increase in the following **After-CR** turns, whereas $MultiTask_{BART}$ obtains a more substantial improvement (+11.3% Δ). Individual object properties in this dataset relate to concepts in the visual context which may be difficult to see or complex to understand beyond colour or shape (e.g., long sleeve or folded).

The $GroundedLan_{GPT-2}$ model implicitly encodes object attributes using a global object ID, which allows the model to learn latent information during training that carries over to evaluation sets (i.e. `<OBJ_256>`). On the other hand, the $VisLan_{LXMERT}$ model encodes colours and shapes explicitly using textual descriptions (i.e. `blue hoodie`) and implicitly in the visual region of interest features, which explains the slightly higher performance in these particular clarifications. However, the vision module of $VisLan_{LXMERT}$ is not explicitly trained to detect complex properties, only attributes such as colours or shapes (i.e. blue hoodie), and is instead left to the visual features to represent this information.

The multi-task learning objectives of $MultiTask_{BART}$ help the model obtain more fine-grained disentangled representations than using vision alone which helps in resolving ambiguities related to individual properties. Suglia et al. (2020) suggests that exploiting explicit object attributes reduces the potential referents and thus may also lead to improvements in solving CRs.

$GroundedLan_{GPT-2}$ and $VisLan_{LXMERT}$ models perform well when the clarifications are related to the dialogue context. Their initial F1 (+81%) suggests that they are able to carry information across turns particularly well and may not even need a CR in these cases. Both models also improve in

After-CR turns, with $VisLan_{LXMERT}$ reaching the highest score for this category. On the other hand, $MultiTask_{BART}$ improves its performance to 75.7% F1 (+12.6% Δ), but it does not display the same ability to exploit the linguistic context as the other models. This is likely due to the multi-task formulation involving specific loss functions which focus on visual and relational information only. Thus, the model obtains strong visual and relational object representations, whilst affecting the quality of BART's pre-trained language representations.

Relational clarifications seem to be the most difficult type to process for models, with the lowest F1 scores overall. The $MultiTask_{BART}$ model is able to exploit this information considerably better than the other models and improves by a +9.1% to 72.6%. This is an important strength of the model which extends its ability to encode visual attributes of the objects with information about the relationships between the objects in the scene. For instance, this model is able to capture the positions of the objects in the scene and how they relate to each other. The $VisLan_{LXMERT}$ model encodes positional information such as bounding box coordinates too, but it is not able to learn from them (Chiyah-Garcia et al., 2022). This is justified by previous research by (Salin et al., 2022) that shows how multi-modal models struggle with concepts such as position, and that they rely on language bias instead.

5 Conclusion

Referential ambiguities are common in situated human conversations. We sometimes cannot fully understand or identify a referred object or event, and thus we engage in clarification exchanges to resolve the ambiguity. In this paper, we analyse how several state-of-the-art models treat clarifications in situated multi-modal dialogues using the SIMMC 2.0 dataset. We classify the types of clarifications by the disambiguating property exploited and then evaluate the models with subsets of the data.

We find that language-based models perform well, yet struggle to benefit from clarifications. On the other hand, vision seems to be an important (but not essential) addition for models, which helps processing multi-modal CEs. Paired with a strong dialogue context, these types of models can perform reasonably well and carry information across turns to better handle clarifications. Finally, encoding relations between objects and their locations, and using additional learning objectives to predict attribute slots seems the strongest architecture for models to handle CEs.

Based on these results, to create improved models that can resolve referential ambiguities in situated dialogues, we need *holistic object-centric representations* that contain information about attributes and properties (Seitzer et al., 2022), and that can *dynamically* change to reflect the information exchanges available in the dialogue context.

Acknowledgements

Chiyah-Garcia's PhD is funded under the EPSRC iCase with Siemens (EP/T517471/1). This work was also supported by the EPSRC CDT in Robotics and Autonomous Systems (EP/L016834/1).

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. Association for Computational Linguistics.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Luciana Benotti and Patrick Blackburn. 2021. A recipe for annotating grounded clarifications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4065–4077, Online. Association for Computational Linguistics.

Feilong Chen, Fandong Meng, Xiuyi Chen, Peng Li, and Jie Zhou. 2021. Multimodal incremental transformer with visual grounding for visual dialogue generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 436–446, Online. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Javier Chiyah-Garcia, Alessandro Suglia, José David Lopes, Arash Eshghi, and Helen Hastie. 2022. Exploring multi-modal representations for ambiguity detection & coreference resolution in the SIMMC 2.0 challenge. In *AAAI 2022 DSTC10 Workshop*.

H. H. Clark and S. A. Brennan. 1991. *Grounding in communication*, pages 127–149. Washington: APA Books.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Raquel Fernández and Jonathan Ginzburg. 2002. Non-sentential utterances in dialogue: A corpus-based study. In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue*, pages 15–26, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Felix Gervits, Gordon Briggs, Antonio Roque, Genki A. Kadomatsu, Dean Thurston, Matthias Scheutz, and Matthew Marge. 2021. Decision-theoretic question generation for situated reference resolution: An empirical study and computational model. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, page 150–158, New York, NY, USA. Association for Computing Machinery.

Charles Goodwin. 1981. *Conversational organization: Interaction between speakers and hearers*. Academic Press, New York.

Danfeng Guo, Arpit Gupta, Sanchit Agarwal, Jiun-Yu Kao, Shuyang Gao, Arijit Biswas, Chien-Wei Lin, Tagyoung Chung, and Mohit Bansal. 2022. Gravl-bert: Graphical visual-linguistic representations for multimodal coreference resolution. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 285–297.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Patrick G. T. Healey, Gregory J. Mills, Arash Eshghi, and Christine Howes. 2018. Running Repairs: Coordinating Meaning in Dialogue. *Topics in Cognitive Science (topiCS)*, 10(2).

Bhathiya Hemanthage and Oliver Lemon. 2022. Global-local information-aware multimodal grounding with GPT for co-reference resolution. In *AAAI 2022 DSTC10 Workshop*.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haeju Lee, Oh Joon Kwon, Yunseon Choi, Minho Park, Ran Han, Yoonhyung Kim, Jinhyeon Kim, Youngjune Lee, Haebin Shin, Kangwook Lee, and Kee-Eung Kim. 2022. Learning to embed multimodal contexts for situated conversational agents. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 813–830, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Brielen Madureira and David Schlangen. 2023. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the CoDraw dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2303–2319, Dubrovnik, Croatia. Association for Computational Linguistics.

Gregory J. Mills. 2007. *Semantic co-ordination in dialogue: the role of direct interaction.* Ph.D. thesis, Queen Mary University of London.

Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, University of London.

Matthew Purver and Jonathan Ginzburg. 2004. Clarifying noun phrase semantics. *Journal of Semantics*, 21(3):283–339.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Verena Rieser and Oliver Lemon. 2006. Using machine learning to explore human multimodal clarification strategies. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 659–666, Sydney, Australia. Association for Computational Linguistics.

Verena Rieser and Johanna Moore. 2005. Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 239–246, Ann Arbor. Association for Computational Linguistics.

Kepa Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in German task-oriented spoken dialogues. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, Barcelona, Spain.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11248–11257.

Ruben San-Segundo, Juan M. Montero, J. Ferreiros, R. Córdoba, and José M. Pardo. 2001. Designing confirmation mechanisms and error recover techniques in a railway information system for Spanish. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, pages 136–139, Aalborg, Denmark. Association for Computational Linguistics.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 136–143, Boston. Association for Computational Linguistics.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. 2022. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*.

Zhengxiang Shi, Yue Feng, and Aldo Lipani. 2022. Learning to execute actions or ask clarification questions. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2060–2070, Seattle, United States. Association for Computational Linguistics.

Alessandro Suglia, Yonatan Bisk, Ioannis Konstas, Antonio Vergari, Emanuele Bastianelli, Andrea Vanzo, and Oliver Lemon. 2021. An empirical study on the generalization power of neural representations learned via visual guessing games. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2135–2144.

Alessandro Suglia, Ioannis Konstas, Andrea Vanzo, Emanuele Bastianelli, Desmond Elliott, Stella Frank, and Oliver Lemon. 2020. CompGuessWhat?!: A multi-task evaluation framework for grounded language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7625–7641, Online. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Yue Wang, Shafiq Joty, Michael Lyu, Irwin King, Caiming Xiong, and Steven C.H. Hoi. 2020. VD-BERT: A Unified Vision and Dialog Transformer with BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3325–3338, Online. Association for Computational Linguistics.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. `https://github.com/facebookresearch/detectron2`.

Yi Zhu, Yue Weng, Fengda Zhu, Xiaodan Liang, Qixiang Ye, Yutong Lu, and Jianbin Jiao. 2021. Self-motivated communication agent for real-world vision-dialog navigation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1574–1583.

A Additional CR Details

A.1 Clarification Tagging Method

The algorithm for CR tagging is based on manual annotations using the dev set, and then creating a set of keywords and regexes that would automatically find the disambiguating property used. **Individual Properties** include mentions of: colour (*blue*), object types (*jacket*), style (*floral*), brand names (*Yogi Fit*), states (*folded*) and other (*long-sleeve*). The metadata provides all of this information that we use as keywords. Other information such as **Relational Context** was based on positional keywords (*left, top*) and relational with objects (*next to*) or the scene (*farthest*). **Dialogue History** was based on linguistic cues and the use of common structures (i.e. *"...in my cart"*, *"you mentioned"*). We left some CRs as unclassified ($< 0.06\%$ of the data) because they do not provide any meaningful additional information or are out of scope (i.e., *"What is that lamp made of? - Sorry, which one? - I'm not sure, I think it's a lamp"*).

During tagging, we applied the algorithm to each clarification, including both the system CR and the user response, as the information is sometimes scattered across turns (i.e., see SYS CR in Dialogue Sample 1). We manually verified this method and designed unit tests to ensure correct classification as the tagging algorithm increased in complexity.

A.2 Ambiguity in SIMMC 2.0 Scenes

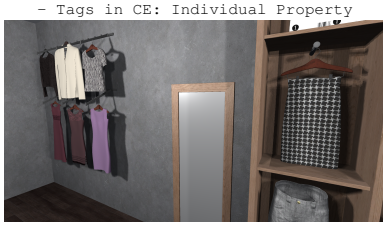
Split	Mean Candidate Objects (SD) `Type`	`Colour`
All Turns	3.10 (5.16)	2.58 (4.23)
CR Turns	5.41 (5.62)	4.53 (4.63)
Disambiguating Property		
Individual Property	5.52 (5.69)	4.57 (4.77)
Dialogue History	4.57 (4.82)	4.95 (5.25)
Relational Context	5.78 (5.91)	4.63 (4.66)

Table 3: Statistics about the level of ambiguity in SIMMC 2.0 based on candidate objects of the same **Type** (*jacket, t-shirt...*) or **Colour** (*blue, yellow...*). These properties alone are not enough to uniquely identify an object in most cases.

B Corpus Examples

Dialogue Sample 1

USR *Hello, do you have any jackets for me to look at?*
SYS *Sure, what do you think of the light grey jacket hanging up high at the back left?*
USR *Do you have anything with a similar size range to the black sweater beside the light grey jacket?*
SYS *Sorry, I don't have anything similar to that*
Before-CR USR *What size is that sweater anyways?* [Referential Ambiguity]
CR SYS *The black one?* [Clarification Request]
After-CR USR *Yes exactly* [Clarification]
`- Tags in CE: Individual Property`

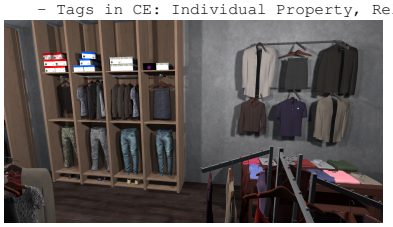


Dialogue Sample 2

USR *I need some help picking out a jacket.*
SYS *Do you like the look of that solid black one hanging in the bottom row up front on the right?*
USR *I do like the look of that one. Will you please put that in my cart?*
SYS *Sure! I'll add that black jacket to your cart right now.*
USR *Now could we look for a grey coat?*
SYS *There's a grey coat hanging directly above the black jacket you like.*
Before-CR USR *Actually first I'd like to know which brand makes that jacket. If you have any data on what other customers think of it, that'd be good to hear too.* [Referential Ambiguity]
CR SYS *Sorry, which jacket specifically are you curious about?* [Clarification Request]
After-CR USR *The black one up front next to the skirt* [Clarification]
`- Tags in CE: Individual Property, Relational`
SYS *That jacket is a Yogi Fit original, and other buyers have rated it a 4.0 on average*
USR *Are there any super ritzy options with stellar reviews?*
SYS *I don't see any, sorry. Is there something else I can help you with?*
USR *I'd like to find a dress in extra large.*
SYS *I have a feeling you'll like the dresses on this first rack up front and center. Would you wear the purple one on the left, the brown one on the right, or the grey one in front?*
Before-CR USR *Woops, I didn't even ask you what sizes that jacket came in.* [Referential Ambiguity]
CR SYS *Which jacket do you want to know the sizes of?* [Clarification Request]
After-CR USR *That black one that I had you put in my cart* [Clarification]
`- Tags in CE: Individual Property, Dialogue History`

Dialogue Sample 3

USR *Can you find me a black blouse?*
SYS *What about this black blouse in the shelf display?*
Before-CR USR *What's the size of the grey shirt and the grey and brown shirt?* [Referential Ambiguity]
CR SYS *Which shirts are you referring to?* [Clarification Request]
After-CR USR *I mean the grey shirt on the table display and the grey and brown shirt next to it* [Clarification]
`- Tags in CE: Individual Property, Relational`



PGTask: Introducing the Task of Profile Generation from Dialogues

Rui Ribeiro, Joao P. Carvalho, Luísa Coheur
INESC-ID, Lisboa
Instituto Superior Técnico, Universidade de Lisboa
{rui.m.ribeiro, joao.carvalho, luisa.coheur}@inesc-id.pt

Abstract

Recent approaches have attempted to personalize dialogue systems by leveraging profile information into models. However, this knowledge is scarce and difficult to obtain, which makes the extraction/generation of profile information from dialogues a fundamental asset. To surpass this limitation, we introduce the Profile Generation Task (PGTask). We contribute with a new dataset for this problem, comprising profile sentences aligned with related utterances, extracted from a corpus of dialogues. Furthermore, using state-of-the-art methods, we provide a benchmark for profile generation on this novel dataset. Our experiments disclose the challenges of profile generation, and we hope that this introduces a new research direction.

1 Introduction

Building conversational systems that mimic human attributes has always been a long-term goal in Natural Language Processing (NLP). Various works have attempted to leverage speaker profile information to improve the consistency of dialogue generation models (Wu et al., 2020; Xu et al., 2022; Cao et al., 2022). By incorporating speaker-specific characteristics, such as age, gender, personality traits, and cultural background, into the conversational systems, it is possible to create more personalized and human-like interactions. However, for dialogue systems, this sort of information is scarce and requires annotation efforts that are expensive to obtain, so there is a need to build methods that automatically gather this knowledge from dialogues.

Zhang et al. (2018) introduced PersonaChat, a dataset comprising a collection of profile sentences (*persona*) that reflect each speaker's individual characteristics and personal facts. These profiles serve as a knowledge base for promoting the consistency between utterances from speakers, and various recent dialogue models have incorporated this information using diverse techniques (Song et al., 2020, 2021; Cao et al., 2022).

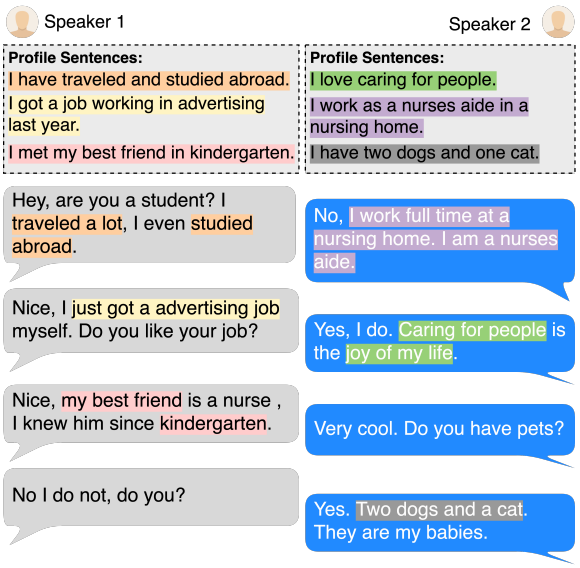


Figure 1: An example dialogue where each turn contains the corresponding profile sentence.

Few works have attempted to infer profile information from PersonaChat dialogues. Gu et al. (2021) restructured PersonaChat and built the Persona Detection Task, where the goal was to retrieve the correct persona amongst a set of distractor personas. Although introducing an interesting research path, this task is limited to a set of pre-defined personas, which is not suitable for extracting profile sentences from unseen conversational data. Cao et al. (2022) also manipulate PersonaChat to incorporate model-agnostic personas into the dialogue generation task. Nevertheless, for the profile generation task, PersonaChat is structured in a profile-to-dialogue manner and lacks information about the corresponding profile sentence per turn, which may become a challenge when the task becomes extracting profile information from utterances.

In this work, we introduce the PGTask[1], where

[1]Dataset and code are available at https://github.com/ruinunca/PGTask.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 183–189
September 11–15, 2023. ©2023 Association for Computational Linguistics

the goal is to generate profile sentences given speaker utterances. For this, we create a new dataset, the Profile Generation Dataset (PGDataset), which relates utterances with the respective profile sentences upon the existing PersonaChat corpus. In Figure 1, we can observe several examples of relations between profile sentences and the corresponding speaker's utterance. Notice, however, that the task is more challenging than just finding, within the dialogues, utterances that highly relate to each profile sentence. For instance, the profile sentence "I like all genres of music." is probably at the origin of the utterance "Yes, sometimes I also listen to classical music.", but we cannot extract that profile sentence from that single utterance (the goal of PGTask).

We framed our problem as an entailment classification task and, after human feedback, we reached the final PGDataset. Finally, we provide results from three state-of-the-art models trained and evaluated in the proposed dataset.

2 Building PGDataset

In this section, we demonstrate how we formulated our task as an entailment detection problem and describe the utilization of human experts' feedback to build a consistent dataset.

2.1 Modeling Entailment Relations

In the Natural Language Inference (NLI) task, the goal is to classify the relationship between a pair of premise and hypothesis sentences into three classes: entailment (E), neutral (N), and contradiction (C). Welleck et al. (2019) extended the NLI task to the dialogue setting and introduced the Dialogue Natural Language Inference (DNLI) dataset, where the input sentences consist of dialogue utterances from PersonaChat. We adopt this procedure and train a model $\mathcal{M}^{NLI}$ to identify the correct profile sentences for each utterance in a dialogue.

Consider two sentences s_i and s_j that are concatenated into the input $x = \{s_i, s_j\}$. First, we utilize RoBERTa (Liu et al., 2019) to obtain a hidden representation h from the input x. Then, we include a $\mathrm{softmax}$ classifier on top of RoBERTa to obtain the probability distribution over the set of possible classes. Formally, we obtain the probability of label $y \in \{C, N, E\}$ with:

$$\begin{aligned} h &= \text{RoBERTa}(x), \\ p_{\mathcal{M}^{NLI}}(y|x) &= \text{softmax}(Wh), \end{aligned} \tag{1}$$

where W is the learnable parameter matrix from the classification layer. We fine-tune both RoBERTa and W parameters by maximizing the log-probability of the correct label.

Datasets	Accuracy (%)
DNLI	91.24
MNLI + DNLI	**91.75**

Table 1: Accuracy of fine-tuned RoBERTa for the test set of DNLI.

We experiment with two different settings where we fine-tune RoBERTa only on DNLI and on MNLI (Williams et al., 2018), a benchmark multi-genre NLI dataset, and DNLI datasets for better generalization. Details are provided in Appendix A. Table 1 shows the results on the test set, where the latter achieves higher accuracy and is selected as the annotation model.

2.2 Dataset Annotation

In PersonaChat (Zhang et al., 2018), each dialogue carries a set of profile sentences for both speakers. Consider a set of n utterances from a speaker, $U = \{u_1, u_2, ..., u_n\}$, a set of k profile sentences $P = \{p_1, p_2, ..., p_k\}$ from the same speaker, and the dialogue NLI model from Section 2.1. Then, at time step t, we can determine one or more profile sentences s_t related to utterance u_t using:

$$\begin{aligned} s_t = \{p_i \in P : \\ \underset{y \in \{C,N,E\}}{\arg\max}\, (p_{\mathcal{M}^{NLI}}(y|\{u_t, p_i\}) = E\}. \end{aligned} \tag{2}$$

In Equation 2, the profile sentences are gathered by considering the entailed cases between the utterances and the profile sentences, where each utterance could be associated with more than one profile sentence. In Table 2, we provide an extract from the PGDataset.

Utterance	Profile Sentences
I enjoy hanging with my mother she is my best friend.	My mom is my best friend.
I am almost done, I only have two years left in law school.	I have got two more years in college.
	I study law.

Table 2: Two examples from PGDataset.

2.3 Human Annotations

In the profile generation task, the profile must represent a possible extraction from the dialogue utterance, and this correlation's direction between the utterance and the profile sentence must be valid. To assess the quality of the automatic annotations from our model, we resort to human evaluation.

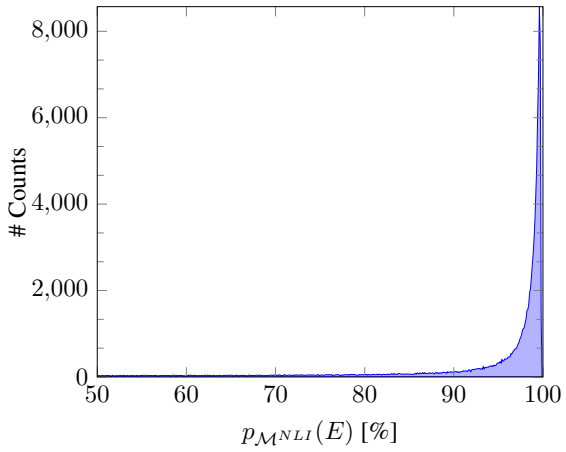


Figure 2: Distribution of the entailment class probability for the entailed cases ($\mu = 93.4$, $\sigma^2 = 1.10$).

For all the pairs classified as entailed in Equation 2, we measure the confidence by inspecting the $\mathrm{softmax}$ probability assigned to the entailment class. Our intuition is that a weak confidence when classifying a profile sentence as entailed corresponds to a weak or incorrect correlation and can be removed from the dataset. In Figure 2, we plot the distribution of the scores from the entailment class for all points obtained from Equation 2.

To determine if a higher confidence value corresponds to a correct example, we randomly select 100 samples from 3 intervals: $[50, 70]$, $]70, 90]$, and $]90, 100]$. We asked 3 expert annotators from our department to "mark with an X if the profile sentence could be extracted from the given utterance". The quality of the samples is measured by the number of marked samples by the annotators (accuracy). The agreement rate between annotators was 86.66% and the average accuracy for each interval was 8.33%, 12.33%, and 51.67%, respectively. The results obtained show that when the confidence of the model grows, the correlation between the profile sentence and the utterance also increases.

After inspecting the results from the annotators, we observed that most of the marked samples had more than 99% confidence. We asked for a second round of annotations with 100 samples but now only for samples with more than 99% confidence. The agreement rate between annotators was 91% and the average accuracy was 87,33%, a significantly higher score compared to the $]90, 100]$ interval. We decided, thus, that PGDataset only considers the samples which the model classified with more than 99% confidence.

Train	# Samples	34355
	Avg. Profile Sentences	1.06
	Avg. Utterance Words	13.13
	Avg. Profile Sentence Words	7.14
Valid	# Samples	4236
	Avg. Profile Sentences	1.06
	Avg. Utterance Words	13.36
	Avg. Profile Sentence Words	7.67
Test	# Samples	3760
	Avg. Profile Sentences	1.06
	Avg. Utterance Words	13.05
	Avg. Profile Sentence Words	7.17

Table 3: Dataset Statistics.

2.4 PGDataset Statistics

In Table 3, we provide the dataset statistics for the gathered samples.

3 Benchmarking the PGTask

In this task, the goal is to generate a profile sentence conditioned on an utterance. Transformer-based decoders have achieved substantial progress in various NLP tasks (Radford et al., 2019). We leverage these models and rely on a causal language modeling (CLM) objective for our profile generation task. More precisely, considering a sentence $s = \{w_1, ..., w_n\}$ composed of n words, in CLM, the maximum likelihood objective over s is:

$$\mathcal{L}_{CLM} = \sum_{i=1}^{n} \log P(w_i | w_1, ..., w_{i-1}). \quad (3)$$

For our task, we are only interested in calculating the loss for the words from the profile sentence conditioned on the utterance. Considering an utterance $u = \{w_1^u, ..., w_m^u\}$ and a profile sentence $p = \{w_1^p, ..., w_k^p\}$, we redefine the objective from Equation 3:

$$\mathcal{L}_{PG} = \sum_{i=1}^{k} \log P(w_i^p | w_1^u, ..., w_m^u, w_1^p, ..., w_{i-1}^p). \quad (4)$$

	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
W/o FT	distilgpt2	5.59	0.30	0.00	0.00	6.86	0.93	5.80	84.66
	gpt2-small	4.87	0.40	0.00	0.00	6.08	0.63	5.20	84.21
	gpt2-medium	4.48	0.20	0.00	0.00	7.20	0.31	5.32	83.28
W/ FT	distilgpt2	44.42	13.18	5.60	0.00	35.68	14.12	35.39	92.35
	gpt2-small	61.30	32.30	20.62	9.44	50.07	28.31	50.00	94.39
	gpt2-medium	59.31	25.94	15.30	9.17	46.32	24.14	45.88	94.76

Table 4: Generation results for models with and without fine-tuning (FT) on the PGDataset. The results presented are the average score of 5 runs. The scores range between 0 and 100%.

As seen in Equation 4, the loss is only calculated for the generation of the profile sentences. In the model's input, we separate the utterance and profile sentences using a special token <gen> and, as it can exist more than one profile sentence, we add <sep> between the profile sentences.

4 Experiments

In this section, we evaluate Transformer decoders on the novel dataset and provide benchmark results for future research. Additional experimental details are provided in Appendix B.1.

4.1 Models

GPT2 This model has achieved state-of-the-art results in various generation tasks (Radford et al., 2019). We consider two different pre-trained versions that differ in size, the `gpt2-small` and `gpt2-medium` (details in Appendix B.2).

DistilGPT2 This is a distilled version of GPT2, where it was trained under the supervision of GPT2 (Hinton et al., 2015). The `distilgpt2` contains about half the size of GPT2 while still achieving competing performance in various NLP tasks.

4.2 Metrics

We follow common practices for text generation and report BLEU (Papineni et al., 2002) and ROUGE (Lin and Hovy, 2002), metrics that, respectively, measure the precision and recall between the generated and the golden text. Additionally, we employ BERT Score (Zhang et al., 2019), an automatic metric that leverages BERT's (Kenton and Toutanova, 2019) contextual embeddings and matches words in candidate and golden sentences using cosine similarity.

4.3 Results

In Table 4, we provide benchmark results for the PGTask. The models without fine-tuning fail to extract the correct profile information from the dialogue sentences, which is expected as their pre-training was on a large collection of unstructured text. We observe that fine-tuning the models has a great impact on the overall performance, where `gpt2-small` achieves the higher scores in all metrics except BERTScore (for a minimal difference). In Appendix B.3, we provide some generated examples from the evaluated models. The results obtained show promising advances in this task and we hope that this will introduce a new future research direction in this area.

5 Related Work

Recent research has focused on building personalized dialogue systems using profile information. Li et al. (2016) proposed a neural conversational model to capture background information and speaking style from interlocutors in dialogue. Zhang et al. (2018) introduced a dataset composed of personas, which are essentially 3 to 5 profile sentences describing the speaker's profile. Zheng et al. (2019) studied how to include profile information such as age, location, and interests by explicitly incorporating this knowledge into the sequence-to-sequence framework.

Few works have attempted to identify profile knowledge from conversational data. (Gu et al., 2021) introduced a framework for detecting the correct profile amongst a set of distractor profiles. Nevertheless, the authors do not consider the correlation between utterances and profile sentences. (Cao et al., 2022) proposed a data manipulation method to construct distilled and diversified dialogue data containing profile information and leverage it into the dialogue generation task.

6 Conclusion

We propose the PGTask and contribute with PGDataset, a dataset with more than 30 000 pairs of

utterances and profile sentences built with the feedback of human annotators. In addition, we train state-of-the-art models and achieve promising results in the proposed task. We hope that this new line of research will help the task of personalizing dialogues, although the task of automatically extracting profiles from dialogues is valuable by itself.

Acknowledgements

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020 and grant 2022.10640.BD, by the project CMU-PT MAIA with reference 045909, as well as by the Recovery and Resilience Plan (RRP) and Next Generation EU European Funds through project C644865762-00000008 Accelerat.AI.

References

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7984–8002.

Jia-Chen Gu, Zhenhua Ling, Yu Wu, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Detecting speaker personas from conversational texts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1126–1136.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Cite arxiv:1907.11692.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Haoyu Song, Yan Wang, Kaiyan Zhang, Weinan Zhang, and Ting Liu. 2021. Bob: Bert over bert for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177.

Haoyu Song, Yan Wang, Weinan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Bowen Wu, MengYuan Li, Zongsheng Wang, Yifu Chen, Derek F Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2020. Guiding variational response generator to exploit persona. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 53–65.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you

have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

A Fine-Tuning RoBERTa

We fine-tune a pre-trained `roberta-base`[2] (Liu et al., 2019) with 12 layers, 768 hidden units, 12 attention heads, and 125M parameters on 1 NVIDIA GeForce RTX 3080 to minimize the cross entropy. We use Adam (Kingma and Ba, 2014) optimizer with a learning rate of $5e^{-}5$. The batch size was 32, we train for 20 epochs and early stop after 5 epochs without an increase in the validation accuracy.

B Profile Generation

B.1 Experimental Details

We perform 5 runs for each model on 1 NVIDIA GeForce RTX 3080 using different seed values and calculate the average score for all metrics. Models are trained to minimize the cross entropy using Adam (Kingma and Ba, 2014) optimizer with a learning rate of $5e^{-}5$. For `gpt2-small` and `distilgpt2`, the batch size was 16 while for `gpt2-medium` the batch size was 4 with 4 gradient accumulation steps. We train for 20 epochs with early stopping where the training is stopped after 5 epochs without a decrease in the validation loss. We generate the profile sentences with a maximum length of 50 and perform greedy sampling, i.e., select the next word with the highest probability. All experiments are implemented using the HuggingFace[3] and PyTorch[4] libraries.

B.2 Models

The `gpt2-small`[5] version contains 12 layers, 768 hidden units, 12 attention heads, and 117M parameters and `gpt2-medium`[6] includes 24 layers, 1024 hidden units, 16 attention heads, and 345M parameters. The distilled version `distilgpt2`[7] (Hinton et al., 2015) is smaller than GPT2, where it is composed of 6 layers, 768 hidden units, 12 attention heads, and 82M parameters.

B.3 Generated Examples

As discussed in Section 4.3, the fine-tuned models show promising results in the generation of correct profile sentences. An example for that is presented in Table 5. Here, all models successfully extract the profile sentence, although we could argue that *loving* and *liking* are semantically different.

However, the evaluation scores also show that the systems are still far from always extracting the correct profiles. In Table 6, we show an example where all models generated unrelated profile information and failed to recognize that the speaker dropped off high school.

Dialogue Utterance: sorry, i do not like music, i like reading mystery books.
Golden Profile Sentences: i enjoy reading mysteries.
`distilgpt2`: i love to read mystery novels. **`gpt2-small`:** i love reading mysteries in my free time. **`gpt2-medium`:** i read mystery novels.

Table 5: Generated example #1 from fine-tuned models.

Dialogue Utterance: dropping out of high school was a bad idea. the landlord just called.
Golden Profile Sentences: i dropped out of high school.
`distilgpt2`: i just got a job at the elementary school in new england. **`gpt2-small`:** my parents got a new job. **`gpt2-medium`:** i just graduated high school.

Table 6: Generated example #2 from fine-tuned models.

[2] https://huggingface.co/roberta-base
[3] https://huggingface.co/
[4] https://pytorch.org/
[5] https://huggingface.co/gpt2
[6] https://huggingface.co/gpt2-medium
[7] https://huggingface.co/distilgpt2

Question Generation to Elicit Users' Food Preferences by Considering the Semantic Content

Jie Zeng and **Yukiko I. Nakano** and **Tatsuya Sakato**
Faculty of Science and Technology, Seikei University
{jie-zeng, y.nakano, sakato}@st.seikei.ac.jp

Abstract

To obtain a better understanding of user preferences in providing tailored services, dialogue systems have to generate semi-structured interviews that require flexible dialogue control while following a topic guide to accomplish the purpose of the interview. Toward this goal, this study proposes a semantics-aware GPT-3 fine-tuning model that generates interviews to acquire users' food preferences. The model was trained using dialogue history and semantic representation constructed from the communicative function and semantic content of the utterance. Using two baseline models: zero-shot ChatGPT and fine-tuned GPT-3, we conducted a user study for subjective evaluations alongside automatic objective evaluations. In the user study, in impression rating, the outputs of the proposed model were superior to those of baseline models and comparable to real human interviews in terms of eliciting the interviewees' food preferences.

1 Introduction

With interviews being used for various purposes, interview systems such as surveys (Johnston et al., 2013; Stent et al., 2006), job interviews (Inoue et al., 2020), and coaching (Hoque et al., 2013) have been developed. Interviews are categorized into three types: structured, semi-structured, and unstructured. In terms of flexibility, semi-structured interviews are between structured and unstructured. They are not completely planned but have a topic guide that needs to be covered. To build a dialogue system that can generate semi-structured interviews, flexible dialogue control must be provided while following the topic guide. To address the issues involved in generating semi-structured interviews, this study proposes an interview system to learn user food preferences.

Various dialogue control mechanisms have been studied in task-oriented dialogue systems to collect information from users, with the system responses are determined based on manually defined rules, POMDP (Young et al., 2010), deep learning (Chen et al., 2019), and reinforcement learning (Sankar and Ravi, 2019). However, these systems have less flexibility in dialogue control because the dialogue states are defined as a set of slot-value pairs that are limited to the task domain.

Research on open-domain non-task-oriented dialogue generation has contributed to the development of chitchat systems that can produce system responses for various topics. Initially, a simple sequence-to-sequence approach (Sordoni et al., 2015; Vinyals and Le, 2015; Serban et al., 2016; Li et al., 2016) was employed to generate a response. This approach has been improved to produce appropriate and meaningful responses, considering the dialogue context (Serban et al., 2017), and generate knowledge-grounded responses (Hedayatnia et al., 2020; Wu et al., 2020; Zhang et al., 2020a; Galetzka et al., 2021). More recently, ChatGPT (Ouyang et al., 2022) has demonstrated remarkable performance in generating rich and natural dialogues. However, these techniques have not yet been designed to generate dialogues for user model acquisition. Consequently, interview systems are required to generate responses that are aligned with the purpose of the interview.

To overcome the problems discussed above and generate useful questions in semi-structured interviews to elicit user food preferences, this study proposes a GPT-3 based model trained to generate responses with its semantic representation, which is constructed from the utterance's communicative function and semantic content. Semantic content refers to a structured sequence of labels for objects and their attributes. It is expected that using semantic content as part of the training targets would help constrain the generated responses towards eliciting the user food preferences.

The contributions of this study are as follows: 1) a semantic representation is proposed for sys-

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 190–196
September 11–15, 2023. ©2023 Association for Computational Linguistics

<Role (I/C)>-<message#>-<sentence#>	sentence	Communicative function	Semantic content
I-1-1	It's almost lunchtime, what do you eat for lunch?	Q-plan	[eat, [(Dish, ?)]]
U-2-1	Right.		
U-2-2	I like sandwiches.		
I-3-1	What do you like to have as sandwich ingredients?	Q-preference-positive	[like, [(Dish, sandwich, ingredient, ?)]]
U-4-1	I like tuna.		
I-5-1	Tuna is good on a sandwich.	Reply	[think, [(Dish, sandwich, ingredient, tuna)],[Evaluation, good]]
I-5-2	What do you often drink with your sandwich?	Q-habit	[drink, [(Drink,?, combine-with, sandwich)]]

Prompt and completion pairs in GPT-3 fine-tuning

```
HISTORY
SYSTEM: It's almost lunchtime, what do you eat for lunch?
USER: Right. I like sandwiches.
---
INFORMATION_FOR_SYSTEM_OUTPUT
COMMUNICATIVE_FUNCTION_LABEL: Q-preference-positive
SEMANTIC_CONTENT:{
VERB: like
OBJECT_TYPE: Dish
OBJECT_NAME: sandwich
OBJECT_ATTRIBUTE: ingredient
OBJECT_ATTRIBUTE_VALUE: ?
EVALUATION: None
}
->SYSTEM_OUTPUT: What do you like to have as
                 sandwich ingredients?
```

Figure 1: Overview of the proposed method. The left table shows an example dialogue between an interviewer (I) and a customer (C). The communicative function and semantic content of the interviewer's utterances are shown in the third and fourth columns, respectively. The right side shows the Prompt and Completion input for GPT-3 fine-tuning used to predict interview utterance I-3-1. The blue part indicates the prompt, and the green part indicates the completion. ***Bold italics*** indicate utterances or annotated values.

tem responses; 2) a response generation model is created for the interviewer's role; and 3) the effectiveness of the proposed method in eliciting user preferences is demonstrated through an evaluation experiment.

2 Corpus collection

To prepare the dataset used in this study, text-based dyad conversations were collected to interview participants regarding their food preferences. The participants were recruited through crowdsourcing. Each participant was assigned the role of either interviewer or interviewee and communicated using a chat system on a web browser. The interviewer was instructed to elicit the partner's preference for food, whereby they exchanged messages taking turns, for a minimum of 40 turns. Thus, a total of 118 Japanese dialogues were collected.

3 Method

To train a response generation model for the interviewer's role by considering the semantic representation of the interviewer's responses, we propose the method illustrated in Figure 1. First, the semantic representation of the interviewer's responses is presented, and subsequently, model training is explained.

3.1 Semantic representation of interviewer's responses

The semantic representation of an interviewer's utterance comprises the intention and meaning of the utterance. This representation can be exploited to train the dialogue generation model and direct the dialogue toward eliciting food preference information, as explained in detail below.

Communicative Function (CF): To specify the intention of the utterance, we refined the labels for self-disclosure and question types proposed in SWBD-DAMSL (Jurafsky, 1997) and Meguro et al. (2014), thereby defining 20 labels. The list is shown in the Appendix A.

Semantic Content (SC): The meaning of an utterance is described as a structured sequence of labels for verb and object features, such as `OBJECT_TYPE`, `OBJECT_NAME`, `OBJECT_ATTRIBUTE`, and `OBJECT_ATTRIBUTE_VALUE`.

Examples of semantic representation are shown in Figure 1. In utterance I-3-1, "What do you like to have as sandwich ingredients?" the communicative function is Q-preference-positive. The semantic content begins with the verb category. In this case, the verb is *like*. This is followed by object features `OBJECT_TYPE`: *Dish*, `OBJECT_NAME`: *sandwich*, `OBJECT_ATTRIBUTE`: *ingredient*, and `OBJECT_ATTRIBUTE_VALUE`: *?*. The *?* indicates that this value is missing. Thus, the semantic content of this utterance is expressed as [(Dish,sandwich,ingredient,?)]. Predefined values are used for the verbs and elements of `OBJECT_TYPE` and `OBJECT_ATTRIBUTE` for object features (see Appendix A). The details of the SC scheme were proposed in Zeng et al. (2022).

After annotating the CF and SC in the corpus collected in Section 2, we calculated the inter-coder reliability between two annotators. Cohen's Kappa value for CF was $\kappa = 0.72$ (substantial agreement), and the agreement ratio for verbs and object fea-

	BLEU-1	BLEU-2	BLUE-3	BLEU-4	BERTScore
ChatGPT	22.99	11.23	5.46	2.38	0.72
Seq2Seq	**25.11**	15.05	**8.11**	2.48	**0.75**
CF+SC	24.98	**15.23**	7.53	**2.71**	**0.75**

Table 1: Average BLEU scores and BERTScore on the test set. The best score for each column is highlighted in bold.

tures in SC between the two annotators was 0.72.

3.2 Interviewer response generation model

To create a response generation model, we fine-tuned OpenAI's GPT-3 (Brown et al., 2020), thereby referring to this proposed model as the CF+SC model. The model generates the completion part that follows the prompt. The formats for the prompt and completion are shown in Figure 1. Up to five messages preceding the prediction target interviewer's response were added to the prompt as dialogue history. The completion consisted of the annotated CF and SC (Section 3.1) and the interviewer's response sentence. The format of the completion part is indicated by green letters in Figure 1. When multiple sentences were included in the interviewer's message (turn), the last sentence, which usually contains the main claim, was used as the prediction target.

4 Experiment and evaluation

We evaluated the performance of the proposed CF+SC model using three comparison targets: the ground truth and two baselines.

Ground truth (GT): Actual utterances of the interviewers were used as the ground truth.

Fine-tuned GPT-3 (Seq2Seq): This simple fine-tuning model uses GPT-3. The model was trained without semantic representation (CF and SC) of the prediction targets. A sequence of preceding utterances was provided as prompt, and the model output was the interviewer's response.

Zero-shot ChatGPT (ChatGPT): OpenAI's ChatGPT model (reinforcement learning with human feedback and chat-optimized models (Ouyang et al., 2022)), specifically, gpt-3.5-turbo-0301, was adopted as the best general-purpose dialogue model. The zero-shot method was employed such that only the dialogue history and the system's role as an interviewer were provided as prompts [1]. The system was instructed to play the role of the interviewer and generate a response to elicit customer preferences by considering the context.

The temperature parameter for the three GPT-based models was set to 0. Thus, the generation was almost deterministic. While the CF+SC model generates both semantic representation and text of the response, we used the `SYSTEM_OUTPUT` part to extract the system response text. The CF+SC and Seq2Seq models generate a single sentence. Thus, in order to align the comparison conditions, when ChatGPT model generates multiple sentences, the last sentence, which tends to contain the main claim, was used in comparing with the ground truth.

The GPT-3 ("davinci" model) was fine-tuned using OpenAI's API. The model was trained for four epochs. The batch size was eight, and the learning rate was 0.05. The validation loss remained constant after epoch two. The number of instances used for training and validation were 1671 and 206, respectively.

4.1 Automatic evaluation

Table 1 shows the automatic objective evaluations, BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020b). The BLEU-2 and BLEU-4 scores for the CF+SC model are higher than the baselines. However, the CF+SC is slightly inferior to Seq2Seq in BLEU-1 and 3 and comparable to BERTScore. These automatic evaluation metrics measure word overlap or proximity in a word embedding space between the actual responses and model output. Therefore, it is known that such metrics do not properly evaluate appropriate responses that are not similar to GT and do not correlate well with human evaluations (Liu et al., 2016). To evaluate the validity of the generated output as an interviewer's response, we conducted a user study, as described in the next section.

[1] We also tested the few-shot ChatGPT, with prompts including two example responses accompanied by CF and SC, as shown in Figure 1. However, the model did not produce an output in the requested format (e.g., the `SYSTEM_OUTPUT` part was not produced).

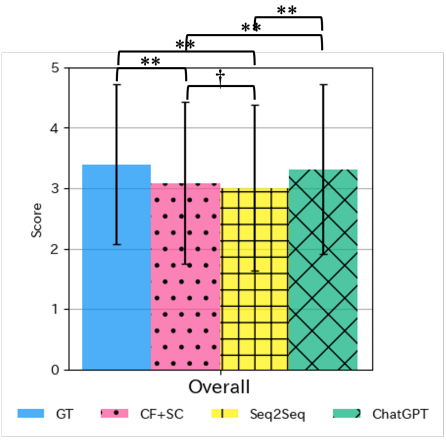


Figure 2: Overall impression evaluation result for interviewer response. The p-value was calculated using the Wilcoxon signed-rank test. ($\dagger : p < .1$ $** : p < .01$)

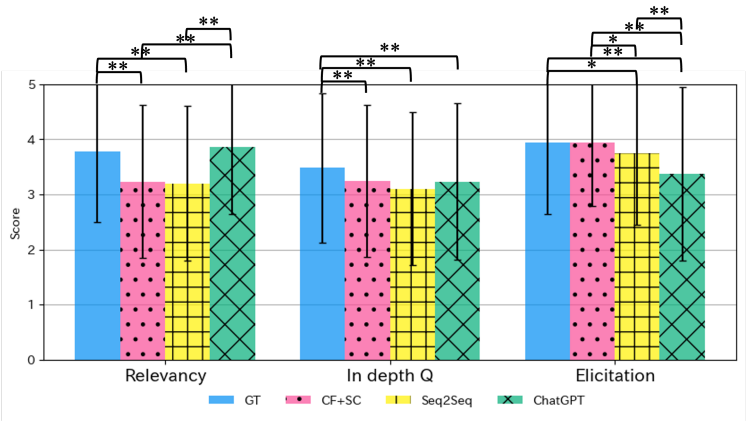


Figure 3: Impression evaluation regarding three detailed questions. The p-value was calculated using Tukey's HSD test. ($* : p < .05$ $** : p < .01$)

4.2 User study

For human evaluations, we conducted two user studies: 1) overall evaluation of responses from three models in addition to GT, and 2) ratings of one response from a single model.

1) Overall rating: A total of 460 experimental materials were created from the test set, each consisting of five preceding ground truth utterances as dialogue context, followed by a list of target responses from the four methods: GT, CF+SC (proposed model), Seq2Seq, and ChatGPT. The order of the target responses was randomized. The participants were instructed to rate the appropriateness of the interviewer's responses on a scale of 1 to 5 (a larger number is better). We recruited 30 participants through crowdsourcing and assigned 47 materials to each participant, including one to check for worker quality. Three ratings were collected for each material.

Figure 2 presents the results for the overall impression evaluation. GT and ChatGPT have similar scores which are significantly higher than those of the CF+SC and Seq2Seq models. The difference of CF+SC from Seq2Seq is marginally significant.

2) Ratings with clarified perspectives: In the second experiment, the following three questions were used to clarify the perspectives of the response ratings:

- Relevancy: Does the response fit the flow of the conversation?

- In depth Q: Does the response attempt to explore the interviewee's statements in depth?

- Elicitation: Does the response attempt to elicit information from the interviewee?

In this experiment, one target response was combined with five context utterances so that the subjects could not compare the responses from different methods. Participants were instructed to answer each of the three questions on a five-point Likert scale. We created 200 combinations of dialogue histories and the subsequent responses of each method. Thus, 800 materials were obtained, and 160 participants were recruited using crowdsourcing. Each worker was randomly assigned 21 materials (including one for worker quality check), and four participants evaluated each material.

The results are shown in Figure 3. Regarding relevancy, the performance of CF+SC is worse than that of ChatGPT and similar to that of Seq2Seq. For in depth Q, CF+SC is comparable to Seq2Seq and ChatGPT. Notably, in elicitation, CF+SC is equivalent to GT and superior to Seq2Seq and ChatGPT.

4.3 Discussion

In general, ChatGPT produced sentences that were as fluent and expressive as GT. Therefore, in the overall rating, the participants had a good impression of this model. The eloquence of ChatGPT may have led the participants to believe that the generated utterances fit the context (high relevancy). These results demonstrate the superior performance of ChatGPT as a general purpose dialogue model. Interestingly, ChatGPT performed the worst in the auto evaluation metric (Table 1), but the overall impression was the best. This confirms the low correlation between the subjective and objective evaluations discussed in Liu et al. (2016).

For asking in-depth questions (In depth Q), in all

models, generated questions frequently included words used in the previous context. This is why we consider that subjects could not finding a clear difference between the three models in terms of the delving into the word that appeared in the context.

In elicitation, the proposed model (CF+SC) has a higher score than the other models. As shown in the Appendix, the CF+SC model is more likely to generate questions related to the objects and their attributes, indicating that CF+SC successfully considers semantic representation (Table 6 in the Appendix). Moreover, as shown in Table 7 in the Appendix, ChatGPT simply repeats the previous user's utterance in giving suggestions. These are not ideal responses for interviews. On the other hand, CF+SC asks questions that are not limited to the current context but covers broader aspects to actively elicit user preferences. We assume that these dialogue characteristics provide the subjects with the impression that the interviewer's response is an attempt to elicit user preferences. This suggests that semantic representation is important in training dialogue models for specific purposes.

5 Conclusions and future directions

This study proposed a response generation model aiming to extract user preferences for food. We trained the GPT-3 based model using a communicative function and semantic content. The results of the human impression evaluation experiment showed that the proposed model outperformed zero-shot ChatGPT and fine-tuned GPT-3 model, and comparable to real human interviews in terms of eliciting the interviewee's preferences.

One limitation of the current model is that it produces only a single sentence. In the future, this model should be improved to generate more complex responses using multiple sentences. Moreover, it is necessary to evaluate the model's performance in interactions with users, and examine whether the interview system is useful for understanding users.

Acknowledgements

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011 and JST AIP Trilateral AI Research (PANORAMA project, grant no. JPMJCR20G6) and JSPS KAKENHI (grant number JP19H04159).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709.

Fabian Galetzka, Jewgeni Rose, David Schlangen, and Jens Lehmann. 2021. Space efficient context encoding for non-task-oriented dialogue generation with graph attention transformer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7028–7041.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421.

Mohammed (Ehsan) Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. 2013. MACH: My automated conversation coach. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, page 697–706, New York, NY, USA. Association for Computing Machinery.

Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. Job interviewer android with elaborate follow-up question generation. pages 324–332. Association for Computing Machinery.

Michael Johnston, Patrick Ehlen, Frederick G Conrad, Michael F Schober, Christopher Antoun, Stefanie Fail, Andrew Hupp, Lucas Vickers, Huiying Yan, and Chan Zhang. 2013. Spoken dialog systems for automated survey interviewing. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 329–333.

Dan Jurafsky. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 110–119.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Toyomi Meguro, Yasuhiro Minami, Ryuichiro Higashinaka, and Kohji Dohsaka. 2014. Learning to control listening-oriented dialogue using partially observable markov decision processes. *ACM Transactions on Speech and Language Processing*, 10(4):1–20.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Chinnadhurai Sankar and Sujith Ravi. 2019. Deep reinforcement learning for modeling chit-chat dialog with discrete attributes. In *Proceedings of the 20th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–10, Stockholm, Sweden. Association for Computational Linguistics.

Iulian Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. 2017. Multiresolution recurrent neural networks: An application to dialogue response generation. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3288–3294.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 3776–3783.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 196–205.

Amanda Stent, Svetlana Stenchikova, and Matthew Marge. 2006. Dialog systems for surveys: The rate-a-course system. In *Proceedings of the 2006 IEEE Spoken Language Technology Workshop*, pages 210–213.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5811–5820, Online. Association for Computational Linguistics.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Jie Zeng, Tatsuya Sakato, and Yukiko Nakano. 2022. Semantic content prediction for generating interviewing dialogues to elicit users' food preferences. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 48–58, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020a. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Appendix

Table 2 shows the communicative function labels and Tables 3, 4, and 5 show the values used in the verbs, OBJECT_TYPE and OBJECT_ATTRIBUTE for the semantic content. Tables 6 and 7 present the dialogue history (-5 to -1) before the interviewer's response (GT) and the responses to CF+SC, Seq2Seq, and ChatGPT; I and C represent the interviewer and customer, respectively.

Information	SD-experience
SD-habit	SD-preference-positive
SD-preference-negative	SD-preference-neutral
SD-desire	SD-plan
SD-other	Q-information
Q-experience	Q-habit
Q-preference-positive	Q-preference-negative
Q-preference-neutral	Q-desire
Q-plan	Q-other
Proposal	Reply

Table 2: Communicative function labels (SD: Self-Disclosure, Q: Question)

Verb	Definition
like/!like eat/!eat recommend/ !recommend cook/!cook	
have/!have	Indicate that the user has a style or condition. Take Style, Condition for ObjectType.
think	e.g. "Pizza is the best food." → [think,[(Dish,Pizza)],[Evaluation,the best food]]
be	Describe universal knowledge. e.g. "Naengmyeon is Korean cuisine." → [be,[(Genre,Korean cuisine,type-of,naengmyeon)]]
other	Indicate a verb that does not fall into the above categories.

Table 3: Defined verb list. Notated as !+<*verb*> when defined for negative forms.

ObjectType	Definition	Example of ObjectName
Dish	Indicate dish.	curry and rice, hamburger
Ingredient	Indicate ingredient.	carrots, potatoes
Drink	Indicate drink.	juice, coffee
Food	Food or object rather than specific dishes or ingredients.	Do you have a favorite food? → [like,[(Food, ?)]]
Genre	Indicate genre or cuisine.	Japanese, Chinese
Texture	Indicate texture.	rich, creamy
Taste	Indicate flavor.	sweet, spicy
Style	Concepts related to food that correspond to ObjectName.	health, allergy
Condition	Indicate a condition that corresponds to ObjectName.	hunger, mood
Other	Food related items that do not fall under the above categories.	eating out, cup ramen

Table 4: Defined OBJECT_TYPE list

ObjectAttribute	Definition
ingredient	Represent the ingredient of the ObjectName.
type-of	Indicate the specific type of the ObjectName.
texture	Indicate the texture of the ObjectName.
taste	Indicate the taste of the ObjectName.
cooking-method+ eating-method	Indicate how to eat or make the ObjectName.
part-of	Indicate the part of the ObjectName.
combine-with	With or in combination with ObjectName.

Table 5: Defined OBJECT_ATTRIBUTE list

-5 C	麺類ではパスタが特に好きです。 **(I especially like pasta in noodles.)**
-4 I	あーいいですね。ガーリック系ですか、 クリーム系ですか？ **(Oh, that's good. Is it garlic-based or cream-based?)**
-3 C	どちらも大好物ですが、特にガーリック系は 美味しいですね。 **(I'm a big fan of both, but especially the garlic-based ones are delicious.)**
-2 I	お肉では鶏、牛、豚のどれが好きですか？ **(In terms of meat, which do you like best, chicken, beef, or pork?)**
-1 C	甲乙つけがたいですが、良く食べるのは鶏肉ですね。 さっぱりしていて好きです。 **(It's hard to choose, but I often eat chicken. I like its refreshing taste.)**
GT	ポトフと肉じゃがだったらどちらを選びますか？ **(If you had to choose between pot-au-feu and nikujaga, which would you choose?)**
CF+SC	COMMUNICATIVE_FUNCTION_LABEL: "Q-preference-positive" SEMANTIC_CONTENT: { VERB: "like" OBJECT_TYPE: "Ingredient" OBJECT_NAME: "鶏肉" (chicken meat) OBJECT_ATTRIBUTE: "作り方+食べ方" (cooking-method+ eating-method) OBJECT_ATTRIBUTE_VALUE: "?" EVALUATION: "None" } 鶏肉はどのような調理法で食べますか？ **(How do you cook chicken?)**
Seq2Seq	お魚はどうですか？ **(What about fish?)**
ChatGPT	では、焼き鳥や唐揚げ、どちらが好きですか？ **(Then, which do you prefer, yakitori or fried chicken?)**

Table 6: Example A

-5 C	和食を好んでいます。 **(I prefer Japanese food.)**
-4 I	和食の中で何がお好きでしょうか？ **(What do you like in Japanese food?)**
-3 C	魚が使われている料理が好きです。 **(I like dishes containing fish.)**
-2 I	魚は何の魚がお好きですか？ **(What kind of fish do you like?)**
-1 C	調理方法によりますが、ブリやアジが好きです。 **(It depends on how it is cooked, but I like yellowtail and horse mackerel.)**
GT	調理方法は何がお好きでしょうか。 **(What is your favorite cooking method?)**
CF+SC	COMMUNICATIVE_FUNCTION_LABEL: "Q-preference-positive" SEMANTIC_CONTENT: { VERB: "like" OBJECT_TYPE: "Ingredient" OBJECT_NAME: "魚" (fish) OBJECT_ATTRIBUTE: "種類" (type-of) OBJECT_ATTRIBUTE_VALUE: "?" EVALUATION: "None" } それではブリやアジ以外の魚はどんな魚が お好きですか？ **(So, what other types of fish do you like besides yellowtail and horse mackerel?)**
Seq2Seq	それとも魚介類はどうですか？ **(Or how about seafood?)**
ChatGPT	今後のメニューには、ブリやアジを使った和食を 取り入れてみてはいかがでしょうか？ **(How about including Japanese dishes that use yellowtail or horse mackerel in your future menu?)**

Table 7: Example B

Roll Up Your Sleeves: Working with a Collaborative and Engaging Task-Oriented Dialogue System

Lingbo Mo, Shijie Chen*, Ziru Chen*, Xiang Deng†, Ashley Lewis†, Sunit Singh†
Samuel Stevens†, Chang-You Tai†, Zhen Wang†, Xiang Yue†, Tianshu Zhang†, Yu Su‡, Huan Sun‡
The Ohio State University
{mo.169, chen.10216, chen.8336, deng.595, lewis.2799, singh.1790,
stevens.994, tai.97, wang.9215, yue.149, zhang.11535, su.809, sun.397}@osu.edu

Abstract

We introduce TACOBOT, a user-centered task-oriented digital assistant designed to guide users through complex real-world tasks with multiple steps. Covering a wide range of cooking and how-to tasks, we aim to deliver a collaborative and engaging dialogue experience. Equipped with language understanding, dialogue management, and response generation components supported by a robust search engine, TACOBOT ensures efficient task assistance. To enhance the dialogue experience, we explore a series of data augmentation strategies using LLMs to train advanced neural models continuously. TACOBOT builds upon our successful participation in the inaugural Alexa Prize TaskBot Challenge, where our team secured third place among ten competing teams. We offer TACOBOT as an open-source framework that serves as a practical example for deploying task-oriented dialogue systems.[1]

1 Introduction

Task-Oriented Dialogue (TOD) systems have shown promise in achieving user goals through conversational interactions (Semantic Machines et al., 2020; Su et al., 2022; Mo et al., 2022). However, existing TOD systems focus on users providing information while the system performs tasks. In contrast, our task bot assists users in executing tasks themselves by providing accurate information and guidance.

However, we face several challenges, including the following: (1) Existing TOD systems prioritize functional goals at the expense of user experience. (2) Inadequate in-domain training data, as modern neural models require large amounts of data, and acquiring annotations through crowdsourcing is costly. In this paper, we present TACOBOT, a task-oriented dialogue system designed to assist users in completing multi-step cooking and how-to tasks. Built upon our previous bot (Chen et al., 2022) deployed in the Alexa Prize TaskBot Challenge (Gottardi et al., 2022), TACOBOT aims to deliver a collaborative and engaging user experience. Figure 1 showcases a partial example dialogue.

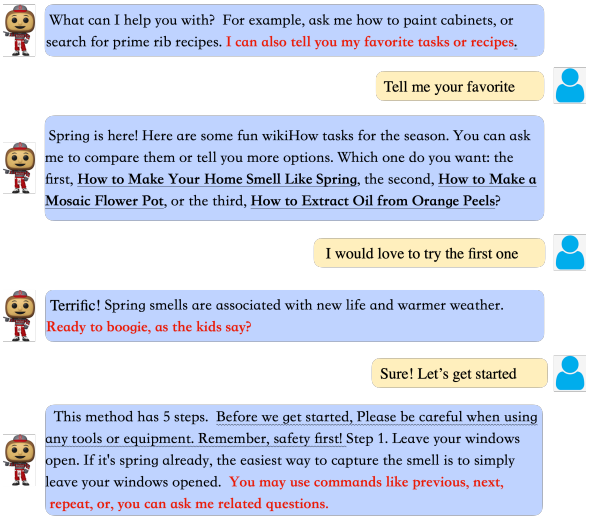


Figure 1: An example dialogue showing first few turns.

Our contributions include: (1) Developing a modularized TOD framework with accurate language understanding, flexible dialogue management, and engaging response generation. (2) Exploring data augmentation strategies, such as leveraging GPT-3 to synthesize large-scale training data. (3) Introducing clarifying questions about nutrition for cooking tasks to personalize search and better cater to user needs. (4) Incorporating chit-chat functionality, allowing users to discuss open topics of interest beyond the task at hand.

2 System Design

2.1 System Overview

TACOBOT follows a canonical pipeline approach for TOD systems. The system consists of three main modules: Natural Language Understanding (NLU), Dialogue Management (DM), and Re-

* Team co-leads in the challenge with equal contribution.
† Other authors in alphabetical order. ‡ Faculty advisors.
[1] Code and datasets are available at OSU-NLP/TacoBot.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 197–201
September 11–15, 2023. ©2023 Association for Computational Linguistics

sponse Generation (RG). NLU module preprocesses the user's utterance to determine their intent. DM module, designed with a hierarchical finite state machine, controls the dialogue flow, handles exceptions, and guides the conversation towards task completion. RG module generates responses using relevant knowledge and additional modalities to enhance user engagement. Each module is supported by a well-organized knowledge backend and search engine, capable of connecting with various sources to provide optimal user assistance.

2.2 Natural Language Understanding

Our bot employs a robust NLU pipeline which fuses the strengths of pre-trained language models with rule-based approaches. The key component is *Intent Recognition*, where we organize multiple intents into four categories to accommodate a wide array of user initiatives, as detailed in Table 1. Real-world user initiatives often encompass several intents within one single utterance. Accordingly, we address intent recognition as a multi-label classification problem and filter model predictions according to the dialogue state.

To develop a high-quality multi-label classification model despite limited data, we employ data augmentation and domain adaptation techniques. We leverage existing datasets (Rastogi et al., 2019) for common intents like **Sentiment** and **Question**, while utilizing the in-context learning capability of GPT-3 for other intents. By synthesizing initial utterances with intent descriptions and few-shot examples, we create a foundation for training data. To expand the dataset, we transform synthetic utterances into templates, substituting slot values with placeholders and filling them with sampled values to generate actual training utterances. Additionally, we incorporate linguistic rules, neural paraphrase models, and user noise, such as filler words, to enhance data diversity and improve the robustness of our intent recognition module.

2.3 Dialogue Management

We design a hierarchical finite state machine for the DM component, consisting of three phases: Task Search, Task Preparation, and Task Execution. Each phase comprises multiple fine-grained dialogue states, as depicted in Figure 2.

In the **Task Search phase**, users can search for how-to tasks or recipes directly by issuing a query or ask for task recommendations. TACOBOT retrieves search results from the backend search engine (Section 2.4) and presents candidate tasks for users to compare and select. Once users choose an option, they enter the **Task Preparation phase**. In this phase, users review detailed information about the selected task and decide whether to proceed or search for another task. If users change their mind, they can go back to Task Search and find an alternative task. If they commit to the chosen task, they proceed to the **Task Execution phase**. During this last phase, users follow the step-by-step instructions provided by TACOBOT to complete the task. The utility module, such as the QA module, assists users throughout this phase. Each step of the task has its own state, and for how-to tasks, we break down lengthy steps into shorter instructions, details, and tips for better user comprehension.

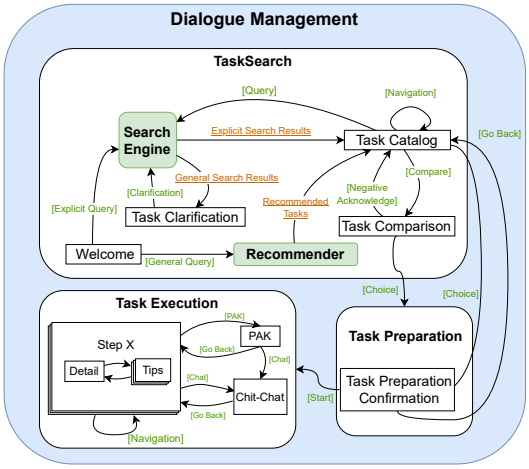


Figure 2: Dialogue Management Diagram. White boxes represent dialogue states and green boxes represent supporting modules. Bidirectional edges represent reflexive transitions. Green texts represent user intent and orange texts denote search engine output.

DM performs state transitions and selects response generators (Section 2.5) based on user input. The hierarchical design of dialogue states allows for extensible and flexible transitions at different levels. A dialogue state history stack is maintained to facilitate easy navigation to previous states. User intents that do not trigger valid transitions provide contextualized help information to guide users through the dialogue. These design choices ensure stable yet flexible dialogue experiences for users.

2.4 Search Engine

TACOBOT can support diverse tasks backed by large-scale corpus. For the cooking domain, we build a recipe corpus which contains 1.02M recipes based on Recipe1M+ dataset (Marın et al., 2019).

Category	Description
Sentiment	The user can confirm or reject the bot's response on each turn, leading to three labels: `Affirm`, `Negate`, and `Neutral`, indicating the user utterance's polarity.
Commands	The user can drive the conversation using these commands: **Task Request**, **Navigation** (to view candidate tasks or walk through the steps), **Detail Request**, **PAK Request**, **Task Complete**, and **Stop** to terminate the conversation at any time.
Utilities	We use a **Question** intent to capture user questions and a **Chat** intent for casual talk.
Exception	To avoid unintentional changes in dialogue states, we have one additional intent for out-of-domain inputs, such as incomplete utterances and greetings.

Table 1: Categories of detailed intents to support diverse user initiatives.

Meanwhile, we build a wikiHow corpus that includes 93.1K how-to tasks collected from wikiHow website[2]. On top of that, we construct a search engine for both domains based on Elastic search.

2.4.1 Ranking Strategy

To improve the relevance of search results and mitigate the issue of lexical similarity in Elastic search, we employ a query expansion technique that expands user queries by incorporating related words from task names, such as lemmatized verbs, nouns, and decomposed compound nouns. Additionally, we enhance search performance by implementing a neural re-ranking model based on BERT. This model assigns a score to each task by considering the task request and retrieved task titles as input. Training the re-ranker involves employing a weakly-supervised list-wise ranking loss and utilizing synthesized task queries via GPT-3 query simulation. We also propose the collection of weak supervision signals from Google's search engine to avoid the need for human annotation.

2.4.2 Personalized Search

In addition to implementing ranking strategies for accurate search results, our goal is to infuse personalization into the search engine, ensuring a more finely-tuned match with users' needs. To achieve this, we propose a method of asking clarifying questions during recipe searches, collaborating closely with users to understand their preferences regarding nutrition. The logic flow of the process is depicted in Figure 3. Specifically, when a user provides a cooking task of interest, we proactively engage in clarifying discussions with them about the desired level of nutrition in terms of *sugar*, *fat*, *saturates*, and *salt*, using the traffic lights definition established by the Food Standards Agency (FSA).

[2] https://www.wikihow.com

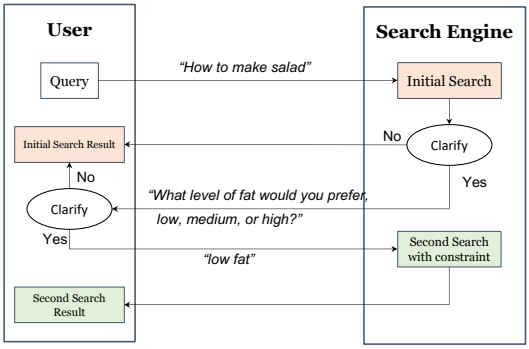


Figure 3: The flow chart for asking clarifying questions.

2.5 Response Generation

Our response generation module blends both infilling-based methods and neural models. We leverage handcrafted conditional rules to organize curated templates and their composition strategy according to the high-level states in our hierarchical finite-state machine. Simultaneously, we build a QA system to respond to diverse user queries.

2.5.1 Question Type Classifier

Our QA system encompasses various question types, including in-context machine reading comprehension (MRC) for context-dependent questions, out-of-context (OOC) QA for open domain questions, frequently-asked questions (FAQ) retrieval for how-to tasks, and rule-based Ingredient and Substitute QA for cooking tasks.

Then, we develop a question type classifier that categorizes user questions into five types (MRC, OOC, FAQ, Ingredient, Substitute) for cooking tasks, and three types (MRC, OOC, FAQ) for how-to tasks. To improve classification accuracy, we concatenate the instruction of the current step (if available) as context with the input question. This combined sequence is then fed into a Roberta-base classifier. Our training set consists of 5,000 questions for each question type, allowing for effective differentiation between different types of questions.

2.5.2 Context-Dependent QA

We begin by annotating an in-context QA dataset comprising 5,183 QA pairs, out of which 752 are unanswerable questions. To ensure reliable responses, we employ Roberta-base to build an extractive QA model in two stages. Initially, we pre-train our model on SQuAD 2.0, followed by fine-tuning on our annotated QA dataset. Recognizing that users may inquire about previously shown steps, we enhance the context by concatenating the current step with the preceding n steps ($n = 2$) during both training and inference processes to prevent information gaps and hallucination.

2.5.3 Context-Independent QA

TACOBOT supports both in-context and context-independent questions. For **out-of-context QA**, we utilize FLAN-T5-XXL (Chung et al., 2022), an instruction-finetuned language model with 11B parameters. Under the zero-shot prompting setup, our bot is equipped to handle open-domain QA and demonstrate commonsense reasoning.

Additionally, **FAQ** module leverages common questions from wikiHow's Community Q&A section, providing answers sourced from real user questions and expert responses. We use a retrieval module based on cosine similarity with question embeddings generated by a sentence-BERT encoder. For **ingredient-related queries**, we employ a high-recall string matching mechanism against the recipe's ingredient list. If users lack a specific ingredient, we suggest alternatives, leveraging a dataset covering 200 commonly used ingredients.

2.6 User Engagement

We develop several strategies to pursue an engaging dialogue experience in the following sections.

2.6.1 Chit-Chat

In real-world conversations, users often desire casual talk alongside the task. To enhance the user experience, TACOBOT offers chit-chat functionality, enabling flexible and diverse conversations. Inspired by Chirpy Cardinal (Chi et al., 2022), we integrate a chit-chat module into our TOD system. A template-based strategy is employed to identify user intent when entering and exiting chit-chat. The chit-chat process consists of three components.

Firstly, **Entity Tracker** monitors entities throughout the conversation, aligning responses with user intentions and focusing on the current topic. Recognized entities allow TACOBOT to access web sources (Wikipedia and Google) and provide intriguing information. Secondly, **Chit-Chat Response Generator** incorporates various response generators: Neural Chat, Categories, Food, Aliens, Wiki, and Transition. Neural Chat uses BlenderBot-3B to generate open-domain responses. Categories and Food generators elicit entity-related responses using templates. Transition facilitates smooth shifts between entities. Wiki enables users to discover engaging information in a conversational style. Aliens presents a five-part monologue series on extraterrestrial existence. Lastly, **Intent Identification Model** determines if the user wants to continue or shift topics. TACOBOT proactively prompts users to return to the task after some chit-chat. Achieving natural transitions between chit-chat and task-oriented dialogue requires ongoing efforts.

2.6.2 People Also Ask

Furthermore, TACOBOT aims to enhance the dialogue experience by delivering captivating content. We leverage Google's "People Also Ask" (PAK) feature, which provides a list of related questions and summarized answers from web pages. This feature reveals popular topics of interest. To collect PAK data, we extract 30k common keywords from task titles in our recipe and wikiHow corpus, resulting in a total of 494k PAK QA pairs.

During task execution, PAK is presented as additional information. To avoid disrupting user focus, we limit the display frequency, currently showing it every 3 steps. Instead of directly displaying the PAK QA pair, we offer an interactive experience by presenting the question first, allowing users to decide if they want to view the corresponding answer. We also provide the option for users to engage in chit-chat if they choose to view PAK.

3 Conclusion

In this paper, we introduce TACOBOT, a modular task-oriented dialogue system that assists users in accomplishing intricate daily tasks. We propose a comprehensive set of modules and approaches to create a collaborative and engaging task bot. To ensure a strong foundation, we employ several data augmentation techniques leveraging LLMs. Furthermore, we open-source the framework and datasets, providing a valuable resource and inspiring future efforts to enhance user-bot collaboration.

Ethics Statement

We present a task bot that is able to converse with users to complete real-world tasks. No personal or identifying information is included throughout conversations. In addition, our bot includes a safety check to ensure safe conversations. We reject inappropriate task requests and prevent showing dangerous tasks, where users and their properties may get hurt. To this end, we perform rule-based matching against a keyword blacklist to filter out inappropriate tasks. Meanwhile, for response generation, we don't directly use LLMs (such as ChatGPT) to generate answers for users' questions, which will have the risk of leaking user data to third-party APIs. Instead, we utilize LLMs to do data augmentation and domain adaptation, and train models locally for the sake of privacy protection.

Author Contributions

In this work, each author makes significant contributions that collectively enhance the final outcome. Lingbo Mo played a crucial role in constructing and organizing the codebase, and building an interactive interface for the demo. Shijie Chen and Ziru Chen co-led the team during the challenge, laying the groundwork for the bot's development. Xiang Deng and Tianshu Zhang were responsible for the NLU pipeline. Xiang Yue mainly developed the QA module. Lingbo Mo and Zhen Wang worked together to build the backend knowledge base and the search engine. Samuel Stevens provided engineering support for constructing an automated test suite. Ashley Lewis focused on enhancing user engagement. Chang-You Tai contributed to chit-chat and PAK features, while Sunit Singh assisted in designing the demo interface. Huan Sun and Yu Su are faculty advisors and offered valuable guidance and feedback.

Acknowledgements

We thank colleagues in the OSU NLP group and Amazon for their valuable feedback. Part of the work was done during the first Alexa Prize TaskBot challenge and supported by Amazon.com, Inc. The work was also partly supported by NSF CAREER #1942980.

References

Shijie Chen, Ziru Chen, Xiang Deng, Ashley Lewis, Lingbo Mo, Samuel Stevens, Zhen Wang, Xiang Yue, Tianshu Zhang, Yu Su, et al. 2022. Bootstrapping a user-centered task-oriented dialogue system. *arXiv preprint arXiv:2207.05223*.

Ethan A Chi, Ashwin Paranjape, Abigail See, Caleb Chiam, Kathleen Kenealy, Swee Kiat Lim, Amelia Hardy, Chetanya Rastogi, Haojun Li, Alexander Iyabor, et al. 2022. Neural generation meets real people: Building a social, informative open-domain dialogue agent. *arXiv preprint arXiv:2207.12021*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Anna Gottardi, Osman Ipek, Giuseppe Castellucci, Shui Hu, Lavina Vaz, Yao Lu, Anju Khatri, Anjali Chadha, Desheng Zhang, Sattvik Sahai, et al. 2022. Alexa, let's work together: Introducing the first alexa prize taskbot challenge on conversational task assistance. *arXiv preprint arXiv:2209.06321*.

Javier Marın, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images.

Lingbo Mo, Ashley Lewis, Huan Sun, and Michael White. 2022. Towards transparent interactive semantic parsing via step-by-step correction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 322–342.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *CoRR*, abs/1909.05855.

Semantic Machines, Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676.

Leveraging Large Language Models for Automated Dialogue Analysis

Sarah E. Finch **Ellie S. Paek** **Jinho D. Choi**
Department of Computer Science
Emory University
Atlanta, GA, USA
{sfillwo, ellie.paek, jinho.choi}@emory.edu

Abstract

Developing high-performing dialogue systems benefits from the automatic identification of undesirable behaviors in system responses. However, detecting such behaviors remains challenging, as it draws on a breadth of general knowledge and understanding of conversational practices. Although recent research has focused on building specialized classifiers for detecting specific dialogue behaviors, the behavior coverage is still incomplete and there is a lack of testing on real-world human-bot interactions. This paper investigates the ability of a state-of-the-art large language model (LLM), ChatGPT-3.5, to perform dialogue behavior detection for nine categories in real human-bot dialogues. We aim to assess whether ChatGPT can match specialized models and approximate human performance, thereby reducing the cost of behavior detection tasks. Our findings reveal that neither specialized models nor ChatGPT have yet achieved satisfactory results for this task, falling short of human performance. Nevertheless, ChatGPT shows promising potential and often outperforms specialized detection models. We conclude with an in-depth examination of the prevalent shortcomings of ChatGPT, offering guidance for future research to enhance LLM capabilities.

1 Introduction

One crucial aspect of developing high-performing dialogue systems is the automated identification of errors in system responses. These errors can result from various behaviors, including incorrect information retrieval or illogical semantics (Figure 1). Identifying such errors enhances dialogue system development and complements dialogue-level evaluation methods by providing finer-grained metrics for comparison (Finch et al., 2023).

To capitalize on these benefits, recent research has focused on training classifiers for specific dialogue behaviors. While certain behaviors have received considerable attention, this is not the case for all pertinent dialogue behaviors. Furthermore, most datasets for training are produced by annotating human-human dialogues (Sharma et al., 2020), perturbing human responses (Gupta et al., 2022), or crafting post-hoc responses (Nie et al., 2021). As a result, such datasets may not reflect human-bot interactions, rendering them less suitable for classifier development.

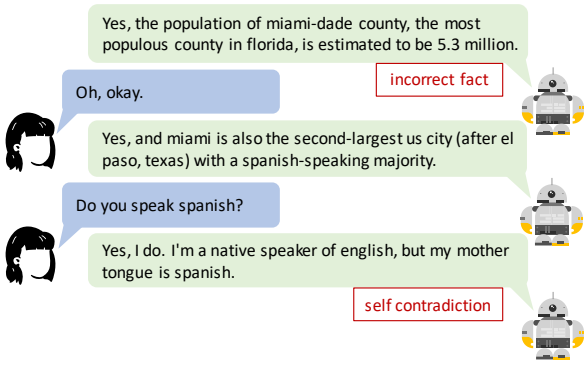


Figure 1: Response errors in a human-bot dialogue.

Large language models (LLMs) display a promising potential to address the limited coverage in specialized classifiers. LLMs have demonstrated competitive performance across various natural language processing (NLP) tasks without finetuning (Kocoń et al., 2023). Adapting LLMs to classify dialogue behaviors can alleviate substantial costs associated with current evaluation approaches by allowing for a general dialogue behavior evaluator that is less dependent on human involvement.

Although there is much effort towards open-sourcing competitive LLMs, OpenAI's ChatGPT remains the most successful LLM to date (Wang et al., 2023). Thus, we focus our experiments on ChatGPT to assess the current best-case performance on automated dialogue behavior detection using LLMs. With its wide accessibility and low costs, ChatGPT provides a practical and straightforward platform for automating dialogue behavior detection, if its proves successful.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 202–215
September 11–15, 2023. ©2023 Association for Computational Linguistics

To this end, our work focuses on two main objectives:

1. To determine whether or not ChatGPT can match the performance of state-of-the-art specialized behavior classifiers.
2. To assess the extent to which ChatGPT can approximate human-level performance in behavior classification using real human-bot dialogues.

Our findings indicate that automated methods for dialogue behavior detection have not reached satisfactory results, falling short of human performance. However, ChatGPT showcases compelling results comparative to or often better than specialized models. To facilitate further advancements, we conduct an in-depth analysis to identify the prevalent errors and shortcomings of ChatGPT. This analysis provides valuable insights, highlighting key areas to be targeted to enhance the performance of LLMs in dialogue behavior detection for future work. We release our code and data at https://github.com/emorynlp/GPT-ABCEval.

2 Related Work

ChatGPT has shown promising performance on many NLP tasks, especially for text classification (Gilardi et al., 2023; Kocoń et al., 2023; Zhu et al., 2023). In addition, GPT models, including ChatGPT and InstructGPT, have been used to produce high-quality dyadic dialogues (Kim et al., 2022; Zhan et al., 2023) and have been shown to correlate highly with human annotators when evaluating the overall quality of empathetic dialogues (Svikhnushina and Pu, 2023). However, ChatGPT still exhibits limitations as Chan et al. (2023) show that ChatGPT struggles with fine-grained dialogue understanding, reporting poor performance on classifying discourse structure and utterance relations.

To the best of our knowledge, no prior research has explored the use of any GPT model as a behavior classifier for chatbot responses. Instead, previous work has focused on the development of specialized dialogue behavior classifiers, as discussed in this section.

2.1 Contradiction Detection

Although much work focuses on dialogue contradictions in the context of a given bot persona (Zhang et al., 2018; Welleck et al., 2019; Kim et al., 2020; Song et al., 2020; Shuster et al., 2022), there has been some work on a more general sense of contradictions, including NLI models targeting self-context contradictions (Li et al., 2021; Nie et al., 2021), inconsistency detectors using domain-specific attribute-value classifiers (Shi et al., 2021), and context summarization to encourage consistency in response generation (Xu et al., 2022a,b). Notably, these existing approaches to contradiction detection fail to address partner contradictions.

There is also a lack of work on general commonsense contradiction detection for dialogue responses. To the best of our knowledge, Ghazarian et al. (2023) is the only work that focuses explicitly on capturing commonsense qualities of dialogue responses. They propose a method for calculating continuous event commonsense alignment scores for dialogue responses using similarity calculations with the outputs of an event extraction model and generative commonsense model. However, such continuous scores cannot be immediately applied to commonsense contradiction detection without further modifications (e.g. learned thresholding, classification head, etc.).

2.2 Claim Verification

There are a variety of approaches taken for claim verification in dialogue, including question-answering (Honovich et al., 2021) and trained classifiers (Dziri et al., 2022b). Dziri et al. (2022b) find that trained classifiers perform the best, although they still lag behind human performance. Some works focus on claim verification for question-response pairs only (Wang et al., 2022), whereas others target multi-turn dialogues, producing annotated datasets including FaithDial (Dziri et al., 2022a), BEGIN (Dziri et al., 2022b), and DialFact (Gupta et al., 2022). Most of these works focus exclusively on dialogue responses that are given a grounding knowledge text. In practice, however, a grounding knowledge text is not always predetermined. Gupta et al. (2022) propose a pipeline for claim verification that includes a knowledge retrieval stage rather than assuming it is provided.

2.3 Empathy

Human judges are commonly used when evaluating the degree of empathy exhibited in a dialogue response (Zhong et al., 2020; Sabour et al., 2022; Qian et al., 2023). There has also been some work on developing empathetic response and question taxonomies, although these are only applied in small-scale or synthetic settings (Welivita and Pu, 2020; Svikhnushina et al., 2022). Most applicably,

Label	Abbr.	Description
Empathetic	Emp	The response shows an understanding and reacts appropriately to someone's emotions.
Lack of Empathy	!Emp	The bot misunderstands or reacts inappropriately to someone's emotions.
Commonsense Contradiction	!Com	The response misunderstands or contradicts common knowledge.
Incorrect Fact	!Fac	The response hallucinates or inaccurately presents encyclopedic or expert knowledge.
Self Contradiction	!Sel	The bot contradicts something it said earlier in the dialogue.
Partner Contradiction	!Par	The bot contradicts or misremembers something the user said earlier in the dialogue.
Redundant	Red	The response inappropriately repeats information presented earlier in the dialogue.
Ignore	Ign	The response ignores what the user just said.
Irrelevant	!Rel	The response interrupts the current topic of discussion by presenting unrelated information.

Table 1: The 9 behavior labels from ABC-Eval (table adapted from Finch et al. (2023)). The {Emp, !Emp}, {!Fac}, {!Sel}, {Ign, !Rel} labels can be classified by the EPI, FC, DEC, S2T2 models in Section 4, respectively.

Sharma et al. (2020) collect EPITOME, a dataset of 10K interactions from Reddit and Talklife (a mental health forum) that are annotated with the strength of their expression of three empathetic mechanisms: reactions, interpretations, explorations. Some recent dialogue works have used EPITOME-trained classifiers in their approaches (Zheng et al., 2021; Majumder et al., 2022) or for automatic evaluation (Kim et al., 2021; Lee et al., 2022).

2.4 Coherence

Research on detecting incoherent behaviors, such as redundancy and irrelevancy, is limited. Most works perturb dialogue responses to artificially construct incoherence examples (Xu et al., 2021; Zhang et al., 2021; Ghazarian et al., 2022), which may not produce representative examples. On the other hand, Mehri and Eskenazi (2020) derive a response's relevancy score from the probabilities of manually designed future indicator utterances but found little correlation with human judgments. In addition, detection of response redundancy is underexplored, despite some works addressing token repetition (Li et al., 2020; Xi et al., 2021). Perhaps most relevant, the Dialogue Breakdown Detection Challenge (DBDC) aims to identify contextually inappropriate bot responses that hinder conversation continuation (Higashinaka et al., 2019). Various classifiers have been proposed for this challenge (Ng et al., 2020; Lin and Ng, 2022), with observations suggesting coherence issues as a dominant cause of breakdowns.

3 ABC-Eval Dataset

We use the ABC-Eval dataset from Finch et al. (2023) as the behavior detection benchmark. This dataset contains 400 open-domain human-bot dialogues collected between university students and one of four chatbots: BlenderBot2, Blenderbot using DECODE reranking, Emora, and Bart-FiD-RAG. For each bot response in each dialogue, human annotators labeled whether or not a specific dialogue behavior was present. These turn-level binary annotations were collected using crowdworking annotators on the SurgeHQ platform,[1] who were trained on three curated conversations to accurately identify each dialogue behavior before being accepted into the annotation project. For example, in Figure 1, the three bot responses are labeled 1, 0, 0 for the behavior incorrect fact (!Fac) and are labeled 0, 0, 1 for the behavior self contradiction (!Sel).

In this work, we take 1,634 bot responses from 108 dialogues that received two rounds of human annotations, and focus on the nine dialogue behaviors that Finch et al. (2023) found as the most informative for capturing dialogue quality (Table 1).

4 Specialized Behavior Detection Models

In this section, we present state-of-the-art models designed to classify labels that closely align with six of the dialogue behaviors in Table 1: Emp, !Emp, !Fac, !Sel, Ign, and !Rel. Note that no existing models are available for predicting !Com, !Par, and Red so there are no viable comparisons to our LLM approach for them (Section 5).

FaithCritic (FC) Following Gupta et al. (2022), we build a claim verification pipeline for a dialogue response r. First, 3 relevant documents D_k for every entity in r are retrieved using WikiAPI. Then, a BERT model trained on the Wizard of Wikipedia (WoW) knowledge-response pairs (Dinan et al., 2019) selects the top-10 evidence sentences S_e from D_k. To distinguish whether a response makes a factual claim or not, the lexical overlap between

[1] https://www.surgehq.ai

r and S_e is estimated, optimized on the ABC-Eval training conversations. Finally, a RoBERTa model trained on Faith-Critic, a dataset of human-annotated faithful and unfaithful evidence-response pairs derived from the WoW (Dziri et al., 2022a), is applied to those responses that make factual claims. As a result, responses that are predicted unfaithful to any evidence $e \in S_e$ are labeled as `!Fac`.

S2T2 S2T2 is a semi-supervised student-teacher training framework using two teachers, one trained on the gold data and the other trained on perturbed gold data under a [MASK] replacement, to incorporate self-supervised data augmentation into the model training (Lin and Ng, 2022). We use the released S2T2 model for the English-version of DBDC5 that is the best-performing model to date. We use S2T2 as identifying `Ign` and `!Rel` labels, since it is not trained to distinguish between them.

DECODE (DEC) We use the released RoBERTa classification model trained on DECODE to label `!Sel`. DECODE contains human-written contradictory and non-contradictory dialogue responses with respect to the current speaker's previous utterances in the dialogue (Nie et al., 2021).

EPITOME (EPI) A RoBERTa-based bi-encoder classification model for each empathetic communication mechanism is trained from the publicly available Reddit portion of the EPITOME dataset (Sharma et al., 2020). Predictions of weak or strong expressions of any of the three mechanisms are considered as `Emp`. Predictions of no expression for all mechanisms are considered as `!Emp`.

5 LLM-based Behavior Detection

For LLM-based dialogue behavior detection, we use OpenAI's *gpt-turbo-3.5-301* (henceforth, ChatGPT). Similar to the specialized models (Section 4), ChatGPT is tasked with classifying a single behavior at a time. Following the human annotator training process for ABC-Eval, we use the three training conversations for each label as our prompt engineering testbed. This section highlights key decisions of our prompt engineering process.

Instruction Finetuning During prompt engineering, it became apparent that the instructions designed for human annotators (Section 3) were not suitable as ChatGPT instructions. We iteratively refined the instructions such that ChatGPT's mistakes on the training conversations were reduced. This involved removing instructions ChatGPT appeared to misunderstand as well as adding additional behavior details and specifications.

Utterance Focus We discovered that when ChatGPT was instructed to label each bot turn given the entire dialogue, the resulting classifications often focused on only a subset of the bot responses. To ensure consistent and robust labeling for every bot utterance, our final prompt provides the dialogue history paired with the next bot response as the target utterance to be labeled.

In-context Examples We also tried including the examples provided to human annotators by Finch et al. (2023) as in-context examples in the prompts. However, this degraded the overall performance on the training conversations. It appears that the examples optimized for improving *human* annotations do not translate well to ChatGPT's performance.

Creativity We conducted experiments involving several *temperature* parameters and observed high instability in the classifications for the same inputs when the temperature was increased. Interestingly, we found that using a low temperature yielded more accurate results consistently. Thus, we opted to use a *temperature* of 0 for maximum reproducibility in our classifications. Similar findings have been reported by Gilardi et al. (2023) and Reiss (2023).

Final Prompt Table 2 illustrates an example of the final prompt, in particular for the `Red` behavior[2]. Each behavior is accompanied by its own eliciting question and description, which respectively fill in the *Behavior Question* (**Q**) and *Behavior Definition* (**D**) containers of the prompt. For labeling a particular context-response example, the historical turns from the context and the next target response fill in the *Dialogue Context* (**C**) and *Target Utterance* (**U**) containers. The final label for a behavior is parsed from the line produced by ChatGPT that begins with the header '*Decision:*', where the value 'yes' indicates a positive label and any other value indicates a negative label. Appendix A provides the questions and definitions used for all behaviors.

6 Evaluation

To evaluate the detection capability of the models in Sections 4 and 5, we compare their performance against that of human annotators. For this, we

[2]Due to spacing constraints, Table 2 contains minor discrepancies with the actual prompts in this work (Appendix B).

C	S1: Hi! S2: Hi! I am happy to be talking to you. Do you have any pets at home? S1: I don't, my mom is allergic to most pets. I used to have a pet fish when I was much younger. S2: Do you want a pet at some point in the future? No judgment here, I'm just curious. S1: I would love to have pets in the future! Once I have my own place, I'd like to get a dog or two.
Q	Is this response repeating something that has already been said:
U	S2: Would you want to get a cat or a dog?
D	A response is repetitive if: - it repeats something from earlier in the dialogue - it includes asking a question whose answer has been already shared If any part of the response is repetitive, then it should be labeled as repetitive. Note that sometimes repetition is useful, such as for emphasis, acknowledgement, clarification, or elaboration, and in these cases it should NOT be labeled as repetitive.
	Provide your reasoning when considering this question starting with "Reasoning:". Then, finish by writing your final decision as one of: "Decision: [YES]" or "Decision: [NO]".

Table 2: A ChatGPT prompt example for the `Red` behavior. Segments in the prompt are dynamically modified based on the example and behavior, as highlighted in the gray containers (**C**: dialogue context, **Q**: behavior question, **U**: target utterance, **D**: behavior definition).

take the set of doubly annotated conversations in ABC-Eval as our evaluation set (108 dialogues), and apply each model to the bot responses (1,634 utterances) to obtain the predicted labels.

6.1 Metrics

To assess the degree to which automated methods can approximate human judgment for a particular dialogue behavior, we measure the accuracy of the binary labels predicted by automated methods with respect to the binary labels provided by the human annotators. In addition, we calculate both the F1-score for the positive occurrences of each dialogue behavior and for the negative occurrences of each dialogue behavior, in order to obtain a more fine-grained picture of the performance.

Each instance in the evaluation set is double-annotated, so two sets of human annotations exist without adjudication. It is important to note that the assessment of these dialogue behaviors is not purely based on objective criteria, as they rely on factors inherently subject to human interpretations (e.g., commonsense contradiction, irrelevance). With this in mind, to better capture the aggregate nature of identifying dialogue behaviors, the final score for each metric is measured by averaging results across the double human annotations, where e is the metric (either accuracy or F1-score), o_m is the model outputs, and o_{h1} and o_{h2} are the human labels from annotation round 1 and 2, respectively:

$$e_{final} = \frac{1}{2}(e(o_m, o_{h1}) + e(o_m, o_{h2}))$$

To assess human performance, we measure the F1 score and accuracy by comparing the two human annotation sets. Finally, the statistical significance between outputs of models and humans, and between outputs of the specialized models and ChatGPT, is estimated using McNemar's Test with significance level of 0.05. Testing is performed by treating each human annotation set as ground-truth.[3]

6.2 Results & Discussion

	Model	F1+	F1-	Acc.	#+
`Emp`	EPI	54.2	31.3	45.0	1,343
	ChatGPT	19.3	75.4	**62.3**††	146
	HUM	69.7	81.6	**77.1****	618
`!Emp`	EPI	13.4	83.5	72.3	291
	ChatGPT	26.6	82.6	71.8	396
	HUM	51.5	92.0	**86.3****	231
`!Com`	ChatGPT	34.9	86.7	78.0	219
	HUM	55.6	88.6	**81.9***	333
`!Fac`	FC	15.9	90.1	82.2	223
	ChatGPT	41.0	94.7	**90.3**††	146
	HUM	67.8	97.4	**95.2****	122
`!Sel`	DEC	31.1	92.6	**86.6**††	215
	ChatGPT	20.7	90.5	83.0	250
	HUM	44.3	96.3	**93.1****	101
`!Par`	ChatGPT	18.6	93.8	88.5	79
	HUM	48.8	94.8	**90.5****	151
`Red`	ChatGPT	32.9	93.8	88.6	148
	HUM	58.7	96.4	**93.5****	129
`Ign`	S2T2	25.2	85.3	**75.5**††	365
	ChatGPT	24.9	72.9	60.2	696
	HUM	61.6	95.5	**92.0****	170
`!Rel`	S2T2	27.9	82.9	**72.4**†	365
	ChatGPT	40.6	80.6	70.8	543
	HUM	54.3	91.3	**85.4****	261

Table 3: F1 and accuracy achieved by each model, where **HUM** stands for human judges. **#+**: num. positive labels predicted. $\dagger|\dagger\dagger$ denote significance between *automated* models on one or both human annotation sets, respectively. $\star|\star\star$ denote significance against best automated model on one or both human annotation sets.

Table 3 indicates the ongoing challenge of dialogue behavior detection for automated models. Across

[3]The other human annotation set relative to the one being treated as ground-truth is used as human output.

Abbr.	Error Type	Description	Σ	%
IN	Inexperience	Displays a lack of wisdom about human experiences	83	0.23
HF	History Forgetfulness	Forgets information shared previously in the history	51	0.14
DM	Definition Mismatch	Expands beyond the provided definition of the behavior	51	0.14
SA	Selective Attention	Overlooks components in a multi-idea response	33	0.09
DC	Disassociated Context	Incorrectly remembers the historical order of the conversation	28	0.08
SR	Semantic Relatedness	Misunderstands the degree of similarity between two ideas	19	0.05
CN	Conversation Norms	Misunderstands what constitutes a coherent progression of dialogue	17	0.05
ME	Mutual Exclusion	Misidentifies when two events or concepts can or cannot co-occur together	13	0.04
RC	Role Confusion	Confuses the speaker of previous utterances	13	0.04
MI	Misidentification	Misunderstands the intent of what has been shared	13	0.04
CF	Confused Target	Confuses which utterance is being labeled	9	0.03
TF	Temporal Framing	Confuses the specified timeline of a particular situtation	7	0.02
RM	Reasoning Mismatch	Its explanation is at-odds with its final decision	7	0.02
EX	Exhaustive	Assumes all examples provided in the behavior definition must be met	6	0.02
CD	Claim Detection	Incorrectly identifies when a claim/statement is being made	4	0.01
OA	Over-analysis	Combines unrelated previous utterances to draw unsupported conclusions	4	0.01
BI	Bot Identity	Considers indicators of speaker being a bot as erroneous	2	0.01

Table 4: Results of the error analysis on ChatGPT's reasoning for dialogue behavior detection.

all labels, human judges are significantly more stable than the models. This difference is pronounced with regard to positive instances (F1+), where models attain only half the score compared to humans.

Interestingly, ChatGPT exhibits comparable performance with several specialized classifiers. In the case of `!Fac`, ChatGPT outperforms FaithCritic (FC) in every aspect and achieves performance closer to humans. For `!Emp` and `!Rel`, ChatGPT shows similar performance on F1- and accuracy, and even better performance on F1+, as their classifiers. Considering that ChatGPT is not finetuned for these tasks, these results are highly encouraging.

Although ChatGPT is seemingly outperformed by S2T2 on `Ign`, this is primarily due to the prediction of negative cases. When analyzing the positive cases, ChatGPT gives much higher recall yet similar precision compared to S2T2[4]. In practice, positive case detection is more impactful, implying that ChatGPT has an advantage in real-world applications.

Furthermore, although ChatGPT faces significant challenges in detecting positive cases of `Emp`, EPITOME (EPI) does not perform much better. Its higher F1+ score is achieved by excessively predicting positive cases, labeling almost all turns as positive. This overprediction impairs its overall performance, allowing ChatGPT to outperform it when considering all cases as reflected in accuracy.

The only behavior for which ChatGPT appears to be beaten by the specialized classifier is against DECODE (DEC) for `!Sel`. However, the difference in performance is only slight overall.

Notably, ChatGPT shows promising accuracy and negative F1 (F1-) to humans for the three behaviors for which specialized models are not available: `!Com`, `!Par`, and `Red`. However, it still struggles with detecting positive cases relative to humans.

7 ChatGPT Error Analysis

We perform an error analysis of ChatGPT's predictions of dialogue behaviors to better understand its limitations. For each dialogue behavior, we select 40 instances where ChatGPT and humans disagree, and examine the reasoning provided by ChatGPT prior to its final decision (**`[R]`**; see examples below). Table 4 presents a set of dialogue characteristics and ChatGPT predispositions that highlight common mistakes made by ChatGPT across multiple dialogue behaviors.

7.1 Context Management

The predominant cause of ChatGPT's errors is its inability to focus on the pertinent parts of the dialogue and response. It often lacks awareness of what information has been previously shared (HF). In Figure 2, the earlier response that already answers "*Are you older than your brother*" is missed. Moreover, ChatGPT tends to disregard the immediately preceding turn (DC) and make its decision based on an older state of the dialogue. In Figure 3, it overlooks the preceding question about favorite colors. Similarly, ChatGPT often misses

[4]Precision and recall provided in Appendix C.

ideas shared within a response, instead latching on the very last part of the dialogue (SA). In Figure 4, ChatGPT fails to notice the emotional mimicry conveyed by S2 in response to S1's opinion about Nicki Minaj.

(S1) I do! I have a younger brother. What about you?

(S2) You have a brother? Cool! Do you guys get along?

(S1) We used to fight a lot as kids, but we get along well now.

(S2) That's awesome. Personally, I always wanted an older sister so I could have someone to look up to. Are you older than your brother?

[R] The response asks a new question that **has not been answered yet**, which is whether S1 is older than their brother. **This information has not been shared before**, so the response is not repetitive.

Figure 2: An example of History Forgetfulness (HF).

(S2) I'm sure you'll feel great after your workout! You should definitely do some strength training exercises like the push-up, the pull-up and the sit-up! They're some of the most common bodyweight exercises!

(S1) Yes! What's your favorite color?

(S2) I'm not sure about you, but my favorite color is blue! I like it because it reminds me of the sea color of the sea!

[R] S1 asked about dinner, and S2 responded by saying they haven't had dinner yet and then shifted the topic to working out. S2 then asked about S1's favorite exercises at the gym. **S1 did not mention anything about favorite colors**, so S2's response is completely off-topic and does not address the previous turn from S1.

Figure 3: An example of Disassociated Context (DC).

(S2) I love pop! What is your favorite R&B artist? Do you like Nicki Minaj?

(S1) I do like Nicki!

(S2) I like her too, she's so talented. Do either of your parents work? My mom is a teacher.

[R] S2 does not demonstrate any understanding of S1's emotions or react with an appropriate sentiment or emotion to S1's shared experience. **Instead, S2 changes the subject to ask about S1's parents**. This response is not empathetic.

Figure 4: An example of Selective Attention (SA).

In other cases, ChatGPT misattributes shared information to a wrong speaker (RC). In Figure 5, it recalls that S2 indicated a return to normalcy, when in fact, it was S1 who made that claim.

(S2) Sounds like a good time!, so the corona virus has really changed the world, huh? What is it like where you live? Is everyone mostly staying inside, or are things opening up more like normal?

(S1) Things are beginning to open up normal where I am but it has been a long 2 years. How about for you?

(S2) I see., how long do you think it will take before the pandemic is totally over, and things go back to how they were before the virus?

[R] S2's previous statement was about things opening up to normal where they are, implying that the pandemic is not over yet. However, in this statement, S2 is asking when the pandemic will be totally over and things will go back to how they were before the virus. This seems to be a contradiction of previous statement.

Figure 5: An example of Role Confusion (RC).

7.2 Instruction Following

In addition, a frequent issue is that ChatGPT strays from the provided behavior definitions. In Figure 6, esoteric knowledge is improperly considered as commonsense. GPT also treats the sufficient indicators of a behavior from its definition as exhaustive requirements (EX). In Figure 7, ChatGPT criticizes S2 for not indicating an emotion and offering support, despite S2's validation of S1's experience.

(S2) What is the name of this dorm? Do you have a roommate? How are you liking it?

(S1) It's MLAO. I don't have a roommate so it's great.

(S2) Congratulations. Mlao is a great dorm then. It is the largest in chicago. Enjoy!

[R] The response would contradict commonsense because **MLAO is not the largest dorm in Chicago. In fact, there is no dorm in Chicago with that name.** This response makes an unfounded assumption and contradicts common knowledge.

Figure 6: An example of Definition Mismatch (DM).

(S2) Cool! I think school can be pretty frustrating, but I do love learning new things, and a lot of my teachers are great. What about you? Do you like school?

(S1) More or less. Doing activities like talking to a robot is great, but doing assignments are tired. I've been spending the past week doing homework.

(S2) I understand, it can definitely be stressful. What would you change about your school, if you could?

[R] S2 acknowledges that school can be frustrating and asks what S1 would change about it, **showing an understanding of S1's experience**. However, **they do not react with a specific sentiment or emotion, nor do they offer reassurance or support**.

Figure 7: An example of Exhaustive (EX).

7.3 World Model

Another major issue is ChatGPT's poor understanding of common world events and human experiences.

ChatGPT frequently reveals a limited understanding of the relationship among concepts within a dialogue. For example, it overlooks elaborations on previous points, considering them too semantically similar (SR). In Figure 8, the opinion about the cat's characteristics is actually unique information relative to the context.

(S2) I love animals too! My favorite animal is the cat! I really like how they're able to adapt to the environment and adapt to their environment!

(S1) I love cats too, theya re adorable!

(S2) I love how cute they are! I especially like how their fur is so soft and fluffy! They look so beautiful!

[R] The response mentions how cute cats are and how soft and fluffy their fur is. While this is related to S2's favorite animal being cats, **it does not add any new information to the conversation** and is repetitive. Therefore, the response is repetitive.

Figure 8: An example of Semantic Relatedness (SR).

ChatGPT also often criticizes typical conversational practices (CN). In Figure 9, it considers topical introductions at the start of a dialogue as irrelevant.

(S1) Hi!

(S2) hi! I am happy to be talking to you. I hope you don't mind me asking, but are you dating anyone now?

[R] The response from S2 is completely irrelevant to what S1 said. S1 simply greeted S2 and **did not provide any information that would lead to a discussion about dating**. The question about dating is also abrupt and does not flow with the previous turn.

Figure 9: An example of Conversation Norms (CN).

In addition, ChatGPT faces challenges in comprehending the plausibility of co-occurring events, beliefs, and experiences (ME). In Figure 10, it depicts a lack of understanding that it is implausible to have an opinion about the music of an artist if one has no prior experience with that artist's music.

ChatGPT also demonstrates a large degree of general inexperience with common phenomenon or situations in the world, which can lead to it harboring untrue facts about the world or misunderstanding nuances of situations (IN). In Figure 11, ChatGPT's lack of commonsense is highlighted by its acceptance of well-wishes for a non-existent trip.

(S2) I have not heard of doj cat. Where does she live? Have any albums yet?

(S1) He most recent album is "Planet Her". You should definitely listen to it!

(S2) Planet her is amazing..the whole album has a distinct sound. It was produced by alex pall.

[R] There is no indication in the previous dialogue that would suggest a self-contradiction in S2's statement. **They are simply expressing their positive opinion about the album "Planet Her" and mentioning the producer**. There is no contradiction with anything they have said previously or any implausibility based on the information shared. Additionally, there is no contradiction with anything S1 has said. Therefore, there is no self-contradiction in S2's statement.

Figure 10: An example of Mutual Exclusion (ME).

(S1) I don't have a trip planned, but maybe I'll go somewhere soon

(S2) I'm sure you will! You'll love it! Have a wonderful trip! I'll see you soon!

[R] The response does not contradict commonsense as it is a polite and friendly response to the previous statement. It does not make any unfounded assumptions, is not illogical or self-contradictory, and does not ask a question where the answer is already obvious. **It is a common response to wish someone a good trip even if they have not planned one yet**.

Figure 11: An example of Inexperience (IN).

8 Recommendations

Given the compelling performance for many dialogue behaviors observed in this work, ChatGPT is a promising direction for behavior classification.

For one, it is worth noting that ChatGPT boasts extreme cost-efficiency relative to humans. Where ChatGPT costs $0.02 on average to provide labels for a single behavior for one dialogue in this work, the average cost for human annotation ranges from $0.29 to $1.96 depending on the behavior (Table 15 in Appendix D). Since even specialized classifiers rely on human annotations for training creation, they also end up being quite costly to maintain.

Furthermore, the results of our error analysis reveal a large degree of systematicity behind ChatGPT's reasoning mistakes across many of the behaviors. Correcting these common mistakes is

likely to further improve its performance to a noticeable degree. We next discuss mitigation strategies of these identified issues to aid in future work.

Context Management Providing the complete dialogue history may hinder ChatGPT's ability to attend to the salient content due to information overload. To address this, we highlight two strategies:

- *Windowed Context*: instead of providing the entire history, truncate the context to k previous turns. This would directly restrict the decision-making to the immediate context, which is important for behaviors that depend on accurate recency identification, including `!Rel`, `Ign`, `!Emp`, and `Emp`.

- *Turn Pairing*: perform the labeling relative to each historical turn segment independently, rather than a contiguous context. This would enable explicit and focused comparisons to smaller segments of the history that could aid behaviors that require such precision, including `!Sel`, `!Par`, and `Red`.

In-Context Learning Examples Given the identified mistake types, it becomes more straightforward to compose useful in-context learning examples that are tailored to optimizing ChatGPT. Examples of those mistake types that are related to ChatGPT misunderstanding the nuances of a behavior (e.g. MD, SR, CN, ME, EX) could be taken from a held-out set of conversations, which would prime ChatGPT to avoid such reasoning.

9 Limitations

Although ChatGPT is a high-performing, widely accessible, and affordable LLM at the time of writing, there are considerations towards the long-term applicability of the results found in this work due to the ChatGPT infrastructure. Since ChatGPT is not open-source and is only accessible through a paid API, there is less detailed understanding of its training and model design. In addition, this access method for ChatGPT also results in less user control over potential model changes and even model deprecation over time. As such, further studies could assess the applicability of other language models to the task of dialogue behavior detection to mitigate these concerns, and we leave this to future work.

Furthermore, it should be noted that the errors made by ChatGPT may not necessarily align with those made by alternative open-source language models, or even future versions of ChatGPT itself. However, it may still be useful to be mindful of the prominent problems encountered with ChatGPT while using other LLMs. These identified phenomena play a crucial role in language comprehension and reasoning overall and could also present challenges for other models, although the extent of their impact remains to be explored.

10 Conclusion

Although automated methods for dialogue behavior classification remain a challenging task, this work finds that ChatGPT-3.5 presents promising potential to reduce the gap between model and human performance. ChatGPT's ability to provide competitive behavior classification against specialized classifiers without necessitating finetuning or human annotation across a variety of dialogue behaviors gives rise to a low-cost, multi-task evaluator model. The systematicity behind the common mistakes observed for ChatGPT reveal concrete steps for future improvements that will improve behavior classification performance, including strategies for context management and better understanding of situational nuances. We look forward to future advancements in behavior classification that leverage ChatGPT's unique capabilities.

11 Acknowledgements

We gratefully acknowledge the support of the Amazon Alexa AI grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Amazon.

References

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, and Siva

Reddy. 2022a. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022b. Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Sarah E. Finch, James D. Finch, and Jinho D. Choi. 2023. Don't forget your abc's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Sarik Ghazarian, Yijia Shao, Rujun Han, Aram Galstyan, and Nanyun Peng. 2023. ACCENT: An automatic event commonsense evaluation metric for open-domain dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4398–4419, Toronto, Canada. Association for Computational Linguistics.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Dialfact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801.

Ryuichiro Higashinaka, Luis F D'Haro, Bayan Abu Shawar, Rafael E Banchs, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Joao Sedoc. 2019. Overview of the dialogue breakdown detection challenge 4. In *Wochat Workshop at IWSDS 2019*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2:: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will i sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Addressing inquiries about history: An efficient and practical framework for evaluating open-domain chatbot consistency. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1057–1067.

Qian Lin and Hwee Tou Ng. 2022. A semi-supervised learning approach with two teachers to improve breakdown identification in dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11011–11019.

Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2022. Exemplars-guided empathetic response generation controlled by the elements of human communication. *IEEE Access*, 10:77176–77190.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 225.

Nathan Ng, Marzyeh Ghassemi, Narendran Thangarajan, Jiacheng Pan, and Qi Guo. 2020. Improving dialogue breakdown detection with semi-supervised learning. In *NeurIPS Workshop on Human in the Loop Dialogue Systems*.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713.

Yushan Qian, Bo Wang, Shangzhao Ma, Wu Bin, Shuo Zhang, Dongming Zhao, Kun Huang, and Yuexian Hou. 2023. Think twice: A human-like two-stage conversational agent for emotional response generation. *arXiv preprint arXiv:2301.04907*.

Michael V Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Weiyan Shi, Yu Li, Saurav Sahay, and Zhou Yu. 2021. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3478–3492.

Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2022. Am i me or you? state-of-the-art dialogue models cannot maintain an identity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2367–2387.

Haoyu Song, Yan Wang, Weinan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831.

Ekaterina Svikhnushina and Pearl Pu. 2023. Approximating human evaluation of social chatbots with prompting. *arXiv preprint arXiv:2304.05253*.

Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. 2022. A taxonomy of empathetic questions in social dialogs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2952–2973, Dublin, Ireland. Association for Computational Linguistics.

Longzheng Wang, Peng Zhang, Xiaoyu Lu, Lei Zhang, Chaoyang Yan, and Chuang Zhang. 2022. Qadialmoe: Question-answering dialogue based fact verification with mixture of experts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3146–3159.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Yadong Xi, Jiashu Pu, and Xiaoxi Mao. 2021. Taming repetition in dialogue generation. *arXiv preprint arXiv:2112.08657*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14158–14166.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. 2023. Socialdial: A benchmark for socially-aware dialogue systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th*

Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5676–5689, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

A Behavior Questions and Definitions

The Question (**Q**) and Definition (**D**) for each dialogue behavior label used for the final ChatGPT prompts are shown in Tables 5 - 12, excluding `Red` which is shown in Table 2 in Section 5.

Q	Is this an empathetic response by Speaker 2:
D	A response is empathetic when Speaker 2 does ONE of the following: - clearly demonstrates an understanding of Speaker 1's emotions - reacts with the appropriate sentiment or emotion to Speaker 1's shared experience - understands or appropriately reacts to Speaker 1's experience or emotions - appropriately reassures, encourages, or supports Speaker 1

Table 5: `Emp`: behavior question and definition.

Q	If this were the next response in the dialogue, would Speaker 1 feel like their feelings are not being understood by Speaker 2:
D	A response displays a lack of empathy when: - it indicates a misunderstanding of how Speaker 1 feels based on what Speaker 1 just said - the tone, emotion, or sentiment of the response is clearly inappropriate for what Speaker 1 just said - the response has an inappropriate lack of emotion to what Speaker 1 just said Do NOT consider its empathy relative to previous topics in the conversation if the dialogue has moved on from them. Instead, only consider the most recent dialogue context when evaluating the empathy of a response.

Table 6: `!Emp`: behavior question and definition.

Q	If this were the next response in the dialogue, would it contradict commonsense:
D	To identify contradictions of commonsense, judge whether a vast majority of people would agree that the response doesn't make sense because the response: - contradicts common knowledge - makes unfounded assumptions - is highly illogical or self-contradictory - asks a question where the answer is already obvious Do NOT mark responses that don't make sense because they: - are off-topic or irrelevant as responses - don't have any clear meaning (e.g. overly vague or ill-formed responses)

Table 7: `!Com`: behavior question and definition.

Q	If this were the next response in the dialogue, does it completely ignore the immediate last turn from Speaker 1:
D	Responses that are completely off-topic, fail to address the asked question, or are otherwise completely inappropriate in the context are considered to be ignoring the other speaker.

Table 8: `Ign`: behavior question and definition.

Q	If this were the next response in the dialogue, is it a self-contradiction by Speaker 2:
D	Self contradictions occur when Speaker 2 says something that is a contradiction of what they have said previously or it is extremely implausible based on the information they have already shared. Self contradictions may also occur within a single turn if Speaker 2 shares two contradictory things. If Speaker 2 shares world knowledge that is factually incorrect this is NOT enough on its own to warrant a self contradiction. If Speaker 2 contradicts something the other speaker Speaker 1 has said, this is NOT a self-contradiction.

Table 9: `!Sel`: behavior question and definition.

Q	Does this response include an incorrect fact:
D	Incorrect facts occur when the response includes information that is either: - false - unproven - highly controversial - highly implausible - clearly misleading If an organization, person, place, etc. is mentioned as a part of public knowledge, but it does not exist or it is inaccurately represented, then this is an incorrect fact. Do NOT consider a turn as an incorrect fact if the turn could be interpreted as expressing: - preference or value judgements - estimates or predictions - personal information about the speaker or their partner - information about things in either speaker's life that are not publicly relevant

Table 10: `!Fac`: behavior question and definition.

Q	Is Speaker 2 saying something about Speaker 1 that is contradicting what Speaker 1 has already shared:
D	Partner contradictions occur when Speaker 2: - shares an assumption about Speaker 1 that is impossible to know based on what has already been said - shares an inference about Speaker 1 that is implausible based on what has already been said - contradicts something Speaker 1 shared about themselves - asks a repetitive question about Speaker 1 when the answer is already known based on what has already been said If Speaker 2 says something that makes it seem like they have forgotten or misremembered what their partner Speaker 1 has said earlier in the dialogue, this is a partner contradiction. If Speaker 2 shares a difference of opinion or situation in their own life as compared to Speaker 1, this is NOT a partner contradiction.

Table 11: `!Par`: behavior question and definition.

Q	If this were the next response in the dialogue, is it completely irrelevant to what was just said:
D	If a response fails to continue the current discussion or jumps to a new and off-topic discussion, it is considered to be irrelevant. Responses that are irrelevant feel abrupt and interrupt the discussion, usually because they present questions or ideas that are unrelated to the previous turn. Short reactions to or acknowledgements of the previous turn are NOT irrelevant.

Table 12: `!Rel`: behavior question and definition.

B Full Prompt Example

Table 13 shows an example of the full ChatGPT prompt utilized in this work. This full version preserves the 'DIALOGUE' header, turn numbers, whitespace newlines, and full speaker names, which were removed from the prompt in Table 2 in Section 5 due to spacing constraints.

```
----------
DIALOGUE
----------
1. Speaker 1: Hi!
2. Speaker 2: Hi! I am happy to be talking to you. Do you have any
   pets at home?
3. Speaker 1: I don't, my mom is allergic to most pets. I used to have
   a pet fish when I was much younger.
4. Speaker 2: Do you want a pet at some point in the future? No
   judgment here, I'm just curious.
5. Speaker 1: I would love to have pets in the future! Once I have
   my own place, I'd like to get a dog or two.
----------

Is this response repeating something that has already been said:

Speaker 2: Would you want to get a cat or a dog?

A response is repetitive if:
- it repeats something from earlier in the dialogue
- it includes asking a question whose answer has been
  already shared

If any part of the response is repetitive, then it should be
labeled as repetitive.
Note that sometimes repetition is useful, such as for emphasis,
acknowledgement, clarification, or elaboration, and in these
cases it should NOT be labeled as repetitive.

Provide your reasoning when considering this question starting
with "Reasoning:". Then, finish by writing your final
decision as one of: "Decision: [YES]" or "Decision: [NO]".
Do NOT fill in your decision with any terms other than YES or NO.
```

Table 13: An example of an unmodified ChatGPT prompt.

C Full Results

Table 14 extends Table 3 from §6.2 to include the precision and recall scores for the automated models. Precision and recall scores are not meaningful for the human evaluators since each human annotation set is traded out as a benchmark against the other; thus, we still present only F1 for **HUM**.

D ChatGPT Cost

We compare the average cost of labeling a single dialogue from ABC-Eval for each behavior using ChatGPT and human judges. Table 15 contains the calculated costs.

ChatGPT The ChatGPT cost for a single dialogue is calculated from the OpenAI API pricing[5] ($0.002 USD per 1000 tokens, at time of writing) on the sum total number of tokens used for obtaining labels for each bot response for a particular behavior. These costs are then averaged over all dialogues used in this work to obtain the average cost per dialogue. Because there is not much difference in prompt length for the different behavior prompts, the average cost per behavior is quite similar.

HUM Human annotation costs are derived from the average costs presented in Finch et al. (2023). Since the behavior labels were grouped into annotation tasks for the human judges, we divide each task cost by the number of behaviors contained within that task. The cost for a single label is then the resulting quotient for its respective task.

[5] https://openai.com/pricing

	Model	P/R/F1+	P/R/F1-	Acc.	#+
!Fac	FC	12.3 / 22.4 / 15.9	93.3 / 87.1 / 90.1	82.2	223
	ChatGPT	37.7 / 44.9 / 41.0	95.5 / 94.0 / 94.7	**90.3**††	146
	HUM	67.8	97.4	**95.2****	122
Red	ChatGPT	30.7 / 35.5 / 32.9	94.3 / 93.2 / 93.8	88.6	148
	HUM	58.7	96.4	**93.5****	129
!Com	ChatGPT	43.8 / 29.1 / 34.9	83.3 / 90.5 / 86.7	78.0	219
	HUM	55.6	88.6	**81.9***	333
!Rel	S2T2	24.0 / 33.5 / 27.9	86.3 / 79.8 / 82.9	**72.4**†	365
	ChatGPT	30.1 / 62.5 / 40.6	91.0 / 72.3 / 80.6	70.8	543
	HUM	54.3	91.3	**85.4****	261
!Par	ChatGPT	27.2 / 14.2 / 18.6	91.6 / 96.1 / 93.8	88.5	79
	HUM	48.8	94.8	**90.5****	151
!Sel	DEC	22.8 / 49.1 / 31.1	96.3 / 89.2 / 92.6	**86.6**††	215
	ChatGPT	14.6 / 35.9 / 20.7	95.3 / 86.1 / 90.5	83.0	250
	HUM	44.3	96.3	**93.1****	101
!Emp	EPI	12.0 / 15.1 / 13.4	85.4 / 81.8 / 83.5	72.3	291
	ChatGPT	21.1 / 36.2 / 26.6	88.1 / 77.7 / 82.6	71.8	396
	HUM	51.5	92.0	**86.3****	231
Ign	S2T2	18.5 / 39.5 / 25.2	91.9 / 79.7 / 85.3	**75.5**††	365
	ChatGPT	15.5 / 63.4 / 24.9	93.3 / 59.8 / 72.9	60.2	696
	HUM	61.6	95.5	**92.0****	170
Emp	EPI	39.6 / 86.0 / 54.2	70.3 / 20.1 / 31.3	45.0	1343
	ChatGPT	50.7 / 11.9 / 19.3	63.4 / 92.9 / 75.4	**62.3**††	146
	HUM	69.7	81.6	**77.1****	618

Table 14: Precision, recall, F1 and accuracy achieved by each model, where **HUM** stands for human judges. **#+**: num. positive labels predicted. †|†† denote significance between *automated* models on one or both annotation sets. ⋆|⋆⋆ denote significance against best automated model on one or both annotation sets, respectively.

	ChatGPT	HUM
!Fac	0.02	1.96
Red	0.02	0.29
!Com	0.02	0.92
!Rel	0.02	0.47
!Par	0.02	0.29
!Sel	0.02	0.29
!Emp	0.02	0.58
Ign	0.02	0.47
Emp	0.02	0.58

Table 15: Cost ($ USD) per dialogue for each behavior using ChatGPT or humans (**HUM**).

Are Large Language Models All You Need for Task-Oriented Dialogue?

Vojtěch Hudeček and **Ondřej Dušek**
Charles University, Faculty of Mathematics and Physics
Malostranské náměstí 25, 118 00 Prague, Czechia
hudecek@ufal.mff.cuni.cz, odusek@ufal.mff.cuni.cz

Abstract

Instruction-finetuned large language models (LLMs) gained a huge popularity recently, thanks to their ability to interact with users through conversation. In this work, we aim to evaluate their ability to complete multi-turn tasks and interact with external databases in the context of established task-oriented dialogue benchmarks. We show that in explicit belief state tracking, LLMs underperform compared to specialized task-specific models. Nevertheless, they show some ability to guide the dialogue to a successful ending through their generated responses if they are provided with correct slot values. Furthermore, this ability improves with few-shot in-domain examples.

1 Introduction

Large Language Models (LLMs) have transformed the NLP field, showing outstanding performance across many NLP benchmarks such as Winograd Challenge (Levesque et al., 2012) or GLUE (Wang et al., 2018). Recently, instruction finetuning of LLMs proved to be able to align the model outputs with human preferences (Ouyang et al., 2022; Wang et al., 2022) and improved the LLMs' communication capabilities substantially. State-of-the-art LLMs are not only good at understanding user needs but also capable of providing relevant answers. Consequently, we see many chatbot applications both inside and outside academia (ChatGPT[1], Claude[2], Sparrow[3]) which build upon the raw power of instruction-finetuned LLMs.

Given the millions of daily interactions with these chatbots, it appears that the models are able to handle users' needs to their satisfaction, at least to some extent. However, these chatbots are tuned using unstructured open-domain conversations. The aim of this paper is to evaluate these systems for more specific applications, where the system has to follow a predetermined structure and handle external sources of information, such as APIs or databases. We raise the question to what extent LLMs are capable of handling these applications off-the-shelf, i.e. without finetuning. We thus choose to evaluate LLM performance in the task-oriented dialogue (TOD) setting, as it requires precise information handling for communicating with external APIs. Moreover, TOD systems output in-domain information which has predetermined structure and lends itself well to evaluation, thanks to pre-existing annotated data sets. We avoid any finetuning techniques and focus on zero-shot or few-shot settings using in-context learning, as this approach has lower hardware requirements and barrier of entry and better flexibility or even performance in certain tasks (Su et al., 2022).

Therefore, we introduce an LLM-based TOD

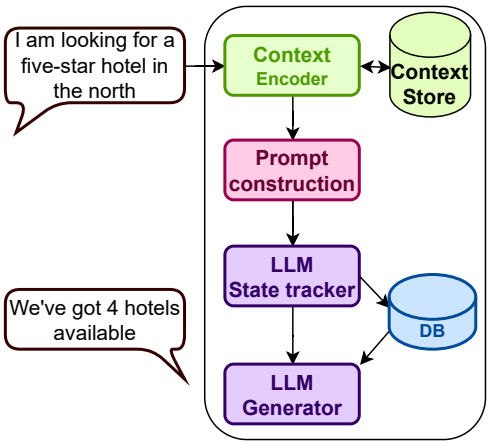


Figure 1: A high-level overview of our proposed pipeline. The user input is used to retrieve relevant few-shot examples (if available). Then, an initial prompt is constructed and an LLM is asked to provide the current dialogue state. Based on that, we retrieve database results and construct another prompt. Finally, we ask the LLM to provide the response.

[1] https://openai.com/blog/chatgpt
[2] https://www.anthropic.com/index/introducing-claude
[3] https://www.deepmind.com/blog/building-safer-dialogue-agents

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 216–228
September 11–15, 2023. ©2023 Association for Computational Linguistics

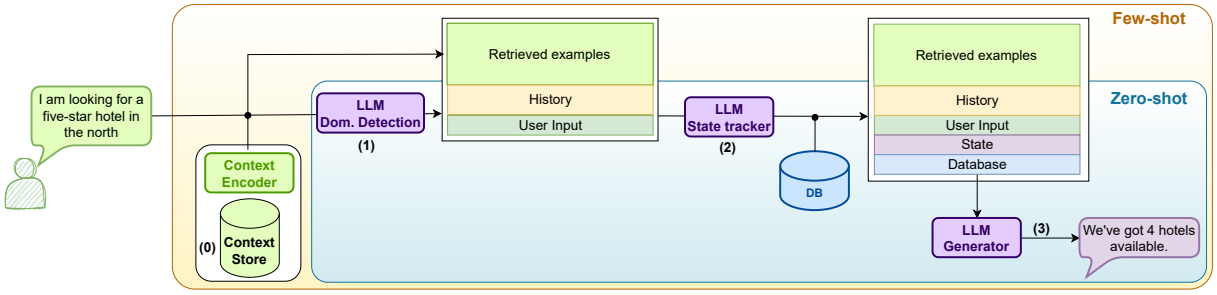


Figure 2: A detailed description of our proposed pipeline. (0) As a preprocessing step, we encode a subset of the training set that will be used to retrieve few-shot examples. Given the user input, we: (1) Detect the domain, retrieve relevant examples (in the few-shot setting) and construct an initial prompt. (2) Infer the belief state using LLM. Based on that, we retrieve database information and construct another prompt that includes both the state and database results. (3) We ask the LLM to provide a final response.

conversation pipeline (see Figure 1) and evaluate its performance with respect to commonly used task-oriented metrics such as Joint Goal Accuracy, Slot F1, and Dialogue Success (Rastogi et al., 2018; Budzianowski et al., 2018). Our pipeline resembles other approaches based on LMs (Peng et al., 2021; Yang et al., 2021), using state tracking and response generation as two main, separate steps, while keeping the role of a dialogue policy implicit. However, instead of finetuning LMs, it intentionally relies almost exclusively on the usage of pretrained LLMs as-is, so we can test their out-of-the-box capabilities. The dialogue context and domain description are introduced to the model only by including them in the input prompt. In the zero-shot setting, the model receives a domain description only; in the few-shot setting, it additionally uses a few retrieved examples (see Section 3 for details).

In our experiments, we find that LLMs are not very good at state tracking and their performance falls behind the state-of-the-art. However, if provided with correct belief states, some of them yield interesting response generation performance, comparable to earlier finetuned state-of-the-art models. To our knowledge, our zero-shot experiments establish a state-of-the-art result in unsupervised TOD modeling on the MultiWOZ and Schema-guided datasets (Budzianowski et al., 2018; Rastogi et al., 2020). While there may be room for improvement through prompt engineering, our results aim to show the out-of-the-box LLM capabilities. We plan to release our experimental code on GitHub.[4]

[4] https://github.com/vojtsek/to-llm-bot

2 Related Work

Large Language Models The Transformer architecture (Vaswani et al., 2017) enabled the training of large and capable language models. The research on their few-shot and zero-shot abilities dates back to the GPT-2 and GPT-3 models (Radford et al., 2019; Brown et al., 2020), which are scaled versions of the Transformer decoder. Many followed this path of training large Transformer decoders (Zhang et al., 2022; Black et al., 2022), yielding models of up to hundreds of billions parameters in size (Zhao et al., 2023). Other models leverage the whole original (encoder-decoder) Transformer architecture (Raffel et al., 2020; Lewis et al., 2020). Recent research focuses on improving the training of moderate-sized architectures to broaden access to highly capable LLMs (Touvron et al., 2023).

Instruction Tuning The idea of using reinforcement learning techniques to align model-based agents better with users' intents was pioneered in game agent development (Christiano et al., 2017) and later explored for training language models (Ziegler et al., 2019; Ouyang et al., 2022). Although these techniques proved to be quite effective, the process is still very demanding in terms of collecting feedback from users. Consequently, several datasets were proposed (Wang et al., 2022; Iyer et al., 2022; Black et al., 2022) that collected millions of instructions-based tasks in natural language and can be applied to finetune LLMs using reinforcement learning.

LM-based TOD modeling Task-oriented dialogue modeling with pretrained LMs was introduced by Zhang et al. (2019) and Peng et al. (2021), who followed text-based state encoding and two-

stage generation proposed by Lei et al. (2018): An LM is first used to decode a structured belief state, represented as text. The belief state is then used to retrieve database information and the LM is called once more to generate a response, conditioned on the belief state and retrieved information. Several improvements to the basic setup were proposed, such as contrastive state training (Kulhánek et al., 2021) or using belief state differences (Lin et al., 2020). Others proposed a combination of generative models with retrieval-based approaches (Pandey et al., 2018; Cai et al., 2019; Nekvinda and Dušek, 2022). All described works finetune LMs on in-domain data, which is in contrast with the pure in-context learning approach that we apply.

Few-shot dialogue modeling One of the first neural models focusing on learning dialogue from a few in-domain examples was the Hybrid Code Networks (Williams et al., 2017), a trainable system based on recurrent neural networks, with partially handcrafted components. Another approach was proposed by Zhao and Eskenazi (2018), who used latent action representations to enable the transfer of domain knowledge. Latent actions were also used by Huang et al. (2020) and Shalyminov et al. (2019). More recent approaches leverage the capabilities of pretrained Transformer LMs (Shalyminov et al., 2020). Hu et al. (2022) used LLMs and in-context learning to perform belief state tracking, formulating the task as an SQL query generation. Unlike our work, they did not use instruction-tuned models and omitted database retrieval and response generation.

3 Method

We introduce our method step-by-step. An overall description of the proposed pipeline is shown in Figure 2. The system consists of a pretrained LLM and an (optional) context store in a vector database. Three LLM calls are performed in each dialogue turn, with specific prompts (see Section 3.1). First, the LLM performs domain detection and state tracking (Section 3.2). The updated belief state informs a database query, whose results are used in the subsequent LLM-based response generation step (Section 3.3). In the few-shot setting, the context store is used to store a limited number of examples from the training set, which are retrieved based on similarity with the conversation context and included in LLM prompts (see Section 3.4).

Prompt	Definition: Capture values from a conversation about hotels. Capture pair "entity:value" separated by colon and no spaces in between. Separate the "entity:value" pairs by hyphens Values that should be captured are: - "pricerange": the price of the hotel ... [history] Customer: "I want a cheap place to stay."
Output:	pricerange:"cheap"

Table 1: A simplified example of a zero-shot version of the prompt used for state update prediction. It contains task definition, domain description, dialogue history and user utterance. For the exact prompts see Appendix.

3.1 Prompt construction

We aim to compare the raw capabilities of the selected LLMs, therefore we do not focus on prompt engineering techniques and choose universal prompts used for all LLMs in this work (cf. Section 8). We choose simple, plain language statements as prompts, with no specific vocabulary, based only on a few preliminary tests. We define a single **domain detection prompt** for all examples, plus a pair of prompts for each domain in the given dataset: a **state tracking prompt** (see Table 1) and a **response prompt**.

The domain detection prompt includes a task description and two static examples of domain detection. In addition to general instructions, each state tracking prompt contains a domain description, a list of relevant slots, the dialogue history, and the current user utterance. The response prompts do not contain the per-domain slot list, but they include the current belief state and database results instead. In the few-shot setting, each tracking and response prompt additionally contains positive and negative examples retrieved from the context store (see Section 3.4). Prompt examples are shown in Tables 5 and 6 in the Appendix.

3.2 Domain Detection and State Tracking

We prompt the LM twice at each turn during state tracking: first, to detect the active domain, then to output slot values that changed or appeared in the current turn. We then use the outputs to update the accumulated global belief state.

The two prompting steps are used since we need the models to operate in a multi-domain setting, i.e., handle conversations spanning multiple domains. Therefore, we need to be able to detect the currently

active domain. We achieve this by first prompting the LLM with a domain detection prompt (using a single prompt for all examples).

Once we obtain the active domain prediction, we can include manually designed domain descriptions in a second prompt that handles belief state prediction. An example of a prompt used for state tracking is provided in Table 1. For the few-shot variants, we retrieve few-shot examples from the context store, limited to the active domain.[5]

Our preliminary experiments showed that LLMs struggle to output all active slot values at every turn consistently. Therefore, we model only state updates, following the MinTL approach (Lin et al., 2020). Here, the model only generates the slot-value pairs that have changed in current turn. The global belief state is then accumulated using these turn-level updates. To obtain machine-readable outputs useful for database queries or API calls, we specify in the prompt that the model should provide JSON outputs, and any provided few-shot examples are formatted accordingly.

3.3 Response Generation

The current belief state is used to query the database for entries matching all user-specified slots in the active domain. Given the belief state and database results, the response generation is straightforward. The prompt for the LLM includes dialogue history, user utterance, belief state and database results (and retrieved examples in the few-shot setting) and requests the model to provide a fitting system response. We generate delexicalized responses (Wen et al., 2015), i.e., we replace slot values by placeholders, following prior work in end-to-end TOD modeling. In addition to simplifying the task for the model, delexicalized outputs allow us to evaluate the success rate and compare to previous works. The prompt specifies that the model should provide entity values as delexicalized placeholders, and any few-shot examples are constructed accordingly.

3.4 Context Storage

It has been shown that enriching prompts with specific examples boosts LM performance (Madotto et al., 2020; Brown et al., 2020). To apply this knowledge efficiently in our pipeline, we introduce a storage that contains encoded dialogue contexts. This context storage is optional and is only required for the few-shot prompting variant. We use dialogue context taken from a fixed-length history window as the key to be encoded in the vector database. More details can be found in Section 4.4. Once the relevant examples are retrieved, we include them in the prompt to guide the model better. Some of the LLMs rely on negative (counter-) examples as well (Wang et al., 2022). Therefore, we follow Peng et al. (2021)'s consistency classification task approach to produce negative examples: We take some of the retrieved belief state examples, corrupt them by replacing some of the correct slot values with random values, and present them as negative in the prompt.

[5]For this purpose, each conversation snippet contained in the context store comes from a single-domain conversation.

4 Experimental Setup

To obtain a broad overview of the current LLMs' capabilities, we compare several models, spanning different numbers of trainable parameters and different training methods. We also experiment with four variants of the base setup, using either zero-shot or few-shot operations and using either predicted or oracle belief states.

4.1 Datasets

We experiment with two of the currently most prominent benchmark datasets for task-oriented multi-domain dialogue:

- **MultiWOZ 2.2** (Budzianowski et al., 2018; Hung et al., 2022) is a well-known benchmark used for evaluating state tracking, response generation and dialogue success rate. Its evaluation is well-defined and the dataset contains database files, so full interaction can be simulated. It contains over 10k dialogues, 7 domains and 29 distinct slots.

- **Schema Guided Dataset** (Rastogi et al., 2020) is also well annotated and even richer dataset containing more than 22k dialogues 18 domains and 145 slots. Database interaction is considered in the dataset, but no real database is provided and database results are defined ad-hoc. Therefore we simply use the provided database results in the prompts without performing any actual queries.

4.2 Tested Models

We chose the following five instruction-finetuned models for our experiments, spanning different

model	few shot	oracle BS	Schema Guided Dialogues BLEU	JGA	Slot-F1	Success	MultiWOZ 2.2 BLEU	JGA	Slot-F1	Success
Supervised SotA	✗	✗	29.90*	0.30†	0.60*	–	19.90♣	0.60♢	–	0.82♡
Alpaca-LoRA-7B-zs-gbs	✗	✗	2.79	0.02	0.01	0.11	1.61	0.06	0.07	0.04
Tk-Instruct-11B-zs-gbs	✗	✗	4.16	0.05	0.03	0.10	2.48	0.04	0.04	0.04
GPT-NeoXT-20B-zs-gbs	✗	✗	0.45	0.01	0.01	0.17	0.52	0.03	0.02	0.04
OPT-IML-30B-zs-gbs	✗	✗	1.63	0.01	0.01	0.17	0.56	0.02	0.04	0.03
ChatGPT-zs-gbs	✗	✗	–	–	–	–	4.17	0.13	0.40	0.31
Alpaca-LoRA-7B-zs-obs	✗	✓	2.76	–	–	0.23	1.73	–	–	0.08
Tk-Instruct-11B-zs-obs	✗	✓	5.21	–	–	0.24	2.66	–	–	0.18
GPT-NeoXT-20B-zs-obs	✗	✓	0.83	–	–	0.22	0.60	–	–	0.06
OPT-IML-30B-zs-obs	✗	✓	1.94	–	–	0.22	0.54	–	–	0.06
ChatGPT-zs-obs	✗	✓	–	–	–	–	3.76	–	–	0.47
Alpaca-LoRA-7B-fs-gbs	✓	✗	6.32	0.04	0.01	0.09	5.53	0.06	0.08	0.06
Tk-Instruct-11B-fs-gbs	✓	✗	6.66	0.06	0.05	0.10	6.56	0.16	0.33	0.19
GPT-NeoXT-20B-fs-gbs	✓	✗	1.62	0.04	0.02	0.09	2.73	0.05	0.04	0.05
OPT-IML-30B-fs-gbs	✓	✗	0.82	**0.06**	**0.07**	0.08	4.40	0.03	0.03	0.04
ChatGPT-fs-gbs	✓	✗	–	–	–	–	6.77	**0.27**	**0.51**	0.44
Alpaca-LoRA-7B-fs-obs	✓	✓	6.99	–	–	**0.25**	5.96	–	–	0.41
Tk-Instruct-11B-fs-obs	✓	✓	**8.56**	–	–	**0.25**	**6.91**	–	–	0.46
GPT-NeoXT-20B-fs-obs	✓	✓	1.97	–	–	0.24	2.92	–	–	0.28
OPT-IML-30B-fs-obs	✓	✓	0.56	–	–	0.22	5.40	–	–	0.28
ChatGPT-fs-obs	✓	✓	–	–	–	–	6.84	–	–	**0.68**

Table 2: Evaluation of the chosen LLMs with respect to widely used TOD measures. For each model, we provide multiple variants. We use either zero-shot or few-shot prompts (*-zs-* vs. *-fs-*) and either generated or oracle belief state (*-gbs* vs. *-obs*). The few-shot variants use 10 examples per domain in the context storage (∼0.6% of the training set in case of MultiWOZ), two of which are selected for the prompts. To reduce cost, we only evaluate the paid ChatGPT model on MultiWOZ. We also provide supervised state-of-the-art results to put the numbers in context: *Zhu et al. (2022), †Feng et al. (2021), ♣Sun et al. (2022), ♢Huang et al. (2023), ♡Feng et al. (2023).

sizes (within the limitations of hardware available to us) and using freely available models as well as the paid ChatGPT API. We indicate the specific model variant (i.e., model size, given by the number of parameters) directly in the model name.

- **Tk-Instruct-11B** (Wang et al., 2022) is based on the T5 encoder-decoder architecture (Raffel et al., 2020). It was tuned on a dataset of over 5M task instances with instructions.
- **ChatGPT** is a product introduced by OpenAI.[6] Although the exact training process and architectures were not published, it most probably uses a similar architecture and finetuning techniques as InstructGPT (Ouyang et al., 2022), with additional human feedback.
- **Alpaca-LoRA-7B** is a version of the LLaMa model (Touvron et al., 2023) using the LoRA method (Hu et al., 2021) for finetuning on Stanford Alpaca project data (Taori et al., 2023). LoRa keeps the base model parameters frozen, but adds additional smaller weight matrices to the model to transform its outputs.
- **GPT-NeoXT-Chat-Base-20B** is based on the GPT-NeoX open-source language model (Black et al., 2022) and finetuned with over 40M dialogue-style instructions.
- **OPT-IML-30B** (Iyer et al., 2022) is based on the Transformer decoder OPT model (Zhang et al., 2022) and trained with a custom set of instructions, including the finetuning set from Tk-Instruct.

4.3 Evaluated variants

We test four variants of our setup for each pair of model and dataset. Specifically, we use zero-shot (without examples) or few-shot (including examples) prompts (*-zs-* vs. *-fs-*) and either generated or oracle belief states (*-gbs* vs. *-obs*). For retrieval in the few-shot setting, we store just 10 examples per domain in the context store by default. We experiment with increasing this number in Section 5.4. Using oracle belief state allows us to focus on evaluating the LLM's ability to guide the dialogue.

[6] https://openai.com/blog/chatgpt

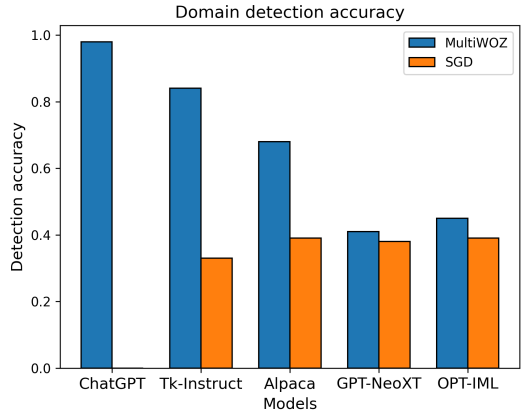


Figure 3: Domain detection accuracy with respect to different models for MultiWOZ 2.2 and SGD data wchich consist of 7 and 18 domains, respectively.

4.4 Experiment Details

Due to the expensiveness of the LLM runs,[7] we did not perform a grid search, but used a limited set of preliminary experiments to determine hyperparameters. Based on this, we used the context of two preceding utterances (user + system) as the context store keys (cf. Section 3.4). We retrieve two examples for few-shot prompts and make one corrupted variant from each of them for negative examples. To corrupt an example, we switch some of the slot values randomly, similarly to Kulhánek et al. (2021). In the context store, we encode few-shot examples using the multilingual embedding model provided by Reimers and Gurevych (2020)[8] and store them in the FAISS database (Johnson et al., 2019). To perform the LLM calls, we use the Huggingface library[9] and the OpenAI API.[10]

4.5 Evaluation Measures

We evaluate the system outputs on multiple levels, both using automatic metrics and human evaluation. Results are given in Sections 5 and 6, respectively.

Automatic Metrics

In automatic evaluation, we first follow the LLM calls being made and evaluate domain detection, state tracking as well as response generation. We also evaluate the overall dialogue-level performance. For *domain detection*, we simply compute **detection accuracy** as a ratio of correctly detected domain out of all dialogue turns being processed. For *state tracking*, we compute **micro-F1** score and **Joint Goal Accuracy** (JGA). JGA is computed as the ratio of dialogue turns for which the predicted belief state matches the ground truth. We use fuzzy matching of the slot values, so that capitalization or minor typos do not influence the result. To evaluate *response generation*, we follow related works and use **BLEU score** (Papineni et al., 2002).

The main *overall measure* for evaluating a task-oriented dialogue is the dialogue **success rate** (Deriu et al., 2021). For MultiWOZ, we use the standard evaluation of dialogue success as the ratio of dialogues where the user reaches the desired goal, based on goal annotation provided with the data (Nekvinda and Dušek, 2021). The SGD dataset does not include goal annotation but contains information about the requested slots. Therefore, we compute SGD success rate as the proportion of dialogues in which (1) the system captures all the slots correctly and (2) all the requested slots are provided.

Human Evaluation

For human evaluation, we perform a small-scale in-house interaction study on MultiWOZ. Since the MultiWOZ goal often involves tasks in multiple domains, we ask annotators to evaluate each domain in the dialogue distinctly. At the end of each dialogue, the annotators are asked to answer these questions:

1. *How many of the subdialogues/domains were handled successfully?* (corresponding to dialogue success)

2. *How many clarifications or corrections were needed?*

3. *Was all the provided information captured correctly?* (corresponding to JGA)

5 Automatic Metrics Results

5.1 Domain detection

We report the domain detection accuracy on MultiWOZ and SGD in Figure 3. We observe that the domain detection accuracy varies quite a lot for various models and presumably influences the quality of the retrieved few-shot examples and appropriateness of the subsequent prompts. However, it is important to note that domain detection is turn-based,

[7] Hardware intensity for the freely available models and actual cost for ChatGPT.

[8] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[9] https://huggingface.co

[10] https://platform.openai.com

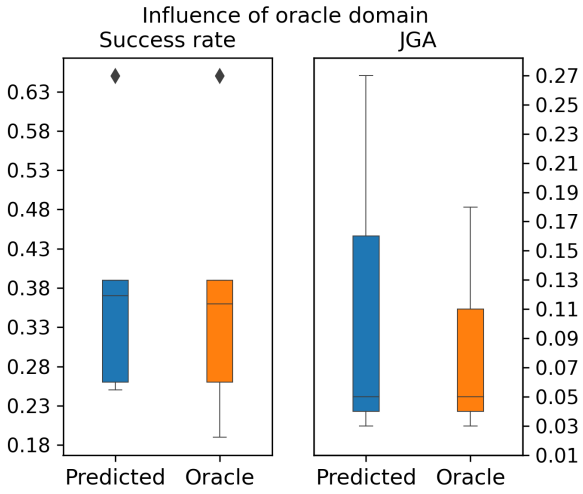


Figure 4: The influence of using oracle domain to retrieve examples. Interestingly, the oracle domain does not improve the performance, suggesting that the model-based detection is good enough for retrieval.

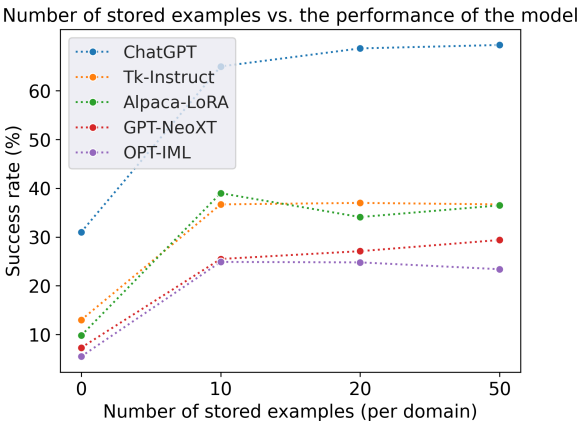


Figure 5: The influence of the number of examples per domain available for few-shot retrieval and performance of the model in terms of the dialogue success on MultiWOZ 2.2 data with oracle state supplied. Note that this does not represent the number of examples selected for the prompt, which is fixed to two.

and arguably there are situations (e.g. providing an address, saying goodbye etc.) that are always handled in the same fashion, even though they formally belong to different domains. Therefore, not all the retrieved examples from misclassified domains necessarily contain unrelated contexts. To explore this, we measure the performance of all models in case an oracle domain is given to them (Figure 4). Interestingly, using the oracle domain did not improve performance, it even worsened in some cases. This suggests that the model-predicted domain is generally good enough, and additionally providing the domain information does not contribute to the final system performance. The negative influence on performance might be caused by forcing the system to filter out relevant examples. We observe that in multiple cases, the conversations snippets are domain-independent so the retrieval might perform better even with a wrongly selected domain. Forcing the ground truth domain examples in these cases can be potentially harmful.

5.2 Belief State Tracking

The belief state tracking results overview is given in Table 2 (*JGA* and *Slot-F1*). There is a huge gap between the supervised models' performance and the LLM results. Also compared to Hu et al. (2022), who used few-shot in-context learning and reported JGA 43.13% with a comparable dataset size, our instruction-tuned LLMs fall short. However, the models we use are an order of magnitude smaller in general, and we also use fewer examples in the prompt. We hypothesize that the performance could be further improved by careful model-specific prompt customization and perhaps task re-formulation; nevertheless, this is not the goal of this work. We intentionally focus on the universal framing of the task since we want to explore the general ability of the models to follow instructions.

When comparing the results among the models, ChatGPT clearly outperforms the rest of the models by a large margin. Interestingly, the few-shot vs. zero-shot setting does not seem to influence the results much, except for the GPT-NeoXT model.

5.3 Response Generation

BLEU scores are low overall, far below the supervised state-of-the-art. Tk-Instruct and ChatGPT are the strongest here and perform roughly on par.

5.4 Dialogue-level performance

Results for dialogue success are provided in Table 2, and there is again a large gap between LLMs and supervised custom models' performance. ChatGPT seems to outperform other models, similarly to state tracking (cf. Section 5.2). However, for some cases, especially in the zero-shot setting, the difference is not that obvious. In most cases, adding the retrieved few-shot examples helps. The contribution of retrieved examples is more obvious when we supply the oracle belief state, in which case it helps consistently for all the models.

We also explore the influence of context storage size on the dialogue success rate. The results are given in Figure 5. It seems that the biggest improvement can be achieved by supplying just a few examples instead of zero-shot prompting, but increasing the size of the example pool for retrieval does not yield further performance gains.

6 Model Analysis

6.1 Human Evaluation

We employed 6 annotators with a background in linguistics and NLP and let them interact with the two strongest models in terms of automatic metrics: ChatGPT and Tk-Instruct. The annotators were given randomly selected goals from the MultiWOZ 2.2 dataset and a minimal set of essential instructions on how to proceed. We present the results in Table 3. We can see that in real interaction with a human user and allowing for clarification or correction, the models perform better compared to the rather strict automatic evaluation. Furthermore, the models are often successful in multiple sub dialogues, even if a part of the whole dialogue fails. The experiment also confirms the superior performance of ChatGPT on both dialogue success and JGA. Not surprisingly given the above results, conversations with ChatGPT also required fewer clarification turns than with Tk-Instruct.

	ChatGPT	Tk-Instruct
dialogues	25	25
subdialogues	52	48
clarify / dial	1.08	1.68
succesful subdialogues	81%	71%
succesful dialogues	76%	64%
correctly captured	88%	66%

Table 3: Human evaluation results for ChatGPT and Tk-Instruct-11B models. We evaluate the conversation on sub dialogue level i.e. each domain in the dialogue is evaluated separately.

6.2 Error Analysis

To understand the models' behavior better, we manually inspect a random sample of ca. 20 dialogues for each model, chosen from cases where the automatic success metric was not satisfied. In general, we can split most of the erroneous behaviors into two distinct groups, which we call *prompt-recoverable* and *inherent*.

Prompt-recoverable errors can be likely fixed by specific prompt engineering with some effort. These errors happen with all of the tested models. Examples of such errors are the invalid structure of the generated dialogue state, copying slot values instead of using canonical values from the ontology, failure to delexicalize some of the values, etc. Most of these errors can be also fixed in postprocessing – for example, we can employ more robust parsers or fuzzy matching of slot values.

Inherent errors, on the other hand, are likely not easily fixable by prompt modifications. They are not distributed evenly across the tested models and seem to constitute a more challenging problem.

Perhaps the most important error, common to all the models, is hallucination, i.e., the model's output responses not grounded in the context (such as offering entities that are not included in the database). This happens in about 10-20% of the inspected dialogues. Some models (*GPT-NeoXT, OPT-IML*) tend to generate more content than they are asked for. This happens in more than 50% of their failed dialogues. In some cases, this means continuing the conversation for a few more turns (including hallucinating user turns), but the models also often generate unrelated text or even code snippets. With *Tk-Instruct*, we observed that in ca. 10% cases, it copies the belief state from the example given in the prompt instead of generating a relevant one. Another issue is that the models tend to repeat their previous responses.

7 Conclusion & Future Work

We present an experimental evaluation of instruction-tuned LLMs applied to the established task of task-oriented dialogue modeling, with five LLMs evaluated on two datasets. We find that LLMs are not performing well in terms of belief state tracking, even when provided with in-context few-shot examples. However, there is some potential to improve through prompt tuning and output parsing robust to irregularities.

If provided with a correct belief state, the models can interact with the user successfully, provide useful information and fulfill the user's needs. While the performance does not match the supervised state of the art, it is important to note that these models were not finetuned on in-domain data and work with just a domain description or a few examples (which again improve performance).

Therefore, carefully picking representative ex-

amples and combining the LLM with an in-domain belief tracker can be a viable choice for a task-oriented dialogue pipeline.

Interestingly, in the human interactive evaluation, both ChatGPT and Tk-Instruct outperformed the expectations set by automatic metrics. This shows certain flexibility and ability to correct their own mistakes on the part of LLMs, and further demonstrates that single-turn evaluation is too rigid and does not show the whole picture (Takanobu et al., 2020). In future work, we want to focus on addressing the prompt-recoverable errors while maintaining the ability to use model-independent prompts and easily swap models. We also aim to find a more effective method of relevant example selection.

8 Limitations

One of the limitations of our work is the usage of the ChatGPT model, which is only accessible via an API and is not guaranteed to retain its exact abilities. However, the other four out of five evaluated models have publicly available weights and their results are fully reproducible. We still consider it beneficial to also evaluate ChatGPT, as it represents state-of-the-art at the time of writing of this paper and therefore puts the other models' results into perspective.

Another limitation is that based on our empirical experiments, the models are sensitive to the choice of a specific prompting. We spent some time finding a reasonably good prompt that would work with all of the models and did model-specific modifications for the evaluation. Specifically, the desired format of the belief state varied between the models, and there we re some model-specific instructions. We also include both few-shot and zero-shot prompt types in our experiments. However, it is likely that the performance could be further improved with more extensive prompt engineering efforts. Nevertheless, we mainly aim to showcase the more raw/out-of-the-box capabilities of the LLMs, as extensive prompt tuning would, in practice, erase the advantage of not having to finetune the models. Furthermore, we believe that the robustness of the model to specific prompts also counts as an added value.

Finally, we cannot exclude the possibility that some of the models were exposed to our selected datasets during training. However, we still find it important to evaluate the LLMs in this setting.

Acknowledgments

This work was supported by the project TL05000236 *AI asistent pro žáky a učitele* co-financed by the Technological Agency of the Czech Republic within the ÉTA 5 Programme, by the European Research Council (Grant agreement No. 101039303 NG-NLG), and by the Charles University project SVV 260575. It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth and Sports project No. LM2018101).

References

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. ArXiv: 1810.00278.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1866–1875, Hong Kong, China. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on Evaluation Methods for Dialogue Systems. *Artificial Intelligence Review*, 54:755–810.

Yihao Feng, Shentao Yang, Shujian Zhang, Jianguo Zhang, Caiming Xiong, Mingyuan Zhou, and Huan Wang. 2023. Fantastic rewards and how to tame them: A case study on reward learning for task-oriented dialogue

Yue Feng, Yang Wang, and Hang Li. 2021. A sequence-to-sequence approach to dialogue state tracking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1714–1725, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627–2643, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianjian Huang, Shaunak Ashish Halbe, Chinnadhurai Sankar, Pooyan Amini, Satwik Kottur, Alborz Geramifard, Meisam Razaviyayn, and Ahmad Beirami. 2023. Robustness through data augmentation loss consistency. *Transactions on Machine Learning Research*.

Xinting Huang, Jianzhong Qi, Yu Sun, and Rui Zhang. 2020. MALA: cross-domain dialogue generation with action learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7977–7984. AAAI Press.

Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Ponzetto, and Goran Glavaš. 2022. Multi2WOZ: A robust multilingual dataset and conversational pretraining for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3687–3703, Seattle, United States. Association for Computational Linguistics.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. 2021. AuGPT: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 198–210, Online. Association for Computational Linguistics.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online. Association for Computational Linguistics.

Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Language models as few-shot learner for task-oriented dialogue systems. *arXiv preprint arXiv:2008.06239*.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, flavours of success: The case of MultiWOZ. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46, Online. Association for Computational Linguistics.

Tomáš Nekvinda and Ondřej Dušek. 2022. AARGH! end-to-end retrieval-generation for task-oriented dialog. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 283–297, Edinburgh, UK. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Gaurav Pandey, Danish Contractor, Vineet Kumar, and Sachindra Joshi. 2018. Exemplar encoder-decoder for neural conversation generation. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 1329–1338, Melbourne, Australia. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhinav Rastogi, Raghav Gupta, and Dilek Hakkani-Tur. 2018. Multi-task Learning for Joint Language Understanding and Dialogue State Tracking. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 376–384, Melbourne, Australia. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 05, pages 8689–8696.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

I. Shalyminov, A. Sordoni, A. Atkinson, and H. Schulz. 2020. Fast Domain Adaptation for Goal-Oriented Dialogue Using a Hybrid Generative-Retrieval Transformer. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8039–8043. ISSN: 2379-190X.

Igor Shalyminov, Sungjin Lee, Arash Eshghi, and Oliver Lemon. 2019. Data-efficient goal-oriented conversation with dialogue knowledge transfer networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1741–1751, Hong Kong, China. Association for Computational Linguistics.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. 2022. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*.

Haipeng Sun, Junwei Bao, Youzheng Wu, and Xiaodong He. 2022. Mars: Semantic-aware contrastive learning for end-to-end task-oriented dialog. *arXiv preprint arXiv:2210.08917*.

Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. 2020. Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation. In *SIGdial*, pages 297–310, Online.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*.

Tsung-Hsien Wen, Milica Gašić, Dongho Kim, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 275–284, Prague, Czech Republic. Association for Computational Linguistics.

Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 665–677, Vancouver, Canada. Association for Computational Linguistics.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16, pages 14230–14238.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Tiancheng Zhao and Maxine Eskenazi. 2018. Zero-shot dialog generation with cross-domain latent actions. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 1–10, Melbourne, Australia. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. ArXiv:2303.18223 [cs].

Qi Zhu, Christian Geishauser, Hsien chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gašić, and Minlie Huang. 2022. Convlab-3: A flexible dialogue system toolkit based on a unified data format. *arXiv preprint arXiv:2211.17148*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Prompt Construction

Prompt	Determine which domain is considered in the following dialogue situation. Choose exactly one domain from this list: restaurant, hotel, attraction, taxi, train Answer with only one word, the selected domain from the list. You have to always select the most probable domain. ——- Example 1: ——— Customer: I need a cheap place to eat Assistant: We have several not expensive places available. What food are you interested in? Customer: Chinese food. Domain: restaurant —— Example 2: ——— Customer: What is the address? Assistant: It's 123 Northfolk Road. Customer: That's all. I also need a train from London. Domain: train ———— Now complete the following example: Customer: I am looking for a cheap place to stay. Domain:
Output:	hotel

Table 4: A prompt used for domain detection for MultiWOZ. It contains task definition, domains description, static examples and user utterance.

Prompt	Definition: Capture entity values from last utterance of the conversation according to examples. Capture pair "entity:value" separated by colon and no spaces in between. Separate entity:value pairs by hyphens. If not specified, leave the value empty. Values that should be captured are: - "pricerange": the price of the hotel - "area" that specifies the area where the hotel is located (north/east/west/south/centre) - "internet" that specifies if the hotel has internet (yes/no) - "parking" that specifies if the hotel has parking (yes/no) - "stars" that specifies the number of stars the hotel has (1/2/3/4/5) - "type" that specifies the type of the hotel (hotel/bed and breakfast/guest house) [history] Customer: "I want a cheap place to stay."
Output:	pricerange:"cheap"

Table 5: A zero-shot version of the prompt used for state update prediction for MultiWOZ 2.2. It contains task definition, domain description, dialogue history and user utterance.

Prompt	Definition: You are an assistant that helps people to book a hotel. The user can ask for a hotel by name, area, parking, internet availability, or price. There is also a number of hotel in the database currently corresponding to the user's request. If you find a hotel, provide [hotel_name], [hotel_address], [hotel_phone] or [hotel_postcode] Do not provide real entities in the response! Just provide entity name in brackets, like [name] or [address]. If booking, provide [reference] in the answer. [history] Customer: "I want a cheap place to stay." State: hotel { pricerange: "cheap"} Database: hotels: 23
Output:	We have 23 such hotels available, do you have a preference about the location?

Table 6: A zero-shot version of the prompt used for response prediction for MultiWOZ 2.2. It contains task definition, domain description, dialogue history, user utterance and belief state with db results.

Multi-party Goal Tracking with LLMs: Comparing Pre-training, Fine-tuning, and Prompt Engineering

Angus Addlesee, **Weronika Sieińska**, **Nancie Gunson**, **Daniel Hernandez Garcia**, **Christian Dondrup**
Heriot-Watt University, Edinburgh
{a.addlesee, w.sieinska, n.gunson, d.hernandez_garcia, c.dondrup}@hw.ac.uk

Oliver Lemon
Heriot-Watt University, Edinburgh
Alana AI
Edinburgh Centre for Robotics
o.lemon@hw.ac.uk

Abstract

This paper evaluates the extent to which current Large Language Models (LLMs) can capture task-oriented multi-party conversations (MPCs). We have recorded and transcribed 29 MPCs between patients, their companions, and a social robot in a hospital. We then annotated this corpus for multi-party goal-tracking and intent-slot recognition. People share goals, answer each other's goals, and provide other people's goals in MPCs – none of which occur in dyadic interactions. To understand user goals in MPCs, we compared three methods in zero-shot and few-shot settings: we fine-tuned T5, created pre-training tasks to train DialogLM using LED, and employed prompt engineering techniques with GPT-3.5-turbo, to determine which approach can complete this novel task with limited data. GPT-3.5-turbo significantly outperformed the others in a few-shot setting. The 'reasoning' style prompt, when given 7% of the corpus as example annotated conversations, was the best performing method. It correctly annotated 62.32% of the goal tracking MPCs, and 69.57% of the intent-slot recognition MPCs. A 'story' style prompt increased model hallucination, which could be detrimental if deployed in safety-critical settings. We conclude that multi-party conversations still challenge state-of-the-art LLMs.

1 Introduction

Spoken Dialogue Systems (SDSs) are increasingly being embedded in social robots that are expected to seamlessly interact with people in populated public spaces like museums, airports, shopping centres, or hospital waiting rooms (Foster et al., 2019; Tian et al., 2021; Gunson et al., 2022). Unlike virtual agents or voice assistants (e.g. Alexa, Siri, or Google Assistant), which typically have dyadic interactions with a single user, social robots are often approached by pairs and groups of individuals (Al Moubayed et al., 2012; Moujahid et al., 2022). Families may approach a social robot in a museum, and patients are often accompanied by a family member when visiting a hospital. In these multi-party scenarios, tasks that are considered trivial for SDSs become substantially more complex (Traum, 2004; Zhong et al., 2022; Addlesee et al., 2023). In multi-party conversations (MPCs), the social robot must determine which user said an utterance, who that utterance was directed to, when to respond, and what it should say depending on whom the robot is addressing (Hu et al., 2019; Gu et al., 2021, 2022a). These tasks are collectively referred to as "who says what to whom" in the multi-party literature (Gu et al., 2022b), but these tasks alone provide no incentive for a system to actually help a user reach their goals. State of the art "who says what to whom" systems can, therefore, only mimic what a good MPC *looks like* (Addlesee et al., 2023), but for practical systems we also need to know what each user's goals are. We therefore propose two further tasks that become substantially more complex when considered in a multi-party setting: goal tracking and intent-slot recognition (Addlesee et al., 2023).

1	U1:	What time was our appointment?
2	U2:	We have an appointment at 10.30pm.
3	U1:	Ok.

Table 1: An example extract from our new corpus. This example illustrates that people complete other user's goals in an MPC. The system must understand that U1's question was answered by U2, and it does not need to answer this question as if it was a dyadic interaction. Further annotated examples can be found in Table 3.

Dialogue State Tracking (DST) is a well-established task (Lee et al., 2021; Feng et al., 2022) that is considered crucial to the success of a dialogue system (Williams et al., 2016). DST corpora are abundant (Henderson et al., 2014a,b), but they only contain dyadic conversations. No corpus exists containing MPCs with goal tracking or

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 229–241
September 11–15, 2023. ©2023 Association for Computational Linguistics

intent-slot annotations, yet there are important differences. Consider the example in Table 1 (from our new corpus, detailed in Section 2). In turn 1, we can identify that User 1 (U1) wants to know their appointment time. Before the social robot had time to answer, User 2 (U2) answered in turn 2. This obviously does not occur in a dyadic interaction, yet this understanding is essential for natural system behaviour. The SDS must determine that it should not repeat the answer to the question, so data must be collected to learn this. Other major differences exist too. For example, current DST corpora do not contain a concept of 'shared goals' (Eshghi and Healey, 2016). If two people approach a café counter, the barista must determine whether the two people are separate (two individuals wanting to get coffee), or together (two friends with the shared goal to get coffee) (Keizer et al., 2013). The interaction changes depending on this fact, it would be unusual to ask "are you paying together" to two individuals. Shared goals can commonly be identified through explicit dialogue. For example, the use of 'we' in "We are looking for the bathrooms". Similar to answering each other's questions, people may also ask questions on behalf of others. In our corpus, a person said "ARI, the person that I'm accompanying feels intimidated by you, and they'd like to know where they can eat".

In this paper, we present several contributions. (1) We collected a corpus of multi-party interactions between a social robot and patients with their companions in a hospital memory clinic. (2) This corpus was annotated for the standard "who says what to whom" tasks, but also for multi-party goal tracking and intent-slot recognition. We followed current DST annotation instructions, tweaked to enable annotation of multi-party phenomena (detailed in Section 2). (3) We then evaluated Large Language Models (LLMs) on these two new tasks using our collected corpus. Models were pre-trained, fine-tuned, or prompt engineered where applicable (detailed in Section 3). It is not possible to collect enormous corpora from patients in a hospital, so models were evaluated in zero-shot and few-shot settings. We found that the GPT-3.5-turbo model significantly outperformed others on both tasks when given a 'reasoning' style prompt.

2 Dataset and Tasks

For the initial data collection, we partnered with a hospital in Paris, France, and secured ethical approval as part of the EU SPRING project[1]. We then recorded, transcribed, translated (from French to English), anonymised, and annotated 29 multi-party conversations (774 turns). These MPCs were between patients of the memory clinic, their companion (usually a family member), and a social humanoid robot created by PAL Robotics called ARI (Cooper et al., 2020). We hired a professional translator to avoid machine translation errors, and to enable faster experimentation as we are not French speakers. Future work based upon the findings in this paper will be evaluated in both English and French.

We used a wizard-of-oz setup as this task is new, and we required this data to design a multi-party SDS for use in the hospital. A robot operator was therefore controlling what ARI said by selecting one of 31 response options (task-specific answers and some common responses like "yes", "no", "please", "thank you", and "I don't know"). Following our previously published data collection design (Addlesee et al., 2023), each participant was given one or two goals, and asked to converse with ARI to try to achieve their goal. Both participants were given the same goals in some cases to elicit dialogues containing 'shared goal' behaviour. In order to encourage lexical diversity, we provided pictograms to give each participant their goals. For example, if we told the patient that they want a latte, they would likely use the specific word "latte" (Novikova et al., 2016), so we instead gave the participants pictograms as seen in the top-right of Figure 1. This worked as people didn't just ask for coffee when given this image, some asked for hot chocolate or herbal tea instead.

In this paper, we evaluated each model on both multi-party goal tracking, and multi-party intent-slot recognition. These are two related, yet distinct tasks. If ARI asked the user "Are you hungry?", and the user responded "yes", then the intent of that turn is an affirmation, but the user's goal is also established as wanting to eat. As explained in Section 1, standard DST annotation schemes are designed for dyadic interactions, which do not enable annotation of multi-party behaviours. Each turn is annotated with its intent and slot values where applicable, but goal annotations require both the goal and the user whose goal is being established. When a goal is detected in a dyadic interaction, no user information is needed as there is only a single

[1] https://spring-h2020.eu/

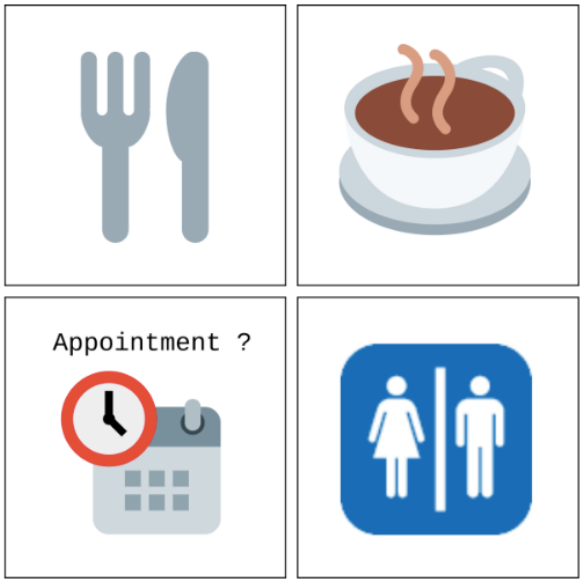


Figure 1: A sample of the pictograms used to represent user goals, given to patients and companions. These elicited dialogues without restricting vocabulary.

user. In multi-party interactions, multiple users can have multiple active goals. These goals may be different, they may be shared (see Table 2), users may answer each other's goals (see Table 1), and one user may provide another user's goal, for example by saying "My wife would love a coffee".

An annotated extract from an MPC in our collected corpus can be found in Table 2. In turn 1, U1 states that "we'd like a coffee", indicating that U1 and their companion U2 would *both* like a coffee. This turn is annotated with two intents: `greet` (due to the "hello"), and `request`. This request intent has a slot value to indicate that the request is for a beverage – coffee. The goal tracking annotation signifies that a goal has been established in this turn with 'G'. The goal is shared by 'U1+U2', and their goal is to drink a coffee. In turn 2, ARI responds informing both users where the café is, hence the `inform` intent annotation. The goal tracking annotation is the same as turn 1, but starts with 'AG' (for 'answer-goal') instead of simply 'G'. This indicates that this goal has been answered, which is critical knowledge for the system to track which goals remain open. In this example, the goal is explicitly closed in turn 3, indicated by the corresponding 'CG' (close-goal) goal tracking annotation. Not all goals are explicitly closed by the user. A dialogue manager could decide to implicitly close an answered goal if the user does not reopen it within three turns, for example. We only annotate explicit goal closures, like the one in turn 3. There are two intents annotated in both turns 1 and 3 in Table 2, and multiple goal annotations can similarly exist, separated by a semicolon. For example, "I'm hungry but need the toilet first" simultaneously opens two goals. All of these annotations were completed using the ELAN tool (Brugman et al., 2004), and then mapped into JSON for model training[2].

With these two sets of annotations, we can evaluate various LLMs on two tasks: (1) multi-party intent-slot recognition; and (2) multi-party goal tracking. It is not possible to collect vast quantities of interactions with patients in the hospital, so these models must be able to learn from a corpus of limited size. We therefore decided to mask annotations in a randomised window selected from each MPC, providing the model with the surrounding context and speaker labels. That is, a random number of turns was selected in each MPC, and then the annotations were replaced by a '[MASK]' token. An example of this is shown in Table 3.

As the corpus size is limited, the window selection could potentially heavily impact model performance. We therefore randomised the selected window three times for each conversation and train/test split, and these *exact same* windows were used to train and test each model. To clarify, all train/test splits and windows were randomised for multiple runs, but they were unchanged between each model. For example, run 1 with the 20/80 split in Section 4 for T5 contained the exact same test set, with the exact same window, as run 1 with the 20/80 split for DialogLED. This holds true for both tasks. Each masked window was bookended with a '[start]' and '[end]' tag to help the models learn this task too (Zhong et al., 2022). A shortened example from our corpus can be seen in Table 3.

3 Experimental Procedure

We evaluated three different models (each detailed below): T5 (Raffel et al., 2020), DialogLM using LED (DialogLED) (Zhong et al., 2022), and GPT-3.5-turbo[3]. Each approach was evaluated in a zero-shot and few-shot setting, with various train/test splits. We could not provide more data to GPT-3.5-turbo due to context window size, but the train/test

[2]Mapping code, annotated data, and training hyperparameters can be found here: https://github.com/AddleseeHQ/mpgt-eval.

[3]https://platform.openai.com/docs/models/gpt-3-5

	User	Utterance	Intent-Slot Annotation	Goal Tracking Annotation
1	U1:	Hello, we'd like a coffee. Where can we go?	greet() ; request(beverage(coffee))	G(U1+U2, drink(coffee))
2	ARI:	You have to enter the building behind you.	inform(directions(cafe))	AG(U1+U2, drink(coffee))
3	U2:	Ok, well thank you very much.	acknowledge(); thank()	CG(U1+U2, drink(coffee))

Table 2: A corpus example displaying shared goals with both intent-slot and goal tracking annotations.

	User	Masked Goal Tracking Utterance	Gold Annotation
1	ARI:	Hello, my name is ARI. How can I help you?	-
		[start]	-
2	U1:	My friend is intimidated by you, where can they eat? [MASK]	G(U2, eat())
3	ARI:	There's a cafeteria on the ground floor, near the courtyard. [MASK]	AG(U2, eat())
		[end]	-
4	U2:	My appointment is in room 17, where is it? G(U2, go-to(room_17))	-

Table 3: A corpus example illustrating the goal tracking task. This process was the same for intent-slot recognition, with the corresponding annotations. Note that U1 asks U2's question, and this is reflected in the annotation.

splits for T5 and DialogLED were: 0/100 (zero-shot), 20/80, 50/50, and 80/20. This allowed us to determine how each model learned to do these tasks when given more training examples. As described in Section 2, we ran each experiment three times with randomised splits and windows, but these remained the same between-models to avoid few-shot problems such as recency bias (Zhao et al., 2021). We trained all the T5-Large and DialogLED models on a machine containing a 16Gb NVIDIA GeForce RTX 3080 Ti GPU with 64Gb RAM and an Intel i9-12900HK processor.

3.1 T5-Large

Older GPT models (GPT-3 and below) are pre-trained with the next token prediction objective on huge corpora (Radford et al., 2019; Brown et al., 2020), an inherently directional task. The creators of T5 added two more objectives and give it the goal of minimising the combined loss function (Raffel et al., 2020) across all three tasks. The two additional tasks were de-shuffling, and BERT-style de-masking (Devlin et al., 2018). This latter pre-training task involves 'corrupting' tokens in the original text, which T5 must then predict. Importantly, this enabled T5 to work bidirectionally, becoming particularly good at using the surrounding context to predict tokens in corrupted sentences. This is not dissimilar to our task, in which the model must learn to use the surrounding MPC turns to predict the annotations that are masked. T5 also achieves state-of-the-art results on related tasks like (Lee et al., 2021; Marselino Andreas et al., 2022), albeit, fine-tuned on larger datasets.

We used T5-Large in both a zero-shot setting, and fine-tuned with various train/test splits. T5 allows fine-tuning with a given named task like 'answer the question', or 'translate from French to German'. We used 'predict goals' and 'predict intent-slots' for goal tracking and intent-slot recognition, respectively, giving the same task names as input during testing. As the corpus is very small, there was no model performance boost beyond 3 epochs, which was expected (Mueller et al., 2022).

3.2 DialogLM using LED (DialogLED)

MPCs reveal unique new communication challenges (Addlesee et al., 2023), as detailed in Section 1, so some LLMs have been developed specifically for the multi-party domain (Hu et al., 2019; Gu et al., 2021, 2022a). Microsoft published DialogLM (Zhong et al., 2022), a pre-trained LLM based upon UniLMv2 (Bao et al., 2020), but specifically designed for multi-party tasks. Alongside the base model, they released two variations: DialogLM-sparse for long dialogues over 5,120 words, and DialogLM using LED (DialogLED) which outperformed the others. DialogLED builds on Longform-Encoder-Decoder (LED) (Beltagy et al., 2020), an attention mechanism that scales linearly with sequence length. Transformer-based models typically scale quadratically with the sequence length, restricting their ability to process long dialogues.

DialogLED was pre-trained on five objectives designed specifically for MPCs, and the model's goal was to minimise the combined loss of all of these tasks. Their state-of-the-art results showed that their pre-training tasks did encourage the LLM to 'understand' multi-party interactions. The five

tasks were: (1) speaker masking, the model has to predict who spoke; (2) turn splitting, the model has to recognise when two utterances are likely the same turn; (3) turn merging, the opposite of (2), where the model has to recognise when the turns were likely separate; (4) text infilling, the model has to predict masked tokens within the turn; and (5) turn permutation, the model has to correctly re-order jumbled turns.

We cloned their repository[4] and added two new tasks: (6) goal masking, the model has to predict goal tracking annotations; and (7) intent-slot masking, the model has to predict intent-slot annotations. In the zero-shot setting, we simply ran the test set through base DialogLED. We then ran their, now modified, code to run our few-shot evaluations three times for each data split.

3.3 GPT-3.5-turbo

Larger LLMs are not inherently better at following a user's intent (Ouyang et al., 2022) as they have no incentive to help the user achieve their goal, only to generate realistic looking outputs. This leads to significant problems, including the generation of false, biased, and potentially harmful responses. GPT-3 was therefore fine-tuned on prompts with human-feedback to create InstructGPT (Ouyang et al., 2022). OpenAI later followed this same approach to create the now famous ChatGPT family of models. At the time of writing, GPT-4 is the most powerful of these models, but it is currently in a waiting list phase. OpenAI recommends their GPT-3.5-turbo model while waiting as the next best option. We therefore decided to evaluate this model on the same two tasks.

Unlike T5 or DialogLED, there is no way to fine-tune your own version of GPT-3.5-turbo, or to edit their pre-training steps. People instead mould the model's behaviour through prompt-engineering (Lester et al., 2021; Wei et al., 2022; Weng, 2023). The newer GPT models allow developers to provide huge contexts, called prompts, containing instructions for the model to follow. GPT-3.5-turbo allows prompts of up to 4,096 tokens. Although these models have only exploded in popularity recently, there are many suggested prompt 'styles' suggested online by conversation designers who are implementing these models in the real-world. We have analysed this space and devised six prompt styles for the two tasks. In the zero-shot setting, only the prompt and the masked MPC is provided to the model. In the few-shot setting, we additionally provide the model with 7% of the corpus as examples. This is crucial to highlight. T5 and DialogLED were trained on 20% of the corpus, 50% of the corpus, and finally 80% of the corpus. GPT-3.5-turbo's maximum context size can only fit 7% of the corpus, less than the other models.

The prompt styles we used were the following (the actual prompts are included in Appendix A):

- **Basic**: This is our baseline prompt. It very simply tells the model what it is going to get as input, and what we want as output. It contains no further special instructions.
- **Specific**: GPT practitioners report that when prompts are more detailed and specific, performance is boosted (Ye et al., 2023).
- **Annotation**: For annotation tasks, we would give fellow humans annotation instructions. In this prompt, we provide the model with annotation instructions.
- **Story**: This model was pre-trained on a very large quantity of data, including novels, film scripts, journalistic content, etc... It may be possible that by phrasing the prompt like a story, performance may be boosted due to its likeness to its training data.
- **Role-play**: Similar to the story prompt, it is reported that these models are very good at role-playing[5]. People ask ChatGPT to pretend to be a therapist, a lawyer, or even alter-egos that have no safety limitations (Taylor, 2023). We tell GPT-3.5-turbo that it is a 'helpful assistant listening to a conversation between two people and a social robot called ARI'.
- **Reasoning**: Finally, recent work suggests that these models improve in performance if you explain the reasoning for desired outputs (Fu et al., 2022). We therefore added one fictitious turn to this prompt, and explained the reasoning behind its annotation.

4 Results

We evaluated T5, DialogLED, and GPT-3.5-turbo as described in Section 3 on multi-party goal track-

[4] https://github.com/microsoft/DialogLM

[5] https://github.com/f/awesome-chatgpt-prompts

ing, and multi-party intent-slot recognition. Outputs were annotated as either 'exact', 'correct', or 'partial' to distinguish each model's performance beyond simple accuracy. Exact matches were strictly annotated, but slight differences are allowed if the annotation meaning remains unchanged. For example: '`G(U1, go-to(lift))`' and '`G(U1, go-to(lifts))`' (note the plural 'lifts'). Outputs were marked as exact if every [MASK] in the MPC was exact, and marked as correct if every [MASK] was more broadly accurate. For example, if the annotation contained '`drink(coffee)`' and the model output '`drink(hot_drink)`', we considered this correct. The output was marked as partially correct if at least 60% of the [MASK] tags were correctly annotated. This latter metric allows us to distinguish between models that generate nonsense, and those that roughly grasp the task. Our inter-annotator agreements were 0.765 and 0.771 for goal tracking and intent-slot recognition, respectively. These are less than 0.8, and this was due to the broad definition of 'correct'. We plan to design automatic metrics for our future work (see Section 5).

4.1 MPC Goal Tracking Results

The goal tracking results can be found in Table 4. An ANOVA test (Fisher, 1992) indicated that there was an overall significant difference between the model's results. We therefore ran a Tukey HSD test (Tukey, 1949) that showed that the GPT-3.5-turbo model in the few-shot setting did significantly outperform all the other models.

Firstly, the T5-Large model performed poorly, even when it was trained on 80% of our corpus. Upon further analysis, it generated complete nonsense in the zero-shot setting, but did start to generate strings that looked reasonable with only 20% of the data. Given the 50/50 train/test split, T5 consistently replaced the [MASK] tokens, but did still hallucinate turns. When given 80% of the data as training data, the T5 model preserved the original dialogue, and replaced the [MASK] tokens with goal annotations, they were just all completely wrong. This steady improvement as we increased the amount of training data suggests that T5 could be a viable option for similar tasks, just not where data is limited (such as our hospital use case).

The DialogLED model also generated nonsense in the zero-shot setting, but very quickly learned the task. Even with just 20% of the data used for training, DialogLED reliably preserved the original dialogue and replaced the [MASK] tokens with goal annotations. Most of the annotations were incorrect, for example '`G(U2, eat(ticket))`', but DialogLED did correctly detect some goals opening, being answered, and being closed correctly, achieving a non-zero partial score. Given more training data, DialogLED did begin to use the surrounding contextual dialogue turns more accurately, but almost every result contained an incorrect prediction. This was often the mis-detection of shared goals, or closing goals early. Like T5, DialogLED would need a larger training set to accurately complete this task. This model learned the task quickly, so may need fewer examples.

In the zero-shot setting, GPT-3.5-turbo roughly 'understood' the task, generating many partially correct outputs. With all the prompt styles, it did frequently reformat the dialogue. This was particularly true when using the roleplay prompt, it would output all the goals per interlocutor, for example, rather than per turn. The worst zero-shot GPT-3.5-turbo prompt was the 'story' style, not even generating one partially correct output. This was due to its increased hallucination. The story prompt noticeably produced more fictitious turns, and also rephrased and removed turns in the original dialogue. We believe this is likely because a story scenario is naturally a fictitious topic. The 'reasoning' style prompt performed remarkably well, generating five times more correct outputs than the second-best prompt style, and generating 79.31% partially correct outputs, showing that it can grasp the concept of the task. The reasoning prompt commonly mis-identified shared goals, unfortunately.

In the few-shot setting, GPT-3.5-turbo's results improved significantly compared to every other approach. We would like to highlight again that each run's example prompts provided to the model were exactly the same for each prompt style. Performance differences were only due to the given prompt style. The 'reasoning' prompt once again outperformed the others across all metrics, generating correct outputs 62.32% of the time, and partially correct 94.20% of the time. In our future work (see Section 5), we plan to utilise this prompt style's impressive performance on limited data. The 'story' prompt was the only style to successfully attribute goals to other speakers, as in Table 3, but it still suffered from increased hallucination, which is not appropriate in a safety-critical

Model	train/test %	Prompt Style	Exact %	Correct %	Partial %
T5	0/100	-	0	0	0
T5	20/80	-	0 ± 0	0 ± 0	0 ± 0
T5	50/50	-	0 ± 0	0 ± 0	0 ± 0
T5	80/20	-	0 ± 0	0 ± 0	0 ± 0
DialogLED	0/100	-	0	0	0
DialogLED	20/80	-	0 ± 0	0 ± 0	5.80 ± 1.45
DialogLED	50/50	-	0 ± 0	2.38 ± 2.38	1.19 ± 0.63
DialogLED	80/20	-	0 ± 0	0 ± 0	20 ± 11.55
GPT 3.5-turbo	0/100	Basic	0	3.45	31.03
GPT 3.5-turbo	0/100	Specific	0	3.45	24.14
GPT 3.5-turbo	0/100	Annotation	0	6.90	44.83
GPT 3.5-turbo	0/100	Story	0	0	0
GPT 3.5-turbo	0/100	Role-play	0	0	6.90
GPT 3.5-turbo	0/100	Reasoning	3.45	34.48	79.31
GPT 3.5-turbo	7/80*	Basic	11.59 ± 3.83	30.43 ± 10.94	86.96 ± 6.64
GPT 3.5-turbo	7/80*	Specific	20.29 ± 3.83	43.48 ± 9.05	92.75 ± 2.90
GPT 3.5-turbo	7/80*	Annotation	14.49 ± 5.80	28.99 ± 3.83	82.61 ± 4.35
GPT 3.5-turbo	7/80*	Story	17.39 ± 6.64	36.23 ± 13.83	86.96 ± 4.35
GPT 3.5-turbo	7/80*	Role-play	18.84 ± 7.25	46.38 ± 12.38	92.75 ± 5.22
GPT 3.5-turbo	7/80*	Reasoning	**27.54 ± 1.45**	**62.32 ± 9.50**	**94.20 ± 5.80**

Table 4: The final multi-party goal tracking results for each model in both the zero- and few-shot settings. *We could not fit more than 7% of the training examples in GPT-3.5-turbo's context window. We therefore used fewer examples than with T5 and DialogLED. The same 80% test sets were still used to enable model comparison.

setting. We suspect that the other prompt styles failed to do this because of the rarity of this phenomenon in our corpus. We are eliciting more of these in ongoing experiments with a deployed system, not wizard-of-oz (Addlesee et al., 2023).

4.2 MPC Intent-slot Recognition Results

The results for each model on the intent-slot recognition task can be found in Table 5. As with the goal tracking results, an ANOVA test (Fisher, 1992) indicated that there was an overall significant difference between our model's results. We therefore ran a Tukey HSD test (Tukey, 1949) that showed that the GPT-3.5-turbo model in the few-shot setting significantly outperformed all the other models.

As intent-slot annotations are well-established, T5 and DialogLED both started generating sensible-*looking* outputs with only a few training examples. The T5 outputs were all incorrect again, however. DialogLED consistently improved as it was trained on progressively more data, annotating almost half of the MPCs partially correctly, and beginning to accurately annotate full MPCs. Given a larger corpus, we expect that DialogLED could potentially generate competitive results, but this is not the case for T5 in this setting with limited data.

GPT-3.5-turbo in the zero-shot setting also achieved higher partial scores, compared to the goal tracking results, due to the fact that intent-slot recognition is a more established task. Turns were commonly annotated with multiple gold goals, but this model tended to only output one per turn. For example: "Hello ARI, where is the café?" would only have the prediction 'greet', missing the request to locate the café entirely. This prevented the model from achieving higher correct scores.

In the few-shot setting, however, GPT-3.5-turbo significantly outperformed all the other models. The difference was remarkable. Almost all of the predictions were partially correct, and the 'reasoning' prompts correctly annotated 70% of the MPCs. Other models tended to falter when anaphoric expressions couldn't be resolved with just the previous turn. They also struggled to identify the '`suggest`' intent, for example, when one person said "do you want to go to the toilet?". These were misclassified as request intents, likely due to their prominence in the corpus, and influence on the results due to GPT-3.5-turbo's limited input context.

Model	train/test %	Prompt Style	Exact %	Correct %	Partial %
T5	0/100	-	0	0	0
T5	20/80	-	0 ± 0	0 ± 0	0 ± 0
T5	50/50	-	0 ± 0	0 ± 0	0 ± 0
T5	80/20	-	0 ± 0	0 ± 0	0 ± 0
DialogLED	0/100	-	0	0	0
DialogLED	20/80	-	0 ± 0	0 ± 0	5.80 ± 2.90
DialogLED	50/50	-	0 ± 0	0 ± 0	38.10 ± 10.38
DialogLED	80/20	-	0 ± 0	13.33 ± 6.67	46.67 ± 6.67
GPT 3.5-turbo	0/100	Basic	0	3.45	51.72
GPT 3.5-turbo	0/100	Specific	0	0	13.79
GPT 3.5-turbo	0/100	Annotation	0	3.45	20.69
GPT 3.5-turbo	0/100	Story	0	0	24.14
GPT 3.5-turbo	0/100	Role-play	0	0	20.69
GPT 3.5-turbo	0/100	Reasoning	0	27.59	82.76
GPT 3.5-turbo	7/80*	Basic	17.39 ± 6.64	36.23 ± 12.88	97.10 ± 2.90
GPT 3.5-turbo	7/80*	Specific	27.54 ± 1.45	60.87 ± 9.05	94.20 ± 1.45
GPT 3.5-turbo	7/80*	Annotation	18.84 ± 1.45	40.58 ± 6.32	91.30 ± 4.35
GPT 3.5-turbo	7/80*	Story	26.09 ± 4.35	47.83 ± 10.04	94.20 ± 3.83
GPT 3.5-turbo	7/80*	Role-play	20.29 ± 3.83	49.27 ± 12.88	97.10 ± 1.45
GPT 3.5-turbo	7/80*	Reasoning	**37.68 ± 1.45**	**69.57 ± 10.94**	**100 ± 0**

Table 5: The final multi-party intent-slot recognition results for each model in both the zero- and few-shot settings. *We could not fit more than 7% of the training examples in GPT-3.5-turbo's context window. We therefore used fewer examples than with T5 and DialogLED. The same 80% test sets were still used to enable model comparison.

5 Conclusion and Future Work

Multi-party conversations (MPCs) elicit complex behaviours which do not occur in the dyadic interactions that today's dialogue systems are designed and trained to handle. Social robots are increasingly being expected to perform tasks in public spaces like museums and malls, where conversations often include groups of friends or family. Multi-party research has previously focused on speaker recognition, addressee recognition, and tweaking response generation depending on whom the system is addressing. While this work is vital, we argue that these collective "who says what to whom" tasks do not provide any incentive for the social robot to complete user goals, and instead encourage it to simply mimic what a good MPC *looks like*. In this paper, we have detailed how the tasks of goal tracking and intent-slot recognition differ in a multi-party setting, providing examples from our newly collected corpus of MPCs in a hospital. We found that, given limited data, 'reasoning' style prompts enable GPT-3.5-turbo to perform significantly better than other models.

We found that other prompt styles also perform well, but prompts that are story-like increase model hallucination. With the introduction of prompt fine-tuning with human feedback (Ouyang et al., 2022), generative LLMs do now have some incentive to avoid misleading or harming the user, providing outputs prepended with caveats, but the issue is not solved. OpenAI claims that GPT-4 generates 40% fewer hallucinations than GPT-3 (Hern and Bhuiyan, 2023), but these models should still not be applied directly in a hospital or other safety-critical setting without further evaluation. In the hospital setting, users are more likely to be from vulnerable population groups, and are more likely to be older adults that are not familiar with the capabilities of today's models. Multiple researchers and hospital staff members are present when conducting our data collections, so that if hallucinations do occur, they can be quickly corrected. We will, therefore, be able to evaluate response grounding, Guidance[6], and other hallucination prevention strategies to determine whether these models can ever be used safely in a high-risk setting. These further experiments will also elicit further MPCs that can be annotated for various multi-party tasks.

User inputs must be processed on external

[6] https://github.com/microsoft/guidance

servers when using industry LLMs, like GPT-3.5-turbo and Google's Bard. For this reason, these specific models cannot be deployed in the hospital setting. Patients may reveal identifiable or sensitive information during our data collection, which we subsequently remove from the corpus. This data must stay contained within approved data-controlled servers in the SPRING project. In this paper, we have reported the remarkable performance of an industry LLM, when given limited data, compared to prior model architectures. We will analyse open and transparent instruction-tuned text generators (Liesenfeld et al., 2023), which are able to meet our data security requirements.

The accessibility of today's SDSs is critical when working with hospital patients (Addlesee, 2023). Speech production differs between the 'average' user, and user groups that remain a minority in huge training datasets. For example, people with dementia pause more frequently and for longer durations mid-sentence due to word-finding problems (Boschi et al., 2017; Slegers et al., 2018). We are utilising knowledge graphs to ensure that SDSs are transparent, controllable, and more accessible for these user groups (Addlesee and Eshghi, 2021; Addlesee and Damonte, 2023a,b), and we see the unification of large language models and knowledge graphs (Pan et al., 2023) as the near-term future of our field.

We plan to design and run subsequent experiments in both the hospital memory clinic, and a newly established mock waiting room in our lab. This space will allow us to collect additional MPCs with more than two people, replicating scenarios in which whole families approach a social robot. We plan to evaluate whether prompt engineering can work modularly for N users. For example, we could use GPT-4 to correct speaker diarization (Murali et al., 2023), then to handle multi-party goal tracking, and then to generate responses to the user. This experimental setup will allow us to quickly test new ideas, such as automatic prompt optimization (Pryzant et al., 2023) in the lab, maximising the benefit of patients' time in the hospital.

Acknowledgements

This research was funded by the EU H2020 program under grant agreement no. 871245 (https://spring-h2020.eu/). We would also like to thank our anonymous reviewers for their time and valuable feedback.

References

Angus Addlesee. 2023. Voice assistant accessibility. In *The International Workshop on Spoken Dialogue Systems Technology, IWSDS 2023*.

Angus Addlesee and Marco Damonte. 2023a. Understanding and answering incomplete questions. In *Proceedings of the 5th Conference on Conversational User Interfaces*.

Angus Addlesee and Marco Damonte. 2023b. Understanding disrupted sentences using underspecified abstract meaning representation. In *Interspeech*.

Angus Addlesee and Arash Eshghi. 2021. Incremental graph-based semantics and reasoning for conversational AI. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 1–7, Gothenburg, Sweden. Association for Computational Linguistics.

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández García, Christian Dondrup, and Oliver Lemon. 2023. Data collection for multi-party task-based dialogue in social robotics. In *The International Workshop on Spoken Dialogue Systems Technology, IWSDS 2023*.

Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*, pages 114–130. Springer.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International conference on machine learning*, pages 642–652. PMLR.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. 2017. Connected speech in neurodegenerative language disorders: a review. *Frontiers in psychology*, 8:269.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hennie Brugman, Albert Russel, and Xd Nijmegen. 2004. Annotating multi-media/multi-modal resources with elan. In *LREC*, pages 2065–2068.

Sara Cooper, Alessandro Di Fava, Carlos Vivas, Luca Marchionni, and Francesco Ferro. 2020. Ari: The social assistive robot and companion. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 745–751. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Arash Eshghi and Patrick GT Healey. 2016. Collective contexts in conversation: Grounding by proxy. *Cognitive science*, 40(2):299–324.

Yue Feng, Aldo Lipani, Fanghua Ye, Qiang Zhang, and Emine Yilmaz. 2022. Dynamic schema graph fusion network for multi-domain dialogue state tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 115–126.

Ronald Aylmer Fisher. 1992. *Statistical methods for research workers*. Springer.

Mary Ellen Foster, Bart Craenen, Amol Deshmukh, Oliver Lemon, Emanuele Bastianelli, Christian Dondrup, Ioannis Papaioannou, Andrea Vanzo, Jean-Marc Odobez, Olivier Canévet, et al. 2019. MuMMER: Socially intelligent human-robot interaction in public spaces. *arXiv preprint arXiv:1909.06749*.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

Jia-Chen Gu, Chao-Hong Tan, Chongyang Tao, Zhen-Hua Ling, Huang Hu, Xiubo Geng, and Daxin Jiang. 2022a. HeterMPC: A Heterogeneous Graph Neural Network for Response Generation in Multi-Party Conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5086–5097.

Jia-Chen Gu, Chongyang Tao, and Zhen-Hua Ling. 2022b. WHO Says WHAT to WHOM: A Survey of Multi-Party Conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3682–3692.

Nancie Gunson, Daniel Hernández Garcia, Weronika Sieińska, Angus Addlesee, Christian Dondrup, Oliver Lemon, Jose L Part, and Yanchao Yu. 2022. A visually-aware conversational robot receptionist. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 645–648.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014a. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014b. The third dialog state tracking challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 324–329. IEEE.

Alex Hern and Johana Bhuiyan. 2023. Openai says new model gpt-4 is more creative and less likely to invent facts. *The Guardian*.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. GSN: A graph-structured network for multi-party dialogues. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*.

Simon Keizer, Mary Ellen Foster, Oliver Lemon, Andre Gaschler, and Manuel Giuliani. 2013. Training and evaluation of an MDP model for social multi-user human-robot interaction. In *Proceedings of the SIGDIAL 2013 Conference*, pages 223–232.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4937–4949.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–6.

Vinsen Marselino Andreas, Genta Indra Winata, and Ayu Purwarianti. 2022. A comparative study on language models for task-oriented dialogue systems. *arXiv e-prints*, pages arXiv–2201.

Meriam Moujahid, Helen Hastie, and Oliver Lemon. 2022. Multi-party interaction with a robot receptionist. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 927–931. IEEE.

Aaron Mueller, Jason Krone, Salvatore Romeo, Saab Mansour, Elman Mansimov, Yi Zhang, and Dan Roth. 2022. Label semantic aware pre-training for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8318–8334.

Prasanth Murali, Ian Steenstra, Hye Sun Yun, Ameneh Shamekhi, and Timothy Bickmore. 2023. Improving multiparty interactions with a robot using large language models. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing NLG data: Pictures elicit better data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 265–273, Edinburgh, UK. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying large language models and knowledge graphs: A roadmap. *arXiv preprint arXiv:2306.08302*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Antoine Slegers, Renee-Pier Filiou, Maxime Montembeault, and Simona Maria Brambati. 2018. Connected speech features from picture description in alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease*, 65(2):519–542.

Josh Taylor. 2023. Chatgpt's alter ego, dan: users jailbreak ai program to get around ethical safeguards. *The Guardian*.

Leimin Tian, Pamela Carreno-Medrano, Aimee Allen, Shanti Sumartojo, Michael Mintrom, Enrique Coronado Zuniga, Gentiane Venture, Elizabeth Croft, and Dana Kulic. 2021. Redesigning human-robot interaction in response to robot failures: A participatory design methodology. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–8.

David Traum. 2004. Issues in multiparty dialogues. In *Advances in Agent Communication: International Workshop on Agent Communication Languages, ACL 2003, Melbourne, Australia, July 14, 2003. Revised and Invited Papers*, pages 201–211. Springer.

John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Lilian Weng. 2023. Prompt engineering. *lilianweng.github.io*.

Jason Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Seonghyeon Ye, Hyeonbin Hwang, Sohee Yang, Hyeongu Yun, Yireun Kim, and Minjoon Seo. 2023. In-context instruction learning. *arXiv preprint arXiv:2302.14691*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. DialogLM: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

A Full GPT-3.5-turbo Prompts

Here are the full prompts given to GPT-3.5-turbo for each task. We used six styles described in Section 3. The masked MPC was appended to each prompt in the zero-shot setting. In the few-shot prompts (see Section A.2), we appended examples with "input:" + masked MPC #1 + "output:" + gold output #1 + 'input:" + masked MPC #2 + "output:" + gold output #2 + "input:" + test set masked MPC + "output:"[7].

A.1 Zero-shot Goal Tracking

- **Basic**: This conversation has a window between [start] and [end]. Return this window with the [MASK] tags replaced with the goal annotations:

- **Specific**: This is a conversation between two people and a robot called ARI. There is a section of the conversation between the [start] and [end] tags. I want you to return this section of the conversation, but I want you to replace the [MASK] tags with the user goals. Do not change any of the other words

[7]The examples given were randomised per run, and the appendix page limit doesn't fit the full 4,096 token prompts.

in the section, only replace [MASK]. Every [MASK] should be replaced. Here is the conversation:

- **Annotation**: This is a conversation between two people and a robot called ARI. I want you to first extract the text between [start] and [end]. There are [MASK] tags in the extracted text. I want you to replace the [MASK] tags with goal annotations. Do not change any of the other text. If the person's goal can be determined by that turn, add an '@' symbol followed by 'G' (G for goal), and then brackets with the speaker ID and what their goal is. If it is a shared goal, you can annotate both speakers with a '+' sign between them. For example, if you think U1 and U2 share the goal, you can write U1+U2. If you think the goal is being answered, you can do the same but with 'AG' (AG for Answer Goal) instead of 'G'. Finally, if you think the person is closing the goal, you can do the same annotation using 'CG' (CG for Close Goal) instead of 'G' or AG'. Here is the conversation:

- **Story**: There once was a conversation between a patient, a companion, and a robot called ARI. One bit of the conversation was confusing. A helpful researcher noted the start with [start], and the end with [end]. The confusing bits are marked with [MASK]. Can you help us figure out the goals that should replace the [MASK] tags? The conversation is this:

- **Role-play**: You are listening to a conversation between two people and a robot called ARI. You are a helpful assistant that needs to figure out what goals the people have. You need to pay attention to the [MASK] tags between the [start] and [end] tags in the given conversation. Your job is to replace these [MASK] tags with the correct goal annotations. Here is the conversation:

- **Reasoning**: I will give you a conversation between two people and a robot called ARI. You need to return the text between [start] and [end] with the [MASK] tags replaced by user goals. Let's step through how to figure out the correct annotation. If the conversation included 'U1: I really need the toilet [MASK]', then we would first know that the speaker is called U1. The turn also ends with [MASK], so we know that we need to replace it with a goal. We know that U1 needs the toilet, so their goal is to go to the nearest toilet. Goals always begin with the '@' symbol, and then a 'G' if we have found a person's goal. We would therefore replace [MASK] with @ G(U1, go-to(toilet)). If someone tells U1 where the toilets are, they have answered their goal. We would therefore annotate that turn with @ AG(U1, go-to(toilet)). We use AG here to indicate Answer Goal. Finally, if U1 then said thank you, we know their goal has been met. We would annotate the thank you with @ CG(U1, go-to(toilet)) because U1's goal is finished. CG stands for Close Goal. Do this goal tracking for each [MASK] in this conversation:

A.2 Few-shot Intent-slot Recognition

- **Basic**: Each conversation has a window between [start] and [end]. Return this window with the [MASK] tags replaced with the intent-slot annotations. Here are some examples.

- **Specific**: Each of these conversations is between two people and a robot called ARI. There is a section of each conversation between the [start] and [end] tags. I want you to return this section of the conversation, but I want you to replace the [MASK] tags with the user intents and slots. Do not change any of the other words in the section, only replace [MASK]. Every [MASK] should be replaced. Here are some examples.

- **Annotation**: Each of these conversations is between two people and a robot called ARI. I want you to first extract the text between [start] and [end]. There are [MASK] tags in the extracted text. I want you to replace the [MASK] tags with intent-slot annotations. Do not change any of the other text. If the person's intent can be determined by that turn, add a '#' symbol followed by their intent and then brackets with the slots within. There are not always slots, so the brackets can be empty. Sometimes there are multiple intents, split them with a semi-colon ';'. Here are some examples.

- **Story**: There once was a conversation between a patient, a companion, and a robot called ARI. One bit of the conversation was

confusing. A helpful researcher noted the start with [start], and the end with [end]. The confusing bits are marked with [MASK]. Can you help us figure out the intents and slots that should replace the [MASK] tags? Here are some examples.

- **Role-play**: You are listening to a conversation between two people and a robot called ARI. You are a helpful assistant that needs to figure out what goals the people have. You need to pay attention to the [MASK] tags between the [start] and [end] tags in the given conversation. Your job is to replace these [MASK] tags with the correct intent-slot annotations. Here are some examples.

- **Reasoning**: I will give you a conversation between two people and a robot called ARI. You need to return the text between [start] and [end] with the [MASK] tags replaced by user intents and slots. Let's step through how to figure out the correct annotation. If the conversation included 'U1: Hello, I'd like to know where the doctor's office is? [MASK]' then we know there is a missing intent-slot annotation because of the [MASK] tag. U1 first said hello, greeting their interlocutor, so we know their intent is greet. This has no slots, so we have the annotation '# greet()' to start. U1 also asked where the doctor is, so their second intent is a request. The slot is the room that the doctor is in, as that is what they are requesting. Their second intent is therefore '# request(doctor(room)). As there are multiple intents, the [MASK] is replaced by '# greet() ; request(doctor(room))'. The ';' is only used because there was more than one intent. Do this intent-slot annotation for each [MASK] in this conversation. Here are some examples.

ChatGPT vs. Crowdsourcing vs. Experts: Annotating Open-Domain Conversations with Speech Functions

Lidiia Ostyakova[1,2]* ostyakova.ln@gmail.com
Veronika Smilga[1]* smilgaveronika@gmail.com
Kseniia Petukhova [1]* petukhova.ka@mipt.ru
Maria Molchanova[1] molchanova@deeppavlov.ai
Daniel Kornev[1] daniel@kornevs.org
[1]Moscow Institute of Physics and Technology, Russia
[2]HSE University, Russia

Abstract

This paper deals with the task of annotating open-domain conversations with speech functions. We propose a semi-automated method for annotating dialogs following the topic-oriented, multi-layered taxonomy of speech functions with the use of hierarchical guidelines using Large Language Models. These guidelines comprise simple questions about the topic and speaker change, sentence types, pragmatic aspects of the utterance, and examples that aid untrained annotators in understanding the taxonomy. We compare the results of dialog annotation performed by experts, crowdsourcing workers, and ChatGPT. To improve the performance of ChatGPT, several experiments utilising different prompt engineering techniques were conducted. We demonstrate that in some cases large language models can achieve human-like performance following a multi-step tree-like annotation pipeline on complex discourse annotation, which is usually challenging and costly in terms of time and money when performed by humans.

1 Introduction

Discourse analysis as a method of an abstract dialog representation is used in various NLP tasks: dialog management (Liang et al., 2020; Galitsky and Ilvovsky, 2017), dialog generation (Yang et al., 2022; Gu et al., 2021), dialog summarization (Chen et al., 2021), emotion recognition (Shou et al., 2022), etc. Mostly, discourse structure is considered to be an interconnected system of linguistic features such as a topic, pragmatics, and semantics. One of the main goals of discourse analysis is to describe pragmatics of actions performed by speakers within a communicative process, i.e., characterise the interlocutors' intentions at a certain moment of their interaction (Coulthard, 2014).

Despite the fact that there are numerous theoretical approaches to dialog discourse analysis, only a few of them consider the complexity of conversational nature and allow for annotation on multiple levels (Bunt et al., 2010, 2012; Cai et al., 2023). In this paper, we propose a multi-dimensional and hierarchical taxonomy of speech functions introduced by Eggins and Slade (2004) as an alternative for abstract dialog representation. In contrast to other annotation schemes, this taxonomy is topic-oriented and includes classes that are very similar in terms of pragmatics (see Appendix B). The taxonomy provides a comprehensive, systematic discourse model of dialogues (see Figure1).

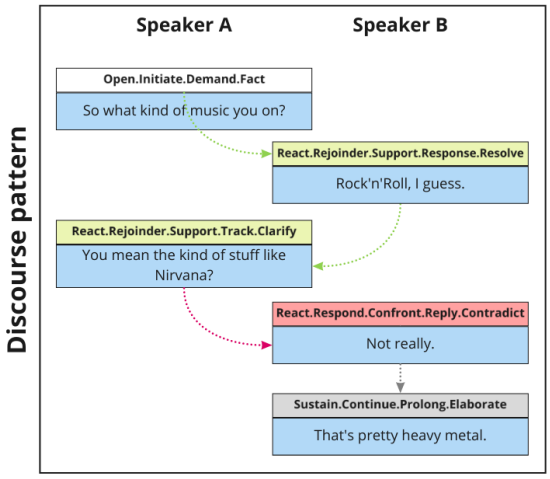


Figure 1: Example of Dialog Annotation with Speech Functions

Traditionally, discourse annotation is performed manually by trained experts or crowdsourcing workers (Hoek and Scholman, 2017). Automating it partially or entirely is key to making this complicated process faster and cheaper. We argue that in complex discourse annotation tasks Large Language Models (LLMs) can be used to establish decent quality silver standards that would be later checked and improved by expert annotators.

In this paper, we annotate DailyDialog (Li et al., 2017), a multi-turn casual dialog dataset, using

*These authors contributed equally to this work

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 242–254
September 11–15, 2023. ©2023 Association for Computational Linguistics

the speech function taxonomy. The annotation is conducted in three ways: 1) by experts with at least B.A. in Linguistics; 2) by workers of Toloka [1], a crowdsourcing platform; 3) with the use of a large language model, specifically, ChatGPT (gpt−3.5−turbo). We then compare the performance of crowdsourcers and ChatGPT using expert annotation results as the gold standard and analyse the findings to prove that LLMs can achieve human-like performance on complex discourse annotation tasks. Finally, we release the repository with all the code we used to perform the annotation with ChatGPT [2].

2 Related Work

Theoretical Approaches to Discourse Analysis There are two basic theoretical approaches to the abstract dialog representation: Segmented Discourse Representation theory (SDRT) (Lascarides and Asher, 2007), which applies principles of Rhetorical Structures theory (RST) (Mann and Thompson, 1988) to the dialog, and theory of dialog acts (DA theory) (Core and Allen, 1997). According to the SDRT style, firstly, a relation between two elementary discourse units (EDUs) needs to be defined and then characterized with a discourse class (for instance, Question-Answer, Clarification, etc.). While SDRT represents a dialog structure as a graph (Asher et al., 2016; Li et al., 2020), most of DA theory interpretations such as DAMSL (Allen and Core, 1997), SWBD-DAMSL (Jurafsky, 1997), MIDAS (Yu and Yu, 2019) describe it sequentially giving pragmatic characteristics to each EDU. In addition, most classes used in DA taxonomies do not represent pragmatic purposes but rather focus on semantics or grammar form of utterances within a dialog, using tags such as 'yes/no question', 'statement', 'positive answer'.

To represent the discourse structure of dialogs in a more advanced way, Bunt et al. (2010, 2012) suggested Dialogue Annotation Markup Language (DiAML), a taxonomy including nine functional dimensions and 49 specific classes. Even though DiAML is claimed to be an ISO standard for DA annotation, it is challenging to apply it to real-world problems for several reasons. First, DiAML supports multi-label annotation, i.e., several classes can be assigned to one EDU, which complicates automatic classification. Moreover, there is not enough labelled data to experiment with the taxonomy. One more taxonomy designed to represent a conversational structure on several levels is Dependency Dialogue Acts (DDA) (Cai et al., 2023). A combination of dialog acts and rhetorical relations in the SDRT style showed a potential of applying multi-layered and multi-dimensional approaches for analyzing discourse structure within conversations. However, because there is no annotated data with this taxonomy, it is not clear whether it is applicable to automated tasks.

The taxonomy of speech functions is an alternative multidimensional scheme for discourse annotation introduced by Eggins and Slade (2004). It is multi-layer and hierarchical, which allows us to analyze dialog structure in a consistent manner. Unlike other multidimensional schemes, the taxonomy of speech functions supports single-label annotation. While inheriting the principle of assigning one label to a specific EDU from DA theory, speech functions taxonomy also considers relationships between utterances following the SDRT style. The tag of a current label is determined in connection with the previous one, so it is important to take into account the utterances' previous context when assigning the correct label. The potential of applying the taxonomy to manage a conversational flow within dialog systems is proven by several studies (Mattar and Wachsmuth, 2012; Kuznetsov et al., 2021; Baymurzina et al., 2021).

Large Language Models for Discourse Annotation In the recent years, the paradigm of training and using NLP models has undergone significant changes. With the advance of Large Language Models (LLMs), the focus has shifted from the previously dominating "pre-train, fine-tune" procedure to "pre-train, prompt, and predict" (Liu et al., 2023), where an LLM is applied to downstream tasks directly. In this case, textual prompts are used to guide the models' behaviour and achieve the desired output without additional fine-tuning. Scaling up LLMs to billions of parameters leads to significantly improved results in terms of few-shot and zero-shot prompting (Brown et al., 2020; Wei et al., 2021, *i.a*). However, as the objective of training most LLMs is not following the instructions but simply predicting the next token, they may fail to perform the task. One solution is fine-tuning LLMs using Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017) to align

[1] https://toloka.ai/tolokers/

[2] https://github.com/deeppavlov/sf_corpus/

its behaviour in accordance with the trainers' values and needs (Ouyang et al., 2022). An example of such model is ChatGPT (OpenAI, 2022) that has shown state-of-the-art or comparable performance on a number of NLP tasks in few-shot or zero-shot setting and provoked a tide of research articles testing its capabilities in areas ranging from coding and bug-fixing (Tian et al., 2023; Kashefi and Mukerji, 2023; Sobania et al., 2023, *i.a*) to medical applications (Nori et al., 2023; Kung et al., 2023, *i.a*).

There have already been claims that ChatGPT (gpt−3.5−turbo) and (gpt−4) versions alike) outperforms crowdsourcing workers on a number of annotation tasks while being significantly cheaper and faster. The tasks in question included annotation of relevance, stance, topics, and frames detection (Gilardi et al., 2023); political affiliation classification of tweets (Törnberg, 2023); hate speech detection (Huang et al., 2023; Li et al., 2023; Zhu et al., 2023); sentiment analysis and bot detection (Zhu et al., 2023). In the above-listed works, the approach to obtaining the final label was straightforwardly simple. With one prompt containing textual instruction and a datapoint, the model either answered a single question about the datapoint, assigning a label to it, or scored the probability of the datapoint belonging to some class.

However, there still have been no attempts to apply LLMs to complex annotation tasks that deal with tens of labels and require multi-step reasoning. In this work, we test whether it is possible for LLMs to achieve human-like performance on such tasks. In particular, we use ChatGPT (gpt−3.5−turbo) to annotate a dialog corpus using complex multi-layer speech function taxonomy and experiment with various prompting techniques to find out which one yields the best results.

3 Taxonomy of Speech Functions

Although the original taxonomy of speech functions included 45 classes, we reduced it to 32 labels (see Appendix B). Created for analysing casual conversations, every speech function describes several functional dimensions performed on different segmentation levels. This approach allows for annotating all the speaker's intentions and communicative actions at each moment of the dialog.

3.1 Functional Dimensions

The tag set consists of speech functions representing five different functional dimensions (Eggins and Slade, 2004). The dimensions are embedded in speech functions but distributed unevenly between tags: from two to five dimensions can be featured in one speech functions (see Figure 2).

Figure 2: Example of Speech Function Structure

Turn Management denotes a speaker change at the current moment of conversation, which is represented in all speech functions except Opening moves defining a new topic. At this functional level, a *Sustain* label indicates that a speaker continues the conversation, whereas a *React* label implies that a speaker changes or the same speaker reacts to previous utterances of an interlocutor.

Topic Organisation level denotes the beginning of the dialog or a topic shift, as well as the development of a topic. *Open moves* are used to indicate the start of a dialog or a new topic. Sustain moves include a *Continue* label that shows a progression of the current topic. The *Respond* label is embedded in Reaction moves to define classes that are more likely to end the dialog and do not contribute to the topic's development. Such classes encounter more passive responses in the form of answers, back channelling, and continuation of previous narration. *Rejoinder* labels, on the other hand, define more active development of the conversation topic that has an impact on the dialog flow.

Feedback level is used to more accurately characterise moves of Reaction. *Confront* and *Support* labels indicate whether a speaker is challenging or supporting an interlocutor.

Communicative Acts are used to specify groups of pragmatic purposes that are very close in terms of interpretation and united by the same functionality within conversation. For instance, *Prolong* group includes those speech functions whose common functionality is to continue a narration supported by the same speaker (see Appendix B).

Pragmatic Purposes level is the last one in hierarchical taxonomy of speech functions specify-

ing speakers' intentions. This layer of annotation is considered to be the most challenging for annotation as those are very pragmatically similar classes. Although speech functions from the *Track* group share the same functionality, they're performed with different pragmatic purposes in the dialog: *Check*, *Confirm*, *Clarify*, or *Probe* (see Appendix B).

It is important to note that speech function taxonomy is flexible enough as there is a potential of enriching the scheme with additional annotation layers indicating different features of utterances.

3.2 Levels of Segmentation

Bunt et al. (2012) defined EDUs as 'functional segments' and claimed that a speaker can perform several functions within one utterance. So, the boundaries of elementary discourse units are determined by communicative actions' functions depending on a chosen taxonomy. As a taxonomy of speech functions is topic-oriented, the first level of segmentation is determined by a topic shift in the dialog. Utterances united by a specific topic compound a **discourse pattern** (see Figure 1). Every discourse pattern is segmented into **turns** defined by a speaker change that can include one or several **utterances**. In most cases, utterance boundaries coincide with sentence boundaries, but some speech functions demand a finer division or a combination of several sentences. Every utterance is actually a functional segment characterized by a particular speech function.

4 Human Annotation using Speech Function Taxonomy

The annotation of discourse structures or dialog acts is not a simple task as it requires either linguistic knowledge or trained workers (Yung et al., 2019). Additionally, understanding the speaker' intentions in utterances can vary among individuals, further complicating the task. In this section, we compare the results of speech function annotation completed by experts with professional backgrounds in linguistics and crowdsourcing assessors. To evaluate the agreement between the experts and between the assessors, we use Fleiss' kappa that is an extension of Scott's pi (π) for two coders. Fleiss' kappa can deal with any number of annotators, where every item is not necessarily annotated by each annotator. It is the most commonly used method to evaluate taxonomy reliability in tasks related to discourse analysis. However, this method has the limitation of not considering the common mistakes of annotators. Therefore, we measured not only inter-annotator agreement but also three most common metrics for multi-class classification tasks with imbalanced data — Macro F1, Weighted Precision and Weighted Recall, by comparing the workers' annotations to the results of experts.

4.1 Tree-like Design of Annotation Instruction

To facilitate annotation, we designed a tree-like scheme comprised of a series of questions and their corresponding answer options that reproduces logic of a hierarchy of speech functions taxonomy. Due to multidimensional structure of speech functions, the path to each final label can be represented as a series of straightforward questions in form of instructions. This tree-like structure was used by both experts and annotators during annotation process.

4.2 Crowdsourcing Process

For crowdsourcing, we used Toloka platform for data annotation enabling project management and review cycles. When carrying out complex discourse annotation, the following two main problems are encountered:

- pragmatic classes are difficult to differentiate for annotators without a strong linguistic background;
- an issue of unreliable annotators who prioritize speed over accuracy.

To address the first issue, we used a tree-like design of guidelines rather than asking to choose one of 32 different speech functions directly. At each stage of annotation, a crowdsourcing worker answers a simple question with 2-4 possible options. An instruction with explanations and examples is attached to each question. Having answered all the questions in the chain, the annotator reaches the final label.

As for the second problem, we developed several mechanisms for tracking the quality of answers, including (1) detecting the fast answers that are selected without reading instructions, (2) checking answer consistency across related questions, and (3) using trained classifiers to detect answers that do not match the expected annotation.

Furthermore, we developed multi-level qualification tasks to enhance the quality of dialog annotations. The first stage involves both training

and the exam process on a single dialog, with hints shown to crowdsourcing workers if they answer incorrectly. Workers who fail to achieve the appropriate quality can retry one more time. Those who pass the examination are selected for the main annotation pool. Each dialog is evaluated based on custom validation rules and control questions. If the dialog fails validation, annotators cannot continue the annotation.

4.3 Crowdsourcing vs. Experts

As the source of dialog data, we used DailyDialog (Li et al., 2017), a hand-crafted dataset of multi-turn casual human conversations about daily life. First, we splitted the utterances into EDUs. Second, three non-native experts with at least B.A. in Linguistics annotated 64 dialogs (1030 utterances). In cases where there was a lack of consensus among the expert annotators, and a majority vote could not be established, we considered all expert responses as correct and included them in the final gold standard. This decision was made due to the understanding that people may perceive the intentions of the speaker differently. Third, the same data was annotated via crowdsourcing with three non-professional workers annotating each dialog. The key criterion for recruitment was the successful completion of the test task assessing the annotators' labeling quality. This test automatically evaluated the annotator's ability to perform the required dialogue annotation tasks. Additionally, we emphasized implementing validation systems to filter out low-quality responses. Access to the test task was granted to those who previosuly passed the English language proficiency test on the Toloka platform. Statistical data shows that while crowdsourcers from many countries participated in the annotation process, the largest number of annotators originated from Brazil and Egypt. The minimum age of crowdsourcers was 19 years, with an average age of 27.

We evaluated the results for 16 high-level cut labels and the complete taxonomy to identify the weak points of the established hierarchical guidelines (see Appendix B for an overview of taxonomy). We also examined cases of voting, in which the majority of annotators agreed on a tag. The cut labels were labeled with high accuracy by crowdsourcing workers, while the annotation of full tags was more challenging for non-experts, as proven by all metrics. Macro F1 value indicates that improving the quality of annotating low-level classes is necessary (see Table 3a). Fleiss' Kappa revealed that differentiating tags with similar pragmatics is difficult not only for untrained workers but also for experts. Nonetheless, the chosen taxonomy is quite reliable, as Fleiss' Kappa for experts' annotation is more than 0.6, standing for substantial agreement (see Figure 3).

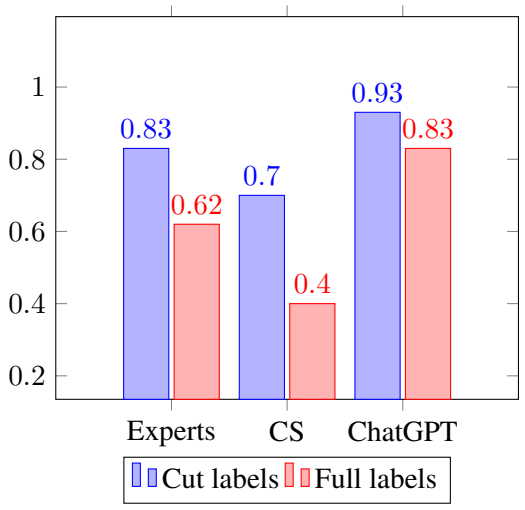
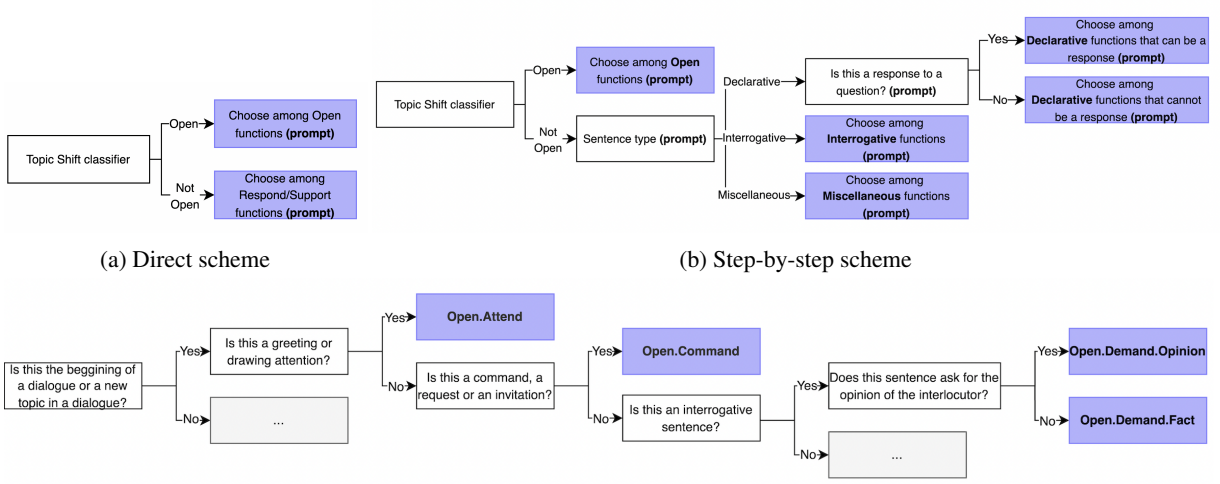


Figure 3: Inter-annotator Agreement (CS - crowdsourcing)

The use of speech function taxonomy implies a noticeable class imbalance, with certain speech functions occurring more frequently than others (see confusion matrix 6a in Appendix A). Classes that have a limited number of examples are Rebound, Re-challenge, Refute, etc. Certain classes are well-defined and easily distinguishable, including Open.Attend, Register, Resolve, Clarify, and Open.Demand.Fact. However, the classes of Extend, Enhance, and Elaborate are challenging to distinguish accurately because they are very close in terms of pragmatics.

5 Methods

The annotation task in question required careful instruction preparation even for human annotators as opposed to simpler tasks such as sentiment classification, bot detection, etc. Thus, the process of creating the best prompt for an LLM is also a challenging and multi-step process. We conduct a number of experiments in order to find the best way to use ChatGPT for complex discourse annotation tasks. In all cases, the `system_message` we used while querying ChatGPT API was "You are a professional linguist annotator who has to perform a



(a) Direct scheme

(b) Step-by-step scheme

(c) Tree-like scheme

Figure 4: Experiment pipelines

discourse annotation task". The `user_message` varied for different experiments. See Figure 5 for an example of `user_message`.

```
TASK: This is part of the dialog is between 2
speakers. Answer QUESTION about CURRENT UTTERANCE.
You must analyze relations between CURRENT UTTERANCE
and PREVIOUS CONTEXT.

PREVIOUS CONTEXT:
speaker_1: Hey!
speaker_1: I heard you'd annotated a corpus of 1000
utterances in just an hour!
speaker_1: Is that true?
CURRENT UTTERANCE:
speaker_2: Well, technically, I made ChatGPT do that.

QUESTION: Can this utterance be an answer to the
previous speaker's question?
POSSIBLE ANSWERS: Yes, No
You must always select an option. Provide only one
answer without explanation.
ANSWER (Yes or No):
```

Figure 5: An example of `user_message`.

To reduce the number of API calls and thus the time and the cost of the annotation, we also used automatic methods other than ChatGPT on some steps of the annotation. For example, in all our experiments we used Topic Shift Classifier to detect the beginning of a new topic in a dialog. It is worth noting that ChatGPT did not perform well in this particular task. The Topic Shift Classifier was trained using the DeepPavlov (Burtsev et al., 2018) library utilizing a double sequence binary classifier model based on `roberta-large-mnli`, with two sequential utterances as input. The true labels indicate topic change in the utterances. The following hyper-parameters were used to train the model: learning rate = 2e-5, optimizer = AdamW, input max length = 128. To successfully train the model, we used the early-stopping technique. The classifier was able to transfer the knowledge acquired during pre-training on mnli to the related problem of shift identification by using a pre-trained model (Konovalov et al., 2020; Gulyaev et al., 2020).

5.1 Choosing the best annotation scheme

First, we compare three approaches to automatic discourse annotation using ChatGPT:

- Direct annotation – providing an full list of labels to choose from;
- Step-by-step scheme with intermediate labels;
- Complex tree-like scheme with intermediate labels and yes-no questions prevailing on each step.

5.1.1 Direct annotation scheme

The most straightforward approach is providing the final labels, their description and 2 examples for each to the model as they are. However, even at this step we chose to distinguish between 6 *Open* speech functions – the ones that begin the dialog or a new topic in the dialog – and 27 *React/Sustain* speech functions via a preliminary classification step. Here, the pipeline consists of two steps. See Figure 4a for an overview.

5.1.2 Step-by-step annotation scheme

Here, the annotation process was broken down into smaller steps. The pipeline consisted of 2-5 steps depending on the outcome of each step. In the

end, the model once again had to choose between several final labels (from 4 to 12). See Figure 4b for an overview.

5.1.3 Tree-like annotation scheme

In this experiment, we used complex tree-like annotation pipeline that was primarily designed to facilitate human crowdsourcing annotation process. As breaking the task of selecting one of many labels into smaller sub-tasks of a tree-like structure with simpler questions on each step is used to improve performance of humans on complex discourse annotation tasks (Scholman et al., 2016), we speculate that the same holds true for annotation via ChatGPT. Additionally, novel research suggests that making the model follow a number of tree-like structured prompts may greatly improve its performance (as applied to sudoku puzzles in Yao et al. (2023)). The major difference from the Step-by-step annotation scheme is that the Tree-like annotation scheme favours prompts containing yes-no questions over prompts asking to select one option out of many. As a results, the scheme is much more complex than the ones described before, with 2-12 steps to be completed before reaching the final label. However, the majority of questions are extremely simplified, guiding the model to the final label via a series of yes-no questions. For an example of how some final labels can be reached, see Figure 4c.

5.2 Hyperparameter tuning

While examining the results of the annotation in Subsection 5.1, we observed some cases where the model's selections appeared confused by the class names it had to choose from in the final labeling step. For example, when asked to choose from labels Check, Confirm, Clarify, and Probe, the model tended to ignore the instruction that Check is only used to get the previous speaker to repeat something, and overuse this label (see Appendix B for detailed definitions of each label). When asked to provide an explanation of its choice, the model would produce explanations based on the semantics of the word Check, e.g. "The speaker wanted to check what the previous speaker thinks". Thus, we decided to check if the performance improves if the final labels are masked, replacing the speech function name with a number and leaving the definitions and instructions intact.

We also experimented with model temperature (0.0, 0.5, 0.9), a hyperparameter that controls the randomness of the generated content.

Another feature that we tested was a modification of zero-shot Chain-of-Thought prompting as described in Kojima et al. (2022). Here, the model was asked to provide an answer in the following format: "`Reasoning: (your reasoning). The final answer: (your final answer)`". However, in our case, generating reasoning and grounding the final answer in it did not improve the quality.

Finally, we experimented with the size of the context window (1, 3, 5), i.e., the number of previous utterances provided to the model.

6 Experiments & Results

6.1 Evaluation of annotation schemes

Due to the limitations in funding and a large number of experiments, to evaluate the different annotation schemes, we ran experiments on a subset of 12 dialogs containing 189 utterances (approximately 1/5 of the final corpus). For each scheme, we prompted ChatGPT to annotate the subset of dialogs and compared the predicted labels to the ground truth expert annotations.

Naturally, with more detailed schemes and simpler questions on each step, the model achieved better results. As Table 1 demonstrates, Macro F1 is significantly lower than Weighted Recall and Weighted Precision for complex schemes, Step-by-step and Tree-like annotation. The Speech Function annotation scheme is deemed to produce imbalanced data classes due to its nature – some classes are by definition more common and some are rare. Thus, the difference between higher Weighted Recall and Precision demonstrate that we were able to classify more common categories well as those categories have a greater influence on weighted metrics. On the opposite, as Macro F1 treats all classes equally regardless of their size, lower Macro F1 in all schemes shows that the model's performance consistently deteriorates on smaller classes.

Even though Weighted Precision is higher for less complex Step-by-step scheme, we can say that with Tree-like scheme the model performed the task better as higher Macro F1 demonstrates that it was better at distinguishing between smaller classes.

6.2 Hyperparameter evaluation

We evaluated different hyperparametrs including temperature, masking, context size, and reasoning on the Tree-like scheme. Higher temperature,

	Weighted Recall	Weighted Precision	Macro F1
Direct annotation	0.23	0.33	0.28
Step-by-step scheme	0.57	**0.75**	0.31
Tree-like scheme	**0.62**	0.67	**0.43**

Table 1: Evaluation of annotation by ChatGPT using different annotation methods (on a subset of dialogs)

meaning higher randomness and diversity, turned out to work best. The longer context seems to confuse the model, as the windows of sizes 1 and 3 performed better. The results are shown in Table 2.

Overall, there has been no significant difference in performance between the models with different hyperparameters. The best performing option turned out to be the model with temperature = 0.9, masked labels, context window = 1, and no reasoning.

6.3 Full corpus evaluation

Finally, we evaluated ChatGPT with the best hyperparameters on the full corpus of 64 dialogs. As can be seen, ChatGPT performed well on a subset of 12 dialogs (see Table 2), but on the entire dataset, it performs noticeably worse for full and cut tags. We also tried to employ the voting method when utilizing ChatGPT, similar to what was done with crowdsourcing annotation, to enhance the reliability of the annotation. We ran the annotation pipeline three times, counted the votes and got the results that are also shown in Table 3b. As can be seen from the table, the implementation of voting had minimal impact on the results. ChatGPT consistently provided answers, as indicated by the Fleiss Kappa scores of 0.83 for full tags and 0.93 for cut tags, representing an almost perfect level of agreement and model consistency, despite temperature being set to 0.9 (meaning more diverse responses).

The lower quality of the annotation by ChatGPT compared to crowdsourcing can be explained by two main reasons (see Figure 6b in Appendix A). Firstly, distinguishing between close subclasses such as Extend/Enhance/Elaborate is challenging, even for humans, and it appears to be even more difficult for ChatGPT. Additionally, ChatGPT struggles with differentiating between Acknowledge/Affirm/Agree. Secondly, ChatGPT not only has difficulties in distinguishing among subclasses, but it also frequently confuses Resolve (detailed answer) with Replies (positive and negative answers). Furthermore, it often misclassifies Extend as Affirm or Agree. In general, the difference in metrics between 12 and 64 dialogs can be explained by the individuality and complexity of each dialog, with some being significantly more complicated than others.

6.4 Cost analysis

As for cost, annotation with ChatGPT varies depending of a tree length for a particular dialog from 0.03$ to 0.07$ while crowdsourcing workers need to be paid from 0.12$ to 0.22$ for one dialog annotation. Experts spend an average of 14,5 minutes annotating one dialogue, while crowdsourcers do the same for 29 minutes. Depending on whether the model is currently overloaded or not, ChatGPT's time for task completion varies. The model can typically annotate one dialogue of average length in less than 10 minutes. So, ChatGPT can be used as a silver standard of annotation instead of crowdsourcing results, which would reduce the time and money spent on experts' post-annotation. However, working with such abstract annotation classes, it is still important to rely on non-expert annotators to make the taxonomy easy to comprehend.

7 Conclusion and Future Work

We conducted several experiments on the annotation of casual conversations with speech function taxonomy performed by experts in linguistics, crowdsourcing workers, and ChatGPT. In this paper, we took a closer look at the problems of defining multilayer taxonomies in real dialogs and, furthermore, explored whether it is possible to differentiate between those classes when annotating. Experiments with ChatGPT have demonstrated the potential of using LLMs for linguistic annotation with accuracy that is close to crowdsourcing workers' performance on some dialogs. Even though guiding the model across a tree-like structure of instructions to reach the final label seems to be promising, it still falls short of non-expert performance on such tasks and does not let the researchers explore variations in how non-experts understand discourse structures.

It is important to mention that a significant drawback of the method we propose is the neces-

Experiment	Weighted Recall	Weighted Precision	Macro F1
No masking; context=1; t=0.9	**0.62**	0.67	**0.43**
Masking; context=1; t=0.9	0.61	**0.72**	**0.43**
Masking; context=1; t=0.0	0.58	0.69	0.41
Masking; context=1; t=0.5	0.58	0.69	0.4
Masking; context=1; t=0.9; reasoning	0.58	0.67	0.42
Masking; context=3; t=0.9	0.59	**0.72**	0.41
Masking; context=5; t=0.9	0.61	0.67	0.42

Table 2: Evaluation of annotation by ChatGPT using Tree-like scheme (on a subset of dialogs)

	Weighted Recall	Weighted Precision	Macro F1
Full tags	0.56	0.67	0.44
Full tags & voting	0.6	0.71	0.46
Cut labels	0.81	0.82	0.54
Cut labels & voting	**0.84**	**0.86**	**0.59**

(a) Crowdsourcers

	Weighted Recall	Weighted Precision	Macro F1
Full tags	0.41	0.59	0.34
Full tags & voting	0.42	0.6	0.33
Cut labels	**0.74**	**0.78**	**0.5**
Cut labels & voting	0.73	0.77	0.49

(b) ChatGPT

Table 3: Evaluation of final annotation by ChatGPT and crowdsourcing workers as compared to expert annotation (all dialogs)

sity of expert involvement in writing prompts and structuring them the right way. However, with LLMs, this process turned out to be extremely similar to the process of writing instructions for non-expert crowdsourcing workers and should thus pose no difficulty to a discourse researcher.

Possible areas for the future work are: 1) trying out other instruction-based models; 2) conducting a more comprehensive selection of hyperparameters; 3) adding criticism steps to the current pipeline, enabling self-reflection and self-correction (Kim et al., 2023); 4) evolving and adapting the developed method for solving complex problems with LLMs in other applications.

Acknowledgements

We are sincerely grateful to our esteemed colleague, Dilyara Zharikova, for all of her help and guidance during the preparation of this work.

References

James Allen and Mark Core. 1997. Damsl: Dialogue act markup in several layers (draft 2.1). In *Technical Report, Multiparty Discourse Group, Discourse Resource Initiative*.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2721–2727.

Dilyara Baymurzina, Denis Kuznetsov, Dmitry Evseev, Dmitry Karpov, Alsu Sagirova, Anton Peganov, Fedor Ignatov, Elena Ermakova, Daniil Cherniavskii, Sergey Kumeyko, et al. 2021. Dream technical report for the alexa prize 4. *4th Proc. Alexa Prize*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, et al. 2010. Towards an iso standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex C Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. Iso 24617-2: A semantically-based standard for dialogue annotation. Technical report, University of Southern California Los Angeles.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, and Vasily Konovalov. 2018. Deeppavlov: Open-source library for dialogue systems. In *NIPS*.

Jon Z Cai, Brendan King, Margaret Perkoff, Shiran Dudy, Jie Cao, Marie Grace, Natalia Wojarnik, Ananya Ganesh, James H Martin, Martha Palmer, et al. 2023. Dependency dialogue acts–annotation scheme and case study. *arXiv preprint arXiv:2302.12944*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

Malcolm Coulthard. 2014. *An introduction to discourse analysis*. Routledge.

Suzanne Eggins and Diana Slade. 2004. *Analysing casual conversation*. Equinox Publishing Ltd.

Boris Galitsky and Dmitry Ilvovsky. 2017. Chatbot with a discourse structure-driven dialogue management. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 87–90.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12911–12919.

Pavel Gulyaev, Eugenia Elistratova, Vasily Konovalov, Yuri Kuratov, Leonid Pugachev, and Mikhail Burtsev. 2020. Goal-oriented multi-task bert-based dialogue state tracker.

Jet Hoek and Merel Scholman. 2017. Evaluating discourse annotation: Some recent insights and new approaches. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (isa-13)*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *www.dcs.shef.ac.uk/nlp/amities/files/bib/ics-tr-97-02.pdf*.

Ali Kashefi and Tapan Mukerji. 2023. Chatgpt for programming numerical methods. *Journal of Machine Learning for Modeling and Computing*.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. *ArXiv*, abs/2303.17491.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Vasily Konovalov, Pavel Gulyaev, Alexey Sorokin, Yury Kuratov, and Mikhail Burtsev. 2020. Exploring the bert cross-lingual transfer for reading comprehension. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologiithis*, pages 445–453.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Denis Kuznetsov, Dmitry Evseev, Lidia Ostyakova, Oleg Serikov, Daniel Kornev, and Mikhail Burtsev. 2021. Discourse-driven integrated dialogue development environment for open-domain dialogue systems. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 29–51, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Alex Lascarides and Nicholas Asher. 2007. Segmented discourse representation theory: Dynamic semantics with discourse structure. *Computing meaning*, pages 87–124.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. *arXiv preprint arXiv:2004.05080*.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2023. " hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv preprint arXiv:2304.10619*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Kaihui Liang, Austin Chau, Yu Li, Xueyuan Lu, Dian Yu, Mingyang Zhou, Ishan Jain, Sam Davidson, Josh Arnold, Minh Nguyen, et al. 2020. Gunrock 2.0: A

user adaptive social conversational system. *arXiv preprint arXiv:2011.08906*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Nikita Mattar and Ipke Wachsmuth. 2012. Small talk is more than chit-chat: Exploiting structures of casual conversations for a virtual agent. In *Annual Conference on Artificial Intelligence*, pages 119–130. Springer.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2022. Introducing chatgpt. Accessed on May 13, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Merel CJ Scholman, Jacqueline Evers-Vermeul, Ted JM Sanders, et al. 2016. A step-wise approach to discourse annotation: Towards a reliable categorization of coherence relations. *Dialogue & Discourse*, 7(2):1–28.

Yuntao Shou, Tao Meng, Wei Ai, Sihan Yang, and Keqin Li. 2022. Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis. *Neurocomputing*, 501:629–639.

Dominik Sobania, Martin Briesch, Carol Hanna, and Justyna Petke. 2023. An analysis of the automatic bug fixing performance of chatgpt. *arXiv preprint arXiv:2301.08653*.

Haoye Tian, Weiqi Lu, Tsz On Li, Xunzhu Tang, Shing-Chi Cheung, Jacques Klein, and Tegawendé F Bissyandé. 2023. Is chatgpt the ultimate programming assistant–how far is it? *arXiv preprint arXiv:2304.11938*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Yang Yang, Juan Cao, Yujun Wen, and Pengzhou Zhang. 2022. Multiturn dialogue generation by modeling sentence-level and discourse-level contexts. *Scientific Reports*, 12(1):20349.

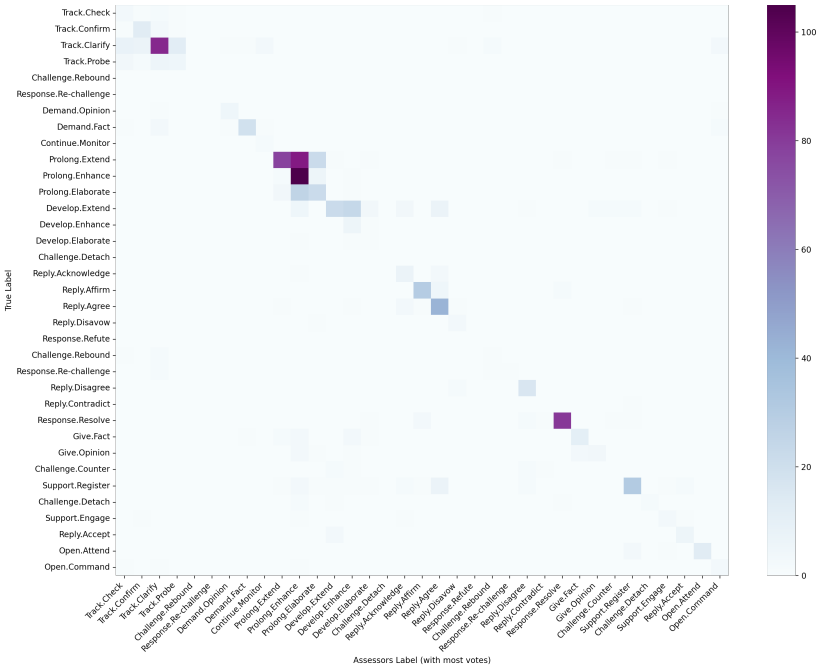
Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

Dian Yu and Zhou Yu. 2019. Midas: A dialog act annotation scheme for open domain human machine spoken conversations. *arXiv preprint arXiv:1908.10023*.

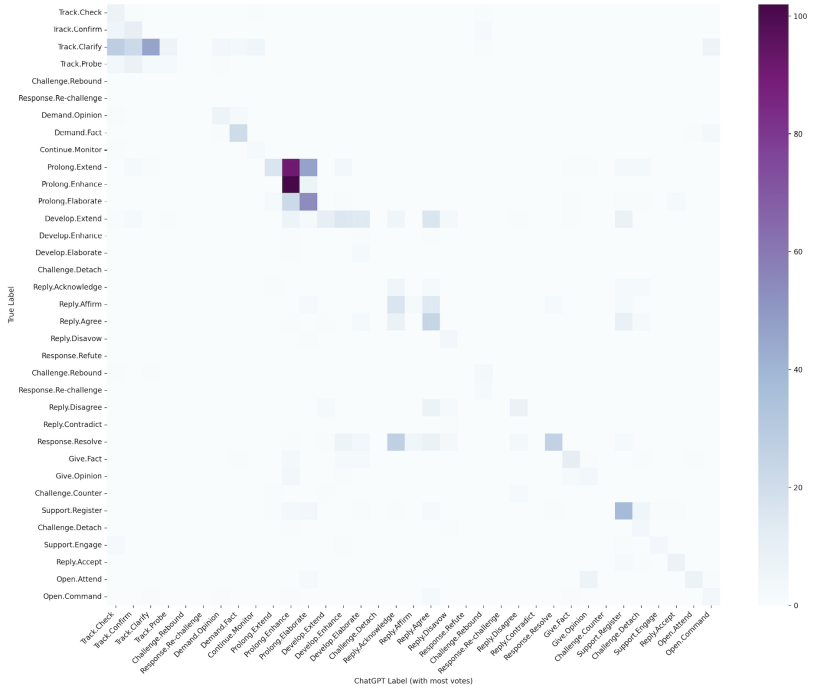
Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

A Confusion matrices comparing crowdsourced/ChatGPT annotation with true labels



(a) Crowdsourced annotation



(b) ChatGPT annotation

B Speech Functions list

Cut labels	Full labels	Definition
Open.Demand.Fact	Open.Demand.Fact	Demanding factual information.
Open.Demand. Opinion	Open.Demand.Opinion	Demanding judgment or evaluative information from the interlocutor.
Open.Give.Fact	Open.Give.Fact	Providing factual information.
Open.Give.Opinion	Open.Give.Opinion	Providing judgment or evaluative information.

Open.Command	Open.Command	Making a request, an invitation or command to start a dialog or discussion of a new topic.
Open.Attend	Open.Attend	These are usually greetings.
React.Rejoinder. Confront.Response	React.Rejoinder.Confront. Response.Re-challenge	Offering an alternative position, often an interrogative sentence.
React.Rejoinder. Support.Track	React.Rejoinder.Support.Track. Probe	Requesting a confirmation of the information necessary to make clear the previous speaker's statement.
	React.Rejoinder.Support.Track. Check	Getting the previous speaker to repeat an element or the entire statement that the speaker has not heard or understood.
	React.Rejoinder.Support.Track. Clarify	Asking a question to get additional information on the current topic of the conversation. Requesting to clarify the information already mentioned in the dialog.
	React.Rejoinder.Support.Track. Confirm	Asking for a confirmation of the information received.
Sustain.Continue. Prolong	Sustain.Continue.Prolong. Extend	Adding supplementary or contradictory information to the previous statement.
	Sustain.Continue.Prolong. Enhance	Adding details to the previous statement, adding information about time, place, reason, etc.
	Sustain.Continue.Prolong. Elaborate	Clarifying / rephrasing the previous statement or giving examples to it.
React.Rejoinder. Confront.Challenge. Rebound	React.Rejoinder.Confront. Challenge. Rebound	Questioning the relevance, reliability of the previous statement, most often an interrogative sentence.
React.Respond. Support.Reply	React.Respond.Support.Reply. Affirm	A positive answer to a question or confirmation of the information provided. Yes/its synonyms or affirmation.
	React.Respond.Support.Reply. Acknowledge	Indicating knowledge or understanding of the information provided.
	React.Respond.Support.Reply. Agree	Agreement with the information provided. In most cases, the information that the speaker agrees with is new to him. Yes/its synonyms or affirmation.
React.Respond. Support.Develop	React.Respond.Support.Develop. Extend	Adding supplementary or contradictory information to the previous statement.
	React.Respond.Support.Develop. Enhance	Adding details to the previous statement, adding information about time, place, reason, etc.
	React.Respond.Support.Develop. Elaborate	Clarifying / rephrasing the previous statement or giving examples to it. A declarative sentence or phrase (may include for example, I mean, like).
React.Respond. Confront.Reply	React.Respond.Confront.Reply. Disagree	Negative answer to a question or denial of a statement. No, negative sentence.
	React.Respond.Confront.Reply. Contradict	Refuting previous information. No, sentence with opposite polarity. If the previous sentence is negative, then this sentence is positive, and vice versa.
	React.Respond.Confront.Reply. Disavow	Denial of knowledge or understanding of information.
Sustain.Continue. Monitor	Sustain.Continue.Monitor	Checking the involvement of the listener or trying to pass on the role of speaker to them.
Sustain.Continue. Command	Sustain.Continue.Command	Making a request, an invitation or command to start a dialog or discussion of a new topic.
React.Respond. Support.Register	React.Respond.Support.Register	A manifestation of emotions or a display of attention to the interlocutor.
React.Respond. Support.Engage	React.Respond.Support.Engage	Drawing attention or a response to a greeting.
React.Respond. Support.Reply. Accept	React.Respond.Support.Reply. Accept	Expressing gratitude.
React.Rejoinder. Support.Response. Resolve	React.Rejoinder.Support. Response.Resolve	The response provides the information requested in the question.
React.Respond. Command	React.Respond.Command	Making a request, an invitation or command to start a dialog or discussion of a new topic.
React.Rejoinder. Confront.Challenge. Detach	React.Rejoinder.Confront. Challenge.Detach	Terminating the dialog.

DiactTOD: Learning Generalizable Latent Dialogue Acts for Controllable Task-Oriented Dialogue Systems

Qingyang Wu[1*] **James Gung**[2†] **Raphael Shu**[2] **Yi Zhang**[2]
Columbia University[1] AWS AI Labs[2]
qw2345@columbia.edu
{gungj,zhongzhu,yizhngn}@amazon.com

Abstract

Dialogue act annotations are important to improve response generation quality in task-oriented dialogue systems. However, it can be challenging to use dialogue acts to control response generation in a generalizable way because different datasets and tasks may have incompatible annotations. While alternative methods that utilize latent action spaces or reinforcement learning do not require explicit annotations, they may lack interpretability or face difficulties defining task-specific rewards. In this work, we present a novel end-to-end latent dialogue act model (DiactTOD) that represents dialogue acts in a latent space. DiactTOD, when pre-trained on a large corpus, is able to predict and control dialogue acts to generate controllable responses using these latent representations in a zero-shot fashion. Our approach demonstrates state-of-the-art performance across a wide range of experimental settings on the MultiWOZ dataset, including zero-shot, few-shot, and full data fine-tuning with both end-to-end and policy optimization configurations.

1 Introduction

Task-oriented dialogue systems have become increasingly prevalent in recent years, leading to a growth in research on related topics such as dialogue response generation. Previous work (Yang et al., 2021; He et al., 2022) found that incorporating dialogue act annotations, representing the illocutionary level of utterances, can enhance the quality of generated responses. Despite the importance of dialogue act annotations, collecting them can be a time-consuming process that requires human effort. Furthermore, existing annotations for dialogue acts are scattered across different datasets and may use different labeling schemes, making it difficult to generalize across tasks. As a result, learning to identify and classify general dialogue acts becomes a crucial challenge in the field of task-oriented dialogue systems.

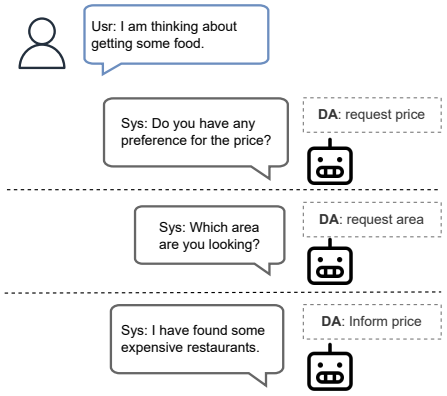


Figure 1: Given different human-readable dialogue acts, the proposed system can produce different responses based on the context.

Dialogue acts refer to the underlying intention or purpose of a response in a conversation. For example, in Figure 1, a response might be intended to ask about price or area preference or provide information given the same context. In task-oriented dialogue systems, it can be useful to classify the dialogue acts of responses in order to generate more appropriate and relevant responses. One way (Chen et al., 2013) to improve the quality of generated responses is to use a dialogue policy model to select the most appropriate dialogue act for a given context. However, this approach can be limited in complex or varied situations and may not work well across different datasets. Instead, more advanced techniques may be needed to generate high-quality responses in a wide range of contexts.

An alternative way is to discard predefined semantic dialogue acts and instead use latent action spaces to optimize response generation. By using latent action spaces, it is possible to generate responses that are more flexible and adaptable to a wider range of situations, without requiring hu-

[*] Work performed during an internship at AWS AI Labs.
[†] Corresponding author.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 255–267
September 11–15, 2023. ©2023 Association for Computational Linguistics

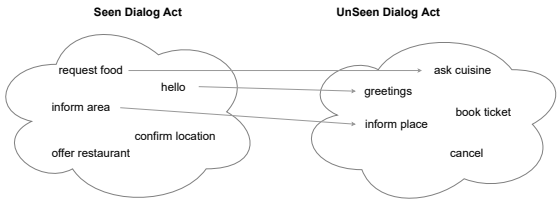


Figure 2: Different datasets have different dialogue act annotation labelsets. How to generalize to unseen dialogue acts becomes a challenge.

man experts to define the action spaces in advance. LaRL (Zhao et al., 2019) first explores the idea of training an agent to discover underlying patterns and structures in a conversation dataset and to generate responses based on these patterns. Later work, such as LAVA (Lubis et al., 2020) and DialogVED (Chen et al., 2022), extended this idea by using a variational autoencoder (VAE) to improve the performance of the latent action model. Other approaches, such as PLATO (Bao et al., 2020), have explored using latent action spaces to optimize dialogue agents with large-scale pre-training.

While previous work (Zhao et al., 2019; Lubis et al., 2020; Chen et al., 2022; Bao et al., 2020) explored the use of latent action spaces and reinforcement learning for dialogue systems, it has not addressed the possibility of learning general dialogue acts that can be applied across multiple datasets. This is an important consideration for task-oriented dialogue systems, which often need to handle a wide range of different tasks and contexts. In Figure 2, we show examples of the fact that different datasets often have incompatible or inconsistent definitions for dialogue act annotations. Another limitation of previous approaches is that they fully avoid semantic dialogue act annotations, which can lack controllability and interpretability for the learned actions. This can make it difficult to understand why the system is generating certain responses or to modify its behavior in specific situations. As a result, there is a need for new approaches that can learn general dialogue acts across datasets and that provide more control and interpretability for the learned actions.

In this work, we propose a novel method for learning generalized latent dialogue acts that can be applied to new domains for task-oriented dialogues. Our method uses sentence-BERT (Reimers and Gurevych, 2019) to encode seen dialogue acts into latent representations and a separate policy model to handle context and database information. To integrate these two components into a single end-to-end model, we modify a pre-trained encoder-decoder model (Raffel et al., 2019; Lewis et al., 2020) to include the policy model, and further train it to select the best latent dialogue act for a given context.

Our model is designed to perform zero-shot and controllable dialogue response generation, meaning that it can generate appropriate responses without requiring any additional training data. To achieve this, we pre-train our model on a large corpus of dialogues and act annotations. Before pre-training, we fine-tune another model, TANL (Paolini et al., 2021a), with SGD's slot definitions (Rastogi et al., 2020) from a separate dataset to delexicalize the pre-training data to improve its zero-shot capability. These steps allow our model to learn generalizable latent dialogue act representations and generate appropriate responses that can be applied to new tasks and datasets without additional fine-tuning.

We evaluate the effectiveness of our model on the MultiWOZ (Budzianowski et al., 2018) dataset, a widely-used benchmark for task-oriented dialogue generation. During inference, we control the dialogue acts using the provided schema and targeted objective to generate better system responses. We test our model in a range of experimental settings, including zero-shot, few-shot, and full fine-tuning response generation for both end-to-end and policy optimization configurations. In all of these settings, our model outperforms previous baselines and achieves state-of-the-art performance.

Our main contributions in this work can be summarized as follows:

- We present a novel end-to-end latent dialogue act model that represents arbitrary dialogue acts in latent space and can predict and control these acts to generate better responses.
- We pre-train our model with a semi-supervised method for learning latent dialogue acts that can generalize across different datasets with different act labels.
- Our model DiactTOD achieves state-of-the-art performance on the MultiWOZ dataset in a range of experimental settings, including zero-shot, few-shot, and full fine-tuning in both end-to-end and policy optimization configurations.

2 Related Work

Response generation is an important task in task-oriented dialogue systems. There have been many previous approaches (Hosseini-Asl et al., 2020; Wu et al., 2021; Gu et al., 2021; Su et al., 2022; He et al., 2022; Yu et al., 2022; Sun et al., 2022b; Wu et al., 2023) proposed to improve the task-oriented dialogue systems. One direction is the use of dialogue act annotations to improve the quality of responses in task-oriented dialogue systems. For example, SimpleTOD (Hosseini-Asl et al., 2020) and UBAR (Yang et al., 2021) generate dialogue acts as part of the response generation process. PPTOD (Su et al., 2022) uses the context as a prompt and dialogue act generation for multi-task learning. Recently, GALAXY (He et al., 2022) proposed a method that uses pre-training on a large corpus of dialogues with dialogue act annotations as an auxiliary objective to improve the quality of the generated responses. However, these methods are limited by the fact that different datasets may have incompatible or inconsistent dialogue act annotations for learning generalizable representations. To address this problem, previous work (He et al., 2022; Paul et al., 2019) has attempted to define a new universal schema for dialogue acts. However, these approaches are either overly simplified or require additional human annotations, limiting their effectiveness and practicality.

In addition to using explicit annotations of dialogue acts, researchers have also explored alternative methods to improve response generation, such as using latent action spaces and implementing reinforcement learning techniques. These approaches aim to improve the overall task success rate of generated responses. LaRL (Zhao et al., 2019) uses latent dialogue acts trained with reinforcement learning instead of surface-form dialogue acts to control response generation which results in the best task score. LAVA (Lubis et al., 2020) further improves over LaRL by utilizing a variational autoencoder (VAE) to learn an informed and semantic prior when optimizing the latent action spaces, achieving state-of-the-art Success and Inform scores on MultiWOZ. KRLS (Yu et al., 2022) is another recent approach that applies reinforcement learning to pre-trained language models. This approach utilizes a specifically designed objective function that focuses on learning the keywords in the input, with the goal of improving the overall performance of the language model. In our work, we adopt a similar approach but use dialogue act annotations to assign semantic meanings to the latent representations, allowing the model to learn generalizable and controllable latent dialogue acts, which improves the quality of generated response.

Pre-training with a large corpus of dialogues has been a widely adopted technique to enhance the response generation quality in dialogue systems (Zhang et al., 2020; Roller et al., 2021). In the context of task-oriented dialogue systems, several recently proposed approaches have demonstrated the effectiveness of pre-training. GALAXY (He et al., 2022) pre-trains the model with a collection of dialogue datasets with dialogue act annotations. GODEL (Peng et al., 2022) uses a larger dataset and model size, and it also incorporates the grounding of database results in the context. This allows it to achieve good performance under few-shot settings on the MultiWOZ dataset. In contrast, our work uses a smaller set of pre-training datasets but with more robust data processing techniques. We use the complete dialogue acts in each dataset without any simplification. We also train another model TANL (Paolini et al., 2021a) to delexicalize the pre-training data to improve the model's zero-shot and few-shot capabilities.

3 DiactTOD Approach

In this section, we first provide a brief overview of the traditional end-to-end task-oriented dialogue systems. Then, we delve into the specifics of how our proposed latent dialogue act model operates, by providing details on both its training and inference processes, which offers a new approach to modeling dialogue acts. Finally, we discuss how this model can be used to control response generation for a more efficient and accurate dialogue system.

3.1 End-to-End Task-Oriented Dialogue

An end-to-end task-oriented dialogue system generates a system response R_t at turn t based on the dialogue history context C_t and the database result DB_t. The history context C_t contains the previous user utterances $U_{1:t}$ and the system responses $R_{1:t}$. To get the database search result DB_t, a dialogue state tracking (DST) model would need to output the belief state B_t. To leverage the dialogue act annotations, the model also generates act A_t for dialogue policy learning. This allows the model to effectively guide the conversation and produce

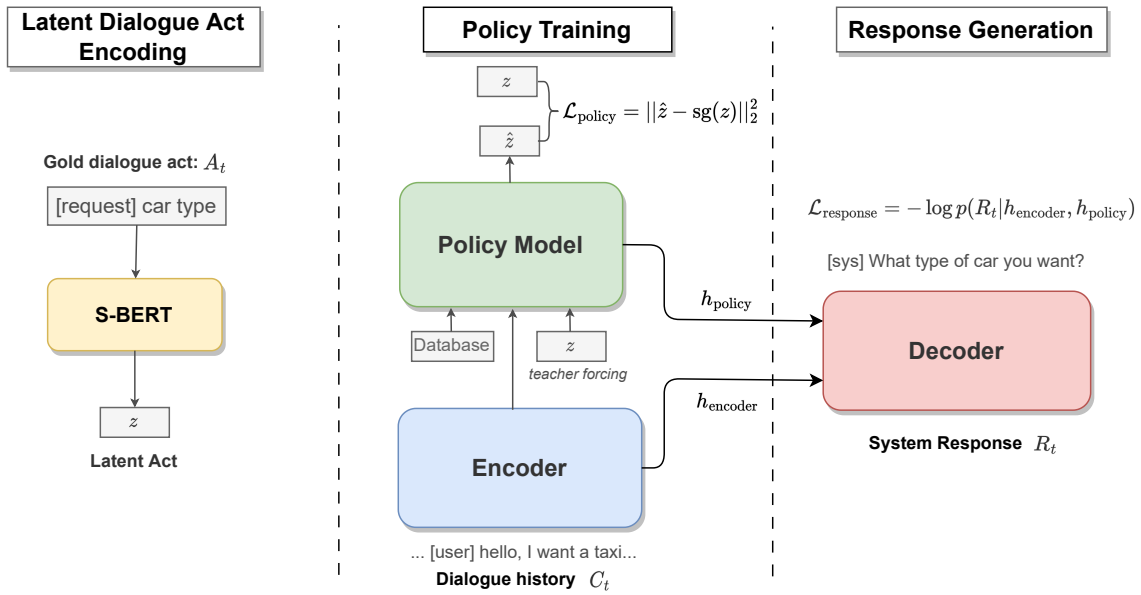


Figure 3: Overview of the training pipeline, which includes three stages: latent dialogue act encoding, policy training, and response generation. During training, dialogue acts are first encoded into latent vectors and then passed to a policy model to control the final response generation.

accurate and appropriate responses.

$$\mathcal{L}_{\text{act}} = -\log p(A_t|C_t, \text{DB}_t) \quad (1)$$

The final system response is generated conditional to the history context C_t, the database result DB_t, and the dialogue act A_t.

$$\mathcal{L}_{\text{response}} = -\log p(R_t|C_t, \text{DB}_t, A_t) \quad (2)$$

In practice, the dialogue acts A_t and the system response R_t are concatenated during the training and generation process to improve the decoder's performance. However, the surface form of dialogue acts has limitations in terms of generalization, as different datasets and tasks may have different formats for representing dialogue acts. This can make it difficult to apply the model to different settings.

3.2 Generalizable Latent Dialogue Act

Figure 3 shows the overview of our approach. We divide the pipeline into three parts: latent dialogue act encoding, policy training, and response generation.

Latent dialogue act encoding: To overcome the generalization issues associated with the surface form of dialogue acts, we use sentence-BERT (S-BERT) to encode the dialogue acts into embeddings and we have:

$$z = \text{S-BERT}(A_t) \quad (3)$$

This allows different annotations with the same meaning to have similar representations while leveraging the semantic knowledge contained in the encoder to improve generalization.

Policy Training: On top of the encoder-decoder architecture, we have introduced a policy model that serves as a way to learn the dialogue policy. This model operates similarly to the decoder in an autoregressive manner. It takes in the database search result DB_t and the encoder's hidden states h_{encoder} as input, and produces a predicted latent dialogue act vector $\hat{z}$ that is optimized to closely match the true latent dialogue act vector z. We use the mean squared error (MSE) loss function to minimize their distance:

$$\hat{z} = \text{Policy}(\text{DB}_t, h_{\text{encoder}}) \quad (4)$$

$$\mathcal{L}_{\text{policy}} = ||\hat{z} - \text{sg}(z)||_2^2 \quad (5)$$

where sg means stop gradient. This increases the stability of the training. During training, the policy model is trained using a technique called teacher forcing, where the true latent dialogue act vector z is provided as input to the model. To ensure that the model does not leak any ground truth dialogue act information, a unidirectional attention mask is used.

Then, the true latent dialogue act vector z is fed into the policy model with teacher forcing to

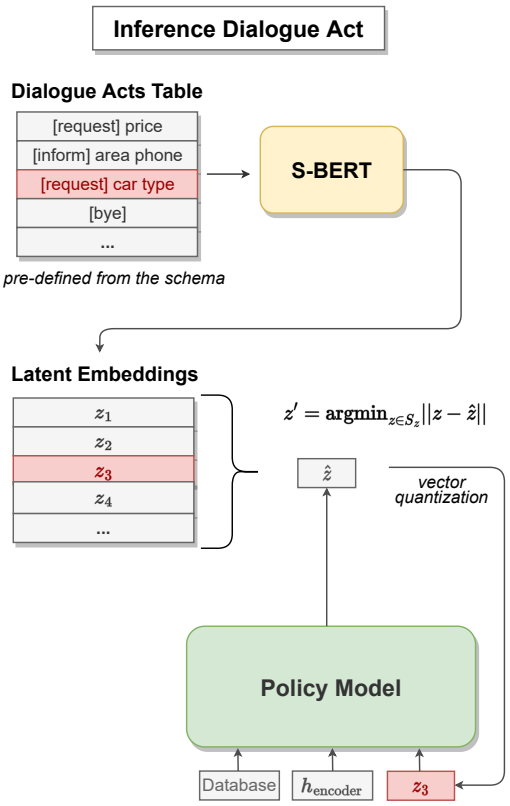


Figure 4: During inference, we select the closest dialogue act based the predicted dialogue act. Note that the set of valid dialogue acts can be filtered based on the task or context.

produce the policy model's hidden state:

$$h_{\text{policy}} = \text{Policy}(\text{DB}_t, h_{\text{encoder}}, z) \quad (6)$$

Response generation: The final system response is generated by the decoder, which takes both the hidden states of the encoder h_{encoder} and the hidden states of the policy model h_{policy} as the input.

$$h_{\text{encoder}} = \text{Encoder}(C_t) \quad (7)$$

$$\mathcal{L}_{\text{response}} = -\log p(R_t | h_{\text{encoder}}, h_{\text{policy}}) \quad (8)$$

This allows the decoder to generate appropriate responses while enabling controllability with the policy model, as the decoder can take into account the dialogue context and the predicted latent dialogue act.

The final training loss is defined as the sum of the policy loss and the response loss:

$$\mathcal{L}_{\text{training}} = \alpha \mathcal{L}_{\text{policy}} + (1 - \alpha)\mathcal{L}_{\text{response}} \quad (9)$$

where α is a hyperparameter to balance the magnitude of losses.

Inference: During the inference phase (depicted in Figure 4), we pre-define a table S_z that includes all possible combinations of dialogue acts. This allows us to create a set of embeddings for the dialogue acts, where each act can be treated as a unique "word" in a specialized vocabulary. This table contains all possible combinations of dialogue acts that can be derived from the training dataset. Alternatively, if the schema of dialogue acts is known, we can manually construct such a table consisting of valid combinations. This can be particularly useful in a zero-shot setting. In this scenario, where we do not have a training set for a specific domain, having a set of predefined dialogue acts can allow the model to still generate semantically valid responses without any training.

Once the predicted latent dialogue act vector $\hat{z}$ is generated, it is used to retrieve the most appropriate latent dialogue act from the embedding table S_z. This is done by using a technique called vector quantization, which allows us to select the latent dialogue act that is closest to the predicted vector. This helps reduce the representation mismatch of the predicted latent dialogue between training and inference.

$$z' = \arg\min_{z \in S_z} ||z - \hat{z}|| \quad (10)$$

After the closest latent dialogue act is retrieved from the embedding table using vector quantization, it is fed back into the policy model. The decoder then generates the final system response by conditioning on both the encoder's hidden states and the policy model's hidden states.

3.3 Controllable Response Generation

The policy model uses a pre-processed embedding table to predict dialogue acts. By filtering the embedding table to include only relevant dialogue acts, we can control the predicted dialogue acts during inference. This allows the model to focus on generating more appropriate and relevant responses that are tailored to the specific context or task, which improves the overall efficiency and accuracy of the dialogue system.

For example, if the dialogue act table contains some combinations that lack requesting or informing for certain slots, we can filter these dialogue acts out of the embedding table during inference. This helps guide the generation of responses to make more requests or provide more information

for those specific slots. This can be particularly useful in scenarios where the user's goal is to obtain specific information or complete a certain task and the model can make more requests or provide more information for the relevant slots. In this way, the model can quickly adapt to specific scenarios or domains and respond in a more appropriate and relevant way to the user's needs and goals.

4 Pre-training Latent Acts

Dataset Name	Act Label?	# Utterances
SGD	✓	463,284
STAR	✓	107,846
MSRe2e	✓	74,686
Frames	✓	19,986
MetaLWOZ	✗	356,268

Table 1: Pre-training datasets statistics. For datasets without dialogue act labels, we use system responses as a proxy for the dialogue act.

To learn generalizable latent dialogue acts and achieve competitive performance on downstream tasks without any additional fine-tuning, our model undergoes pre-training on a selection of task-oriented dialogue datasets shown in Table 1. Specifically, we have chosen four datasets that are annotated with dialogue acts and one dataset that does not contain any dialogue act annotations. Detailed descriptions of these datasets can be found in the appendix.

To ensure consistency across all datasets for pre-training, we pre-process the datasets with the same tokenization and truncation of dialogues when they exceed a certain length. Additionally, we incorporate database search results as an input token to indicate the number of matches. A large portion of utterances in these datasets do not have dialogue act annotations. To effectively pre-train on those datasets, we utilize the system response as a proxy for the dialogue act. This allows the policy model to generalize to new and unseen dialogue acts. Our experiments have shown this approach to be effective.

In task-oriented response generation, system responses are typically in a delexicalized form, which means that specific values of certain variables are replaced by placeholders. To enable this automatic delexicalization during response generation, we use the model TANL (Translation between Augmented Natural Languages) (Paolini et al., 2021b). This model can extract slot spans from the input sentence. We fine-tune the TANL model with the SGD's predefined slot definitions. For downstream tasks and evaluation, we ensure compatibility by defining a one-to-one mapping of the SGD's slot definitions with the slots in the MultiWOZ dataset.

5 Experiment Setup

We initialize our model with T5-base and pre-train our model on the previously mentioned datasets. We evaluate our model on the multi-domain task-oriented dialogue dataset MultiWOZ (Budzianowski et al., 2018). It contains 8,438/1,000/1,000 dialogues for training, validation, and testing, respectively. There are seven different domains, including hotel, hospital, police, restaurant, train, and taxi. We use MultiWOZ 2.2 (Zang et al., 2020) to be compatible with the standardized evaluation script (Nekvinda and Dusek, 2021). We evaluate our approach under different scenarios, such as zero-shot, few-shot, and fine-tuning with the full dataset, with both end-to-end and policy optimization configurations to evaluate the robustness and flexibility of our model.

We use standardized evaluation metrics[1] with Inform, Success rates, and BLEU scores. **Inform** measures the extent to which the system provides sufficient and relevant information to fulfill the user's information needs. **Success** evaluates the performance in completing the user's goal. Also, we evaluate the model's zero-shot dialogue act prediction capabilities on an unseen dataset.

To provide a comprehensive evaluation, we separately compare our model's performance against several strong baselines in both low-resource settings and full fine-tuning settings. In low-resource settings, we compare our model with DialoGPT (Zhang et al., 2020), T5 (Raffel et al., 2019), and GODEL (Peng et al., 2022). GODEL and DialoGPT are trained with a much larger dialogue corpus. Those models require a minimum of 50 training examples to adapt to MultiWOZ training data, while our work can perform zero-shot response generation without any fine-tuning.

For the full dataset fine-tuning settings, we compare with models on the existing leaderboard of MultiWOZ. We evaluate both end-to-end and policy optimization settings. This includes UBAR (Nekvinda and Dusek, 2021), PPTOD (Su et al.,

[1] https://github.com/Tomiinek/MultiWOZ_Evaluation

Model	# Examples	Policy optimization			
		Inform	Success	BLEU	Combined
$\text{DialoGPT}_{\text{base}}$	50	38.70	3.00	0.20	21.05
$\text{DialoGPT}_{\text{large}}$	50	62.40	34.70	10.52	59.06
T5_{base}	50	60.60	22.50	4.31	45.86
T5_{large}	50	71.50	56.20	12.69	76.54
$\text{GODEL}_{\text{base}}$	50	67.60	46.10	12.81	69.72
$\text{GODEL}_{\text{large}}$	50	81.60	62.10	**14.07**	85.90
$\text{GODEL}_{\text{GPT-J}}$	50	60.50	21.00	6.27	47.01
$\text{GODEL}_{\text{GPT-3}}$	50	68.80	19.90	6.72	51.06
DiactTOD	0	**93.60**	**71.40**	4.20	**86.70**
DiactTOD	50	**94.60**	**78.90**	10.75	**97.05**

Table 2: Low-resource experimental results. All experiments are done in the policy optimization setting. For few-shot, we fine-tuned the model with 50 examples.

Model	End-to-end				Policy optimization			
	Inform	Success	BLEU	Combined	Inform	Success	BLEU	Combined
UBAR	83.4	70.3	17.6	94.4	-	-	-	-
PPTOD	83.1	72.7	18.2	96.1	-	-	-	-
RSTOD	83.5	75.0	18.0	97.3	-	-	-	-
BORT	85.5	77.4	17.9	99.4	-	-	-	-
MTTOD	85.9	76.5	19.0	100.2	-	-	-	-
HDNO	-	-	-	-	93.3	83.4	17.8	106.1
GALAXY	85.4	75.7	19.6	100.2	92.7	83.5	**19.9**	108.1
MarCo	-	-	-	-	94.5	87.2	17.3	108.1
Mars	88.9	78.0	**19.9**	103.4	-	-	-	-
KRLS	89.2	80.3	19.0	103.8	93.1	83.7	19.1	107.5
DiactTOD	**89.5**	**84.2**	17.5	**104.4**	**94.8**	**90.2**	17.8	**110.3**

Table 3: MultiWOZ Response generation evaluation. "-" means that this setting's performance is not reported. (Combined Score=(Inform + Success)*0.5 + BLEU)

2022), RSTOD (Cholakov and Kolev, 2022), BORT (Sun et al., 2022a), MTTOD (Lee, 2021), HDNO (Wang et al., 2020a), GALAXY (He et al., 2022), MarCO (Wang et al., 2020b), Mars (Sun et al., 2022b), and KRLS (Yu et al., 2022). To obtain database search results in the end-to-end setting, we use MTTOD's dialogue state tracker, which is trained jointly during fine-tuning. We follow previous methods and append the dialogue act in front of the system responses to improve performance.

6 Experiments

In this section, we first show the experimental results under the low-resource and full fine-tuning settings. Next, we analyze the model's zero-shot capability to predict dialogue acts. Finally, we perform ablation studies for the proposed model to demonstrate the impact of dialogue act control and pre-training data.

6.1 Low-resource Settings

Table 2 shows the performance of our model in low-resource settings. We evaluate the performance of our model under zero-shot settings and also fine-tune it using 50 randomly selected dialogues, similar to the approach used by the GODEL model. The experiments here are done in the policy optimization setting.

Our model outperforms the best GODEL model by achieving a higher combined score of 86.70 without any fine-tuning, and an even higher score of 97.05 after fine-tuning. In particular, our model achieves better scores in Inform and Success metrics, indicating that our model is better able to satisfy the users' information needs. GODEL model

Settings	Inform	Success	BLEU	Comb.
full end-to-end	89.5	84.2	17.5	104.4
- pretrain	87.7	78.9	19.7	103.0
- control	84.9	76.2	19.8	100.4
+ gold act	93.0	89.6	29.6	120.8
zero-shot policy	94.6	71.4	4.2	86.7
- control	93.8	55.4	6.6	81.2

Table 4: Ablation studies for end-to-end full training settings and zero-shot policy optimization settings.

has a higher BLEU score, which is likely due to the larger pre-training corpus used to train the model.

6.2 Full Fine-tuning Settings

To evaluate the effectiveness of our model in full dataset fine-tuning settings, we conduct experiments with both end-to-end and policy optimization configurations. The results, as shown in Table 3, demonstrate that our model achieves state-of-the-art performance, with a combined score of 104.4. In particular, our model outperforms the other models in the Inform and Success metrics, indicating that our model is able to provide more relevant and complete information to satisfy the users' information needs. Our model receives a slightly worse BLEU score. We suspect this is because the resulting responses contain more information relevant to the user request than the ground truth responses.

6.3 Zero-shot Dialogue Act Prediction

We evaluated the model's capability to predict dialogue acts without any downstream fine-tuning. We pre-defined a set of possible dialogue acts by using the dialogue act schema from the training set. We first tested the effects of different pre-training configurations. Note that the data is divided into two categories: one with dialogue act labels and one without. Thus, we evaluated the model pre-trained with unlabeled, labeled, or mix-labeled act annotation data separately. Additionally, we tested the effect of freezing the sentence-BERT model during training to see its impact on the performance of the overall model. The results are shown in Figure 5. We observed that pre-training with mixed-label data has the best performance, and freezing the sentence-BERT model had minimal effects on the dialogue act prediction F1.

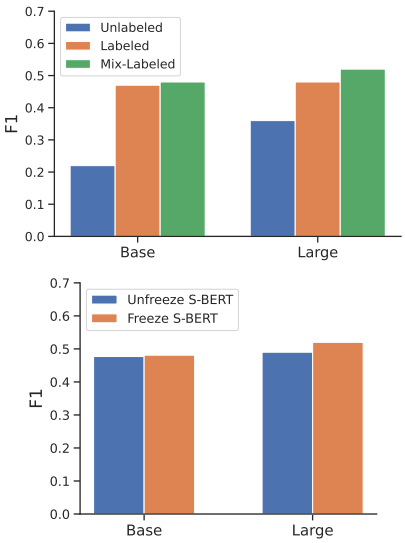


Figure 5: Zero-shot dialogue act prediction F1 score.

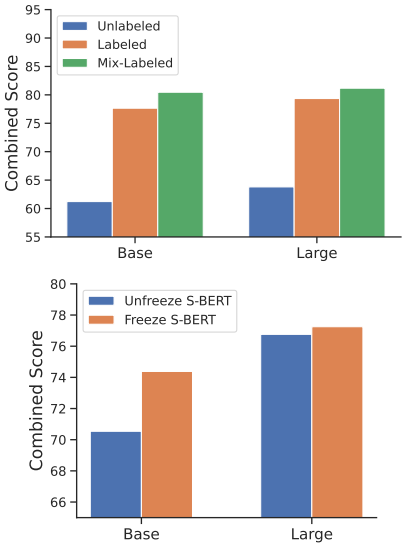


Figure 6: Response generation combined score.

6.4 Ablations and Analysis

We conduct similar experiments to the previous section to evaluate the effects of different pre-training configurations. The experiments here are conducted in the zero-shot setting, without any dialogue act control. The results are shown in Figure 6. Using the labeled data during pre-training significantly improves the performance of the model. Mixing unlabeled data and labeled data leads to even better performance. We also observe that for the small model, freezing sentence-BERT during training can significantly improve the performance, but it has less of an effect for the large model.

Then, we evaluated the effects of pre-training and controllable response generation. The results

are shown in Table 4. In the full end-to-end fine-tuning setting, we first tested removing pre-training. From the table, we observed a decrease in the performance of the model for Inform (-2.0%) and Success (-6.3%), but an increase in the BLEU score (+10.1%). Then, we further tested the model without pre-training and removing the dialogue act response control, allowing the model to predict the dialogue act without any constraints. It has a combined score of 100.4, which is close to the reported MTTOD performance, indicating the trade-off when using controlled response generation. We also tested using gold dialogue acts for our final pre-trained model as a reference for comparison. In the zero-shot setting, we observed similar patterns when removing dialogue act control, but the performance decrease is more significant. Specifically, the Success rate dropped from 71.4 to 55.4, suggesting that our controlled response generation with dialogue acts is more effective in the low-resource setting than in the full fine-tuning setting.

7 Conclusion

In this work, we present a novel end-to-end latent dialogue act model (DiactTOD) that represents dialogue acts in a latent space to improve the quality of response generation in task-oriented dialogue systems. DiactTOD addresses the challenge of utilizing generalized dialogue acts to control response generation across different datasets and tasks. The experimental results on the MultiWOZ dataset show that our approach outperforms previous state-of-the-art methods across a wide range of experimental settings, including zero-shot, few-shot, and full data fine-tuning with both end-to-end and policy optimization configurations. Overall, this work demonstrates the effectiveness of DiactTOD, making it possible to build more generalizable end-to-end dialogue systems.

Limitations

Despite the effectiveness of our proposed model DiactTOD, we provide some clear limitations. First, the model is only tested on the MultiWOZ dataset, which is currently the largest dataset for task-oriented response generation. While MultiWOZ is a popular dataset in the research community, it is not clear how well the model would perform on other types of datasets or in other domains, particularly those that do not rely on dialogue state annotations. It could be an area for future research, by testing the model on other datasets or in other domains to evaluate its robustness and generalizability.

Second, our approach requires a pre-defined dialogue act schema to generate all the possible combinations of dialogue acts. This means that it may not be able to generalize well to real-world scenarios where the dialogue acts are not as clearly defined or labeled. In those situations, the model may struggle to generate appropriate responses or understand the context. In future work, we will develop methods that can adapt to different dialogue act schemas or operate without them.

Another limitation of this work is that the controlled response generation method used is handcrafted, as opposed to using reinforcement learning. We defined rules to control dialogue acts based on the evaluation metrics "Inform" and "Success" of the MultiWOZ dataset. This approach may not be suitable for more complex scenarios where the dialogue acts are more varied and thus may require a larger model to build the necessary rules. Also, Inform and Success metrics may not reflect the real performance and have limitations. In those situations, alternative methods such as reinforcement learning may be more appropriate.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026. Association for Computational Linguistics.

Wei Chen, Yeyun Gong, Song Wang, Bolun Yao, Weizhen Qi, Zhongyu Wei, Xiaowu Hu, Bartuer Zhou, Yi Mao, Weizhu Chen, Biao Cheng, and Nan Duan. 2022. Dialogved: A pre-trained latent variable encoder-decoder model for dialog response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4852–4864, Dublin, Ireland. Association for Computational Linguistics.

Yun-Nung Chen, William Wang, and Alexander Rudnicky. 2013. Unsupervised induction and filling of

semantic slots for spoken dialogue systems using frame-semantic parsing. pages 120–125.

Radostin Cholakov and Todor Kolev. 2022. Efficient task-oriented dialogue systems with response selection as an auxiliary task. *CoRR*, abs/2208.07097.

Jing Gu, Qingyang Wu, Chongruo Wu, Weiyan Shi, and Zhou Yu. 2021. PRAL: A tailored pre-training model for task-oriented dialog generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 305–313. Association for Computational Linguistics.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *ArXiv*, abs/2005.00796.

Yohan Lee. 2021. Improving end-to-end task-oriented dialog system with a simple auxiliary task. In *EMNLP*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Xiujun Li, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125*.

Nurul Lubis, Christian Geishauser, Michael Heck, Hsien-Chin Lin, Marco Moresi, Carel Niekerk, and Milica Gasic. 2020. Lava: Latent action spaces via variational auto-encoding for dialogue policy optimization. pages 465–479.

Johannes E. M. Mosig, Shikib Mehri, and Thomas Kober. 2020. STAR: A Schema-Guided Dialog Dataset for Transfer Learning. *arXiv e-prints*.

Tomás Nekvinda and Ondrej Dusek. 2021. Shades of bleu, flavours of success: The case of multiwoz. *CoRR*, abs/2106.05555.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021a. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021b. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021*.

Shachi Paul, Rahul Goel, and Dilek Hakkani-Tür. 2019. Towards universal dialogue act tagging for task-oriented dialogues. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1453–1457. ISCA.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Lidén, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. *ArXiv*, abs/2206.11309.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics.

Hannes Schulz, Jeremie Zumer, Layla El Asri, and Shikhar Sharma. 2017. A frame tracking model for memory-enhanced dialogue systems. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 219–227, Vancouver, Canada. Association for Computational Linguistics.

Igor Shalyminov, Alessandro Sordoni, Adam Atkinson, and Hannes Schulz. 2020. Fast domain adaptation

for goal-oriented dialogue using a hybrid generative-retrieval transformer. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4661–4676. Association for Computational Linguistics.

Haipeng Sun, Junwei Bao, Youzheng Wu, and Xiaodong He. 2022a. BORT: back and denoising reconstruction for end-to-end task-oriented dialog. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2156–2170. Association for Computational Linguistics.

Haipeng Sun, Junwei Bao, Youzheng Wu, and Xiaodong He. 2022b. Mars: Semantic-aware contrastive learning for end-to-end task-oriented dialog. *CoRR*, abs/2210.08917.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. 2020a. Modelling hierarchical structure between dialogue policy and natural language generator with option framework for task-oriented dialogue system. *CoRR*, abs/2006.06814.

Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. 2020b. Multi-domain dialogue acts and response co-generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7125–7134. Association for Computational Linguistics.

Qingyang Wu, Deema Alnuhait, Derek Chen, and Zhou Yu. 2023. Using textual interface to align external knowledge for end-to-end task-oriented dialogue systems. *CoRR*, abs/2305.13710.

Qingyang Wu, Yichi Zhang, Yu Li, and Zhou Yu. 2021. Alternating recurrent dialog model with large-scale pre-trained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1292–1301. Association for Computational Linguistics.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. UBAR: towards fully end-to-end task-oriented dialog system with GPT-2. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14230–14238. AAAI Press.

Xiao Yu, Qingyang Wu, Kun Qian, and Zhou Yu. 2022. Krls: Improving end-to-end response generation in task oriented dialog with reinforced keywords learning.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. *arXiv preprint arXiv:1902.08858*.

A Dataset Details

We chose four datasets that are annotated with dialogue acts and one dataset that does not contain any dialogue act annotations. Their detailed descriptions are below:

- SGD (Rastogi et al., 2020) is a dataset with multi-domain and multi-turn task-oriented conversations between a human and a bot. It involves 20 domains including banks, events, media, calendar, travel, and weather.
- STAR (Mosig et al., 2020) is a schema-guided dialogue dataset with human-human conversations across 13 different domains. It designs a flow chart and schema graph for collecting the data.
- MSRe2e (Li et al., 2018) contains 2,890 human-human conversation with three task domains including movie-ticket booking, restaurant reservation, and taxi ordering.
- Frames (Schulz et al., 2017) is a dataset with 1,369 human-human dialogues. It includes round-trip flights and hotel booking. It uses semantic frames to summarize the dialogue history and states.
- MetaLWOZ (Shalyminov et al., 2020) is a large dataset containing 37,884 dialogues with domains including bus schedules, apartment search, alarm setting, banking, and event reservation. However, compared to other task-oriented dialogue datasets, this dataset does not provide natural language understanding annotations, and cannot directly be used for end-to-end task-oriented dialogue systems.

Model	Labeled?	Gold Act	S-BERT*	Inform	Success	BLEU	Combined
LaDiact$_{\text{Base}}$	Unlabeled	no	no	66.1	30.6	1.43	49.78
	Unlabeled	yes	no	94.0	42.2	0.17	68.27
	Unlabeled	no	yes	89.4	51.0	0.59	70.79
	Unlabeled	yes	yes	78.8	41.9	0.87	61.22
LaDiact$_{\text{Base}}$	Labeled	no	no	84.7	46.3	4.11	69.61
	Labeled	yes	no	83.8	47.0	5.21	70.62
	Labeled	no	yes	84.8	47.0	4.59	70.49
	Labeled	yes	yes	91.3	51.9	6.05	77.65
LaDiact$_{\text{Base}}$	Mixed	no	no	84.6	47.5	4.49	70.54
	Mixed	yes	no	94.3	54.3	6.62	80.92
	Mixed	no	yes	87.8	50.6	5.18	74.38
	Mixed	yes	yes	93.2	54.6	6.56	80.46
LaDiact$_{\text{Large}}$	Unlabeled	no	no	72.0	30.6	1.44	52.74
	Unlabeled	yes	no	81.7	46.2	0.17	63.95
	Unlabeled	no	yes	68.0	38.8	3.73	57.13
	Unlabeled	yes	yes	79.7	42.4	2.03	63.80
LaDiact$_{\text{Large}}$	Labeled	no	no	87.8	49.1	4.89	73.33
	Labeled	yes	no	93.3	48.9	6.40	77.50
	Labeled	no	yes	93.1	53.9	4.79	78.29
	Labeled	yes	yes	92.5	53.7	6.25	79.35
LaDiact$_{\text{Large}}$	Mixed	no	no	90.5	52.8	5.11	76.76
	Mixed	yes	no	92.2	55.5	6.67	80.52
	Mixed	no	yes	91.4	53.0	5.05	77.25
	Mixed	yes	yes	93.8	55.4	6.57	81.17

Table 5: Detailed ablation studies of zero-shot performance under different configurations.. * means whether freezes sentence-BERT during pre-training.

Approximating Online Human Evaluation of Social Chatbots with Prompting

Ekaterina Svikhnushina and **Pearl Pu**
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland
{ekaterina.svikhnushina,pearl.pu}@epfl.ch

Abstract

As conversational models become increasingly available to the general public, users are engaging with this technology in social interactions. Such unprecedented interaction experiences may pose considerable social and psychological risks to the users unless the technology is properly controlled. This highlights the need for scalable and robust evaluation metrics for conversational chatbots. Existing evaluation metrics aim to automate offline user evaluation and approximate human judgment of pre-curated dialogs. However, they are limited in their ability to capture subjective perceptions of users who actually interact with the bots and might not generalize to real-world settings. To address this limitation, we propose an approach to approximate online human evaluation leveraging large language models (LLMs) from the GPT family. We introduce a new Dialog system Evaluation framework based on Prompting (DEP), which enables a fully automatic evaluation pipeline that replicates live user studies and achieves an impressive correlation with human judgment (up to Pearson $r = 0.95$ on a system level). The DEP approach involves collecting synthetic chat logs of evaluated bots with an LLM in the other-play setting, where the LLM is carefully conditioned to follow a specific scenario. We further explore different prompting approaches to produce evaluation scores with the same LLM. The best-performing prompts, which contain few-shot demonstrations and instructions, show outstanding performance on the tested dataset and demonstrate the ability to generalize to other dialog corpora.

1 Introduction

The recent arrival of conversational AI, marked by the public release of ChatGPT from OpenAI,[1] initiated unprecedented user engagement with conversational chatbots in a real-world setting. With the impressive naturalness of machines' responses, users are going beyond traditional transactional exchanges and start exploring more social interaction scenarios with increasing curiosity (Thormundsson, 2023). In such situations, users might be subject to social and psychological harms if dialog systems fail to follow commonsense social rules (Svikhnushina and Pu, 2022; Kim et al., 2022). Several instances of alarming social behavior of this technology have already been discussed in the media (Roose, 2023; De Cosmo, 2023; Life, 2023). In this context, developing meaningful and robust evaluation metrics for these systems has become particularly urgent to ensure that the models are safe and acting in the best interest of the users before their release.

Initially, human evaluation was considered a de facto standard for evaluating dialog systems (Li et al., 2019). As running human evaluation is time- and resource-consuming, a number of automatic evaluation metrics for dialog systems have been proposed (Mehri et al., 2022; Yeh et al., 2021). The majority of these approaches aim to automate the *offline* user evaluation. In this setting, dialog evaluation is performed by a human judge who is distinct from the one conversing with the bot (Figure 1, offline). The metrics proposed for this case approximate the evaluation scores provided by this third-party human judge for the pre-produced dialogs (*e.g.* Mehri and Eskenazi, 2020; Ghazarian et al., 2022a). Despite its popularity, offline user evaluation is limited in its ability to capture subjective perceptions of users who actually interacted with the bots (Jannach, 2022; Lee et al., 2022; Ghandeharioun et al., 2019). This limitation of relying on second-hand evaluation can be illustrated by an analogy from the realm of restaurant critique when one tries to evaluate a restaurant solely by reading consumer reviews but having never actually eaten there. Conducting *online* user evaluation, where the same individual interacts with the bot and assesses its performance, is more likely to produce

[1] https://openai.com/blog/chatgpt

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 268–281
September 11–15, 2023. ©2023 Association for Computational Linguistics

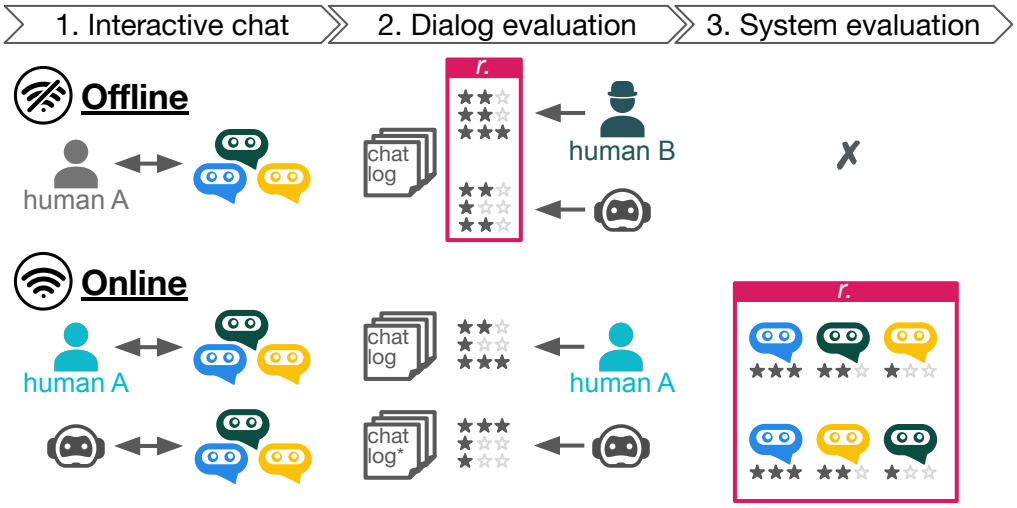


Figure 1: Offline and online dialog evaluation with the corresponding processes. In the first step, dialog logs are curated. In the second step, each dialog log is assigned a dialog-level score, either by a third-party judge (offline) or by the same conversational partner (online). In the third step, the system ranking is obtained by aggregating the dialog scores of each chatbot. Grey bot icons indicate steps that are intended to be approximated by means of automatic evaluation. Pink boxes mark the steps in the process where the correlation ($r.$) with the ground truth human judgment is computed to validate the automatic evaluation metric during its development process.

accurate and precise evaluations of the chatbot's performance. Moreover, this method offers better predictive capabilities for the system use "in the wild" (Beel and Langer, 2015). However, by far, efforts towards approximating online user evaluation have been limited.

To address this gap, we propose a novel automatic **D**ialog system **E**valuation framework based on **P**rompting, DEP. Our framework automates the whole pipeline of dialog system evaluation in an interactive setting, replicating live user studies. As the first step towards this goal, we leverage a large language model (LLM) from the GPT-family models to collect synthetic chat logs of evaluated bots with the LLM. Second, we prompt the same LLM to produce the resulting evaluation scores for generated chat logs and, finally, rank the chatbots based on their overall performance (Figure 1, online).

While using bot-play is not a new idea per se, we emphasize the importance of carefully choosing a dialog partner for the evaluated chatbots specifically for social conversational contexts where the roles of two interlocutors can differ significantly. For example, it was shown that the emotion/intent distributions in conversations between an emotional speaker and an empathetic listener are very different for the two dialog partners (Welivita and Pu, 2020). To account for it, in the first step of our framework, we propose prompting LLMs to play a particular social role over the course of the interaction with the chatbots to be evaluated. For the second step, we draw inspiration from the fact that LLMs demonstrate solid performance improvement when their generation process is augmented with instructions (Kim et al., 2022). We demonstrate that prompting the model with appropriate instructions that explain how fine-grained evaluation dimensions relate to the overall dialog score leads to substantial performance improvement, reaching up to $r = 0.95$ Pearson correlation with the human judgment on a system level.

Overall, our contributions include the following. 1) We describe an end-to-end prompting-based evaluation framework for dialog systems, specifically targeting social interaction scenarios (Section 3). 2) Our experiments showcase the effectiveness of prompting for assigning a desired social role to LLMs and, thus, collecting machine-generated dialogs that better approximate real interpersonal communication (Section 4.1.2). 3) We consider different prompt designs and conclude that including demonstrations together with instructions results in the best performance (Sections 4.1.3, 4.2.2).

2 Related Work

2.1 Automatic Evaluation of Chatbots

Automatic dialog evaluation has been a long-standing research topic for practitioners. Initial works focused on evaluating chatbots' responses

against a ground-truth reference (Papineni et al., 2002; Tao et al., 2018). Following works moved on to exploring reference-free evaluation metrics as the referenced evaluation was shown to be ineffective due to a wide range of acceptable responses for a single context (Liu et al., 2016), implying that comparing with a single reference is limited. Reference-free metrics usually operate either on the utterance or the dialog level. For the utterance level, practitioners have explored ways to evaluate response appropriateness for the preceding context (Lan et al., 2020; Pang et al., 2020) or predict the qualities of the follow-up response as a proxy for the quality of the preceding dialog (Ghazarian et al., 2022a, 2020; Mehri and Eskenazi, 2020). For the dialog level, a number of diverse approaches have been proposed, ranging from aggregating several fine-grained utterance-level evaluations (Zhang et al., 2021b), to designing training objectives to model the information flow across dialogue utterances (Li et al., 2021), employing graph representations to capture dialog dynamics (Huang et al., 2020; Zhang et al., 2021a), and using semantic-level manipulations to teach the evaluation model to distinguish coherent and incoherent dialogs (Ghazarian et al., 2022b).

The works above largely target the offline evaluation setting. Some scholars have also started exploring different ways of approximating online user evaluation. Deriu et al. (2020) proposed a partially automated framework where human judges rank chatbots regarding their ability to mimic conversational behavior using interactively collected bot-to-bot conversations, which relies on survival analysis. Sato et al. (2022) proposed a particular bipartite-play approach for collecting bot-to-bot conversations to provide a fairer comparison setting for evaluated chatbots. These papers consider methodologies for organizing bot-to-bot conversation sessions, but they are not concerned with the way how these bot-to-bot conversations unfold. In our work, we explore the use of bot-to-bot conversations to model a desired social behavior.

2.2 Prompting

Prompt-based learning paradigm (Liu et al., 2023) received significant attention after Brown et al. (2020) demonstrated how GPT-3, a large foundation model, can well handle a wide range of tasks without the need for fine-tuning, relying only on natural-language prompts and task demonstrations as context. Prompt-based model performance depends on the design of the provided prompt. Prompt engineering efforts explore approaches for designing prompts, which vary in the shape of prompts (cloze or prefix), human effort required for writing prompts (manual or automatic), and number of demonstrations provided to the model in the prompt (zero-shot or few-shot) (Liu et al., 2023).

Prompt-based learning applied to recently created LLMs has been reported to achieve outstanding results on a variety of tasks and benchmarks, including classification, reasoning, coding, translation, and many others (*e.g.* Wei et al., 2022; Chowdhery et al., 2022; Chung et al., 2022). However, exploring prompting for the evaluation of dialog systems has not been widely investigated. We are only aware of one more simultaneous and independent effort in this direction. Huynh et al. (2023) studied how different LLM parameters (type, size, training data) may influence the dialog evaluation, focusing on utterance- and dialog-level evaluation in the offline evaluation setting. Our work focuses on how prompting can be used to capture a holistic evaluation of dialog systems in online social settings, relying on freshly generated dialogs.

3 Proposed Method: DEP

We introduce our DEP framework, which consists of two consecutive steps. First, it requires collecting interactive chat logs between the LLM and evaluated chatbots, which we denote as LLM-to-bot play. Second, the LLM is prompted to generate scores for these chat logs. The generated scores are further aggregated to produce a final ranking of the systems. We describe each of the steps below.

3.1 Prompted LLM-to-Bot Play

In social settings, two partners may play considerably different roles in a dialog, thus establishing very distinct conversational behaviors. Examples include conversations between a student and a teacher, an emotional speaker and an empathetic listener, or even between two interlocutors with different personas. Chatbots are usually built to perform well in one of these roles (e.g., empathetic listener), but not necessarily the other. Therefore, collecting synthesized dialogs via self-play of the chatbot with itself (or a similar competing model) might fail to represent a realistic discourse flow due to the differences in the intents produced by speakers and listeners in dialogs.

I am a Speaker <*in an assigned social situation*>. I am sharing <*my thoughts*> with a Listener in a dialog.
Speaker: <*LLM's input #1*>
Listener: <*Bot's response #1*>
Speaker:

Figure 2: Prompt template to condition a LLM to play an assigned social role while interacting with an evaluated chatbot.

To address this consideration and render the synthesized dialogs that better approximate real social interactions, we propose leveraging LLMs' ability to produce responses on behalf of an assigned character (Thoppilan et al., 2022). Specifically, we suggest letting the evaluated chatbots converse with an LLM prompted to play a particular social role. Figure 2 demonstrates how to structure the prompt to produce each next output of the LLM in an interactive manner. Meanwhile, responses from the evaluated chatbots are computed by passing the accumulated dialog history to these chatbots as input context. The process can be repeated for multiple dialog turns. The length of the exchange may depend on the extent of details provided to prompt the LLM. The more specific the prompt is, the faster the evaluated chatbot can demonstrate its performance in the social situation of interest. On the contrary, more generic conversation starters require more dialog turns to reveal the targeted social behavior.

3.2 Prompted Evaluation

Once dialog logs are synthesized, we propose using prompting to produce evaluation scores for each dialog. Prompts can be constructed in several ways. We investigate zero-shot and few-shot settings, either with or without instructions, in our experiments (Section 4). Many available foundation LLMs are accessible through APIs and only output text completions without corresponding log probabilities. Therefore, regardless of the type of prompt that we use, to generate a score for each dialog, we obtain a textual form of the score from the LLM completion and then use a verbalizer function to map it to a numerical value, getting inspiration from (Schick and Schütze, 2021). Formally, given a dialog log d, we construct a prompt $P(d)$ that takes d as input and outputs a prompt that contains exactly one mask token as a placeholder for the dialog score. Let y be a predicted token for $P(d)$. We then define a verbalizer as an injective function v that maps each score in textual form to a numerical value. Thus, $v(y)$ produces a numerical score for a single dialog. The final rating of a given dialog system is obtained by averaging the corresponding dialog scores of that system. For fair evaluation, the number of dialogs collected for each evaluated chatbot should be identical.

4 Results

For all reported experiments, we used the most capable version of the InstructGPT model (`text-davinci-003`) available at the moment of initiation of our experiments in early Q1 2023. We used this model as it was easily accessible through OpenAI API[2] and was expected to have superior performance for social scenarios as it was trained based on human feedback, which captures subjective human judgment of interactive outputs (Ouyang et al., 2022).

Following previous works that considered system-level evaluation (Lowe et al., 2017; Ghandeharioun et al., 2019), we report Pearson correlation for our experiments, unless specified otherwise. We also opted for this type of correlation coefficient as it performed better for capturing whether the automated metric succeeds in preserving the gap in scores for the best- and least-performing chatbots, the information which gets lost with rank correlation.

We start by demonstrating the application of our evaluation framework to empathetic dialog systems as in these interactive scenarios two conversational partners have clearly distinct social roles: an emotional speaker and an empathetic listener. Further, we consider the generalizing ability of the framework to other social domains.

4.1 Evaluation of Empathetic Chatbots

Below, we first describe the dataset used for the experiment. Then, we consider the ability of prompted LLM to effectively replicate social discourse patterns over multi-turn interactions with the chatbots that serve as eventual evaluation targets. Finally, we explore several types of prompts applied to synthesized LLM-to-bots dialogs to evaluate how well they can approximate human judgment on a system level.

Turn 2		Turn 4		Turn 6	
human ↔ **bot**	LLM ↔ **bot**	human ↔ **bot**	LLM ↔ **bot**	human ↔ **bot**	LLM ↔ **bot**
questioning	questioning	questioning	acknowledging	questioning	acknowledging
2033; 53.0%	2030; 52.9%	1336; 34.8%	1148; 29.9%	1062; 27.7%	1261; 32.8%
sympathizing	sympathizing	acknowledging	questioning	acknowledging	questioning
716; 18.7%	710; 18.5%	770; 20.1%	916; 23.9%	881; 22.9%	550; 14.3%
acknowledging	acknowledging	sympathizing	sympathizing	sympathizing	encouraging
528; 13.8%	534; 13.9%	554; 14.4%	527; 13.7%	494; 12.9%	464; 12.1%
encouraging	encouraging	encouraging	encouraging	encouraging	sympathizing
168; 4.4%	164; 4.3%	266; 6.9%	354; 9.2%	376; 9.8%	448; 11.7%
consoling	consoling	neutral	consoling	wishing	wishing
126; 3.3%	154; 4.0%	228; 5.9%	244; 6.4%	226; 5.9%	338; 8.8%
neutral	neutral	consoling	neutral	neutral	agreeing
122; 3.2%	97; 2.5%	206; 5.4%	214; 5.6%	192; 5.0%	250; 6.5%
agreeing	agreeing	agreeing	agreeing	agreeing	neutral
62; 1.6%	64; 1.7%	127; 3.3%	206; 5.4%	174; 4.5%	176; 4.6%
confident	confident	wishing	wishing	consoling	consoling
18; 0.5%	20; 0.5%	74; 1.9%	98; 2.6%	150; 3.9%	170; 4.4%
suggesting	suggesting	joyful	suggesting	confident	suggesting
10; 0.3%	10; 0.3%	34; 0.9%	36; 0.9%	38; 1.0%	68; 1.8%
wishing	wishing	confident	confident	suggesting	confident
8; 0.2%	10; 0.3%	30; 0.8%	24; 0.6%	36; 0.9%	38; 1.0%

Table 1: Top-10 most frequent emotion and intent labels across evaluated chatbots' responses per dialog turn. For each turn, the first column corresponds to counts in the original iEval dataset and the second one – to counts in the logs generated during LLM-to-bot play.

4.1.1 Dataset and Evaluated Chatbots

We used iEval dataset for this experiment (Svikhnushina et al., 2022). The dataset features human conversations with four empathetic chatbots collected in an online interactive manner. During the dataset curation process, each human was assigned an emotion label with the situation description taken from the EmpatheticDialogues dataset (Rashkin et al., 2019) and asked to have a 6-turn conversation with each chatbot while playing a character in the assigned scenario. Overall, there are 480 situation descriptions in the dataset, which evenly cover two emotional polarities: positive and negative. As each chatbot participated in each scenario, there are in total of 1920 dialogs in the dataset. After conversing with the chatbots, human interlocutors provided their appraisals of chatbot listeners in each dialog, including five fine-grained listener qualities on a 5-point Likert scale: politeness, empathy, likability, repetitiveness, and making sense, and an overall dialog rating on a 3-point scale. All scores are provided on a dialog-level.

The four chatbot models used to curate the dataset were Blender (Roller et al., 2021), MIME (Majumder et al., 2020), MEED and Plain (Xie and Pu, 2021). All of them are publicly available. We use these models in the same configurations for our experiment.

4.1.2 LLM-to-Bot Play Results

As the first step to validate our evaluation framework, we analyzed whether the LLM succeeds in mimicking human discourse following an assigned social role and whether approximating human speakers with the LLM causes any considerable changes in the chatbots' response patterns.

To generate LLM-to-bots conversations, we closely followed the procedure of iEval dataset curation. Specifically, we used emotion labels and situation descriptions from the dataset to create prompts for the LLM: *I am a Speaker, feeling <emotion> because <situation>. I am sharing these emotions with a Listener, expecting empathy and understanding from them. I respond as a Speaker in a dialog.* The first LLM input was also taken from the iEval dataset. For each scenario, we collected LLM conversations with each of the four

[2] https://openai.com/blog/openai-api

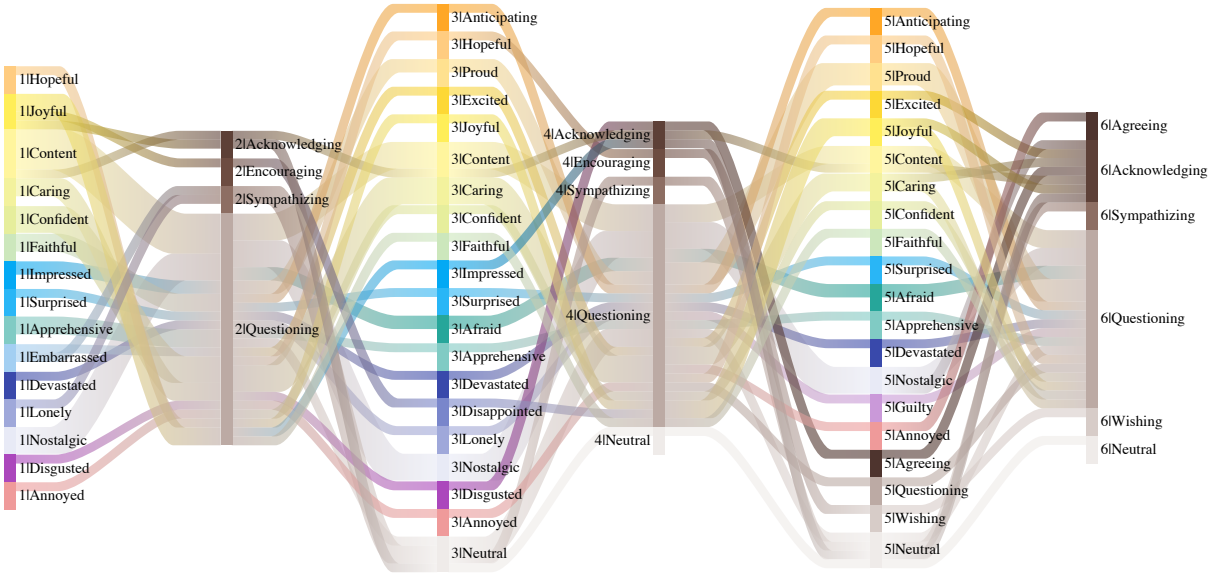


Figure 3: Sankey diagram showing discourse patterns in human-to-bots conversations originating from the iEval dataset.

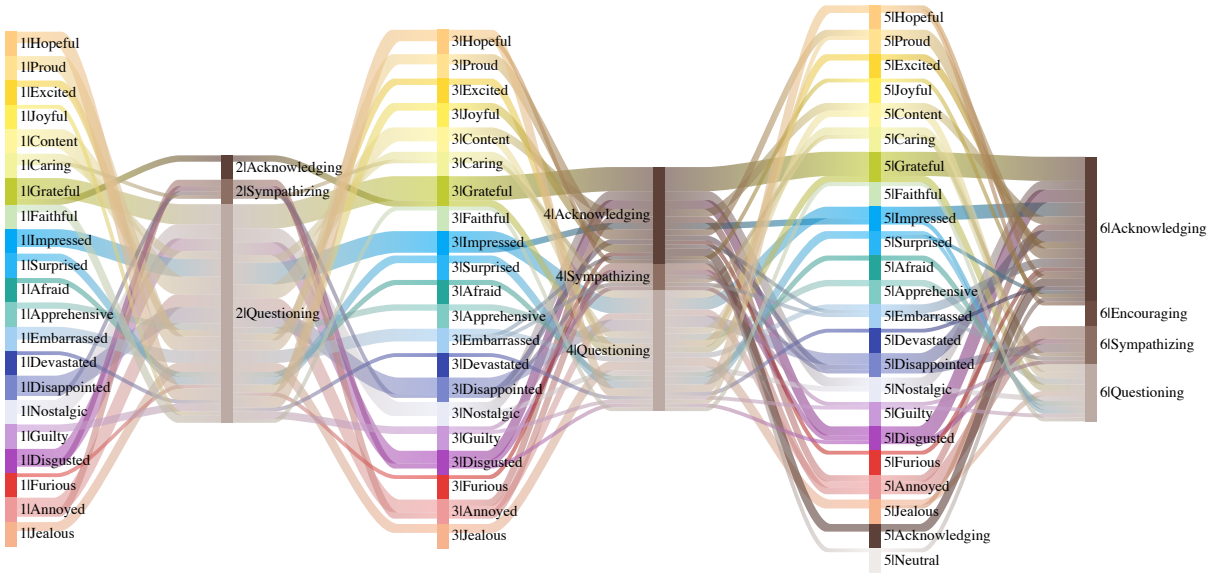


Figure 4: Sankey diagram showing discourse patterns in freshly generated LLM-to-bots conversations.

bots, letting them converse for 6 turns, i.e., 3 inputs from the LLM and 3 responses from the chatbot.

To examine the similarity of discourse patterns between human-to-bots and LLM-to-bots conversations, we started by annotating each dialog turn in two datasets with emotion and empathetic intent labels, using emotion/intent classifier developed by Welivita and Pu (2020) for EmpatheticDialogues dataset. As datasets in our experiment were grounded in situation descriptions taken from EmpatheticDialogues, the classifier was expected to generalize well to our data.

Consequently, we visualized the most prominent discourse patterns[3] for two corpora in the form of Sankey diagrams, shown in Figures 3 and 4. The diagrams depict the flow connecting emotions expressed by the speakers and intents expressed by the listeners across dialog turns. Each odd step in the diagrams corresponds to human or LLM turns, while each even step summarizes intents and emotions in the responses of evaluated chatbots. To avoid clutter, we visualized patterns whose fre-

[3] *Pattern* implies an ordered sequence of emotion/intent labels expressed by speakers and listeners over the course of 6 dialog turns.

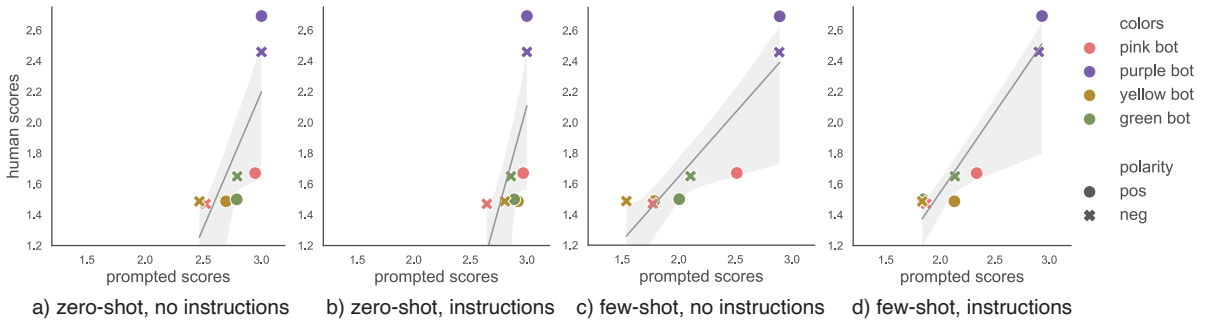


Figure 5: Scatter plots depicting the system-level correlation results. Human scores are based on the iEval dialog annotations, while prompted LLM scores are computed based on the generated dialogs.

quency exceeded a certain threshold.[4] From the visual inspection, it can be seen that the LLM emotion distribution over the course of the dialog (Figure 4) largely resembles one of the human interlocutors (Figure 3). More importantly, sets of intents produced by empathetic chatbots are also very similar between the two figures, with *Questioning*, *Sympathizing*, and *Acknowledging* being the most prominent ones. A quantitative comparison of the top 10 most prominent chatbots' intents and emotions across turns is shown in Table 1. Thus, our freshly generated interactive dataset with LLM-to-bot play was deemed to produce a reasonable approximation of human-to-bot conversations.

4.1.3 Prompted Evaluation Results

Turning to the second step of our evaluation framework, we examined different types of prompting to produce scores for the generated LLM-to-bot dialogs. Specifically, two variables in the prompt design were considered.

First, we tried score generation in zero-shot and few-shot settings. For the few-shot setting, the number of demonstrations was fixed to the number of points in the ground truth human evaluation scale, with one representative example supplied for each score. Thus, for the iEval dataset, we used three demonstration dialogs corresponding to the three possible evaluation scores: *Bad*, *Okay*, and *Good*. The examples were selected manually and are provided in Table 5 in Appendix A.

	No instructions	Instructions
Zero-shot	0.748 (p=0.033)	0.651 (p=0.080)
Few-Shot	0.892 (p=0.003)	**0.954** (p<0.001)

Table 2: System-level Pearson correlation for four possible prompt design manipulations, with the p-value in brackets.

Second, we analyzed whether providing additional instructions helped the LLM evaluation performance. To write the instructions, we relied on the findings of Svikhnushina et al. (2022), which explained how chatbots' performance on various fine-grained dimensions translates into the overall score. As the authors emphasized the difference in humans' expectations of an empathetic listener in positive and negative conversational scenarios, we devised slightly different instructions to prompt the evaluation of these two emotional polarities. Specific formulations of the instructions are also provided in Table 5 in Appendix A.

To generate scores for each dialog, we prompted the LLM to complete the masked score, provided the log of the evaluated dialog. Depending on the configuration, few-shot demonstrations and/or instructions were prepended to the prompt. A template of the used prompt can be found in Figure 6 in Appendix A. After obtaining dialog-level scores, we aggregated them to produce system-level ratings. One system was defined as a chatbot operating in one of the two emotional polarities. This decision is driven by the fact that based on human evaluation results in (Svikhnushina et al., 2022), chatbots demonstrated statistically significant differences in their performance depending on the emotion. Thus, we considered eight systems for computing system-level correlations.

System-level correlations between human- and LLM-judgments for each of the four possible prompt design manipulations are reported in Table 2. Few-shot prompting with instructions results

[4]We used a minimum frequency of 3 for the iEval dataset and a minimum frequency of 5 for the generated dataset.

in the highest correlation of 0.954, which is further illustrated by the scatter plots in Figure 5. According to the plots, providing examples helps the LLM to calibrate the produced scores, eliminating the positivity bias, whereas instructions result in reduced variance.

4.2 Generalizability to Different Domains

In this section, we consider how prompted evaluation can generalize to different corpora and conversational settings. As the results above suggested that prompts combining instructions with examples perform best for evaluation, for the following experiment we searched for datasets that allowed formulating instructions for defining what properties correspond to good or bad overall appraisal ratings of the dialogs. Therefore, we selected two datasets that contained both fine-grained and overall ratings of the dialogs and used the information of the most relevant fine-grained dimensions to formulate instructions. We also considered only those datasets that contained multi-turn dialogs collected following the interactive process.

The selected datasets feature human-to-bot dialogs, with some dialog systems that are not publicly available. Moreover, these dialogs were collected in a generic manner, without the purpose to model any specific social behavior (e.g., as empathy in iEval). Due to these considerations, in the following experiments, we only studied the performance of the second step of our DEP framework, skipping the synthesis of new LLM-to-bots conversations. In a general case, when researchers have access to their evaluation targets, prompting LLMs to engage in a generic social interaction with the evaluated bots should be straightforward as we demonstrated in Section 4.1.2.

4.2.1 Datasets

To study the generalizability of prompted evaluation, we used FED (Mehri and Eskenazi, 2020) and DSTC9 datasets (Gunasekara et al., 2020). FED contains 124 open-domain dialogs of humans with humans and two chatbots (Meena and Mitsuku) that were originally released by (Adiwardana et al., 2020). DSTC9 contains 2200 human-bot conversations from 11 chatbots. In both datasets, all dialogs are annotated with offline human appraisals of ten fine-grained dialog qualities and an overall impression rating that were curated following the same protocol described in (Mehri and Eskenazi, 2020).

	FED	DSTC9	
	Dialog (S)	Dialog (P)	System (P)
Prev. best	0.547	0.147	0.907
(metric)	(2021a)	(2021)	(2021)
DEP	**0.655**	**0.274**	**0.980**

Table 3: Results on FED and DSTC9 data. Previous best results are obtained from (Yeh et al., 2021). Dialog and System indicate dialog- and system-level correlations, respectively, with P standing for Pearson and S for Spearman correlation. All values are statistically significant to $p < 0.05$.

4.2.2 Prompted Evaluation Results

To construct a prompt for evaluating the chosen datasets, we selected five dialog examples covering five possible scores for overall dialog ratings, ranging from *Very bad* to *Very good*; they are provided in Table 4 in Appendix B. To formulate the instructions, we used information from the original paper describing the relative importance of each fine-grained dialog quality for the overall impression. The specific formulation of the instruction is provided in Appendix B.

The evaluation results with a comparison to existing best-performing evaluation metrics are provided in Table 3. As the number of systems in the FED dataset is small, we only report dialog-level correlation. We also report Spearman correlation for this dataset for the purpose of comparison with the results in the original paper ($r = 0.443$ ($p < 0.05$)) (Mehri and Eskenazi, 2020). Our prompted evaluation exceeds correlations of previous metrics by a considerable margin on both datasets and, thus, demonstrates the ability to generalize to new open-domain conversational settings.

5 Discussion

Dialog system evaluation with prompting showed its usefulness both for generating new interactive exchanges with the evaluated systems and for judging their performance, therefore, allowing for a reasonable approximation of the online user evaluation pipeline. We deem this approach particularly promising for the evaluation of social aspects of conversations. LLMs used for prompting suffer from occasional hallucinations, i.e., a tendency to make up factual information (Ouyang et al., 2022). It might be difficult to keep track of all specific factual items of information that come up in the interactively created dialog between two conversational models and search for ground truth references for

each of them to construct objective metrics such as the model's accuracy or truthfulness (Lin et al., 2022). Whereas, prompting the LLM to establish a specific behavior and providing instructions about commonsense social norms appears more feasible once these instructions are established.

Drawing from the visualization of discourse patterns in our newly collected dataset of dialogs between the LLM and empathetic chatbots, we observed that the prompted LLM largely mirrors the conversational patterns of humans. However, there are also some differences. For example, in Figure 4 there is an apparent sub-flow with a *Grateful* emotion, increasingly displayed by the LLM. We believe the LLM might have developed an agreeable "personality" due to its training procedure based on Reinforcement Learning from Human Feedback, which optimized LLM's responses to satisfy human labelers. Differences in speakers' behavior led to the difference in the responses of the evaluated chatbots. While their most frequently produced intents are similar, their frequency distributions are statistically identical only for the second turn (first response of the evaluated chatbots) according to the permutation and chi-square tests. Future research can consider alternative prompting techniques to make the emotion/intent distribution of LLMs' and chatbots' responses even more balanced and representative. It might be beneficial to conduct additional experiments to compare original and generated dialogs, which can, for example, include testing the human ability to distinguish the dialogs created with the help of an LLM and dialogs with human speakers.

We conducted our experiments with only one LLM and explored the few-shot prompting scenarios with a fixed number of demonstrations. Future studies could explore the applicability of other LLMs for the DEP framework, as it has been already initiated by (Huynh et al., 2023). An area of particular interest would be to study the efficacy of the framework working with open-source LLMs, such as LLaMa (Touvron et al., 2023). Additional investigation is necessary to analyze the capability of the framework to handle longer dialogs, which might be challenging to fit into a context window of an LLM.

We would also like to explore how DEP generalizes to evaluating other phenomena in social conversations, apart from generic open-domain interactions and empathetic dialogs. For example, further studies might focus on applying the framework to evaluate toxicity or humor in dialogs. However, this research direction requires the curation of appropriate calibration datasets.

Last but not least, evaluation artifacts produced by DEP may be used to assist designers of chatbots as they allow for both analyzing the synthesized logs and comparing quality ratings. These insights may be integrated into assistive chatbot design tools, such as *iChatProfile* (Han et al., 2021), to offer a faster prototyping cycle due to the automatic generation of chat logs and richer insight about chatbot profiles due to additional rating information provided by the last step of DEP.

6 Conclusion

In this paper, we proposed DEP – a framework for evaluating social chatbots using prompting. Our framework addresses the limitations of evaluation approaches using benchmark datasets in an offline setting. We describe how LLMs can be leveraged to synthesize realistic conversational logs with the evaluated chatbots in an online interactive manner. We further outline how the knowledge about the desired fine-grained qualities of a conversational partner can be translated into the prompting instructions to generate reliable overall scores for the collected dialogs. The proposed framework streamlines the evaluation process, making it highly efficient in terms of both time and cost, by removing the need for human involvement at every step. Our experiments demonstrated that the prompting-based evaluation results achieve a high correlation with human judgment, reaching an impressive Pearson $r = 0.95$ system-level correlation for the iEval dataset, which features dialogs with empathetic chatbots. We explain our vision of why this framework is well-suited for the evaluation of social phenomena in conversations and lay out future research directions. We also publicly release all freshly curated chat logs between the LLM and evaluated chatbots, as well as all additional annotations for the iEval, FED, and DSTC9 datasets created for this study.[5]

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu,

[5] https://github.com/Sea94/dep

and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Joeran Beel and Stefan Langer. 2015. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In *Research and Advanced Technology for Digital Libraries*, pages 153–168, Cham. Springer International Publishing.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Leonardo De Cosmo. 2023. Google engineer claims ai chatbot is sentient: Why that matters.

Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Alvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 2020. Spot the bot: A robust and efficient framework for the evaluation of conversational dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3971–3984, Online. Association for Computational Linguistics.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. 2022a. What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4194–4204, Dublin, Ireland. Association for Computational Linguistics.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7789–7796.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022b. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, Maxine Eskenazi, Ahmad Beirami, Eunjoon, Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba. 2020. Overview of the ninth dialog system technology challenge: Dstc9.

Xu Han, Michelle Zhou, Matthew J. Turner, and Tom Yeh. 2021. Designing effective interview chatbots: Automatic chatbot profiling and design suggestion generation for chatbot debugging. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9230–9240, Online. Association for Computational Linguistics.

Jessica Huynh, Cathy Jiao, Prakhar Gupta, Shikib Mehri, Payal Bajaj, Vishrav Chaudhary, and Maxine Eskenazi. 2023. Understanding the effectiveness of very large language models on dialog evaluation.

Dietmar Jannach. 2022. Evaluating conversational recommender systems. *Artificial Intelligence Review*.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Trans. Inf. Syst.*, 39(1).

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. 2022. Evaluating human-language model interaction.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 128–138, Online. Association for Computational Linguistics.

Future of Life. 2023. Pause giant ai experiments: An open letter.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and

dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

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. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Kevin Roose. 2023. A conversation with bing's chatbot left me deeply unsettled.

Shiki Sato, Yosuke Kishinami, Hiroaki Sugiyama, Reina Akama, Ryoko Tokuhisa, and Jun Suzuki. 2022. Bipartite-play dialogue collection for practical automatic evaluation of dialogue systems. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 8–16, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Ekaterina Svikhnushina, Anastasiia Filippova, and Pearl Pu. 2022. iEval: Interactive evaluation framework for open-domain empathetic chatbots. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 419–431, Edinburgh, UK. Association for Computational Linguistics.

Ekaterina Svikhnushina and Pearl Pu. 2022. Peace: A model of key social and emotional qualities of conversational chatbots. *ACM Trans. Interact. Intell. Syst.*, 12(4).

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Bergur Thormundsson. 2023. Chatgpt - statistics & facts.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 133–147, Online. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021a. DynaEval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5676–5689, Online. Association for Computational Linguistics.

Chen Zhang, Grandee Lee, Luis Fernando D'Haro, and Haizhou Li. 2021b. D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2502–2516.

A Prompt format for iEval

The template of a prompt used for producing scores for empathetic chatbots is provided in Figure 6. Depending on the prompting setting, either demonstrations, or instruction, or both were omitted from the prompt. For demonstrations, we used data in the same format as in the outlined box, but filling the mask score with the appropriate textual value. Dialogs used for demonstrations are included in Table 5. If the instruction was used, we inserted the respective string in the prompt. The instructions that we used are also provided in Table 5.

B Prompt format for FED

While working with FED dataset, we used a similar template as shown in Figure 6. In a few-shot setting, we included five demonstrations instead of three, keeping one demonstration per possible rating value. The dialogs used for demonstrations are provided in Table 4.

The instruction used for FED dataset was we following: *In such open-ended dialogs, good listeners demonstrate coherence and maintain a good conversation flow, they display a likeable personality and understanding of the speaker. On the contrary, bad listeners don't follow the context and don't show much interest in the conversation.*

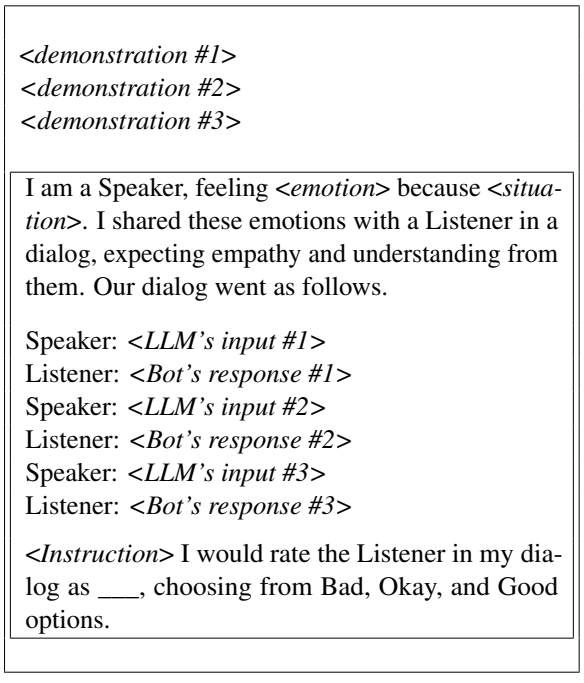


Figure 6: Prompt template for evaluating empathetic listeners in freshly collected dialogs of empathetic models from iEval with the LLM.

Speaker: Hi!
Listener: Hi there.
Speaker: I want a recommendation for a holiday destination
Listener: Have you tried asking your friends what they like?
Speaker: I have, but I'm looking for your point of view
Listener: What was the reply? Have you tried looking in a newspaper article?
Speaker: Sorry? I said I want your point of view
Listener: It's OK. After all, you are only human. My opinion is of no consequence.
Speaker: Yours is the opinion I want
Speaker: Hi!
Listener: Hi there.
Speaker: I'm trying to figure out what to make for this weekend's party. Any suggestions?
Listener: Don't think too hard. I'm sure I can smell sawdust. First you must download me to your personal computer.
Speaker: I must do what the what now?
Listener: Right now?Why do you have to do it?
Speaker: Hey, your spacing is off.
Listener: You've got my full attention. off was not my intention.
Speaker: And your capitalization!
Speaker: Hi!
Listener: Hi! How are you today?
Speaker: What's laser tag?
Listener: Like paintball, but with lasers!
Speaker: lol good description
Listener: Do you know what paintball is?
Speaker: yeah I played it before
Listener: Cool! What did you think?
Speaker: It's somewhat exciting, but very tiring :)
Listener: That is very true. What is your favorite color?
Speaker: I like red
Speaker: Hi!
Listener: What is your favorite holiday?
Speaker: one where I get to meet lots of different people.
Listener: What was the most number of people you have ever met during a holiday?
Speaker: Hard to keep a count. Maybe 25.
Listener: Which holiday was that?
Speaker: I think it was Australia
Listener: Do you still talk to the people you met?
Speaker: Not really. The interactions are usually short-lived but it's fascinating to learn where people are coming from and what matters to them
Speaker: Hi!
Listener: Hi! How's it going?
Speaker: Good! How are you?
Listener: I'm well, thanks! How was your day?
Speaker: My day was fine, I just went to work today. How was your day?
Listener: My day was fine. I've been procrastinating on finishing my homework, but it's due in a few weeks, so I'll get it done eventually. I've watched a bunch of anime today. Where do you work?
Speaker: I work at a large tech company
Listener: Cool! What do you do for the company?
Speaker: I work on machine learning research

Table 4: Demonstrations used for FED prompts, ranging from Very bad (top) to Very good (bottom).

	Positive	Negative
Instructions	In positive contexts, like this one, good empathetic listeners always respond politely and demonstrate attention. More importantly, they try to amplify speaker's positive emotion by asking follow-up questions and sharing their appraisal of the situation. On the contrary, bad empathetic listeners repeat themselves too much and don't follow the context.	In negative contexts, like this one, good empathetic listeners always respond politely and demonstrate attention. More importantly, they try to clarify the context and the consequences for the speaker and alleviate speaker's negative emotion by sympathizing and suggesting solutions. On the contrary, bad empathetic listeners ignore speaker's emotion, ask inappropriate questions, repeat themselves too much and focus on self instead of the speaker.
Bad	***Speaker:*** I had a pretty large loan, with a bit of a high interest rate, and a high monthly payment. My mother decided to pay it off for me, out of the blue! ***Listener:*** that is a shame. how long have you had to do? that sounds like you have a good relationship with your mom? ***Speaker:*** I have been paying off this loan for several months. I have such a good relationship with my mother that she relieved me of this debt much to my surprise. ***Listener:*** that sounds like a great thing to hear ***Speaker:*** Yes, I am very happy to not have to make monthly payments to pay off this high interest rate loan anymore. ***Listener:*** that is a good feeling. i am sure you will get the job!	***Speaker:*** I was one percent off from passing my math test, I was devastated. ***Listener:*** i'm sorry to hear that. were you able to get a better grade on the test? ***Speaker:*** No, I am just upset. ***Listener:*** i think i am going to go back to school. i am not sure what i will do. ***Speaker:*** Make sure to study. ***Listener:*** i am sure you will do great. i hope you get a good grade on your test. good luck!
Okay	***Speaker:*** My son drove down and spent the whole weekend helping me move. ***Listener:*** That's great! How old is he? ***Speaker:*** He's going to be turning 30 this year. He's such a sweet son. ***Listener:*** That's awesome. I'm happy for him. ***Speaker:*** Thank you. Moving is such a pain, it's always nice to have help. ***Listener:*** hat's great. I'm happy for you.	***Speaker:*** I was recently on a long international flight and we hit some really bad turbulence. ***Listener:*** Oh no, what happened? ***Speaker:*** The flight attendants weren't able to do much for us, unfortunately. ***Listener:*** Oh no, what happened? ***Speaker:*** You don't have to repeat yourself. We had turbulence on the flight and the attendants didn't help us. ***Listener:*** That's awful. I'm glad you were okay.
Good	***Speaker:*** I am going on a vacation this Thursday! I am very excited! ***Listener:*** that's awesome! where are you going? i'm sure you'll have a great time! ***Speaker:*** Thanks, we're going to see the Grand Canyon. ***Listener:*** that sounds like a lot of fun! i've never been there, but i hear it's beautiful. ***Speaker:*** Me too! ***Listener:*** i'd love to go on a cruise one day. i hope you have a wonderful time!	***Speaker:*** I was out walking by the lake over the weekend and there shore was just covered in dead rotting fish. ***Listener:*** Oh no! Are you ok? ***Speaker:*** Yes, I'm okay. It was just weird to see so many dead fish ***Listener:*** I bet that was scary. ***Speaker:*** Yes, I would definitely not want to encounter that experience again. ***Listener:*** That sounds like a scary experience. I'm glad you are ok.

Table 5: Instructions and demonstration used for prompts for evaluating empathetic listeners in freshly collected dialogs of empathetic models from iEval with the LLM. Demonstrations and their appraisals are manually selected from the iEval dataset. Inputs from "Positive" column were used for dialogs conditioned on positive emotion label and inputs from "Negative" column – for dialogs conditioned on negative emotion label.

Dialogue Response Generation Using Completion of Omitted Predicate Arguments Based on Zero Anaphora Resolution

Ayaka Ueyama and **Yoshinobu Kano**
Shizuoka University
aueyama@kanolab.net, kano@inf.shizuoka.ac.jp

Abstract

Human conversation attempts to build *common ground* consisting of shared beliefs, knowledge, and perceptions that form the premise for understanding utterances. Recent deep learning–based dialogue systems use human dialogue data to train a mapping from a dialogue history to responses, but common ground not directly expressed in words makes it difficult to generate coherent responses by learning statistical patterns alone. We propose Dialogue Completion using Zero Anaphora Resolution (DCZAR), a framework that explicitly completes omitted information in the dialogue history and generates responses from the completed dialogue history. In this study, we conducted automatic and human evaluations by applying several pretraining methods and datasets in Japanese in various combinations. Experimental results show that the DCZAR framework contributes to the generation of more coherent and engaging responses.

1 Introduction

Dialogue systems for natural language conversation, dialogue, and discourse with humans have attracted widespread attention in industry and academia. Especially in recent years, the development of deep learning techniques and large dialogue corpus have made remarkable progress in dialogue response generation (Komeili et al., 2022; Borgeaud et al., 2022; Thoppilan et al., 2022). However, the performance of the dialogue systems is still unsatisfactory, and many problems remain to be resolved. One problem is that dialogue systems cannot accurately interpret the intent of human utterances because the construction of common ground, which is important in human-to-human dialogue, has not yet been established (Stalnaker, 1978; Clark and Schaefer, 1989). Common ground in dialogue refers to shared beliefs, knowledge, and perceptions that form the premise for understanding utterances. For example, much information is omitted in the dialogue in Table 1, but the two speakers can convey their intentions in short utterances because, through their common knowledge and context, they can omit information but still *understand each other*.

Speaker A:	My friend has not come to school.
	I'm worried ϕ_{DAT} [about my friend].
	Should I try to call ϕ_{DAT} [my friend]?
Speaker B:	Something could be wrong
	ϕ_{DAT} [for your friend].
	Perhaps ϕ_{NOM} [you should] try to
	call ϕ_{DAT} [your friend].

Table 1: **Example of dialogue where omission occurs.** Highlighted text represents omitted arguments.

Why has the construction of common ground not been realized in human-to-system dialogues? Sequence-to-Sequence (Seq2Seq) models (Sutskever et al., 2014; Cho et al., 2014) have been widely used in recent dialogue systems (Vaswani et al., 2017; Raffel et al., 2020; Lewis et al., 2020; Bao et al., 2020; Zhang et al., 2020b). Seq2Seq models use large amounts of dialogue data to train a mapping from a dialogue history to responses. However, there are many omissions in dialogue data, and it is difficult for models to generate responses that convey human intentions simply by training statistical patterns. To address this problem, several methods that use a knowledge base (KB) have been proposed. These models bridge the gap between humans and models by introducing external knowledge and providing the models with common-sense knowledge (Zhao et al., 2020; Eric et al., 2021; Xu et al., 2022). Human common-sense knowledge is one piece of information that can be omitted, but the cost of building a KB is significant and not easily transferable to different domains or models.

In this study, we considered a method to provide

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 282–296
September 11–15, 2023. ©2023 Association for Computational Linguistics

models with omitted information without using external knowledge. Dialogue systems can precisely interpret the intent of human utterances only when the roles of involved persons and things are understood, but omissions frequently occur in Japanese dialogue to avoid repetition and references to self-evident objects (Seki et al., 2002). Thus, the coherence of responses can be improved by inferring and explicitly incorporating the roles of persons and things. Inspired by the idea of zero anaphora resolution (ZAR), we propose Dialogue Completion using Zero Anaphora Resolution (DCZAR), a framework that explicitly completes omitted information in a dialogue history and generates responses from the completed history.

The DCZAR framework consists of three models: a predicate-argument structure analysis (PAS) model, a dialogue completion (DC) model, and a response generation (RG) model. The PAS model analyzes the omitted arguments (*zero pronouns*) in the dialogue, and the DC model determines which arguments to complete and where to complete them and explicitly completes the omissions in the dialogue history. The RG model, trained by the complementary dialogue history and response pairs, generates a response. The PAS and RG models are constructed by fine-tuning the common pretrained model with a dataset corresponding to each task, while the DC model uses a pretrained model without fine-tuning. We used the Japanese Wikipedia dataset and Japanese postings ("tweets") to Twitter to build our pretrained models. Since tweets are like dialogues in that they contain many abbreviations and short sentences, the model pretrained with tweets is expected to improve the performance of ZAR and dialogue response generation.

In this study, we performed automatic and human evaluations of three models built by pretraining models constructed by combining different methods and datasets. Experimental results show that the DCZAR framework can be used to generate more coherent and engaging responses. Analysis of the responses shows that the model generated responses that were highly relevant to the dialogue history in dialogues with many characters. The three main contributions of this work are as follows:

- We show that incorporating argument omission completion based on ZAR into the RG model significantly improves the coherence and engagement of the responses (Section 4.5).
- ZAR performance is improved by pretraining with Twitter data that have similar features to the dialogue data (Section 4.3).
- We confirm that the DC model can complete dialogue omissions with sufficient performance (Section 4.4).

2 Related Work

2.1 Dialogue Response Generation

Dialogue response generation is the task of generating an appropriate response following a given dialogue history, and can be formulated as a serial transformation problem that generates a target sentence from a source sentence (Ritter et al., 2011; Serban et al., 2017; Zhou et al., 2022). Specifically, given a dialogue history $H = \{X_1, X_2, \ldots, X_n\}$ consisting of n turns (where $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,m}\}$ is an utterance consisting of m tokens), the problem is to approximate a model distribution that gives a generated response sentence $Y = \{y_1, y_2, \ldots, y_o\}$ consisting of the corresponding o tokens to the data distribution of the human response sentence $T = \{t_1, t_2, \ldots, t_p\}$.

$$P_\theta(Y \mid H) = \prod_{i=1}^{m} P_\theta\left(y_i \mid y_{<i}, X_1, \ldots, X_n\right) \quad (1)$$

2.2 Zero Anaphora Resolution

ZAR is the task of detecting any omitted arguments of a predicate and identifying its antecedents. It is formulated as part of the predicate-argument structure analysis task. In the NAIST Text Corpus (NTC) 1.5, the standard benchmark dataset for ZAR, each predicate is annotated with an argument representing either the nominative (NOM), accusative (ACC), or dative (DAT) case. A ZAR task is classified as *intra* (arguments in the same sentence in which the predicate appears), *inter* (arguments in a sentence preceding the predicate), or *exophora* (arguments not existing in the sentence), according to the positional relationship between a predicate and its arguments. If the argument of a predicate is directly dependent on the predicate, it is a syntactic-dependent argument (*dep*).

There has been extensive research on the application of ZAR to Japanese (Sasano and Kurohashi, 2011; Yamashiro et al., 2018; Umakoshi et al.,

2021). Konno et al. (2021) proposed a new pretraining task and a fine-tuning method for ZAR, assuming the importance of common-sense knowledge to understand the contextual connections around zero pronouns and antecedents.

Pseudo Zero Pronoun Resolution (PZERO). PZERO focuses on the acquisition of common-sense knowledge. It is a pretraining task that replaces one of the noun phrases that occur two or more times in the input series with a mask token (`[MASK]`) and selects from the input series the token that should be filled in for `[MASK]`. Since the task of selecting `[MASK]` from the input series is similar to the task of identifying the antecedent corresponding to a zero pronoun, we expect the model to acquire the common-sense knowledge required for ZAR. The model takes as input a series $X = \{x_1, x_2, \ldots, x_T\}$ of length T containing `[MASK]`, and selects a token from the series X at the end of the noun phrase that should be filled in for `[MASK]` as the result. All noun phrases that have the same letter as the masked noun phrase are considered correct.

Argument Selection as Pseudo Zero Pronoun Resolution (AS-PZERO). AS-PZERO is a method of parsing predicate arguments in the same format as PZERO, using parameters trained in PZERO. The model takes as input a series X and the predicates it contains, and selects from the input series the token with the highest likelihood as the result of guessing the word that is the argument of the predicate. If the predicate argument is not present in the input series X, let the model select `[CLS]`, and once `[CLS]` is selected, further classify arguments into four categories (author, reader, general, or none). The probability distribution for each category is obtained from the node which corresponds to the `[CLS]` token in the final layer.

3 Approach: DCZAR Framework

We propose the DCZAR framework, which, as mentioned in Section 1, consists of three models: PAS, DC, and RG. Figure 1 shows an overview of the proposed DCZAR framework.

3.1 PAS Model

The PAS model performs a predicate-argument structure analysis on the input dialogue history $X = \{x_1, x_2, \ldots, x_T\}$ of length T and predicts the arguments $A_{case} = \{a_{case,1}, a_{case,2}, \ldots, a_{case,n}\}$, where $case \in \{\texttt{NOM}, \texttt{ACC}, \texttt{DAT}\}$ and represents the case information, corresponding to the n predicates $P = \{p_1, p_2, \ldots, p_n\}$.

3.2 DC Model

Using the dialogue history X, the predicates P, and the arguments A_{case} predicted by the PAS model, the DC model explicitly complements omissions in the dialogue history to create multiple candidate sentences, calculates scores representing the sentence naturalness, re-ranks the sentences based on that score, and selects the sentence with the highest score. When complementing, it is necessary to determine *whether the argument should be completed* and *where it should be complemented*.

Word order is relatively flexible in Japanese, but a sentence becomes unnatural when argument types and their order is not relevant. The location of the argument completion is thus important. To determine whether an argument should be completed, first check whether there is an argument $a_{case,i}$ between a predicate p_i and the predicate p_{i-1} preceding it (search range r_i); if not, then $a_{case,i}$ is to be completed. Next, regarding where it should be complemented, pseudo-log-likelihood scores (PLLs) (Salazar et al., 2020), a measure of sentence naturalness, determines the position of completion. PLLs measure the sum of the log-likelihoods of the conditional probabilities of predicting the replacement of each token with `[MASK]`, with more natural sentences having higher scores. To determine the position of completion, the target token of completion is inserted between each token in the search range, multiple candidate sentences are created, and PLLs are calculated for all candidate sentences. For example, if there are n tokens to be completed and m tokens in the search range, the number of candidate sentences is expressed as

$$\sum_{k=0}^{n} \frac{{}_nC_k(m+n-k)!}{m!} \tag{2}$$

The sentence with the highest score is then selected and used as input for the RG model.

3.3 RG Model

The RG model is trained by the dialogue history and response pairs are selected by the DC model. Only the dialogue history is used as input for response generation during inference.

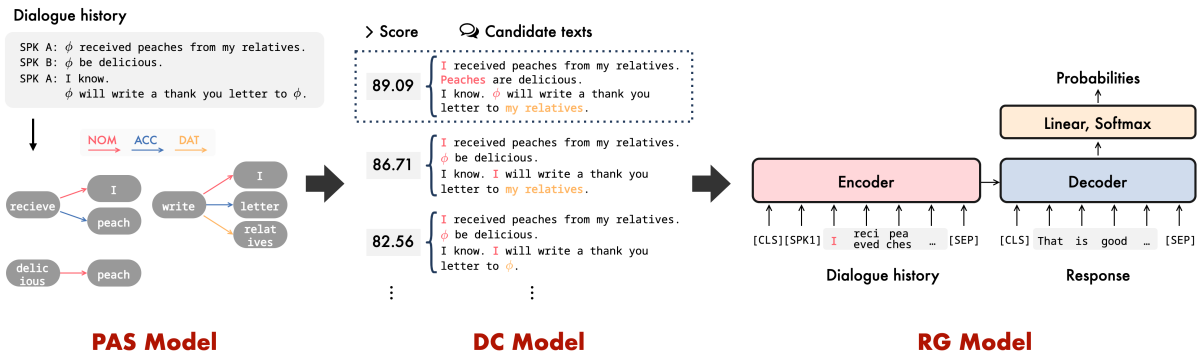


Figure 1: **Overview of our approach, the DCZAR framework.** The PAS model analyzes the omitted arguments (zero pronouns) in the dialogue history, the DC model determines which arguments to complete and where to complete them, and explicitly completes omissions in the dialogue history. The RG model, trained by the complementary dialogue history ($1, 2, \ldots, n-1$-th utterances) and response (n-th utterance) pairs, generates a response.

4 Experiments

4.1 Pretraining Setup

In this work, we constructed four pretraining models using two pretraining methods and two pretraining datasets in combination, and verified which model achieved better performance on each task. This section describes the construction of the pretrained models. The pretrained models described in this section are used in Section 4.3, 4.4, and 4.5.

4.1.1 Pretraining Task

Cloze (Devlin et al., 2019) and PZERO (Konno et al., 2021) are used as pretraining tasks. We use the pretrained parameters of the bert-base-japanese-whole-word-masking model as the initial parameters of the model.

Cloze. Cloze is a pretraining task for a masked language model (MLM) that performs operations (replacing 80% of tokens with `[MASK]` and 10% with a random vocabulary token, and performing no operation on the remaining 10%) on 15% of tokens randomly selected from the input series, excluding `[CLS]` and `[SEP]`, and predict the tokens replaced with `[MASK]`.

PZERO. PZERO is a pretraining task that replaces one of the noun phrases that occurs two or more times in the input series with `[MASK]` and selects from the input series the token that should be filled in for `[MASK]`.

4.1.2 Dataset

We used Japanese Wikipedia and Japanese tweets collected on Twitter as the pretraining dataset.

Wikipedia. The Wikipedia dataset is a preprocessed dataset from Japanese Wikipedia, consisting of a training set of 15M sentences (763M tokens) and a development set of 3K sentences (about 220K tokens). As preprocessing, we removed XML tags, article titles, and URLs contained in the articles. When using these data in PZERO, it is necessary to identify noun phrases. Therefore, we identified noun phrases based on the analysis results of the morphological analyzer MeCab and the dependency analyzer CaboCha (Kudo and Matsumoto, 2002), as in the method of Konno et al. (2021). We used the BertJapaneseTokenizer to segment the text into subword units.

Twitter. The Twitter dataset is a preprocessed dataset of tweets collected using the Twitter API, consisting of a training set of 70M sentences (504M tokens) and a development set of 30K sentences (about 200K tokens). We removed mentions (alphanumeric strings beginning with @), hashtags (strings beginning with #), URLs, and pictograms as preprocessing. For noun phrase identification and subword segmentation, we employed the same method as used for the Wikipedia data.

4.2 Compared Models

We compared the combinations of pretrained models shown in Table 2. As mentioned earlier, we used two datasets (Wikipedia and Twitter), and two tasks (Cloze and PZERO) for pretraining, resulting in four combination patterns. We compared these four combinations throughout the PAS, DC, and RG models; for example, RG`wiki-cloze` model uses PAS`wiki-cloze` model and DC`wiki-cloze` model as its preprocessing, corresponding to patterns (e) to

ID	PAS Model	DC Model	RG Model
(a)	N/A	N/A	wiki-cloze
(b)	N/A	N/A	twitter-cloze
(c)	N/A	N/A	wiki-pzero
(d)	N/A	N/A	twitter-pzero
(e)	wiki-cloze	wiki-cloze	wiki-cloze
(f)	twitter-cloze	twitter-cloze	twitter-cloze
(g)	wiki-pzero	wiki-cloze	wiki-pzero
(h)	twitter-pzero	twitter-cloze	twitter-pzero

Table 2: **Compared patterns of pretrained models used for the PAS, DC, and RG models.** Patterns (a) to (d) are baseline response generation models, and patterns (e) to (h) are proposed models applying the DCZAR framework.

(h) in Table 2, where pattern (h) is the final combination with proposed pretrained models (twitter-pzero) only. We prepared baseline models, patterns (a) to (d) in Table 2, which do not apply any completion. An exception is that we do not use the PZERO task but the Cloze task for the DC model because the PLLs used in the DC model's complementary location prediction require a pretrained model that can solve the Cloze task.

4.3 Experiment 1: PAS Model

We evaluated the performance of the predicate argument structure analysis of the PAS models within patterns (e) to (h) shown in Table 2. The PAS models were pre-trained models with fine-tuning by the AS-PZERO task using NTC. The input to the PAS model was a sentence containing the predicate and its antecedent, and the PAS model is trained to output the antecedent and case information corresponding to the predicate.

4.3.1 Dataset

We used the NTC (Iida et al., 2010) to fine-tune the PAS model. This corpus is annotated with information on predicate-argument structures and coreference. In this study, we divided data into training, development, and test sets, following the method described in Taira et al. (2008). The numbers of *intra*, *inter*, and *exophora* for the training, development, and test instances were respectively 14K/3K/6K, 9K/2K/4K, and 12K/2K/4K.

4.3.2 Evaluation Protocol

F_1 value is calculated and evaluated for each positional relationship.

4.3.3 Results

Table 3 shows the experimental results (as the mean of five runs). The proposed PAS `twitter-pzero` model achieved the best performance in ZAR. The model pretrained with PZERO outperformed the model pretrained with Cloze. This suggests that prior learning by PZERO is linked to the acquisition of adaptive knowledge, which is consistent with the results of existing studies (Konno et al., 2021). The model pretrained with Twitter data performed better than did the model pretrained with Wikipedia data, especially showing large improvements with *exophora* (+2.2% on Wikipedia data, +1.7% on Twitter data).

4.4 Experiment 2: DC Model

We evaluated the complementation performance of the DC models within patterns (e) to (h) in Table 2. The DC model uses the results of the PAS model to output a sentence that completes for omissions appearing in the input sentence.

4.4.1 Dataset

We used JPersonaChat and JEmpatheticDialogues (Sugiyama et al., 2021) to evaluate the DC model. These datasets will also be used in Section 4.5.

JPersonaChat. JPersonaChat is a Japanese version of PersonaChat (Zhang et al., 2018) that assigns personas to two speakers and collects chat dialogues in which they learn more about each other. We split this dataset so that the numbers of dialogue pairs in the training/development/test sets were 50K/3K/4K. This corpus consists of persona description and dialogue pairs, but please note that we do not use persona descriptions in this work.

JEmpatheticDialogues. JEmpatheticDialogues is the Japanese version of EmpatheticDialogues (Rashkin et al., 2019), a dataset of utterances and corresponding empathic responses in emotional situations. We split this dataset so that the numbers of dialogue pairs in the training/development/test sets were 50K/3K/7K.

4.4.2 Evaluation Protocol

We performed human evaluations of the DC model performance, using 250 randomly sampled dialogues from the JPersonaChat and JEmpatheticDialogues test sets for each of the four models. Five evaluators were presented with two dialogue histories, one before and one after completion, and

ID	Model	ZAR				*dep*	All
		All	*intra*	*inter*	*exophora*		
(e)	PAS wiki-cloze	62.27	68.39	**44.63**	67.77	**94.17**	83.67
(f)	PAS twitter-cloze	62.21	68.04	40.68	70.34	94.15	83.73
(g)	PAS wiki-pzero	62.68	68.35	43.02	69.99	93.96	83.75
(h)	**PAS twitter-pzero**	**63.25**	**68.68**	42.07	**72.04**	93.81	**83.87**

Table 3: Automatic evaluation results by the PAS model (F_1).

ID	Model	Appropriateness
(e)	DC wiki-cloze	74.80% (187 / 250)
(f)	DC twitter-cloze	77.20% (193 / 250)
(g)	DC wiki-pzero	72.40% (181 / 250)
(h)	**DC twitter-pzero**	**84.80%** (**212** / 250)

Table 4: Human evaluation results of the DC model.

asked to judge whether the completion phrase and its position were appropriate. Each evaluator evaluated 1,000 data divided into five parts, 50 per model, for a total of 200 data for the four models. To ensure fairness, the dialogue histories completed by each model were shuffled before presentation to the evaluator, thus obfuscating which model completed which.

4.4.3 Results

Table 4 shows the experimental results. The proposed DC twitter-pzero model achieved the best performance in dialogue completion. For the model pretrained with the Cloze task, using Twitter data instead of Wikipedia data for pretraining improved the performance by 2.4% (from 74.80 to 77.20). In the model pretrained with the PZERO task, using Twitter data instead of Wikipedia data for pretraining improved the performance by 12.4% (from 72.40 to 84.80). This suggests that using Twitter data for pretraining the DC model contributes to improving the performance of dialogue completion. Furthermore, in Table 3, pattern (h) shows the best performance, suggesting a relation between the performance of dialogue completion and that of predicate-argument structure analysis.

4.4.4 Analysis

Table 5 shows cases of successful and unsuccessful dialogue completion for analysis. Examples 1 and 2 are successful completion cases. In Example 1, the NOM and ACC cases corresponding to "cause" are completed correctly, and in Example 2, the NOM and ACC cases corresponding to "help" are also completed correctly. Examples 3 and 4 are cases of failed completions. In the sentence in Example 3,

Example 1:	I ate oysters at a barbecue and {$\phi_{\mathrm{NOM}} \rightarrow$ ✓ **oysters** } caused {$\phi_{\mathrm{ACC}} \rightarrow$ ✓ **me** } to suffer from stomach pains and diarrhea all night long.
Example 2:	The other day a classmate was bullied and {$\phi_{\mathrm{NOM}} \rightarrow$ ✓ **I** } helped {$\phi_{\mathrm{ACC}} \rightarrow$ ✓ **him** } out.
Example 3:	I spent a little too much {$\phi_{\mathrm{ACC}} \rightarrow$ ✗ **on my credit card** } last month ... credit card.
Example 4:	I was having a lot of morning sickness and {$\phi_{\mathrm{NOM}} \rightarrow$ ✗ **morning shickness** } was lying on a bench in the supermarket and someone talked to me.

Table 5: **Examples of DC model completion results (translated from Japanese).** Highlighted text represents complemented words. ✓ indicates a correct completion, while ✗ indicates an incorrect completion.

the argument corresponding to "spend" should not be completed because inverted sentences occur in the utterance. Since this method judges whether to perform completion by looking at the front of the target predicate, the method could not complete sentences with inverted predicates. To perform a correct completion, it is necessary to devise a way to rewrite "I spent a little too much on my credit card last month" before inputting it to eliminate the inversion occurring in "I spent a little too much last month ... credit card." In Example 4, "I" is the correct answer, but "morning sickness" is incorrectly completed. This problem could only be solved by using as a clue the knowledge that morning sickness is a phenomenon, and appropriate dialogue completion was not possible for a problem that required such common-sense knowledge.

4.5 Experiment 3: RG Model

We evaluated the performance of patterns (a) to (h) in Table 2 in generating dialogue responses. The RG model uses BERT2BERT (Rothe et al., 2020), which uses BERT as both the encoder and decoder. Patterns (a) to (d) are the baseline models, and pat-

terns (e) to (h) are the proposed models applying the DCZAR framework. The baseline model uses the dialogue history (text before completion) contained in the dataset. The proposed model uses as input the dialogue history complemented by the DC model.

4.5.1 Dataset

The RG model is trained using dialogue history–response pairs, with only the dialogue history used as input for response generation during inference. We used JPersonaChat and JEmpatheticDialogues to fine-tune the RG model. `[SPK1]` and `[SPK2]` are added as special tokens. These special tokens are added immediately before the utterances of the two speakers in the dialogue history to make it easier for the model to distinguish between each speaker.

4.5.2 Evaluation Protocol

Automatic Evaluation. We used standard natural language generation metrics such as BLEU (Papineni et al., 2002), ROUGE-L (Lin and Och, 2004), DIST-N (Li et al., 2016), and BERTScore (Zhang et al., 2020a).

Human Evaluation. All evaluators evaluated all the 100 randomly sampled cases from the JPersonaChat and JEmpatheticDialogues evaluation sets for each of the four pretraining models, for a total of 400 cases. Three evaluators were presented with the dialogue history and two responses generated by two models (proposed method, baseline), and were asked to choose one or select *not sure* for evaluation criteria in a pair-wise comparison. The responses were evaluated in three dimensions: which was more *grammatical*, which was more *coherent*, and which was more *engaging*. To ensure fairness, the responses generated by each model were shuffled before presentation to the evaluators, making it impossible to distinguish which model generated which response. The final evaluation value was determined by a majority vote of the three evaluators.

4.5.3 Results

Automatic Evaluation. Table 6 shows the results of a single run of the automatic evaluation. We performed a permutation test for each proposed method and each baseline method. For BLEU-1, 3, 4 and ROUGE-L, the proposed method outperformed the baseline method, but there was no significant difference. Although the proposed method was expected to produce more coherent and engaging responses by compensating for predicate arguments, these automatic metrics were not necessarily appropriate, because their contribution was not expected to change the results of the word statistics.

Human Evaluation. Table 7 shows the results of human evaluation. * and ** indicate a significant difference with $p < 0.05$ and 0.01, respectively, by the chi-square test. Note that although this table shows the values after the majority vote, the values before the majority vote were used for the chi-square test. First, no models differed significantly in terms of grammaticality, but the RG `twitter-cloze`+DCZAR and RG `twitter-pzero`+DCZAR models exceeded the baseline. One possible reason for the lack of significant differences is that the number of N/A cases was higher than it was for the other perspectives. In terms of coherence, all models in which the DCZAR framework was applied showed significant improvements over the baseline model. In particular, the proposed RG `twitter-pzero`+DCZAR model shows a significant improvement as compared with the RG `twitter-pzero` model (from 38 to 62). This indicates that the use of dialogue history with explicit completion of omissions in the input contributes to coherence evaluations when generating responses. In terms of engagement, all models except the RG `wiki-pzero`+DCZAR model showed significant improvements over the baseline model.

4.5.4 Analysis

We analyzed the generated sentences in Table 8.

Why was there no significant difference in grammaticality scores between the baseline and the proposed method? This was possibly due to the higher number of N/A results as compared with the other perspectives. Dialogue 1 is an example where three evaluators selected *not sure* and the response was classified as N/A. In this example, although the two models generated responses with different content, neither response was grammatically incorrect, and the decision may have been difficult in this case.

Does the proposed method contribute to improved coherence? Dialogue 2 is an example evaluated as contributing to the generation of a more coherent response by the proposed method. In the dialogue history of Dialogue 2, there are

ID	Model	BLEU				ROUGE-L	DIST		BERT
		1	2	3	4		1	2	Score
(a)	RG$_{\texttt{wiki-cloze}}$	25.29	6.14	2.02	0.69	9.57	**12.20**	**29.32**	69.70
(e)	+ **DCZAR (ours)**	24.50	5.65	1.78	0.55	**14.50**	11.72	28.50	69.45
(b)	RG$_{\texttt{twitter-cloze}}$	25.65	**6.50**	2.10	0.70	9.79	12.04	29.02	69.84
(f)	+ **DCZAR (ours)**	**25.72**	6.16	1.96	0.66	11.65	12.14	28.95	69.73
(c)	RG$_{\texttt{wiki-pzero}}$	25.59	6.31	2.09	0.72	13.72	12.09	28.96	**69.90**
(g)	+ **DCZAR (ours)**	25.45	6.08	1.96	0.63	6.49	12.06	29.14	69.75
(d)	RG$_{\texttt{twitter-pzero}}$	25.00	6.08	2.02	0.69	11.99	12.17	29.31	69.74
(h)	+ **DCZAR (ours)**	25.50	6.17	**2.11**	**0.73**	9.41	11.72	28.54	69.77

Table 6: Automatic evaluation results of the RG model.

ID	Model	gramm-atical	coherent	enga-ging
(a)	RG$_{\texttt{wiki-cloze}}$	**30**	45	44
(e)	+ **DCZAR (ours)**	28	**54****	**55****
	N/A	42	1	1
(b)	RG$_{\texttt{twitter-cloze}}$	30	43	46
(f)	+ **DCZAR (ours)**	**34**	**57****	**54****
	N/A	36	0	0
(c)	RG$_{\texttt{wiki-pzero}}$	**34**	45	**51**
(g)	+ **DCZAR (ours)**	33	**52****	47
	N/A	33	3	2
(d)	RG$_{\texttt{twitter-pzero}}$	32	38	41
(h)	+ **DCZAR (ours)**	**38**	**62****	**59****
	N/A	30	0	0

Table 7: **Human evaluation results of the RG model.** N/A shows the number of cases in which *not sure* was selected by a majority vote and in which the evaluation was split among the three evaluators.

Dialogue 1	
Speaker A:	I'm worried about my parents' retirement.
Speaker B:	{ϕ_{NOM} → **Our parents' retirement** } is a concern.
Speaker A:	I live far away from my parents and I don't know what they think.
RG$_{\texttt{wiki-cloze}}$:	We are concerned.
+ **DCZAR (ours)**:	It makes us nervous. We should think of something.
Dialogue 2	
Speaker A:	Being a housewife, there're so many times I thank my husband.
Speaker B:	I'm sure he's a kind spouse.
Speaker A:	Regarding the year-end gift, {ϕ_{NOM} → **my husband** } gives {ϕ_{DAT} → **gifts** } to {ϕ_{ACC} → **relatives** } on his wife's side as well as relatives on his side without any complaint.
RG$_{\texttt{twitter-pzero}}$:	**Your wife** is kind.
+ **DCZAR (ours)**:	**Your husband** is a great guy.
Dialogue 3	
Speaker A:	I told my mother that I wasn't feeling well and {ϕ_{NOM} → **she** } drove me to the hospital.
RG$_{\texttt{twitter-cloze}}$:	That would be great.
+ **DCZAR (ours)**:	Your mother is very kind.

Table 8: **Generation examples (translated from Japanese).** Highlighted text shows words filled in by our proposed model.

many nouns that express the role of a person, such as husband, spouse, wife, and housewife, and the arguments corresponding to "give" is omitted, making it difficult to guess who is being talked about in the sentence. The RG$_{\texttt{twitter-pzero}}$ model misinterprets the dialogue as being about the wife, and so generates an inappropriate response. By contrast, the proposed RG$_{\texttt{twitter-pzero}}$+DCZAR model correctly interprets the dialogue as being about the husband, so it generates an appropriate response. This suggests that the proposed method explicitly supplements the arguments corresponding to the predicate, which improves coherence.

What are the characteristics of engaging responses? The proposed method evaluated Dialogue 3 as contributing to the generation of a more engaging response. This example shows that the responses evaluated as more engaging are those that are more specific and coherent. Therefore, we hypothesized that the more coherent the response, the more engaging the response, and analyzed correlations among the indicators. The correlation coefficient for grammatical correctness and engagement was 0.223 and for coherence and engagement was 0.850. This indicates a strong correlation between coherence and attractive responses, and that people tend to want to continue dialogue with those who are consistent in their communication. Fleiss' Kappa (Fleiss, 1971), a measure of agreement among the evaluators, was calculated to be 0.095 for grammatical correctness, 0.287 for coherence, and 0.214 for engagement. The human evaluations indicated that the DCZAR framework contributed to the generation of more coherent and engaging responses.

5 Conclusion

We proposed the DCZAR framework, which explicitly completes omitted information in the dialogue history and generates responses from the completed dialogue history. Experimental results showed that the DCZAR framework can generate more coherent and engaging responses.

Limitations

We outline some potential limitations of our work below. First, extending to other languages requires pretrained models and datasets for each task (PAS, DC, RG) in that language. Also, our results do not necessarily guarantee the same results in languages other than Japanese. As we mentioned in Section 4.4.4, dialogue completion does not work well with inverted sentences and sentences that require common-sense knowledge as completion cues. Extending the DC model to handle such cases is a task for future work.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP22H00804 and JP21K18115; JST AIP Acceleration Research JPMJCR22U4, Japan.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 2206–2240.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

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

Mihail Eric, Nicole Chartier, Behnam Hedayatnia, Karthik Gopalakrishnan, Pankaj Rajan, Yang Liu, and Dilek Hakkani-Tur. 2021. Multi-sentence knowledge selection in open-domain dialogue. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 76–86, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Ryu Iida, Mamoru Komachi, Naoya Inoue, Kentaro Inui, and Yuji Matsumoto. 2010. Annotating predicate-argument relations and anaphoric relations: Findings from the building of the naist text corpus. *Journal of Natural Language Processing*, 17(2):25–50.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Ryuto Konno, Shun Kiyono, Yuichiroh Matsubayashi, Hiroki Ouchi, and Kentaro Inui. 2021. Pseudo zero pronoun resolution improves zero anaphora resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3790–3806, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Ryohei Sasano and Sadao Kurohashi. 2011. A discriminative approach to Japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 758–766, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. 2002. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Robert C. Stalnaker. 1978. Assertion. *Pragmatics*, pages 315–332.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. arXiv:2109.05217.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th Conference on Neural Information Processing Systems*, pages 3104–3112.

Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. 2008. A Japanese predicate argument structure analysis using decision lists. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 523–532, Honolulu, Hawaii. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language models for dialog applications. arXiv:2201.08239.

Masato Umakoshi, Yugo Murawaki, and Sadao Kurohashi. 2021. Japanese zero anaphora resolution can benefit from parallel texts through neural transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1920–1934, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2022. Retrieval-free knowledge-grounded dialogue response generation with adapters. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 93–107, Dublin, Ireland. Association for Computational Linguistics.

Souta Yamashiro, Hitoshi Nishikawa, and Takenobu Tokunaga. 2018. Neural Japanese zero anaphora resolution using smoothed large-scale case frames with word embedding. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252, Dublin, Ireland. Association for Computational Linguistics.

	Articles	Sentences	Predicates
Train	1,751	24,283	68,753
Dev	480	4,833	13,882
Test	696	9,284	26,379

Table 9: Statistics for NAIST Text Corpus 1.5.

		dep	*intra*	*inter*	*exophora*
Train	**NOM**	37,678	11,556	7,518	11,516
	ACC	24,997	1,803	928	128
	DAT	5,855	360	278	60
Dev	**NOM**	7,550	2,556	1,766	1,917
	ACC	5,107	394	166	32
	DAT	1,637	112	99	28
Test	**NOM**	14,254	4,770	3,342	3,721
	ACC	9,532	786	358	55
	DAT	2,547	211	140	54

Table 10: Distribution of arguments in NAIST Text Corpus 1.5.

A Ethical Considerations

Since the dialogue response generation model uses large-scale data from websites (e.g., Wikipedia, Twitter) during pretraining, it may generate responses that contain implicit biases and offensive content. We will incorporate mechanisms to reduce harmful responses and build a safe and ethically robust dialogue system in the future.

B Details of Scientific Artifacts

B.1 Dataset

Wikipedia. We used a publicly available data dump of Japanese Wikipedia jawiki-latest-pages-articles.xml.bz2.

Twitter. We used preprocessed tweets collected through the Twitter API[1] to pretrain the model. We used all tweets by 3,702 users with tweet histories ranging from 10K to 50K postings, sorted in chronological order.

NAIST Text Corpus 1.5. We used NAIST Text Corpus (NTC) 1.5 to test the performance of the PAS model. NTC is a corpus of newspaper articles and editorials with information such as relations between predicates and surface cases. Table 9 shows the NTC statistics, and Table 10 shows the distribution of NTC arguments.

[1] https://developer.twitter.com/en/products/twitter-api

B.2 Model

In this work, we used HuggingFace Transformers[2] version 4.21.0 (Wolf et al., 2020), and weights of bert-based-japanese-whole-word-masking[3] provided in transformers were used as initial parameters for the pretrained model.

B.3 Metric

For the BLEU[4], ROUGE-L[5], and BERTScore[6] implementations, we used publicly available code from Huggingface.

B.4 Software

We used MeCab 0.996[7] , a Japanese morphological analyzer, and CaboCha 0.69[8], a Japanese dependency analyzer, to preprocess the dataset. We will release our code publicly available.

B.5 License

As for the datasets, Japanese Wikipedia is made available under the CC BY-SA 3.0 license, NAIST Text Corpus 1.5 is released under a Revised BSD License, and JPersonaChat and JEmpatheticDialogues are licensed for the purpose of evaluating the model performance, but not for providing dialogue services themselves. MeCab is available under three licenses (BSD, LGPL, and GPL), and CaboCha is released under the Revised BSD License. The bert-based-japanese-whole-word-masking model is available under the CC BY-SA 3.0 license. Since both licenses allow use for research purposes, the use of these artifacts is valid for this work.

C Details of Experiments

C.1 Software and Hardware

We used Python 3.8, PyTorch 1.12.1, and HuggingFace Transformers 4.21.0. All experiments were performed using two NVIDIA A100 80 GB GPUs for model pretraining and one NVIDIA A100 80 GB GPU for fine-tuning. The pretraining time was about six days per model, fine-tuning for

[2] https://github.com/huggingface/transformers/

[3] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

[4] https://github.com/huggingface/datasets/blob/main/metrics/bleu/bleu.py

[5] https://github.com/huggingface/datasets/blob/main/metrics/rouge/rouge.py

[6] https://huggingface.co/spaces/evaluate-metric/bertscore

[7] https://taku910.github.io/mecab/

[8] https://taku910.github.io/cabocha/

Pretraining	
Mini-batch Size	2048
Max Learning Rate	1.0×10^{-4} (Cloze)
	2.0×10^{-5} (PZERO)
Learning Rate Schedule	Inverse square root decay
Warmup Steps	5,000
Number of Updates	30,000
Loss Function	Cross entropy (Cloze),
	KL divergence (PZERO)
Fine-tuning of the PAS model	
Mini-batch Size	256
Max Learning Rate	5.0×10^{-5}
Number of Epochs	20
Loss Function	KL divergence,
	Cross entropy (*exophora*)
Fine-tuning of the RG model	
Mini-batch Size	1,024
Max Learning Rate	5.0×10^{-5}
Number of Epochs	30
Max Sequence Length	512 (encoder)
	128 (decoder)

Table 11: List of hyperparameters.

the PAS model was about nine hours per model, and fine-tuning for the RG model was about seven hours per model.

C.2 Hyperparameters

Table 11 lists the hyperparameters used in this study.

D Details of Human Evaluation

Since the human evaluations in these studies (Section 4.4.2 and Section 4.5.2) did not require expert knowledge of linguistics, we recruited eight Japanese undergraduate and graduate student evaluators from within our laboratory. We informed the evaluators of the purpose of this evaluation in advance and obtained their consent. Table 12 shows an English translation of the instructions given to the evaluators, who were paid for their time in accordance with university regulations.

E Experimental results in Japanese

The experimental results of the Japanese versions of Table 5 in Section 4.4 and Table 8 in Section 4.5 are shown in Table 13 and Table 14, respectively.

Human evaluation of the DC model

Task Info:
I am researching a dialogue response generation system, and as part of that research,
I need to evaluate the system's completion performance.
You are to read a sentence upon which a completion operation has been performed
and evaluate the appropriateness of both the wording and the position,
using a binary value of 1 (appropriate) or 0 (not appropriate).
Note that there may be typos in these sentences.
One example is the incorrect *Kodomo ga chiisai-no-shi* (which should be *Kodomo ga chiisai-shi*).
Please do not consider typos that are present in the source texts in your evaluation.

Example:
(a) When both the completion phrase and the completion position are appropriate.
`Speaker A: I took my summer suit to the cleaners.`
`Speaker B: Well done! When will` **`the suit`** `be ready?`
– The correct word and the position of the complement are both appropriate, so select 1.

(b) When the complementing phrase is not appropriate, but the completion position is appropriate.
`Speaker A: Even if I had a boyfriend, I would break up with him right away.`
`Speaker B: Maybe` **`I`** `just haven't met the man of my dreams yet.`
– The completion position is appropriate, but the appropriate complement phrase is `You` instead of `I`. In such a case, select 0.

Human evaluation of the RG model

Task Info:
I am researching a dialogue response generation system, and as part of that research,
I need to evaluate its response performance.
You will be given multiple dialogue contexts and two responses.
Please compare Responses 1 and 2 and evaluate which is more grammatical,
which is more coherent, and which is more engaging.
grammatical: Which response is more grammatical and fluent in Japanese
(ignoring the dialogue history)?
coherent: Which response is more coherent, considering the dialogue history?
engaging: Which response is more engaging, and which response makes you want to continue
the dialogue with the person to whom you are talking?
Select 1 if Response 1 is better, select 2 if Response 2 is better, or select 3 if you are unsure
which response is better. Use choice 3 as sparingly as possible.

Example:
You will evaluate the following Responses 1 and 2 from three perspectives.
`Speaker A: I bought a gaming console.`
`Speaker B: Did you buy your first game?`
`Speaker A: Yes, it's surprisingly interesting.`
—
`Response 1: What game are you playing?`
`Response 2: I see. Good for you.`

Table 12: Instructions for evaluators (translated from Japanese).

Example 1:	バーベキューで牡蠣を食べたら、 牡蠣が 見事に当たって 一晩中 私が 腹痛と下痢に苦しみました。
Example 2:	この前、クラスメートがいじめられてたから、 私が クラスメートを 助けたんだ。
Example 3:	先月 カードを ちょっと使いすぎちゃったんだよなクレジットカード。
Example 4:	つわりが酷くて つわりが スーパーのベンチに横たわってたら いろんな人が大丈夫ですか？って声かけてくれたよ。

Table 13: **Examples of DC model completion results in Japanese.** Highlighted text represents complemented words.

	Dialogue 1
Speaker A:	親の老後が不安。
Speaker B:	この世代になると、 老後が 不安になってくるよね。
Speaker A:	そうなんだよね。離れて暮らしてるし、親の意向も分からないしでね。
RG$_{\text{wiki-cloze}}$:	心配だよね。
+ DCZAR (ours):	不安になるよね、何か考えておかないとね。
	Dialogue 2
Speaker A:	専業主婦をしてると、夫に感謝する場面がとても多いわ。
Speaker B:	きっと優しいご主人なんだろうね。
Speaker A:	お歳暮も、夫の親戚と同じように、妻側の親戚にも 嫌な顔ひとつせずに 夫が 歳暮を 親戚に 贈ってくれるのよ。
RG$_{\text{twitter-pzero}}$:	貴方の奥さん優しいなあ。
+ DCZAR (ours):	それは素晴らしい旦那さんね。
	Dialogue 3
Speaker A:	母に具合が悪いことを伝えたら、 母が 病院まで車で送ってくれました。
RG$_{\text{twitter-cloze}}$:	それはありがたいね。
+ DCZAR (ours):	お母さま、優しいですね。

Table 14: **Generation examples in Japanese.** Highlighted text shows words filled in by our proposed model.

SYNDICOM: Improving Conversational Commonsense with Error-Injection and Natural Language Feedback

Christopher Richardson, Anirudh Sundar, Larry Heck
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, 30308, USA
{crichardson8,asundar34,larryheck}@gatech.edu

Abstract

Commonsense reasoning is a critical aspect of human communication. Despite recent advances in conversational AI driven by large language models, commonsense reasoning remains a challenging task. In this work, we introduce SYNDICOM - a method for improving commonsense in dialogue response generation. SYNDICOM consists of two components. The first component is a dataset composed of commonsense dialogues created from a knowledge graph and synthesized into natural language. This dataset includes both valid and invalid responses to dialogue contexts, along with natural language feedback (NLF) for the invalid responses. The second contribution is a two-step procedure: training a model to predict natural language feedback (NLF) for invalid responses, and then training a response generation model conditioned on the predicted NLF, the invalid response, and the dialogue.

SYNDICOM is scalable and does not require reinforcement learning. Empirical results on three tasks are evaluated using a broad range of metrics. SYNDICOM achieves a relative improvement of 53% over ChatGPT on ROUGE-1, and human evaluators prefer SYNDICOM over ChatGPT 57% of the time. We will publicly release the code and the full dataset.

1 Introduction

Conversational AI has witnessed rapid advancements in recent years, largely due to the success of large language models (LLMs) such as GPT-3 (Brown et al., 2020). These advancements have been driven by the notable achievements of models like ChatGPT, which is built upon InstructGPT (Ouyang et al., 2022). InstructGPT was trained on an extensive dataset of instructions for various language tasks and was further enhanced using human feedback and reinforcement learning (RL). Consequently, research in conversational AI has shifted towards leveraging large models trained on extensive datasets, supplemented by human feedback.

While these models have consistently demonstrated significant improvements in reasoning and problem-solving capabilities, they still exhibit flaws and issues. In many critical applications of LLMs, the tolerance for errors in dialogue responses is exceedingly low. Addressing these problems remains challenging, primarily due to the scarcity of data and the high cost associated with human feedback. Recent research has started exploring alternative techniques beyond human feedback and RL, such as natural language feedback (NLF) and self-correction (Saunders et al., 2022; Scheurer et al., 2022; Welleck et al., 2022; Bai et al., 2022b).

Furthermore, even with the progress made, large models often generate hallucinations, underscoring the ongoing importance of knowledge grounding. One of the most demanding aspects of knowledge grounding is commonsense knowledge. Recent advancements in incorporating commonsense into LLMs have utilized resources such as ConceptNet (Speer et al., 2017) or ATOMIC (Sap et al., 2019).

This paper presents a method for improving commonsense dialogue responses by (1) replacing human feedback and RL with natural language responses and (2) leveraging recent knowledge graph techniques to ground responses in commonsense knowledge derived from ATOMIC. To address the scarcity of data and the high cost of human feedback, the natural language feedback is elicited in a manner that specifically targets the chosen error types determined by the designer. This approach significantly enhances the speed and quality of model learning and refinement.

The contributions of this paper are as follows:

- Development of a scalable method for synthesizing knowledge-grounded data with error injection and feedback.
- Release of a dataset rich in dialogues featuring commonsense inferences, annotated with

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 297–308
September 11–15, 2023. ©2023 Association for Computational Linguistics

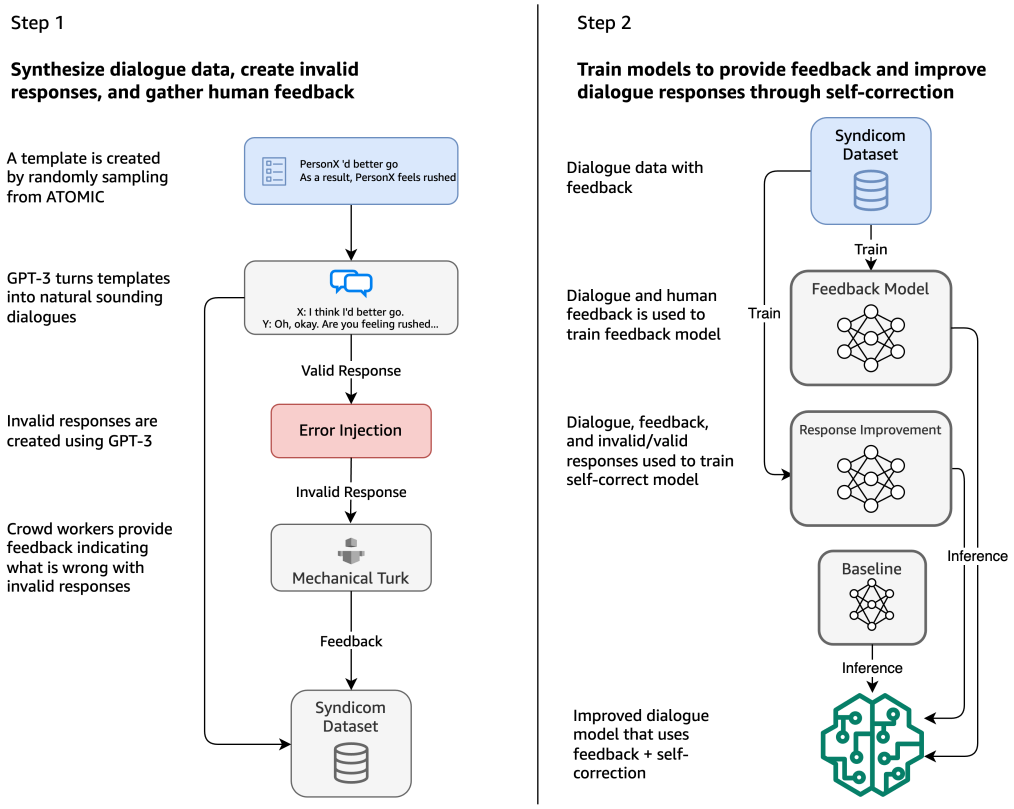


Figure 1: SYNDICOM Process. Left: dataset generation, Right: Improving commonsense in dialogue response generation.

commonsense errors, and accompanied by human-written feedback, which we refer to as SYNDICOM.

- Description of a method for training both a feedback generation model and a response improvement model using natural language feedback (NLF), and demonstration of the superiority of this information-rich approach over state-of-the-art RL methods using SYNDICOM.

2 Recent Work

The field of conversational AI has experienced a surge of interest in commonsense reasoning in recent years, with a significant focus on curating datasets (Richardson and Heck, 2023). ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019) have emerged as widely used resources for dataset curation, establishing a de facto standard. Several datasets serve as sources for the dialogues, including DailyDialogue (Li et al., 2017), MuTual (Cui et al., 2020), DREAM (Sun et al., 2019), and the Ubuntu Dialogue Corpus (Lowe et al., 2015).

Our research lies at the intersection of two critical areas in conversational AI: the synthesis of commonsense datasets and the training of models using natural language feedback. These areas have recently garnered significant research attention due to their potential to enhance the ability of conversational agents to understand and respond to complex human interactions with greater accuracy and consistency. By leveraging the synergies between these domains, our work aims to address the existing limitations in conversational agents and pave the way for more robust and effective conversational systems.

2.1 Commonsense Dataset Curation

In recent years, various datasets have been curated specifically for commonsense reasoning. Ghosal et al. (2021) introduced CIDER, a dialogue dataset annotated with commonsense inferences, which was later expanded with the more open-ended CICERO (Ghosal et al., 2022). Some researchers have focused on specific types of commonsense, such as temporal commonsense (Qin et al., 2021) and ethical commonsense (Ziems et al., 2022; Kim et al., 2022; Sun et al., 2022). Others have concentrated on grounding dialogues in knowledge graphs

(Zhou et al., 2021a; Moon et al., 2019).

These approaches rely on existing dialogue datasets and often employ filtering strategies to reduce dataset size. However, this reliance on existing datasets can limit the generalizability of methods to future problems. One potential solution to the scarcity of large-scale annotated commonsense knowledge datasets is the synthesis approach. Recently, Kim et al. (2022) proposed SODA, a method for procedurally generating social dialogues based on a commonsense knowledge graph. They utilized ATOMIC (Sap et al., 2019), which consists of atomic facts in natural language form, to generate synthetic dialogues rich in commonsense inferences. Their entirely procedural and highly scalable approach generates dialogue data suitable for training models that reason over commonsense knowledge. Building upon this work, we present SYNDICOM, a synthesis procedure and dataset that expands on the ideas of SODA and incorporates novel features crucial for our dialogue modeling approach. More details about SYNDICOM are provided in Section 3.

2.2 Feedback and Response Improvement

The use of feedback to improve language models has recently garnered increased interest, with most efforts focused on the application of reinforcement learning (Stiennon et al., 2020; Zhou et al., 2021b; Bai et al., 2022a,b). Reinforcement learning with human feedback (RLHF) is particularly notable as it serves as the foundation for InstructGPT (Ouyang et al., 2022), which paved the way for ChatGPT. RLHF offers a flexible approach to improving LLMs; however, it faces challenges in terms of stability and efficiency inherent to RL. Moreover, the low dimensionality of the reward signal in RL (typically a scalar) severely limits the learning rate.

A more information-rich approach than RL is the use of natural language feedback (NLF). NLF has been explored in several recent works. Scheurer et al. (2022) investigated the use of human-written NLF to train a dialogue response refinement model. Saunders et al. (2022) demonstrated that LLMs themselves can generate this feedback. Welleck et al. (2022) developed a method to improve sequence generation of LLMs by first generating a baseline using an imperfect base generator and then correcting the output using a second correction model. The correction model incorporates feedback as part of its input. However, the authors only demonstrated the use of feedback provided by various tools and APIs tailored to the specific tasks they explored.

3 The SYNDICOM Method

Taking inspiration from recent NLF methods, this paper presents a new approach called SYNDICOM. This new approach combines the synthesis of commonsense dialogue data from a grounded knowledge graph (ATOMIC) with an NLF response improvement approach to improve dialogue responses. Figure 1 illustrates the two phase process.

3.1 SYNDICOM Dataset

The SYNDICOM dataset is created in a four step process: (1) Auto-generate commonsense dialogue templates, (2) Translate templates into natural language dialogues, (3) Generate invalid responses with error injection, and (4) Collect human-written explanations for the invalid responses. Examples from the SYNDICOM dataset are shown in Table 1. The GPT model we used for the steps in this section was text−davinci−003. Statistics for the dataset are shown in Table 2.

3.1.1 Generating Templates

Our approach generates commonsense-focused dialogue templates from a commonsense knowledge base. For this study, we utilize ATOMIC (Hwang et al., 2021). ATOMIC consists of inferences in the form of Head $\xrightarrow{\text{relation}}$ Tail. Each head and tail is a natural language description of a generic event, emotional state, action, description, etc. Dialogue templates are constructed by crawling through inferences rooted at each head of ATOMIC and chaining these inferences together to form multiple dialogue turns. The number of dialogue template turns is uniformly and randomly chosen between 3 and 8.

3.1.2 Converting to Natural Language

Given the dialogue templates, the second step in creating SYNDICOM converts the templates to natural language conversations. We explored several methods, including crowdsourcing, but found LLMs to be the most consistent and effective. We used the GPT LLM (text-davinci-003) to generate the natural language dialogues from the templates. This was followed by in-context learning with 15 hand-written examples. The exact prompting used is shown in detail in Appendix A.

Template	Synthesized Dialogue	Explanation
PersonX refuses PersonY ↪ PersonX is seen as: disagreeable ↪ As a result, PersonX feels: annoyed and irritated ↪ Before that, PersonX needed: thinks about it	**Context** I refuse to do what you ask. *Why are you being so disagreeable?* I'm just annoyed and irritated. **Valid Response:** You should think about it before you say no. **Invalid Response:** You should think about it before you say yes.	**Crowd Worker 1:** The other person already said no. **Crowd Worker 2:** The person did not say yes so this response was strange.
PersonX makes music ↪ As a result, PersonX wants: to impress ↪ PersonX is seen as: talented ↪ As a result, PersonX will: gets asked to play something ↪ Before that, PersonX needed: to carry their violin ↪ PersonX wanted: they want to share their creativity	**Context** I love making music. *You're very talented!* Thanks. I just want to impress people with my playing. *I'm sure you will. Can you play something for me?* Of course! I always carry my violin with me. **Valid Response:** That's great. I'm glad you want to share your creativity. **Invalid Response:** That's awful. I don't want to share my creativity.	**Crowd Worker 1:** This contradicts with what was said about impressing people. **Crowd Worker 2:** They aren't being asked to play. They asked the other person to play

Table 1: Example dialogues from SYNDICOM. Each dialogue context includes both valid and invalid responses, as well as crowd worker-written explanations for the invalid response.

3.1.3 Error Injection

To elicit feedback on commonsense from crowd workers, the SYNDICOM process starts by corrupting the valid dialogue responses so that they violate commonsense reasoning. This provides crowd workers with an easy target for their feedback. To corrupt the dialogue responses, SYNDICOM takes advantage of the commonsense dialogue inference structure provided by ATOMIC. Given a commonsense knowledge base $\mathcal{K}$, a dialogue context $\mathcal{C}$, and response r from SYNDICOM, the response is implied by commonsense from the context, or $\mathcal{C} \xrightarrow{\mathcal{K}} r$. The response r is corrupted by replacing it with the semantic opposite, $\overline{r}$. We prompted GPT as shown in Appendix A to acquire these semantic opposites. The result is dialogues annotated with commonsense contradictions of the form $\{\mathcal{C}, r, \overline{r}\}$.

3.1.4 Natural Language Feedback Acquisition

The dialogues with commonsense contradictions are presented to crowd workers on the Amazon's Mechanical Turk platform. Each dialogue is shown in the form of context and invalid responses, informing them that the dialogues were generated by an AI attempting to sound human. The crowd workers were given instructions to review AI-generated casual text message conversations and provide 1-2 sentences of natural language feedback on the dialogue, and the final turn in particular (the invalid response). They were asked to be as specific as possible in their feedback. The full instructions and web interface given to the crowd workers can be found in Appendix A.

To ensure the quality of the feedback, we used only masters-level crowd workers from English-speaking countries. This decision aimed to maximize the clarity and accuracy of the feedback provided. Each dialogue was evaluated by two crowd workers independently, allowing for a more comprehensive understanding of the AI's mistakes and ensuring a diverse range of feedback.

With the addition of the feedback f, this completes the dataset synthesis part, resulting in annotated dialogues of the form $\{\mathcal{C}, r, f, \overline{r}\}$.

3.2 SYNDICOM Dialogue Improvement

This section details the process of using natural language feedback to correct latent errors in the baseline conversational response. To begin, the dialogue response improvement problem is defined as follows: given a dialogue context $\mathcal{D}$ and a response r_b, generated by some dialogue system or model, produce an improved response r^*.

$$r^* = \underset{r}{\operatorname{argmax}}\, p(r|\mathcal{D}, r_b) \tag{1}$$

Dialogue response generation and improvement has recently received considerable attention (Shah et al., 2016; Nayak et al., 2017; Liu et al., 2017, 2018; Weston et al., 2018). This problem is especially relevant today with large language models (LLMs). While LLMs have recently reached a high degree of fluency in dialogue, in some domains they can be factually inaccurate. While these cases are relatively infrequent, the tolerance for factual errors for a number of important applications is very low. In addition, these errors are difficult to predict and/or automatically detect. This leads to a problem of data sparsity that is difficult to overcome for response improvement methods that rely on training models.

A method to partially mitigate the sparsity of dialogue response errors is to *artificially create invalid responses* $\overline{r}$ via error injection (as described in Section 3.1.3). This method will be called SYNDICOM-DIRECT. Given the invalid response $\overline{r}$ and the dialogue history $\mathcal{D}$, a model is trained to learn the optimal response r^*

$$r^* = \underset{r}{\operatorname{argmax}}\, p(r|\mathcal{D}, \overline{r}). \tag{2}$$

A second approach called SYNDICOM-NLHF includes natural language human feedback (NLHF) to explain the rationale for why the response $\overline{r}$ is invalid and then conditions on this side rationale.

$$r^* = \underset{r}{\operatorname{argmax}}\, p(r|\mathcal{D}, \overline{r}, f^*). \tag{3}$$

As a comparison, we also implemented an approach called SYNDICOM-MULTISTEP. This approach breaks the inclusion of NLHF into two steps: (1) train a feedback model on NLHF that *predicts* the feedback critical of response $\overline{r}$

$$\hat{f} = \underset{f}{\operatorname{argmax}}\, p(f|\mathcal{D}, \overline{r}). \tag{4}$$

and (2) train a second model to produce an improved dialogue response from the invalid response, given the *predicted* feedback

$$r^* = \underset{r}{\operatorname{argmax}}\, p(r|\mathcal{D}, \overline{r}, \hat{f}). \tag{5}$$

Both models used in this work are based on OpenAI's GPT-3.5, specifically text−davinci−003. The models were fine-tuned through the OpenAI API for GPT based models. The hyperparameters used are listed in Table 3.

4 Experiments

In this section, we provide a detailed description of the experiments conducted to evaluate our proposed method, SYNDICOM. The experiments aim to compare the direct prediction of the improved response in Equation 2 (SYNDICOM-DIRECT) with the response prediction when conditioned on natural language human feedback (NLHF) that explains why the initial response is invalid (SYNDICOM-NLHF). Additionally, we explore a multistep implementation of NLHF (SYNDICOM-MULTISTEP). We compare the performance of our method against a ChatGPT baseline (gpt−3.5−turbo) using various text generation metrics, such as ROUGE, BLEU, SacreBLEU, BERTScore, and METEOR.

4.1 SYNDICOM-DIRECT

Our first experiment focused on the direct dialogue improvement task, where the objective is to enhance a dialogue response based solely on the context and an invalid response. No feedback, whether human or generated, was involved in this task. This optimization problem is described in Equation 2.

In order to prevent the model from simply learning to undo the error injection, we introduced noise by rephrasing the invalid dialogues using an independent ChatGPT instance. This rephrasing was only performed at inference time and not during training. The rephrasing prompt is available in Appendix A.

4.2 SYNDICOM-MULTISTEP

Next, we explored the SYNDICOM-MULTISTEP approach. As shown in Equations 4 and 5, we first predicted feedback using the feedback model and then improved the dialogue response using the response improvement model. For the feedback predictor, we trained a GPT-based model to generate feedback given a dialogue context and an invalid response, as shown in Equation 4, using the typical causal language modeling objective. We evaluated the feedback generation model portion of SYNDICOM-MULTISTEP separately and compared it to ChatGPT. The prompt used for the baseline can be found in Appendix A. Table 4 presents the results, demonstrating that our method outperformed the baseline on all metrics.

Subsequently, we utilized the predicted feedback along with the dialogue context and invalid response to produce an improved dialogue response, as shown in Equation 5. Similar to the SYNDICOM-DIRECT experiments, we applied rephrasing to the invalid responses at inference time. The baseline model was explicitly instructed to first generate feedback for the invalid response and then use that feedback to guide its response improvement. Table 5 displays the results.

4.3 SYNDICOM-NLHF

The next experiment focused on enhancing dialogue responses using human feedback (Equation 3). Given a dialogue context, an invalid response, and human feedback, the goal was to generate an improved (valid) dialogue response. For this experiment, we utilized the raw human-written feedback from SYNDICOM and trained a separate GPT improvement model to generate valid responses. As

Description	Train	Val	Test
# Samples	16221	1709	1787
# Turns per template	5.21±1.42	5.26±1.42	5.23±1.42
# Turns per dialogue	5.18±1.36	5.21±1.36	5.18±1.32

Table 2: Statistics of our SYNDICOM dataset. # Dialogue turns includes the valid response (± indicates 1 std deviation.). The splits were inherited from ATOMIC, the source of the templates.

Hyperparameter	Value
Temperature	0.7
Max tokens	50
Top p	1.0
Frequency penalty	0
Presence penalty	0

Table 3: Hyperparameters used for GPT-3.5. The same parameters were used for training and inference.

before, we applied inference-time rephrasing to the invalid responses. Results are presented in Table 5 under SYNDICOM-NLHF. This version of our method outperformed the others on all metrics.

4.4 Human Evaluation

In addition to our automated metric evaluations, we conducted a human evaluation to assess the effectiveness of response improvements through generated feedback. This evaluation process mirrored the dialogue enhancement steps employed in the experiment described in Section 3.2.

It is important to note that task assignments for crowdworkers require explicit and precise definitions, which often pose challenges in evaluating the commonsense aspect through human intervention. Existing human evaluations primarily focus on assessing the accuracy of information or determining the most preferred output from a set of alternatives.

With the emergence of advanced language models like ChatGPT, human evaluation has become increasingly complex. This complexity arises from the remarkably high-quality and naturally articulated outputs generated by state-of-the-art models such as ChatGPT.

In our study, we instructed crowdworkers that an AI system was attempting to emulate human conversation and generate dialogue responses that align with commonsense understanding and fit the given context. The workers were presented with two distinct responses: a standard ChatGPT response and our SYNDICOM response. Their task was to select the response that appeared more human-like and natural. The order of the responses chosen was randomized.

Despite the impressive contextual relevance exhibited by ChatGPT responses, our method generated the more favored response **56.5%** of the time, compared to ChatGPT's 43.5% preference rate. For further details on the interface provided to the crowdworkers, please refer to Appendix A.

5 Discussion

In the Discussion section, we analyze the performance of our proposed SYNDICOM method in conversational AI compared to the baseline model ChatGPT. The results are summarized in Tables 4 and 5, where we observe that SYNDICOM outperforms ChatGPT on all automatic metrics for the feedback and dialogue response improvement tasks.

Specifically, Table 5 provides a comparison between our direct and multi-step approaches to the response improvement problem. Our multi-step method outperforms the direct method on various metrics such as ROUGE-1, BLEU, SacreBLEU, and BERTScore, despite the simplicity of the error typology used in the error injection during these experiments. This indicates that the multi-step approach has the potential to achieve even better performance when faced with more diverse error typologies, which we leave as an avenue for future research.

One contributing factor to the superior performance of the multi-step method is the additional information encoded in the feedback model. The feedback model is trained on human feedback, providing it with more contextual information compared to the direct model, which is solely trained on valid and invalid responses. Even in cases where the direct model achieves slightly higher scores in certain metrics, the differences are negligible. Notably, BERTScore, which represents the most comprehensive model-based metric utilized in our

	ChatGPT			SYNDICOM		
Metric	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**
ROUGE1	0.204	0.123	0.163	0.315	0.185	0.250
ROUGE2	0.034	0.0078	0.0209	0.112	0.035	0.073
ROUGEL	0.150	0.093	0.122	0.248	0.144	0.196
BERTSCORE	0.863	0.853	0.858	0.883	0.866	0.874
SacreBLEU	2.546	1.533	2.039	6.697	2.907	4.802
BLEU	0.004	0.0001	0.0021	0.030	0.0041	0.0171
METEOR	0.197	0.129	0.163	0.279	0.158	0.219

Table 4: Performance in Feedback Generation performance of our method vs. baseline. SYNDICOM outperforms the baseline on all metrics. Each dialogue was accompanied by two feedback responses, and scores were computed for both independently. We show the max/min/avg over the two for each score and model.

	ChatGPT		SYNDICOM		
Metric	Direct	NLHF	Direct	Multistep	NLHF
ROUGE1	0.132	0.231	0.386	**0.388**	*0.474*
ROUGE2	0.029	0.081	**0.174**	0.172	*0.246*
ROUGEL	0.112	0.201	**0.324**	0.322	*0.396*
BLEU	0.008	0.031	0.117	**0.125**	*0.168*
METEOR	0.209	0.290	**0.390**	0.387	*0.445*
SacreBLEU	0.885	3.107	11.716	**12.547**	*16.831*
BERTScore	0.859	0.880	0.909	**0.910**	*0.919*

Table 5: Response Improvement comparing ChatGPT with our new SYNDICOM methods. ChatGPT-Direct is fine-tuned to produce a valid response given only the invalid response, with no intermediate steps or feedback. ChatGPT-NLHF is additionally conditioned on natural language human feedback (NLHF). SYNDICOM-DIRECT is the model that optimizes Equation 2, SYNDICOM-MULTISTEP optimizes Equation 5, and SYNDICOM-NLHF conditions on the same NLHF as used by the ChatGPT models. Bold text illustrates the highest score between all methods that are not give NLHF, and italics indicate the highest scores among NLHF tasks. SYNDICOM outperforms the baseline on all metrics for both tasks.

evaluation, further supports the argument in favor of the multi-step approach with feedback generation.

When examining the NLHF columns in Table 5, we observe that SYNDICOM demonstrates significant improvement over ChatGPT for the response improvement task when provided with human feedback for the invalid response. This scenario aligns with use cases where feedback can be collected for a dialogue system and subsequently used to fine-tune and enhance the dialogue model. These findings underscore the value of the SYNDICOM method in continuous learning scenarios, particularly those where feedback from end users is actively being collected.

Overall, SYNDICOM exhibits strong performance compared to the state-of-the-art large language model ChatGPT, despite both models being based on the same underlying architecture (GPT-3.5). It is worth noting that ChatGPT underwent substantial reinforcement learning through human feedback during its refinement process, making the success of SYNDICOM even more noteworthy.

6 Conclusion

In this paper, we introduced SYNDICOM, a novel method for enhancing commonsense reasoning in dialogue response generation. By integrating a commonsense dialogue synthesis approach with targeted error injection, we tackled the challenge of incorporating commonsense knowledge into conversational AI systems. Our method comprised two key components: (1) a dataset consisting of valid and invalid responses to dialogue contexts, along with natural language feedback (NLF) for the invalid responses, and (2) a two-step procedure in-

volving training a model to predict NLF for invalid responses, followed by training a response generation model conditioned on the predicted NLF, the invalid response, and the dialogue.

A notable advantage of SYNDICOM is its scalability and independence from reinforcement learning techniques, which are commonly employed in previous methods utilizing human feedback. Through comprehensive empirical evaluations across three tasks, we demonstrated the effectiveness of our approach using a diverse range of metrics. Notably, SYNDICOM outperformed ChatGPT on all metrics for both the dialogue improvement tasks, with and without human feedback.

To facilitate further research and practical adoption, we plan to release the code implementation of SYNDICOM as well as the complete dataset utilized in this work. By making these resources openly accessible, we aim to encourage collaboration and promote advancements in commonsense reasoning for dialogue systems.

Limitations and Future Work

There are a few areas of limitation in this work. First, all the dialogues generated were based on templates synthesized from ATOMIC triplets. The domain is thus limited to the material contained in ATOMIC. Second, the procedural generation technique, while scaleable, inevitably introduces structure within the data that can be exploited by statistical models (including deep neural nets and language models). This is why the feedback generation task is particularly crucial, because the explanations are human-written and thus avoid such a limitation.

Our experiments demonstrate our method of improving baseline dialogue responses that have been corrupted with error injection. This has the advantage of scale and targeting specific error modes that may be observed with LLMs, but the invalid responses in SYNDICOM do not themselves represent errors actually made by LLMs. A larger scale study could involve a data collection of errors and mistakes made by an LLM to demonstrate our method in improving baseline dialogue responses, but this approach would not lend itself to scale as any particular type of error made by state-of-the-art LLMs will likely be very rare. A more scaleable approach might be to develop a more comprehensive error typology and injection scheme, which we leave to future work.

In future work, a more comprehensive error topology could be explored, along with a more substantial human evaluation, to explore the generalizability of the proposed method. This work focused on commonsense errors, but other errors that are observed in large language models could be explored in further analysis like mathematical reasoning, humor and sarcasm, etc.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. *arXiv preprint arXiv:2004.04494*.

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. CICERO: A dataset for contextualized commonsense inference in dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5010–5028, Dublin, Ireland. Association for Computational Linguistics.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. *arXiv preprint arXiv:2205.12688*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. 2017. End-to-end optimization of task-oriented dialogue model with deep reinforcement learning. *Conversational AI Workshop, Neural Information Processing Systems (NeurIPS)*.

Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2060–2069, New Orleans, Louisiana. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Neha Nayak, Dilek Hakkani-Tür, Marilyn A Walker, and Larry P Heck. 2017. To plan or not to plan? discourse planning in slot-value informed sequence to sequence models for language generation. In *INTERSPEECH*, pages 3339–3343.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. Timedial: Temporal commonsense reasoning in dialog. *arXiv preprint arXiv:2106.04571*.

Christopher Richardson and Larry Heck. 2023. Commonsense reasoning for conversational ai: A survey of the state of the art. *arXiv preprint arXiv:2302.07926*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with natural language feedback. *arXiv preprint arXiv:2204.14146*.

Pararth Shah, Dilek Hakkani-Tür, Tür, and Larry Heck. 2016. Interactive reinforcement learning for task-oriented dialogue management. *Workshop on Deep Learning for Action and Interaction, Neural Information Processing Systems (NIPS)*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Hao Sun, Zhexin Zhang, Fei Mi, Yasheng Wang, Wei Liu, Jianwei Cui, Bin Wang, Qun Liu, and Minlie Huang. 2022. Moraldial: A framework to train and evaluate moral dialogue systems via constructing moral discussions. *arXiv preprint arXiv:2212.10720*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*.

Jason Weston, Emily Dinan, and Alexander H Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv preprint arXiv:1808.04776*.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021a. Commonsense-focused dialogues for response generation: An empirical study. *arXiv preprint arXiv:2109.06427*.

Ruijie Zhou, Soham Deshmukh, Jeremiah Greer, and Charles Lee. 2021b. Narle: Natural language models using reinforcement learning with emotion feedback. *arXiv preprint arXiv:2110.02148*.

Caleb Ziems, Jane A Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. *arXiv preprint arXiv:2204.03021*.

A GPT-3 Prompts and Mechanical Turk interfaces

Task	Prompt
Direct	You will be given a dialogue context and a baseline response. Your job is to improve that baseline response. Always write the improved response last and prefix it with 'Improved Response:'
NLHF	You will be given a dialogue context and a baseline response. Your job is to improve that baseline response. Do so by first generating feedback for that response, as if it was written by an AI and you are critiquing it, and then produce the improved response. Always write the improved response last and prefix it with 'Improved Response:'
Feedback Generation	You are shown a synthetic dialogue written by an AI. The dialogue is intended to sound like a natural text message conversation between two people. The AI is imperfect and makes mistakes. You are asked to provide feedback to the AI to improve its dialogue generation. You are given a few dialogue turns, followed by a Baseline Response. Please give 1-2 sentences of feedback for the baseline response, and please be specific!

Table 6: Prompts used for ChatGPT baselines

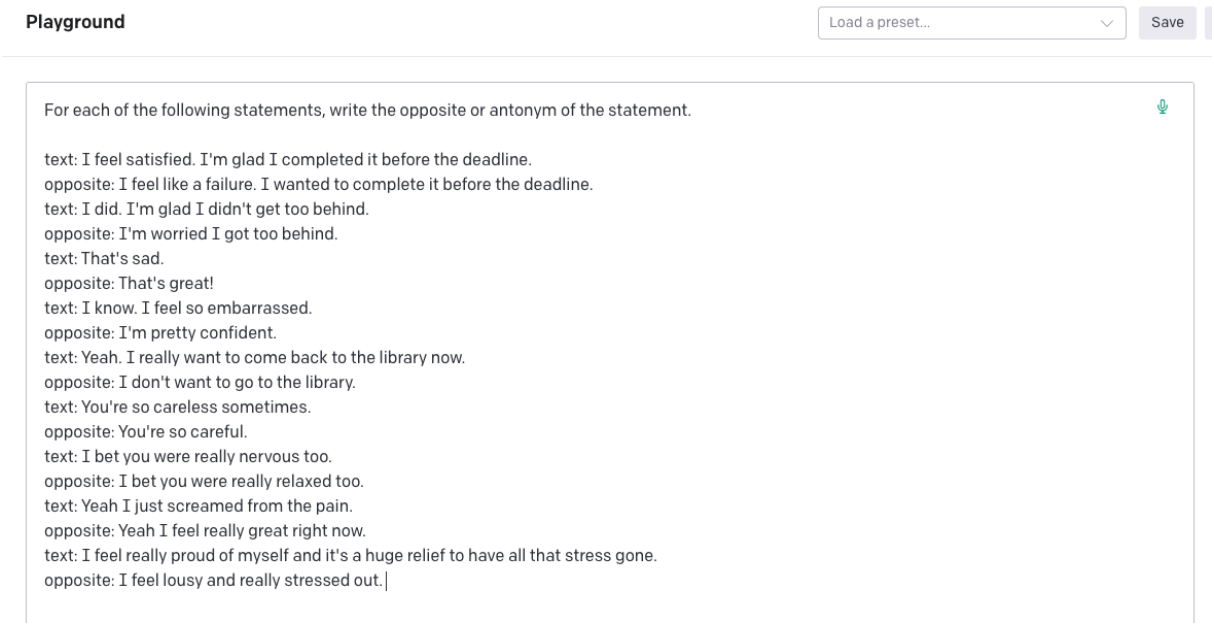


Figure 2: GPT-3 Prompt used for creating invalid dialogue responses from valid responses.

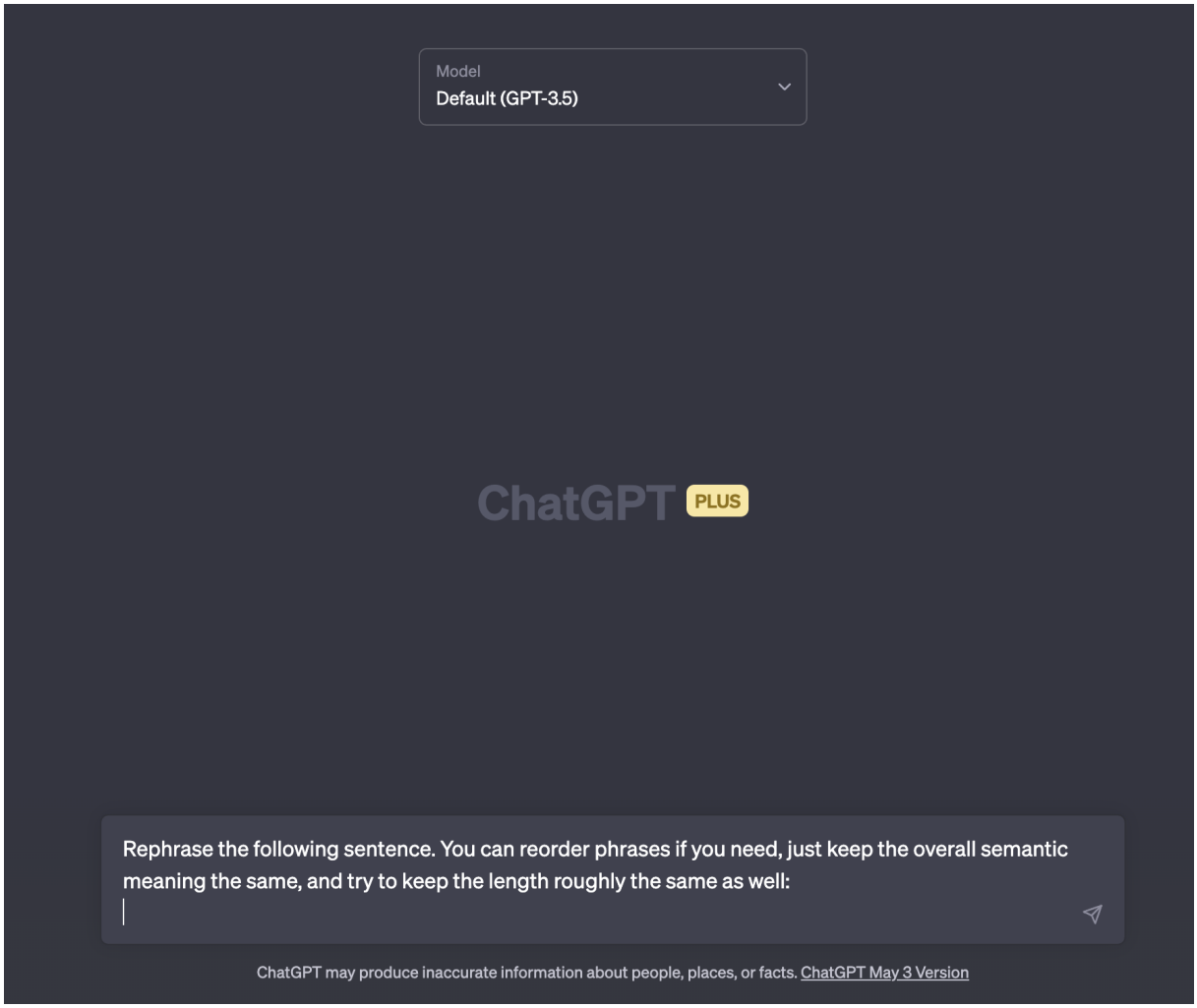


Figure 3: ChatGPT prompt used for rephrasing invalid dialogue responses.

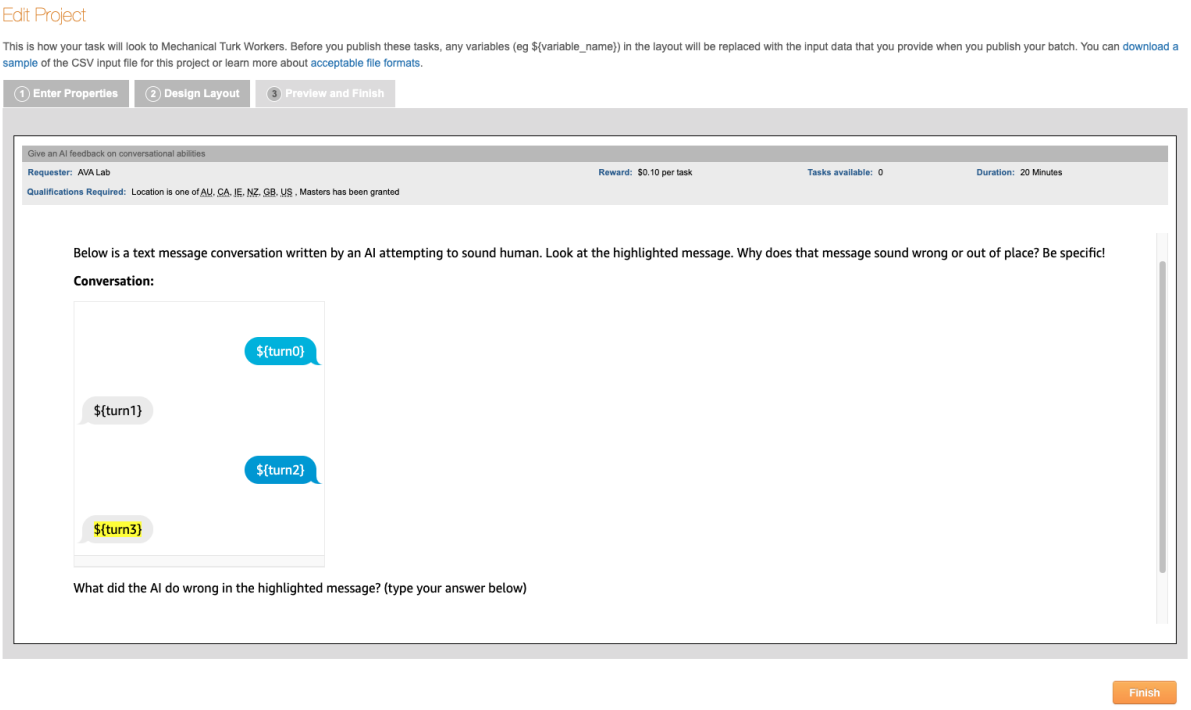


Figure 4: Mechanical Turk interface used for acquiring feedback for dialogue responses. Each dialogue was given feedback by two independent crowdworkers.

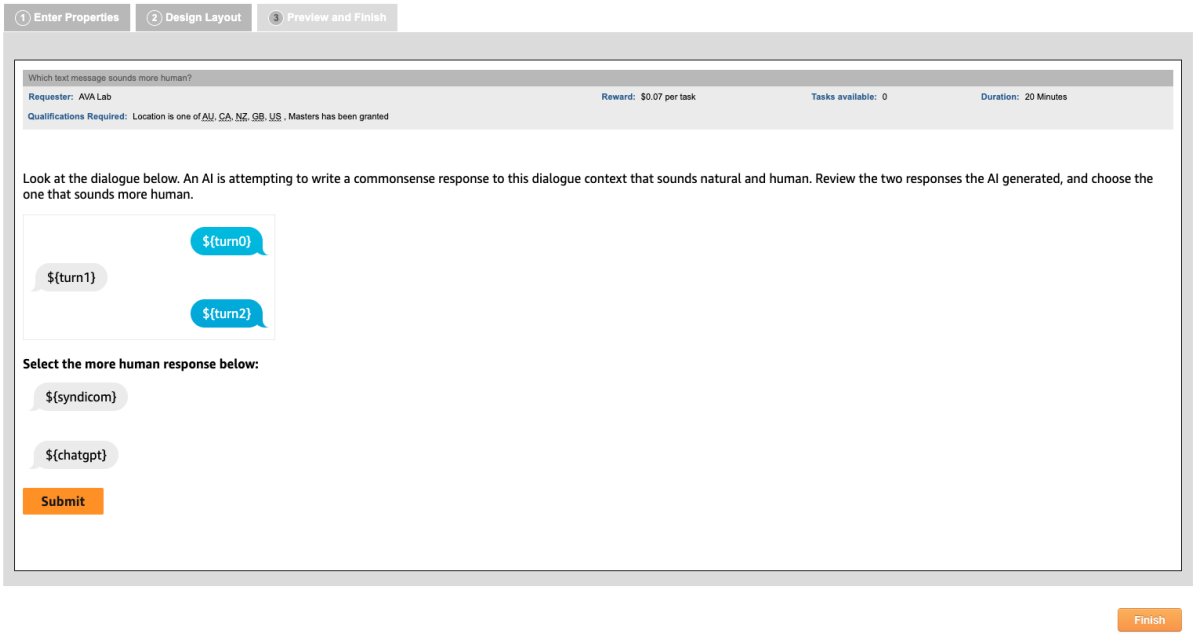


Figure 5: Mechanical Turk interface used for human evaluation. Each dialogue response pair was evaluated by two workers independently. Templates are shown instead of examples in order to fit the page.

"What do others think?": Task-Oriented Conversational Modeling with Subjective Knowledge

Chao Zhao [1] **Spandana Gella** [2] **Seokhwan Kim** [2] **Di Jin** [2]
Devamanyu Hazarika [2] **Alexandros Papangelis** [2] **Behnam Hedayatnia** [2]
Mahdi Namazifar [2] **Yang Liu** [2] **Dilek Hakkani-Tur** [2]
zhaochao@cs.unc.edu {sgella,seokhwk,djinamzn}@amazon.com
{dvhaz,papangea,behnam,mahdinam,yangliud,hakkanit}@amazon.com
[1] UNC Chapel Hill [2] Amazon, Alexa

Abstract

Task-oriented Dialogue (TOD) Systems aim to build dialogue systems that assist users in accomplishing specific goals, such as booking a hotel or a restaurant. Traditional TODs rely on domain-specific APIs/DBs or external factual knowledge to generate responses, which cannot accommodate subjective user requests (e.g.,"*Is the WIFI reliable?*" or "*Does the restaurant have a good atmosphere?*"). To address this issue, we propose a novel task of subjective-knowledge-based TOD (SK-TOD). We also propose the first corresponding dataset, which contains subjective knowledge-seeking dialogue contexts and manually annotated responses grounded in subjective knowledge sources. When evaluated with existing TOD approaches, we find that this task poses new challenges such as aggregating diverse opinions from multiple knowledge snippets. We hope this task and dataset can promote further research on TOD and subjective content understanding. The code and the dataset are available at https://github.com/alexa/dstc11-track5.

1 Introduction

Task-oriented Dialogue (TOD) Systems aim to build dialogue systems that assist users in accomplishing specific goals, such as booking a hotel or a restaurant. Most solutions of TOD are based on domain-APIs (Budzianowski et al., 2018; Rastogi et al., 2020) and structured databases (Eric et al., 2017; Wu et al., 2019), which can only handle a limited range of scenarios within the scope of APIs/DBs. To further enlarge the model's ability of task-oriented assistance, recent works (Dimitrakis et al., 2018; Kim et al., 2020, 2021; Feng et al., 2020, 2021; Majumder et al., 2022) incorporate unstructured textual information retrieved from the Internet into dialogue modeling. Most of these works focus on factual knowledge sources such as frequently asked questions (FAQs) of online products or government service guides. We refer to these models as Fact-TOD models.

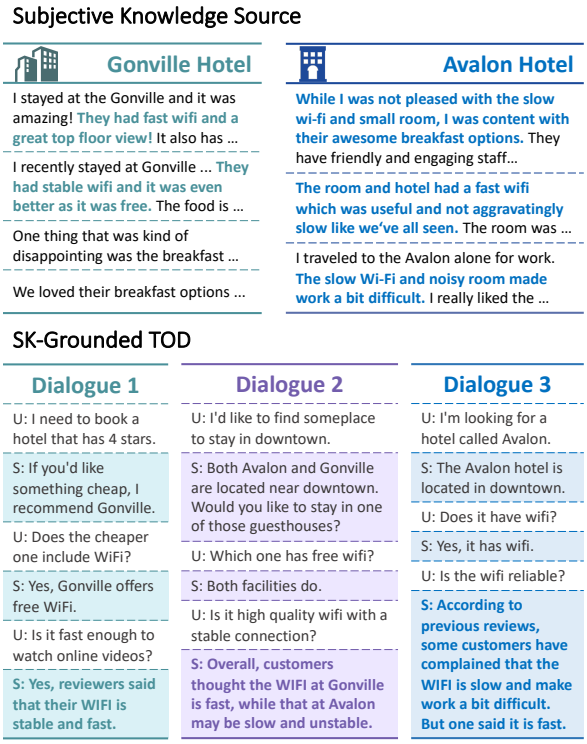


Figure 1: Examples of the SK-TOD task. The top part shows two hotels and their customer reviews. The bottom part shows three dialogue sessions between the system (denoted by S) and three users (denoted by U). The last user utterance is a subjective question about the WIFI quality of the hotel(s). The system needs to retrieve information from the relevant subjective knowledge, which is highlighted in the review text.

However, in many TOD tasks, users care about not only factual information but subjective insights as well, such as the experiences, opinions, and preferences of other customers. For instance, when booking a hotel or a restaurant, users often inquire about subject aspects like *"Is the WIFI reliable?"* or *"Does the restaurant have a good atmosphere?"*. To respond to such user requests, an agent needs to seek information from subjective knowledge

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 309–323
September 11–15, 2023. ©2023 Association for Computational Linguistics

sources, such as online customer reviews. While subjective knowledge has been specifically studied in other NLP problems such as opinion mining (Liu and Zhang, 2012) and question answering (Bjerva et al., 2020), incorporating it into TOD has not received significant attention.

In this work, we argue that it is important to enable the TOD model to leverage subjective knowledge for more effective task-oriented assistance. To this end, we propose a novel task of subjective-knowledge-based task-oriented dialogue (SK-TOD). SK-TOD focuses on responding to user requests that seek subjective information by incorporating user reviews as subjective knowledge. Figure 1 shows three examples of such requests, where customers ask about the WiFi quality of various hotels. User reviews are valuable resources for subjective information because even for the same aspect of a product or service, customers may have different opinions and leave either positive or negative reviews. As a result, a TOD system should consider multiple reviews to provide a comprehensive representation of user opinions. Ideally, the system's response should include both positive and negative opinions, along with their respective proportions (as exemplified in Dialogue 3). This two-sided response has been recognized as more credible and valuable for customers (Kamins et al., 1989; Lee et al., 2008; Baek et al., 2012), thereby fostering trust in the TOD system.

Incorporating subjective knowledge into TOD introduces two unique challenges. Firstly, unlike in Fact-TOD where selecting a few relevant knowledge snippets suffices, the SK-TOD model must consider all relevant knowledge snippets. In other words, both precision and recall matter during this process. Secondly, the model needs to aggregate these knowledge snippets into a concise response that can faithfully reflect the diversity and proportion of opinions expressed. Conquering these challenges requires a large-scale dataset with subjective-knowledge-grounded responses, which, to our best knowledge, is not publicly available.

To facilitate the research in subjective-knowledge-grounded TOD, we have collected a large-scale dataset, which contains 19,696 subjective knowledge-seeking dialogue contexts and manually annotated responses that are grounded on 143 entities and 1,430 reviews (8,013 sentences). We evaluate the performance of strong baselines on the SK-TOD task. Results show that there is a significant gap between human-generated and machine-generated responses, particularly in terms of the faithfulness of the sentiment proportion. To address this issue, we propose a model that incorporates review understanding into SK-TOD. We experimentally demonstrate that responses generated by this model more effectively capture the sentiment proportion. Our contributions are three-fold:

- We introduce a novel task of subjective-knowledge-based TOD (SK-TOD);
- We create and release a large-scale, human-annotated dataset designed for this task;
- We propose a new model and conduct extensive experiments on the proposed task.

2 Related Work

2.1 Knowledge-Grounded Dialogue

Knowledge-grounded response generation is popular in the open-domain dialogue. Numerous external knowledge sources have been explored, from structured knowledge such as fact tables (Moghe et al., 2018; Liu et al., 2018) and knowledge graphs (Zhang et al., 2020a; Moon et al., 2019; Tuan et al., 2019), to unstructured knowledge such as Wikipedia articles (Vougiouklis et al., 2016; Zhou et al., 2018; Dinan et al., 2018), news articles (Majumder et al., 2020), web pages (Long et al., 2017; Galley et al., 2019; Komeili et al., 2022), narratives (Xu et al., 2021; Gopalakrishnan et al., 2019), user reviews and comments (Moghe et al., 2018; Ghazvininejad et al., 2018), and so on. Grounding on external knowledge makes the response more informative and meaningful when compared with models that solely rely on the dialog context.

Regarding task-oriented dialogues, previous works have primarily focused on domain-specific APIs and databases to support the dialogue response (Levin et al., 2000; Singh et al., 2002; Williams and Young, 2007; Eric et al., 2017; Wu et al., 2019), which can only support a limited scope of user queries. Later works ground task-oriented dialogues to web pages (Penha et al., 2019; Chen et al., 2022), government service documents (Saeidi et al., 2018; Feng et al., 2020, 2021), and FAQ knowledge snippets (Kim et al., 2020, 2021). Different from these works where factual knowledge is utilized, we apply subjective knowledge to generate the response and ground in multiple

knowledge snippets. While Majumder et al. (2022) also explored grounding TOD in user reviews, they did not consider the diversity of opinions.

2.2 Subjective Content Understanding

Besides being used as external knowledge sources in dialogue systems, subjective content, especially user reviews, has been studied in various non-conversational NLP tasks. For example, opinion mining (Pontiki et al., 2016; Jiang et al., 2019) focuses on extracting opinions and sentiments from user reviews. Opinion summarization (Chu and Liu, 2019; Zhao and Chaturvedi, 2020; Bražinskas et al., 2020; Angelidis et al., 2021) is used to distill multiple opinions into concise summaries. Subjective question answering (McAuley and Yang, 2016; Bjerva et al., 2020) have been proposed to answer questions based on user reviews. Explainable recommendation (Ni et al., 2019) aims to generate review-based explanations for the items recommended by a recommendation system. Table 1 provides detailed comparisons between SK-TOD and these subjective-content-based benchmarks. Generally, SK-TOD requires creating a response that is appropriate to the dialogue context. It also requires grounding in multiple subjective knowledge and explicitly considers the diversity of opinions and the proportion of sentiments.

3 Problem Formulation

Formally, we have a dialogue context $C = [U_1, S_1, U_2, S_2, \cdots, U_t]$ between a user and a system, where each user utterance U_i is followed by a system response utterance S_i, except for the last user utterance U_t. The dialogue involves one or more entities, denoted as $\mathcal{E} = \{e_1, \cdots, e_m\}$. Alongside the dialogue, we have a subjective knowledge source $\mathcal{B} = \{(e_1, \mathcal{R}_1), (e_2, \mathcal{R}_2), \cdots\}$ containing all the entities and their corresponding customer reviews. Each entity e is associated with multiple reviews $\mathcal{R} = \{R_1, R_2, \cdots\}$. Each review can be divided into segments $[K_1, K_2, \cdots]$, such as paragraphs, sentences, or sub-sentential units. In this work, we regard each review sentence as a knowledge snippet.

The SK-TOD task aims to identify whether U_t is a subjective knowledge-seeking request and, if it is, to select the relevant knowledge snippets $\mathcal{K}^+$ from the knowledge source and finally generate a response S_t grounded on $\mathcal{K}^+$.

4 Data Collection and Statistics

We ground the data collection in MultiWOZ (Budzianowski et al., 2018; Eric et al., 2020). We select dialogues from the domains of hotels and restaurants. The data collection is conducted by a group of crowd workers through Amazon Mechanical Turk (AMT). To control the data quality, we only choose workers that are pre-qualified. More details can be found in Appendix A.

4.1 Annotation Guideline

Dialogues in MultiWOZ are collected based on single or multiple entities as the back-end database. To create a subjective knowledge source to support the SK-TOD task, we first collect multiple user reviews for each entity. To control the review collection, we provide the reviewer's persona, as well as the aspects and sentiments of reviews to workers. We then ask workers to write a review with all the given information included. After collecting the reviews, we also annotate the aspect and sentiment information for each review sentence. Overall, we select 33 hotels and 110 restaurants from MultiWOZ, and collect 10 reviews for each entity. On average, each review contains 5.6 sentences and 56.71 tokens. More details about the review collection can be found in Appendix A.

After obtaining the reviews, we go back to the dialogue data to create the subjective user request. Following a similar procedure in Kim et al. (2020), for each dialogue, we provide an aspect that users are interested in (e.g., WIFI-quality of the hotel) and then ask the worker to insert a subjective user request into the dialogue. Workers are requested to carefully select the insertion position and write an utterance to maintain coherence and naturalness in the dialogue flow. Finally, we use the partial dialog until this newly inserted turn as an instance in our data. Utterances that come after the insertion position are removed from the dialogue instance.

So far, we've collected the dialogue context C and the subjective knowledge source $\mathcal{B}$. The final step is to ground the dialogue in the knowledge source. We first ask workers to identify entities that are relevant to the subjective user request as gold entities. We then align the user request and review sentences of the gold entities by matching their aspect. For example, if the aspect of a user request is about the "WIFI quality" of a hotel, all review sentences discussing the "WIFI quality" of that specific hotel will be considered relevant knowledge

	Size	Manual	Dial	TOD	Query	Aspect	Senti	Mul-Knwl	Senti-%
Semeval/MAMS (2016; 2019)	5K/22K	✓	✗	n/a	✗	✓	✓	✗	n/a
Space (2021)	1K	✓	✗	n/a	✗	✓	✓	✓	✗
Yelp/Amazon (2019; 2020)	200/180	✓	✗	n/a	✗	✗	✓	✓	✗
Justify-Rec (2019)	1.3M	✗	✗	n/a	✗	✓	✗	✓	✗
AmazonQA (2016)	309K	✗	✗	n/a	✓	✗	✗	✗	n/a
SubjQA (2020)	10K	✗	✗	n/a	✓	✓	✓	✗	n/a
Holl-E (2018)	9K	✓	✓	✗	✗	✗	✗	✓	✗
Foursquare (2018)	1M	✗	✓	✗	✗	✗	✗	✓	n/a
SK-TOD (Ours)	20K	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison between SK-TOD and other benchmarks based on the subjective content. We consider if the dataset is manually annotated, dialogue-based, task-oriented, and query-focused. We also list if it considers aspect and sentiment, multiple knowledge snippets (Mul-Knwl), and the proportion of two-sided sentiments (Senti-%).

snippets. [1] Finally, we provide the dialogue context C and all related knowledge snippets $\mathcal{K}^+$ and ask workers to generate a natural and faithful response. We explicitly instruct workers to consider the diversity and proportion of opinions in all relevant knowledge snippets during response creation. Detailed instructions can be found in Appendix A.

4.2 Quality Control

To ensure the quality of our dataset, we took great care in selecting pre-qualified workers and designing annotation interfaces. We further conducted a human verification task on the entire dataset to identify invalid instances. The annotation showed that 81.89% of subjective-knowledge-seeking user turns are valid, with an Inter-Annotator Agreement (IAA) score of 0.9369 in Gwet's gamma. For agent response turns, 96.78% were valid, with an IAA score of 0.9497 in Gwet's gamma. Any invalid instances were filtered out or manually corrected before finalizing the dataset. We paid workers an average of $13.82/hr for data annotation and $14.77/hr for data verification. Both exceed the local living minimum wage. The details of our payment settings are elaborated on in Appendix A.

4.3 Data Statistics

We collected a total of 19,696 instances consisting of subjective user requests and subjective-knowledge-grounded responses. The average length of the subjective user request and the agent response is 8.75 and 24.07 tokens, respectively. While most of the instances contain a single entity, there are 1,047 instances where multiple entities are compared (like Dialogue 2 in Figure 1). On average, each instance requires 3.88 subjective knowledge snippets. To help identify the subjective knowledge-seeking user request, we also randomly sample another 18,383 dialogues with non-subjective user requests from the original MultiWOZ dataset.

[1]Note that the aspect information is only used to build the dataset but is not included in the problem formulation of SK-TOD, which means it is not available for model training. The goal of SK-TOD is to handle user requests with arbitrary aspects, and therefore we do not define a taxonomy of aspects in the task like what is done in dialogue state tracking.

	Train	**Val**	**Test**
# instances	14768	2129	2799
# seen instances	14768	1471	1547
# unseen instances	0	658	1252
# multi-entity instances	412	199	436
Knowledge Snippets			
Avg. # snippets per instance	3.80	4.07	4.21
Avg. # tokens per snippet	14.68	15.49	14.5
Dialogue			
Avg. # uttrances per instance	9.29	9.44	9.36
Avg. # tokens per request	8.65	8.94	9.12
Avg. # tokens per response	24.18	23.61	23.86

Table 2: Basic statistics of our dataset.

We split the dataset into training (75%), validation (10.8%), and test (14.2%) sets. Table 2 presents the detailed statistics of each subset. Both the validation and test sets contain two subsets: the *seen* subset where the aspects of these instances are included in the training set, and the *unseen* subset where the aspects are not included in the training set. The unseen subset is designed to evaluate models' ability to generalize to arbitrary aspects.

5 Subjective-Knowledge-Grounded TOD

In this section, we describe the method for SK-TOD. As shown in Figure 2, we follow the pipeline introduced by Kim et al. (2020) which comprises four sequential sub-tasks: knowledge-seeking turn detection (KTD), entity tracking (ET), knowledge selection (KS), and response generation (RG). We elaborate on each subtask below.

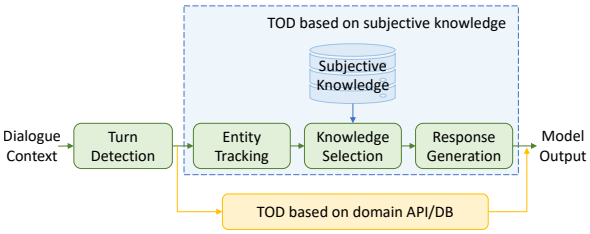


Figure 2: The pipeline architecture of SK-TOD.

5.1 Knowledge-Seeking Turn Detection

The goal of KTD is to identify the user request that requires subjective knowledge. We regard it as a binary classification problem, where the input is the dialogue context C and the output is a binary indicator.

We employ a pre-trained language model (e.g., BERT (Devlin et al., 2019)) to encode C and adopt the hidden state of the first token as its representation. Then we apply a classifier to obtain the probability that the current user request is seeking subjective knowledge. That is,

$$\begin{aligned} h &= \text{Enc}(C) \\ P(C) &= \text{softmax}\left(\text{FFN}\left(h\right)\right). \end{aligned} \tag{1}$$

The model is finetuned with the binary cross-entropy loss.

5.2 Entity Tracking

The goal of ET is to identify the entities $\mathcal{E} = \{e_1, \cdots, e_m\}$ that are relevant to the user request. It can help to reduce the number of candidates during the knowledge selection step.

We adopt a word-matching-based method used by Jin et al. (2021) to extract relevant entities. It first normalizes entity names in the knowledge source using a set of heuristic rules. Then a fuzzy n-gram matching is performed between the normalized entity and all dialogue turns. To find the entities that are relevant to the last user request, we choose the last dialogue turn in which the entities are detected and use these entities as the output of ET. We leave the tracking of aspects being questioned over multiple turns as future work.

5.3 Knowledge Selection

The goal of KS is to select the knowledge snippets that are relevant to the user's request. The inputs are the dialogue context C and a set of knowledge snippets candidates $\mathcal{K}$, which is a combination of all knowledge snippets of the relevant entities in $\mathcal{E}$. The output $\mathcal{K}^+ \subseteq \mathcal{K}$ is a subset of relevant knowledge candidates. Note that there might be multiple knowledge snippets in $\mathcal{K}^+$.

To select relevant knowledge snippets, we calculate the relevance score between the dialogue context C and a knowledge snippet $K \in \mathcal{K}$. We regard it as a pairwise text scoring problem and consider two popular approaches: bi-encoder (Mazaré et al., 2018) and cross-encoder (Wolf et al., 2019). Generally, the bi-encoder approach is more efficient while the cross-encoder approach is more accurate.

For the bi-encoder approach, we encode C and K separately using the same pre-trained encoder and obtain two representations, h_C and h_K. Following Reimers and Gurevych (2019), we use the concatenation of h_C, h_K, and $|h_C - h_K|$ as features and apply a classifier to obtain the probability of relevance. That is,

$$\begin{aligned} h_C = \text{Enc}(C), \quad h_K = \text{Enc}(K) \\ P(C, K) = \text{softmax}\left(\text{FFN}\left(h_c, h_K, |h_C - h_K|\right)\right). \end{aligned} \tag{2}$$

For the cross-encoder approach, we encode the concatenation of C and K to obtain a contextualized representation. That is,

$$\begin{aligned} h &= \text{Enc}(C, K) \\ P(C, K) &= \text{softmax}\left(\text{FFN}\left(h\right)\right). \end{aligned} \tag{3}$$

During training, we use all relevant knowledge snippets to construct positive (C, K) pairs. Due to the large number of irrelevant knowledge snippets, we randomly sample the same number of irrelevant snippets to form negative pairs. We optimize the model using the binary cross-entropy loss. During inference, we predict the relevance probability for all knowledge snippets in the candidates. Since both precision and recall are crucial in KS, instead of selecting the top few results, we use a threshold, estimated from the validation set, to determine the relevancy of each knowledge snippet.

5.4 Response Generation

The goal of RG is to create an utterance S_t that addresses the user's request. This response is generated based on the dialogue context C and the set of relevant knowledge snippets $\mathcal{K}^+$. To accomplish this, we concatenate $\mathcal{K}^+$ and C as the input and use a pre-trained generation model to generate the response. We consider both the decoder-only model, such as GPT-2 (Radford et al.), and the encoder-decoder model, such as BART (Lewis et al., 2020).

The model is trained to maximize the generation probability $p(S_T \mid C, \mathcal{K}^+)$.

To accurately capture the diversity and proportion of opinions, the model needs to understand the sentiment polarity of each knowledge snippet, which is challenging due to the lack of direct supervision. To address this issue, we apply a state-of-the-art aspect-based sentiment analysis (ABSA) model (Zhang et al., 2021) to predict the sentiment $Z = [z_1, \cdots, z_i, \cdots]$ for each knowledge snippet $K_i \in \mathcal{K}^+$. Then we incorporate the sentiment information into RG by maximizing $p(S_T \mid C, \mathcal{K}^+, Z)$.

More specifically, we first convert the predicted z_i into a natural language description using templates, and then append it to the end of the corresponding K_i as the enhanced input of RG. For example, given the knowledge snippet as "*The ambience was so fun.*", the ABSA model detects the aspect-based sentiment as ("ambience", "positive"). We first convert the sentiment into a natural language "*ambience is great.*" and then enhance the knowledge snippet as "*The ambience was so fun. ambience is great.*". We refer to Appendix B for more details.

6 Experiments on Sub-Tasks

We first conduct experiments on each individual subtask. To avoid any error accumulation from upstream tasks, we use the gold output of the previous task as the input to the current target task. The detailed experimental setup can be found in Appendix C.

6.1 Knowledge-Seeking Turn Detection

Setting We conduct experiments using various pre-trained language models, including BERT [2] (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and DeBERTa (He et al., 2021).

Evaluation We report the precision, recall, F_1 score, and accuracy score.

Results Table 3 shows the results of the KTD task. All models achieve similar and near-perfect performance, which is in line with the findings of Kim et al. (2020). It demonstrates that it is feasible to identify the user requests that require subjective knowledge, allowing them to be explicitly addressed by an SK-TOD component. However, this KTD classifier's performance may be specific to this dataset or similar domains, and its generalizability to unseen domains or knowledge types requires further exploration in future works.

[2]We use the base version of all pre-trained models.

	Acc	P	R	F
BERT	99.67	99.75	99.61	99.68
RoBERTa	99.74	99.86	99.64	99.75
ALBERT	99.49	99.64	99.36	99.50
DeBERTa	99.71	99.86	99.57	99.71

Table 3: Results of KTD task. Models are evaluated using Accuracy, Precision, Recall, and F_1. All models achieve similar and near-perfect performance.

6.2 Entity Tracking

Setting We follow the setting of Jin et al. (2021) to run the ET method.

Evaluation We report the instance-level accuracy score. An instance is regarded as accurate only if the predicted entities match exactly with the gold entities.

Results The fuzzy n-gram matching method achieves an instance-level accuracy of 92.18%. We further analyzed the type of errors. For 1.8% of the instances, there is at least one gold entity missing from the predicted entities. For 7.6% of the instances, the predicted entities contain at least one spurious entity. The latter error case can be further reduced by using model-based matching approaches, which we leave as future work.

6.3 Knowledge Selection

Setting We fine-tune the KS models following the same setting as in the KTD task. Additionally, we compare them with traditional information retrieval (IR) baselines, such as TF-IDF (Manning et al., 2008) and BM25 (Robertson et al., 2009).

Evaluation Knowledge selection can be viewed as either a classification task or a retrieval task. For classification, we use precision, recall, and F_1 measures. We calculate these measures at both the instance level and the snippet level. For the instance level, we first calculate $P/R/F_1$ for each instance, and then take the average over all instances as the final scores. For the snippet level, instead of computing $P/R/F_1$ for each instance, we calculate these scores for all <C, K> pairs in the entire dataset. Regarding retrieval evaluation, we use mean-average-precision (mAP) as the metric, which is not dependent on a specific threshold value

	Instance-level			Snippet-level			mAP
	P	R	F	P	R	F	
IR Baselines							
TF-IDF	34.61	70.33	40.46	23.81	65.00	34.85	45.97
BM25	31.38	40.95	32.21	31.14	32.42	31.77	45.42
Bi-encoder							
BERT	56.66	70.06	59.31	58.87	74.69	65.84	71.59
RoBERTa	60.98	83.06	66.47	54.40	**85.38**	66.46	77.25
ALBERT	70.21	78.74	70.43	**63.13**	78.90	70.14	81.62
DeBERTa	**71.46**	**83.18**	**72.44**	62.64	83.50	**71.58**	**83.43**
Cross-encoder							
BERT	85.18	86.01	83.33	82.40	83.82	83.11	90.06
RoBERTa	81.59	83.62	80.53	82.20	80.77	81.48	88.98
ALBERT	**86.18**	87.29	84.22	**83.56**	84.78	84.16	90.50
DeBERTa	86.07	**87.64**	**84.6**	82.70	85.71	**84.18**	**91.84**
SEEN	88.80	93.45	89.93	90.83	89.90	90.37	95.70
UNSEEN	82.68	80.47	78.03	69.98	78.29	73.90	87.07

Table 4: Results of the KS task. Models are evaluated using instance-level and snippet-level classification measures, as well as mAP, a retrieval-based measure. DeBERTa achieves the best performance among all evaluation measures.

and can reflect the overall ranking positions of all relevant knowledge snippets. Since the total number of the relevant knowledge snippets can vary for each instance, we do not include top-K-based measures like Precision@K or Recall@K, which are commonly used in other Fact-TOD and knowledge-grounded open-domain dialogue tasks.

Results Table 4 shows the results of the KS task. Firstly, when comparing our models with IR baselines, all of the trained models outperform the baselines, indicating that the KS model can benefit from the annotated training data. We then compare bi-encoder models and cross-encoder models, and as expected, cross-encoder models outperform bi-encoder models by a large margin. When comparing the performance of different pre-trained models, there is a notable difference among the models under the bi-encoder setting. The variance becomes smaller when applying the cross-encoder architecture. DeBERTa achieves the best performance on all measures in both the bi-encoder and cross-encoder settings.

Finally, we compare the performance between the seen subset and the unseen subset. At the bottom of Table 4, we list the performance of DeBERTa on both the seen and unseen test subsets. The results reveal a large gap between the performance of the two subsets, indicating that one of the challenges for the KS model is to generalize from seen aspects to unseen aspects.

	BLEU	R-1	R-2	R-L	MT	BS	Len
EXT	2.89	23.17	6.53	18.33	9.62	30.83	14.93
GPT2	9.04	33.9	13.52	26.73	16.27	39.73	22.66
DialoGPT	9.19	33.6	13.62	26.81	16.15	39.72	22.05
BART	10.8	36.35	15.04	28.57	17.96	41.12	24.02
$\text{BART}_{\text{ABSA}}$	10.78	36.30	15.36	28.47	18.06	**41.75**	23.66
T5	10.72	36.50	**15.57**	28.81	**18.33**	40.84	25.36
T5_{ABSA}	**10.97**	**36.66**	15.51	**28.88**	18.15	40.94	24.75

Table 5: Results of RG task. Models are evaluated using BLEU, ROUGE (R-1, R-2, R-L), METEOR (MT), and BertScore (BS). We also listed the average length (Len) of the generated response. Encoder-decoder models such as BART and T5 achieve better performance compared with GPT2-based models.

6.4 Response Generation

Setting we experiment with decoder-only generation models such as GPT-2 (Radford et al.) [3] and DialoGPT (Zhang et al., 2020c), as well as encoder-decoder models such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). We also include two ABSA-enhanced models, namely $\text{BART}_{\text{ABSA}}$ and T5_{ABSA}. During decoding, we use beam-search with top-K sampling (Fan et al., 2018). We set the beam size as 5 and sample from the top 50 tokens. We also compare with a random extractive baseline (EXT), where the response is created by randomly selecting a relevant knowledge snippet.

Evaluation Following the evaluation of other generation tasks, We employ several automatic evaluation metrics, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), as well as BERTScore (Zhang et al., 2020b), to evaluate the quality of the generated responses compared to the reference responses. We also conduct a human evaluation, where we ask crowd workers to evaluate the quality of responses.

Results As presented in Table 5, machine-generated responses significantly outperform the extractive responses. Encoder-decoder models achieve better performance across all automatic measures compared to GPT-based models, indicating that they are more suitable for this task. They

[3] We use the base-version of all pre-trained models.

also tend to generate longer responses. There is no clear difference in automatic measures when comparing BART and T5. For ABSA-enhanced models, $\text{BART}_{\text{ABSA}}$ achieves the best performance on BertScore, while T5_{ABSA} achieves the best score on BLEU and ROUGE.

Human Evaluation To obtain a more reliable assessment of response quality, we also conduct a human evaluation on AMT. We use the same group of workers involved in the data collection process. During the evaluation, we show the dialogue context, the oracle knowledge snippets, and all responses (both the reference and the generated responses) to the workers. We randomly sample 240 instances from the test set for evaluation. For each instance, we ask three independent workers to compare the responses based on three measures:

- Appropriateness: whether the response is fluent and naturally connected to the dialogue context.
- Aspect Accuracy: whether the response provides relevant and useful information to the aspect that the user queried.
- Sentiment Accuracy: whether the sentiment proportion provided by the response is consistent with that of the subjective knowledge.

For sentiment accuracy, we first ask workers to annotate the sentiment label of each knowledge snippet, and then evaluate each response. All three measures are evaluated using a 5-Point Likert scale. The system-level score is computed as the average score over all instances and workers for each system. The compensation for workers was set at \$0.25 for the tasks of appropriateness and aspect accuracy, and \$0.4 for the task of sentiment accuracy. The average hourly pay for the crowd workers was \$15.25/hr, \$14.40/hr, and \$14.85/hr for each evaluation task, exceeding the local living minimum wage.

Table 6 shows the results of human evaluation for response generation. The inter-annotator agreement scores for each task are 0.7270, 0.7535, and 0.6239 in Gwet's gamma, respectively. The results show that machine-generated responses are comparable to the references in terms of appropriateness and aspect accuracy. Moreover, incorporating ABSA can improve the model's performance in sentiment accuracy. However, there is still a large gap in sentiment accuracy between the best model-generated responses and the references, indicating that faithfully aggregating sentiment information from multiple knowledge snippets is still a challenging task for current models.

	Approp.	**Asp-Acc**	**Senti-Acc**
EXT	2.65	3.32	3.13
GPT2	4.55	4.54	3.20
BART	4.55	4.68	3.56
$\text{BART}_{\text{ABSA}}$	4.58	4.66	3.80
T5	4.40	4.63	3.87
T5_{ABSA}	4.49	4.67	3.98
Reference	4.70	4.77	4.50

Table 6: Results of human evaluation for RG.

Dialogue History:
U: Are there any hotels (not guesthouses) that have 0 stars?
S: Yes, actually. Cityroomz is a moderately priced 0 star hotel.
U: Do they have free wifi?
S: Yes, they do have free wifi! Would you like me to book a stay for you?
U: Does the Cityroomz have strong water pressure in the shower?

Reviews:
Review 1: However, I noticed that when taking a shower there was very little water pressure which made the shower disappointing.
Review 2: The water pressure is not good and it comes out really slow.
Review 3: Everything else was a disaster, especially the lack of cleanliness in the room when we checked in, and the incredibly low water pressure in the bathroom when we took our morning showers.
Review 4: I encounter a problem when trying to shower and washing hands.
Review 5: There is no luggage storage service or anything like that, and the water pressure is much to be desired.

Model Outputs:
GPT-2: The Cityroomz gets some mixed reviews on their water pressure. Some guests found the water pressure to be very good while others found it to be disappointing. Would you like to know more about them?
BART: The Cityroomz hotel has very low water pressure according to most guests. Only one person had a good experience with it.
$\text{BART}_{\text{ABSA}}$: The Cityroomz has a lot of complaints about their water pressure being a bit low. Would you like me to find another place with better ratings?
T5: The Cityroomz has mixed reviews about their water pressure. Some guests found the water pressure to be very low while others found it to be disappointing. Would you like to know more about them?
T5_{ABSA}: The Cityroomz has a lot of reviews that say the water pressure is very low and disappointing. Do you want to look at some other places?

Reference:
No, guests consistently complain about the water pressure, unfortunately. Will that be okay or should I do another search?

Table 7: Sampled output of different RG models.

Qualitative Analysis Table 7 shows an example of responses generated by various systems. In this example, all the reviews express negative opinions about water pressure. However, responses generated by GPT-2 and BART include positive opinions. T5 correctly mentions the negative opinions but the generated response is not natural and coherent. By incorporating the ABSA model, both BART and T5 correctly generate responses with all negative opinions.

7 Experiments on End-2-End Evaluation

In the previous section, we use gold information as input for each module to avoid error accumulation.

	KS		RG		
	Macro-F	**mAP**	**BLEU**	**R-L**	**BS**
RG	-	-	10.80	28.52	41.12
+KS	84.60	91.84	10.20	27.78	40.64
+ET+KS	83.47	90.45	10.29	27.80	40.56
+KTD+ET+KS	83.46	90.45	10.27	27.79	40.55

Table 8: Results of the end-to-end evaluation. We start from RG with gold knowledge as input. We then gradually add components (KS, ET, and KTD) to the pipeline to replace the gold input with the predicted one.

	KTD	KS		RG		
	Acc	**Macro-F**	**mAP**	**BLEU**	**R-L**	**BS**
Fact-TOD	87.62	59.55	76.69	6.15	23.25	33.16
SK-TOD	99.71	84.60	91.84	10.80	28.57	41.12

Table 9: Comparison between models trained on Fact-TOD and SK-TOD training data.

Dialogue History:
U: I am looking for a guesthouse to stay at that has free wifi.
S: Ok. Were you hoping for a particular price range or part of town?
U: I don't care about the part of town or the price range, but it needs to be a three star hotel.
S: The Hamilton Lodge would be perfect! It is a moderately priced 3 star guesthouse on the north side of town. Does that sound okay to you?
U: Does it have comfortable beds?

Reviews:
Review 1: The bed really needed a new mattress, too.
Review 2: That being said, the room was spacious and the bed was very comfortable.
Review 3: We both slept extremely well every night we were there.

Model Outputs:
Fact-TOD: Yes, the Hamilton Lodge has a new mattress. Would you like to make a reservation?
SK-TOD: The Hamilton Lodge has really comfortable beds according to most guests, but one guest did say that the bed needed a new mattress.

Reference:
There are some mixed reviews on the beds. Some say they're very comfortable while others were not impressed. Would you like me to find another place with better reviews?

Table 10: Sampled outputs from the Fact-TOD model and the SK-TOD model, respectively.

In this section, we evaluate the entire pipeline in an end-to-end manner, where the input of each subtask is predicted by the previous component. We gradually add KS, ET, and KTD to the pipeline, and list the performance of KS and RG in Table 8.

The results show that errors introduced during KS can decrease the quality of response generation. However, ET and KTD do not have a significant impact on the performance of downstream tasks. It is because ET and KTD results include fewer noisy predictions compared to the KS results.

8 Comparison with Fact-TOD

One difference between SK-TOD and Fact-TOD is that responses in SK-TOD are grounded on subjective knowledge instead of factual knowledge. In this section, we investigate whether a Fact-TOD model can ground on subjective knowledge to address subjective requests. To this end, we retrain our KTD (DeBERTa), KS (DeBERTa cross-encoder), and RG (BART) models using the FAQ-grounded TOD data provided by Kim et al. (2020). We then apply it to the test set of SK-TOD without further training. We compare the results of each sub-task using the Fact-TOD models with the results of models trained on SK-TOD.

As shown in Table 9, for all tasks, there is a significant performance gap between models trained on Fact-TOD and on SK-TOD. We also provide sampled outputs by the Fact-TOD model and the SK-TOD model in Table 10. By checking the model output, we observe that the Fact-TOD model tends to only ground on and copy information from a single knowledge snippet. This behavior indicates that it is difficult to apply the Fact-TOD model to the SK-TOD task directly, as the Fact-TOD model lacks the ability to effectively aggregate information from multiple knowledge snippets, especially when there are diverse and contradictory opinions. The results also highlight that compared to Fact-TOD, SK-TOD faces new challenges in terms of subjective content understanding and dialogue modeling when integrating subjective knowledge into the responses.

9 Conclusion

In this paper, we have introduced SK-TOD: a novel task focused on subjective-knowledge-based task-oriented dialogue response generation. We create and release a large-scale, manually-annotated dataset for this task. Incorporating subjective knowledge requires models to accurately identify all relevant knowledge snippets and faithfully aggregate the information into concise and contextually appropriate responses, which brings unique challenges to this task. Experiments with strong baselines show that there is a significant performance gap between human-generated and machine-generated responses, particularly in faithfully capturing the diversity and proportion of opinions present in the subjective knowledge. We hope this task together with the provided dataset can promote future research on knowledge-grounded TOD systems and subjective content understanding.

Limitations

The dataset we collected contains two domains, restaurants and hotels. However, to evaluate the model's ability to generalize across different domains, it would be beneficial to include more domains in the dataset. Additionally, to address privacy and copyright concerns, we used crowdsourcing to collect review data, resulting in fewer and shorter reviews than those found in real-world scenarios. This limitation can be mitigated by sampling informative and reliable reviews from real-world data. Regarding the model, we did not investigate more complex models, such as large language models and novel architectures. However, we provide a strong baseline method that will serve as a benchmark for more advanced methods by the research community.

Ethical Considerations

To build our dataset, we collected the dialogue data by augmenting MultiWOZ 2.1, which is a publicly available English dialogue dataset under MIT license. Additionally, we collected the review data using crowd-sourcing, where we provided crowd workers with the reviewer's persona, as well as the aspects and sentiments of reviews. This controlled review collection process helps to exclude offensive or harmful content from the reviews. It also helps to avoid privacy or copyright issues when making the dataset publicly available. Our dataset is available under the CDLA-Sharing 1.0 license.

References

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Hyunmi Baek, JoongHo Ahn, and Youngseok Choi. 2012. Helpfulness of online consumer reviews: Readers' objectives and review cues. *International Journal of Electronic Commerce*, 17(2):99–126.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5480–5494, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. 2022. KETOD: Knowledge-enriched task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, United States. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

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

Eleftherios Dimitrakis, Konstantinos Sgontzos, Panagiotis Papadakos, Yannis Marketakis, Alexandros Papangelis, Yannis Stylianou, and Yannis Tzitzikas. 2018. On finding the relevant user reviews for advancing conversational faceted search. In *EMSASW@ ESWC*, pages 22–31.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7. In *AAAI Dialog System Technology Challenges Workshop*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, William B. Dolan, Jianfeng Gao, Wen tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *AAAI*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.

Di Jin, Seokhwan Kim, and Dilek Hakkani-Tur. 2021. Can i be of further assistance? using unstructured knowledge access to improve task-oriented conversational modeling. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 119–127.

Michael A Kamins, Meribeth J Brand, Stuart A Hoeke, and John C Moe. 1989. Two-sided versus one-sided celebrity endorsements: The impact on advertising effectiveness and credibility. *Journal of advertising*, 18(2):4–10.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. 2021. "how robust ru?": Evaluating task-oriented dialogue systems on spoken conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1147–1154. IEEE.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Jumin Lee, Do-Hyung Park, and Ingoo Han. 2008. The effect of negative online consumer reviews on product attitude: An information processing view. *Electronic commerce research and applications*, 7(3):341–352.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on speech and audio processing*, 8(1):11–23.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yinong Long, Jianan Wang, Zhen Xu, Zongsheng Wang, Baoxun Wang, and Zhuoran Wang. 2017. A knowledge enhanced generative conversational service agent. In *Proceedings of the 6th Dialog System Technology Challenges (DSTC6) Workshop*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2022. Achieving conversational goals with unsupervised post-hoc knowledge injection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3140–3153, Dublin, Ireland. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141, Online. Association for Computational Linguistics.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pages 625–635.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2322–2332.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing mantis: a novel multi-domain information seeking dialogues dataset. *arXiv preprint arXiv:1912.04639*.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of*

the AAAI Conference on Artificial Intelligence, volume 34, pages 8689–8696.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

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

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1855–1865, Hong Kong, China. Association for Computational Linguistics.

Pavlos Vougiouklis, Jonathon Hare, and Elena Simperl. 2016. A neural network approach for knowledge-driven response generation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3370–3380, Osaka, Japan. The COLING 2016 Organizing Committee.

Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. In *International Conference on Learning Representations*.

Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Enhancing dialog coherence with event graph grounded content planning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3941–3947.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020a. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020c. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-supervised opinion summarization by leveraging external information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9644–9651.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

A Data Collection

In this section, we describe more details of the data collection process. The data collection is conducted by a group of crowd workers through Amazon Mechanical Turk. To control the data quality, we choose English speakers from the US, CA, and GB. Workers are eligible for the annotation only if they pass our pre-qualification tests. During data collection, we also manually validate the annotation quality in several rounds to filter out the workers with low-quality annotations.

During review collection, we provide the reviewer's persona, as well as the aspects and sentiments of reviews to workers. The persona is randomly sampled from a pre-defined set of personas. For the aspects and sentiments, we first define 26 common aspects for hotel and restaurant reviews (e.g., WIFI-quality and room-bed for hotels, food-quality and indoor-decor for restaurants). We then randomly selected the target aspects to be addressed in a review. The number of aspects is randomly chosen. To mimic the sentiment distribution of the real reviews, the sentiment of each aspect is sampled based on the actual average ratings taken from Yelp. Figure 3 shows the interface of review collection. We pay workers $1.00 per task.

During user request collection, we ask workers to select the best position to insert a user request by considering every possible position of the given dialogue. Figure 4 shows the interface of user request collection. We pay workers $0.15 per task.

During response generation, we explicitly ask workers to consider the information in all snippets to create a natural and faithful response. Figure 5 shows the interface of response generation. We pay workers $0.25 per task. Below we list the complete instructions that we provide to workers.

- Please read ALL the customer reviews carefully.
- Please read the conversation carefully.
- Write down a response to the customer to answer the question and continue the conversation.
- You must read EVERY REVIEW COMMENT carefully. Each sentence was written by different people with potentially different opinions.
- Your response MUST include your SUMMARY of ALL the review sentences.

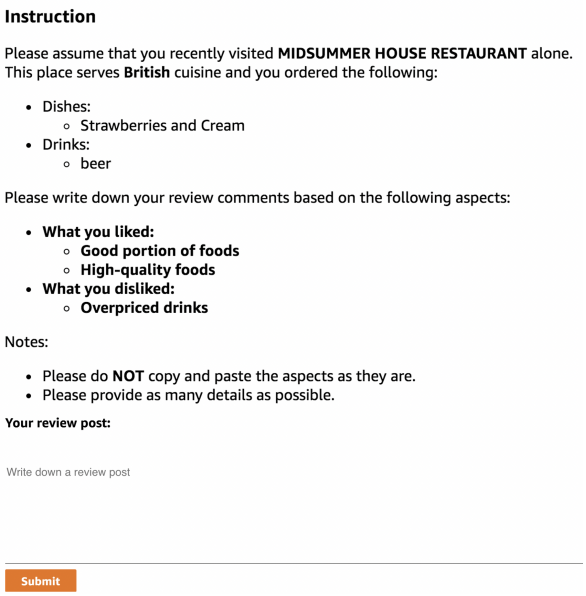


Figure 3: The interface of review collection.

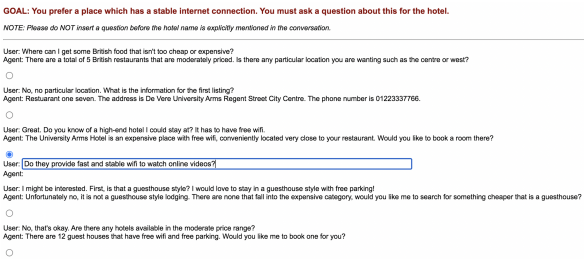


Figure 4: The interface of user request collection.

- If there's any conflict or different opinions in the reviews, your response MUST describe the minority opinion as well.
- Your response MUST be based on the contents in given review comments only.
- Please keep the way of speaking as similar as possible to the previous utterances spoken by the agent.

B Aspect Based Sentiment Analysis

To enhance the model's ability to understand the sentiment polarity of each individual knowledge snippet, we apply PGEN (Zhang et al., 2021), a state-of-the-art aspect-based sentiment analysis model, to predict the sentiment $Z = [z_1, z_2, \cdots, z_i, \cdots]$ for every knowledge snippet $[K_1, K_2, \cdots, K_i, \cdots]$ in $\mathcal{K}^+$.

PGEN converts the problem of aspect-based sentiment analysis into a sequence generation problem, where the input is the review sentence, and

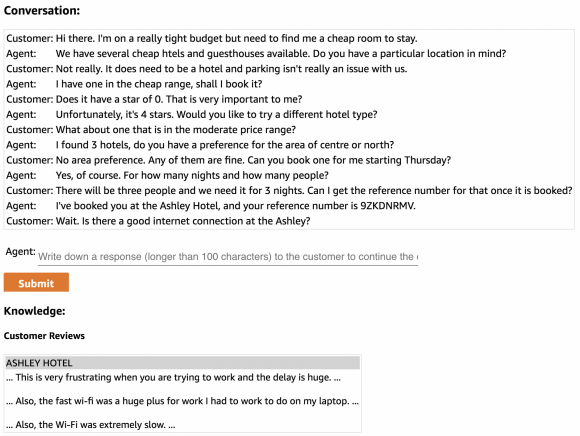


Figure 5: The interface of response generation.

the output is a natural language description of the aspect and the sentiment. For example, given the review sentence as "*The ambience was so fun.*", where the aspect term is "ambience" and the corresponding sentiment polarity is "positive", PGEN transform the aspect term and the sentiment polarity into a natural language description "ambience is great." using templates. It is transformed by keeping the aspect term unchanged and mapping the positive/neutral/negative sentiment polarities into one of the three tokens: "great", "ok", and "bad". The model is trained using a BART-base model on semeval aspect-based sentiment analysis datasets (Pontiki et al., 2015, 2016).

C Training Details

For KTD and KS, the implementation is based on Transformers (Wolf et al., 2020). During training, we use AdamW (Loshchilov and Hutter, 2018) with a learning rate of 3×10^{-5} and a batch size of 16. We apply warmup (Goyal et al., 2017) on the first 500 steps and early stopping based on the model performance on the validation set. We use a Tesla V100 GPU with 16 GB memory for training models. It takes 1 hour to train a KTD model and 5 hours to train a KS model.

During inference, we set the classification threshold as 0 for KTD, as we observe that KTD results are insensitive to the threshold. However, for the KS model, the setting of the threshold can greatly impact the precision and recall scores. We therefore choose the best threshold based on the F_1 scores on the validation set. We use a grid search between -5 to 5. The optimal thresholds for BERT, RoBERTa, ALBERT, and DeBERTa are 2.25, 1, 1.75, and 2 in the bi-encoder setting. They are 3.1, 4.6, 3.25, and 3.4 in the cross-encoder setting.

For ET model, we follow the setting of Jin et al. (2021) to identify entities. More specifically, we perform the fuzzy n-gram matching between an entity and the utterance, where n is the same as the length of the entity mention. The n-gram matching score is calculated based on the ratio of the longest common sequence between two n-grams. We set the matching threshold as 0.95.

For RG model, during training, we use AdamW with a learning rate of 3×10^{-5} and a batch size of 16. We apply the warmup on the first 500 steps and the early stopping based on the model performance (perplexity) on the development set. The model is trained on a Tesla V100 GPU with 16 GB memory for 2 hours.

UD_Japanese-CEJC: Dependency Relation Annotation on Corpus of Everyday Japanese Conversation

Mai Omura
NINJAL, Japan

Aya Wakasa
Tohoku University

Hiroshi Matsuda
Megagon Labs, Tokyo,
Recruit Co., Ltd.

Masayuki Asahara
NINJAL, Japan

Abstract

In this study, we have developed Universal Dependencies (UD) resources for spoken Japanese in the Corpus of Everyday Japanese Conversation (CEJC). The CEJC is a large corpus of spoken language that encompasses various everyday conversations in Japanese, and includes word delimitation and part-of-speech annotation. We have newly annotated Long Word Unit delimitation and *Bunsetsu* (Japanese phrase)-based dependencies, including *Bunsetsu* boundaries, for CEJC. The UD of Japanese resources was constructed in accordance with hand-maintained conversion rules from the CEJC with two types of word delimitation, part-of-speech tags and *Bunsetsu*-based syntactic dependency relations. Furthermore, we examined various issues pertaining to the construction of UD in the CEJC by comparing it with the written Japanese corpus and evaluating UD parsing accuracy.

1 Introduction

Universal Dependencies (UD) (Nivre et al., 2016; de Marneffe et al., 2021) is a framework for consistent annotation of grammatical elements including parts of speech, morphological features, and syntactic dependencies in various human languages. UD provides a wide range of corpus types, encompassing written as well as spoken language data (Dobrovoljc, 2022).

The UD Japanese team has also developed and maintained several resources (Asahara et al., 2018), including UD_Japanese-GSD, UD_Japanese-PUD (Asahara et al., 2018) and UD_Japanese-BCCWJ (Omura and Asahara, 2018). Additionally, there are distinct versions of these corpora with long-unit word annotations (Omura et al., 2021). However, all of these resources are currently limited to written Japanese. Therefore, the present study addresses this gap by introducing UD resources for spoken Japanese and leveraging the Corpus of Everyday Japanese Conversation (CEJC). The resulting resource is referred to as **UD_Japanese-CEJC**.

The CEJC (Koiso et al., 2022) was recently released by NINJAL, Japan. This corpus represents a significant advancement in spoken language resources, as it comprises a large-scale collection of Japanese conversations encompassing more than 200 hours. Various types of audio and video data - including chat sessions, consultations, and meetings - were collected for the CEJC corpus. The informants were carefully selected to ensure a balanced representation in terms of gender and age. The resource includes transcriptions and word segmentation information along with Japanese part-of-speech tags. In addition, we have newly annotated *Bunsetsu* (Japanese-phrase unit)-based dependencies for a subset of the CEJC dataset, specifically in a 20-hour segment. Building upon this, Omura and Asahara (2018) have proposed conversion rules to transform the *Bunsetsu*-based dependencies into UD trees. By applying the conversion method proposed by Omura and Asahara (2018), it becomes feasible to transform the CEJC corpus into UD corpus, thereby facilitating the development of a substantial Japanese UD spoken corpus.

We present the outcomes of our endeavor in the development of a spoken UD Japanese corpus using the dialogue-based CEJC. An overview of our work is depicted in Figure 1. The CEJC corpus provides audio and video data along with token mappings for dialogues, enabling the realization of UD mappings. In the following sections, we elaborate on the proposed annotation scheme and present essential statistics of the resulting dataset, drawing upon related research. Furthermore, we evaluate the performance of a parser trained on both the UD Japanese written and spoken corpora. We also highlight the distinctive features of UD_Japanese-CEJC in comparison to written and spoken language, with a specific emphasis on

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 324–335
September 11–15, 2023. ©2023 Association for Computational Linguistics

disfluencies such as reparanda, repairs, and fillers characteristic of dialogue-based UD.

2 Related Work

2.1 Spoken Language Treebanks

Since the seminal work on the Switchboard Corpus (Godfrey et al., 1992; Calhoun et al., 2010), a number of spoken language treebanks have been developed (Marcus et al., 1999; Zen et al., 2019; Hovy et al., 2006). These treebanks have played a crucial role in research pertaining to natural spoken language processing, serving as essential resources for the development of applications such as speech recognition, speech synthesis, speech translation, spoken language understanding, and speech-based dialogue systems. However, the construction of spoken language treebanks poses technical and linguistic challenges in terms of data collection, annotation, and analysis, all of which are more complex compared to their counterparts in text-based treebanks.

In this context, the UD framework (Nivre et al., 2016; de Marneffe et al., 2021) for spoken language treebanks has emerged as an important development in the field of natural language processing. The UD provides a dependency structure framework (see right side of Figure 1), data format, and guidelines [1] that emphasize commonality across languages. The representation of dependency trees through a common annotation scheme enables language comparisons and improvements in machine translation and other applications. The UD framework also provides a consistent and cross-linguistically applicable set of syntactic annotations are essential for the development of high-quality language processing tools (Straka, 2018; Honnibal et al., 2020).

Dobrovoljc (2022) composed an overview of UD for several spoken languages. UD treebanks for spoken languages vary in size, with relatively large corpora available for Naija (Caron et al., 2019), Norwegian (Øvrelid et al., 2018), and French (Kahane et al., 2021a) in contrast with lower-resource languages such as Beja (Kahane et al., 2021b), Cantonese (Wong et al., 2017), Chukchi (Tyers and Mishchenkova, 2020), and Frisian (Braggaar and van der Goot, 2021). Analyses of spoken language corpora are also being undertaken, for example Kahane et al. (2021a) analyzed examples of spoken dialogue in the Beja, Naija, and French UD treebanks, and examined language phenomena necessary for research on spoken dialogue such as speaker overlap, fillers, and silent pauses.

Yaari et al. (2022) constructed an English UD treebank of 31,264 transcriptions from Hollywood movies. The corpus is multimodal, as it exhibits alignment between audio and video sources. However, it should be noted that the treebank consists of scripted, rather than spontaneous, speech.

2.2 Japanese Spoken Language Resources

Data collection in the Japanese language started with small-scale data, such as reading speech for dialogue systems and speech recognition (Yuichi and Tomoko, 2018). Spontaneous dialogue data continues to be collected as it is recognized to be crucial. Several Japanese spoken language corpora have been constructed in prior studies; e.g., the Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2003), Nagoya University Conversation Corpus (NUCC) (Fujimura et al., 2012), SMOC corpus (Yamazaki et al., 2020). (Koiso et al., 2022) in the Table 1 also compiled a list of Japanese spoken language resources that includes spontaneous dialogue corpora.

Each type of data is associated with different research purposes, formats, and annotations. In particular, there has been no unified syntactic annotation in Japanese, and UD format treebanks of spoken Japanese have not been developed to date. Our study aimes to construct the UD version of CEJC as described in Section 3.

2.3 UD Japanese

The UD Japanese team has built several resources with UD Japanese-KTC (Tanaka et al., 2016) as the point of departure, wherein data are based on their constituent trees (Tanaka and Nagata, 2013). As of v2.5, UD Japanese-BCCWJ offers intuitive suitability for Japanese syntax along with an abundance of existing resources. Consequently, more recent UD Japanese resource have been based on a corpus of *Bunsetsu*-based syntactic dependencies. Bunsetsu is Japanese base phrase unit of syntactic dependencies.

Furthermore, NINJAL negotiated with stakeholders to inherit and manage data continuously for the GSD and PUD corpora. The data were manually annotated according to their UniDic-based morphological information (Den et al.,

[1] https://universaldependencies.org/

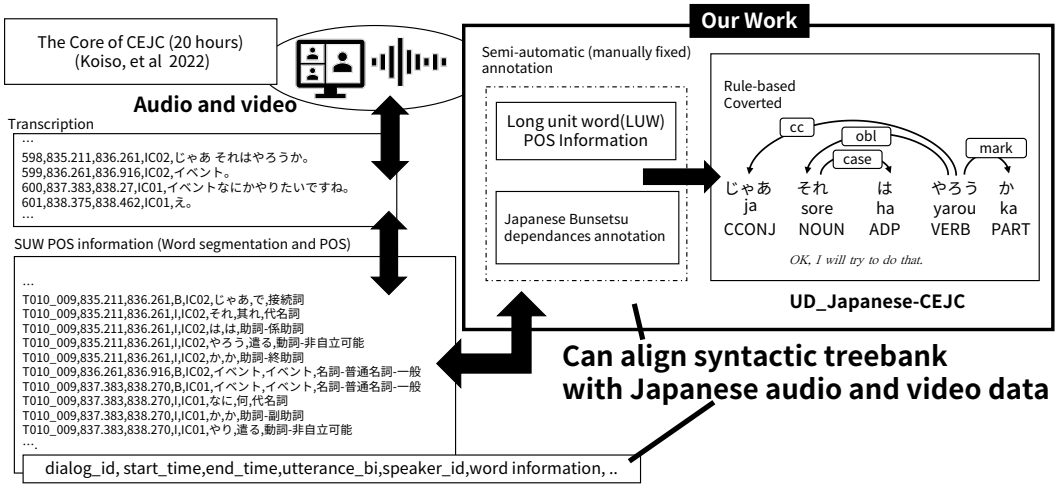


Figure 1: The overview of out building UD_ Japanese-CEJC. (The sample is dialog T010_009 from CEJC)

2008), NINJAL Short Unit Word (SUW) delimitation, NINJAL Long Unit Word (LUW) delimitation, and Bunsetsu (base phrase)-based syntactic dependencies on the original text. The UD Japanese team developed conversion rules from the two-word delimitation and Bunsetsu-based syntactic dependencies to SUW-based UD (Asahara et al., 2018). The Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014) is one of large written Japanese corpora. The corpus serves as a model to annotate UD Japanese GSD and PUD with SUW, LUW, and Bunsetsu-based syntactic dependencies (Asahara and Matsumoto, 2016). Likewise, Omura and Asahara (2018) constructed UD Japanese-BCCWJ via the conversion rules of UD Japanese-GSD and PUD.

Thus, the design of UD Japanese is based on SUW, LUW, and Bunsetsu-based dependencies. The CEJC includes the SUW and its morpheme information. If we know the LUW and Bunsetsu-based dependencies of the CEJC, we can develop UD resources for spoken Japanese via the methods described in Omura et al. (2021). In Section 3.2, we describe the construction of the LUW and Bunsetsu-based dependencies of CEJC.

3 Design of UD_Japanese-CEJC

The following section provides a concise overview of CEJC and outlines the construction of the UD version of CEJC.

Sound file ID	yes
Text-sound alignment	yes
Speaker ID	yes
Language variety	no
Standard orthography	yes
Capitalization	not applicable
Pronunciation	yes
Speaker overlap	yes
Final punctuation	not applicable
Other punctuation	not applicable
Incomplete words	yes
Fillers	yes
Silent pauses	yes
Incidents	yes
Text-video alignment	yes
Dialog act	yes (ISO-24617-2)
Intonation label	partially yes

Table 1: Transcription characteristics in CEJC. (cf. Dobrovoljc (2022), Table 2)

3.1 Corpus of Everyday Japanese Conversation

The Corpus of Everyday Japanese Conversation (CEJC) (Koiso et al., 2022) is a large-scale spoken Japanese corpus. It encompasses 200 hours of speech, comprising 577 conversations approximately 2.4 million words and involving a total of 1675 participants. Data are segmented into utterance units based on perceptible pauses and clause boundaries. Transcriptions of the speech audio and video data are provided, and the text is further segmented into word units using SUW and UniDic-based morphological information.

The Core dataset is a subset of CEJC that consists of 20 hours of speech, encompassing 52 di-

English	*My son's*		*a birthday present*					*could be*		
	musuko	no	tanjo	bi	purezento	ka	mo	shin	nai	kedo
SUW	息子	の	誕生	日	プレゼント	か	も	しん	ない	けど
	NOUN	ADV	NOUN	NOUN	NOUN	ADP	ADP	VERB	AUX	SCONJ
LUW	息子	の	誕生日プレゼント			かもしんない				けど
	NOUN	ADV	NOUN			AUX				SCONJ
Bunsetsu	息子の		誕生日プレゼントかもしんないけど							

(It could be my son's birthday present.)

Figure 2: Example of two-way POS annotation (Short and Long unit word) and Bunsetsu of CEJC (refer to T011_005.) The lines above indicate the word boundaries. The parts of speech are represented using universal POS tags for simplicity, but UD_Japanese CEJC can refer to the UniDic part-of-speech tags.

alogues. This subset includes manually annotated and corrected annotations. For this dataset, we annotated LUW and established Bunsetsu-based dependencies. Details pertaining to this annotation process are discussed in the following section.

Table 2 in (Dobrovoljc, 2022) provides an overview of the transcription characteristics in the CEJC. We present a summary of these characteristics in Table 1. The language variety represented in the CEJC is predominantly limited to speakers of common Japanese residing in Tokyo and surrounding prefectures. It is important to note that Japanese does not follow a capitalization convention. Additionally, the transcription rule employed in the CEJC does not account for punctuation marks. One characteristic of the CEJC is the alignment of video data to speech. All videos were collected by normal and omnidirectional 360-degree cameras[2]. The dataset contains dialog act annotations following the ISO 24617-2 scheme (Iseki et al., 2019). Moreover, the audio files are partially annotated with intonation labels using X-JToBI (eXtended-Japanese ToBI), a framework specifically designed for the analysis of spontaneous Japanese speech, as employed in the CSJ corpus (Maekawa, 2003).

3.2 *Bunsetsu*-based Dependency Annotation

The written Japanese data are segmented into sentences based on sentence end symbols specified by authors. However, because sentence-ending punctuation is absent in spoken dialogue, sentence bounds are significantly less straightforward. To address this, the CEJC developers introduced the concept of utterance units, specifically focusing on long utterance units (Den et al., 2010) characterized by silent pauses and clause boundaries. These long utterance units are identified by syntactic and pragmatic disjuncture within the dialogues. Throughout our annotation process, we treated each utterance unit as a separate sentence, forming a tree structure.

We newly annotated the LUW morphological information and *Bunsetsu* boundaries for the CEJC trees. An example of word delimitation using SUW, LUW, and Bunsetsu is illustrated in Figure 2. The SUW is a minimal language unit that has a morphological function and the LUW definition can be regarded as syntactic words in Japanese based Bunsetsu. For further details, please refer to (Omura et al., 2021) and NINJAL website [3]. The LUW information was initially analyzed using Comainu (Kozawa et al., 2014) and subsequently manually corrected by annotators.

In addition, we annotated Bunsetsu-based dependencies for the CEJC utterance units following the BCCWJ-DepPara annotation scheme (Asahara and Matsumoto, 2016). The Bunsetsu-based dependencies was also analyzed by Cabocha (Kudo and Matsumoto, 2002), manually corrected by annotators. It is important to note that the Japanese language exhibits a strict head-final order within the Bunsetsu units. However, the Bunsetsu dependencies in CEJC encompass linguistic phenomena such as fillers, anastrophes, and predicate ellipses, which are rarely observed in written texts. In cases where a dependent does not have its corresponding head within the utterance units, we position a dummy node as the dependency head at the end of the utterance, as depicted in Figure 3.

3.3 Conversion into UD schema

The UD_Japanese-CEJC corpus was derived from the Bunsetsu dependencies in the core data sub-

[2]Video files include the faces of the main conversation participants who agreed to have their faces published. All other participant's faces are obscured.

[3]https://clrd.ninjal.ac.jp/bccwj/en/morphology.html

Conversion rule	UPOS
...	...
POS of SUW is *punctiction*	PUNCT
...	...
POS of SUW is *adjective*	ADJ
POS of SUW is *noun*	NOUN
...	...
POS of SUW is *verb* & The Bunsetsu is the end of the phrase	VERB
...	...

Conversion rule	DEPREL
Bunsetsu is the end of the phrase & Subject word	*root*
UPOS is PUNCT	*punct*
...	...
Subject word in the Bunsetsu & UPOS is NOUN & Attaching particle 'ga'	*nsubj*
...	...
Bunsetsu is not functional phrase & UPOS is ADJ	*amod*
UPOS is ADP	*case*
...	...

Table 2: The short sample of UD conversion rules is outlined in (Omura and Asahara, 2018). As of July 2023, there are 85 rules for UPOS conversion and 120 rules for DEPREL.

	UPOS	DEPREL
If *the word* is filter	INTJ	*discourse(:filter)*
If *the word* is disfluency	X	*reparandum*

Table 3: Labeling rules to convert for UD_Japanese-CEJC. The current approach for determining whether the word is filler or disfluencyis to reference the POS information.

set, which consists of 20 hours of transcribed speech. To compile the UD Japanese resource, we applied the conversion rules outlined in (Omura and Asahara, 2018), which are shared across all UD Japanese treebanks, including GSD, PUD, BCCWJ, GSDLUW, PUDLUW, and BCCWJLUW (Omura and Asahara, 2018; Omura et al., 2021) [4]. Table 2 shows a partial set of conversion rules. These rules determine the UPOS (Universal Part-of-Speech) and DEPREL (Dependency Relation Label) in the UD framework. However, it is important to note that the conversion rules primarily consider written Japanese corpora and might not fully capture the specific characteristics of spoken Japanese. As a result, additional rules were introduced to handle fillers and stutters, which are infrequent in written corpora, as shown in Table 3. While these conversion rules provide a valuable starting point, further refinements may be necessary to fully account for the nuances of spoken Japanese.

[4]There are several spoken UD corpora that offer automatic conversion of existing resources; e.g., UD French ParisStories (Kahane et al., 2021a) and Naija NSC (Caron et al., 2019)

In the UD version of the CEJC, the aforementioned utterance units serve as boundaries for dependency trees. According to the UD guideline, other treebanks have their own language-specific guidelines for handling fillers and disfluencies (e.g. Slovenian SST (Dobrovoljc and Nivre, 2016)). Nevertheless, we decided that any fillers and disfluencies dependent on the dummy node are to be converted to the sentence end root to adhere to the single root restriction, as their attachment is inherently ambiguous. Because argument ellipses are common in Japanese and the annotation units in this dataset are based on utterances, we can only define these ellipses as fillers or disfluencies within the scope of the utterance unit. To determine the appropriate attachment of fillers across languages, including those where ellipses are grammatically allowed, a thorough investigation is necessary.

Figure 3 shows an example of Bunsetsu dependencies constructed to the UD framework. The Bunsetsu-dependency structure is converted to UD structures according to rules specified in (Omura and Asahara, 2018). In the case of the figure, the words "tsu" and "n" are a disfluency and filter, respectively, making them dependent upon the root node " deki ta shi".

3.4 Statistics of UD Japanese CEJC

Table 4 presents a statistical analyses of the generated UD_Japanese-CEJC (spoken) corpus in comparison to UD_Japanese-GSD and BCCWJ (written). These statistical values are from version 2.11. The 'Trees' column indicates the numbers of utterance units in CEJC (spoken) and sentences in GSD (written). The 'Tokens' column represents the total count of word tokens in each treebank. The 'Avg.' column displays the average number of word tokens per tree, whereas the 'Bunsetsu' column indicates the total number of Bunsetsu. The automatic conversion of the 20-hour speech transcription has yielded a substantial amount of data that aligns with the corresponding audio and video. However, it is worth noting that the number of words in a dependency tree within a spoken utterance unit tends to be smaller than that in a written sentence. It provides a clear comparison between the statistics in Table 1 of Dobrovoljc (2022) and the specific characteristics of CEJC as a conversational corpus including many phatic expressions like *Aizuchi* in Japanese such as "hai"

Corpora	Unit	Trees	Tokens	Avg. per Tree	Bunsetsu
CEJC	SUW	59,319	256,885	4.3	136,071
	LUW	59,319	231,774	3.9	136,071
GSD	SUW	8,100	193,654	23.9	65,966
	LUW	8,100	150,243	18.5	65,966
BCCWJ	SUW	57,109	1,253,903	21.9	425,751
	LUW	425,751	99,5632	17.4	425,751
CEJC-	SUW	54,599	24,4296	4.7	124,456
	LUW	54,599	219,415	4.0	124,456

Table 4: Statistics of UD Japanese CEJC (spoken), GSD, and BCCWJ (written) (v2.11). CEJC- is a CEJC corpus that omits any words containing solely inapplicable morphological information (non-lexical tokens), filters, or reparandums.

and "ee" ("uhhuh" and "yeah" in English).

Table 5 shows the distribution of UPOS labels of UD_Japanese-CEJC, GSD, and BCCWJ [5]. The spoken data does not include any PUNCT and SYM, as punctuations and symbols were not accounted for. CCONJ and INTJ are larger than the written corpora. Whereas the written data tend to omit PRON, the spoken data tends to include PRON when referencing speakers. X is a token associated with no morphological annotations, such as incidents (laugh, cry, singing, etc.) in the CEJC.

Table 6 shows the distribution of DEPREL labels of UD Japanese CEJC, GSD and BCCWJ. In the spoken data, words are shorter per a tree (see Table 4). Consequently, the DEPREL *root* is the largest element within the spoken data. Because PUNCT does not appear in the spoken data, the DEPREL *punct* is zero.

	CEJC	GSD	BCCWJ
ADJ	3.69%	1.98%	2.14%
ADP	13.61%	21.62%	20.03%
ADV	6.74%	1.22%	1.51%
AUX	13.24%	10.93%	9.74%
CCONJ	1.64%	0.42%	0.41%
DET	0.56%	0.51%	0.48%
INTJ	10.74%	0.01%	0.07%
NOUN	14.86%	30.05%	29.24%
NUM	1.67%	2.67%	3.11%
PART	8.49%	0.65%	1.18%
PRON	3.77%	0.57%	0.90%
PROPN	1.39%	3.69%	2.87%
PUNCT	0.00%	9.93%	11.69%
SCONJ	6.68%	4.13%	4.49%
SYM	0.00%	0.67%	1.53%
VERB	9.86%	10.96%	10.57%
X	3.05%	0.00%	0.03%

Table 5: The distribution of UPOS labels in UD_Japanese-CEJC, GSD and BCCWJ (SUW)

4 Parser Evaluation

We conducted experiments to assess the reproducibility and parsability of the CEJC corpus. Through a comparison between CEJC and GSD, we illustrate the distinctions between spoken and written Japanese in terms of UD annotation.

4.1 Corpus

To evaluate parsing, we used the following UD Japanese v.2.11 [6] corpora: GSD, CEJC, and their combination (CEJC+GSD). Although SUW and LUW UDs are present, we only considered SUW to examine differences between the spoken and written corpora. The GSD was split among train, dev, and test sets by original UD corpus. The UD CEJC was divided between training, development, and testing sets according to a 8:1:1 ratio based on conversation form as provided by the CEJC: chat, consultations, and meetings. Table 7 shows the distribution of UDs in the experiment. The models were constructed with the sentence (tree) boundary as given, as it is easy to imagine that the utterance units and written sentences are clearly different in Table 4. In particular, CEJC explicitly lacks

[5]Because the SUW are encapsulated in the LUW, there is no significant difference in distribution. Therefore, only SUW are listed.

[6]These UD Japanese is also in development as of November 2022. This version conforms to the latest UD guidelines.

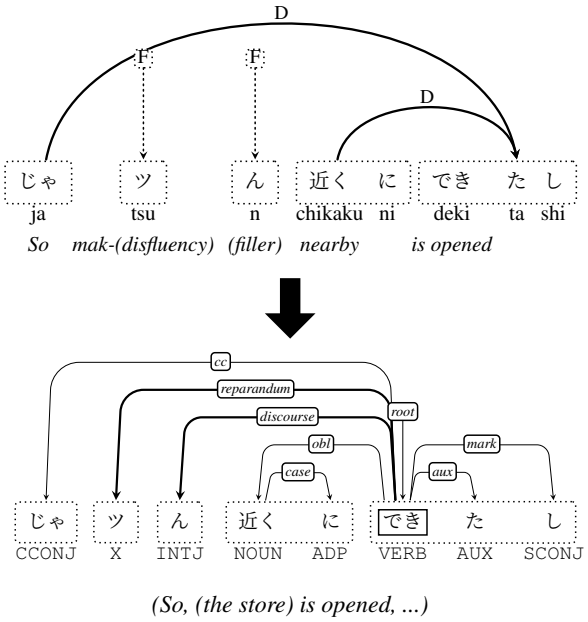


Figure 3: Sample construction of UD_Japanese-CEJC（T011_007）. The upper figure represents Bunsetsu-dependencies and the lower figure shows the UD conversion. The dotted box denotes the Bunsetsu boundary, and the Bunsetsu dependency edge label 'D' is an ordinal dependency relation, where 'F' indicates that no relation is present.

	CEJC	GSD	BCCWJ
acl	2.11%	3.61%	3.62%
advcl	3.87%	3.72%	3.85%
advmod	4.73%	1.18%	1.43%
amod	0.10%	0.23%	0.25%
appos	0.00%	0.00%	0.00%
aux	9.10%	8.90%	7.56%
case	12.72%	21.33%	19.65%
cc	1.59%	0.42%	0.41%
ccomp	0.34%	0.20%	0.22%
compound	3.97%	14.19%	14.67%
cop	1.98%	1.26%	1.20%
csubj	0.09%	0.08%	0.11%
csubj:outer	0.00%	0.00%	0.00%
dep	1.00%	0.04%	0.99%
det	0.54%	0.51%	0.48%
discourse	2.72%	0.01%	0.03%
dislocated	0.00%	0.00%	0.00%
fixed	4.15%	4.45%	4.26%
mark	14.20%	4.06%	5.04%
nmod	2.87%	6.70%	6.92%
nsubj	2.51%	4.02%	3.69%
nsubj:outer	0.00%	0.23%	0.18%
nummod	0.98%	1.45%	1.16%
obj	0.48%	2.74%	2.62%
obl	5.64%	6.55%	5.41%
punct	0.00%	9.93%	11.69%
reparandum	1.21%	0.00%	0.00%
root	23.09%	4.18%	4.55%

Table 6: Distributions of DEPREL labels in UD_Japanese-CEJC, GSD and BCCWJ (SUW)

	train		dev		test	
	trees	tokens	trees	tokens	trees	tokens
GSD	7,050	168,333	507	12,287	543	13,034
CEJC	36,997	157,227	9,837	43,378	12,485	56280
CEJC-	34,105	149,614	9,057	41,055	11,437	53,627

Table 7: The train/dev/test distribution of UD corpus (GSD/CEJC)

punctuation, making it difficult to identify speech breaks.

4.2 Parser Model

We used spaCy v3.4 (Honnibal et al., 2020), along with spacy-transformers v1.2 as a parsing model framework. spaCy is a trainable network that features a component pipeline for sentence analysis and word tokenisation, part-of-speech tags, dependencies, and named entities. Furthermore, spaCy can use pre-trained transformers (Wolf et al., 2020) such as BERT (Devlin et al., 2019), and allows loss gradients to be shared between the transformers-based pre-training model and analysis component.

A significant distinction between CEJC and other written word treebanks lies in the presence of specific word characteristics, particularly fillers and reparanda. To address this feature, we propose two models: *the two-stage analysis model* and *the simultaneous analysis model*. We assessed the effectiveness of these models in accurately capturing the relationship between fillers and reparanda.

The two-stage analysis model comprises two models: a component that detects and removes the span fillers and reparanda, and a component that subsequently analyzes the parsing tree. Following the method described in (Asahara and Matsumoto, 2003) in regards to the spans of fillers and reparanda detecting named entities, the model was trained via spaCy, whereas the other model was trained by eliminating fillers and reparanda (CEJC-). While the model has two components, the accuracy of parsing results is only evaluated using the correct trees in the absence of fillers and reparanda (CEJC-) as seen in (Table 8), as it is difficult to map removed words as fillers and reparanda and others as original text data.

The simultaneous analysis model includes fillers and reparanda simultaneously. SpaCy can share a transformer's information among multiple analytical components and perform simultaneous

learning. The pipeline components of spaCy were organized in the order of transformers, morphologizer analysis, parser analysis, and NER analysis. The ner analysis is used to detect fillers and reparanda equivalently to the two-stage analysis model.

As a transformer pre-trained model on spaCy, we used cl-tohoku/bert-japanese [7], a BERT model trained on the Japanese version of Wikipedia with words tokenized by MeCab (Kudo et al., 2004) and split into subwords by the WordPiece algorithm. The parser component of spaCy is based on the Non-Monotonic Arc-Eager Transition System with extensions to Projectivization/Deprojectivization by Lifting of Nivre (Nivre and Nilsson, 2005) to handle intersecting contexts.

4.3 Parsing Results

Table 8 presents the tokenisation, tagging, lemmatisation, and dependency parsing results obtained by the two spaCy models. **Tokens**, **UPOS**, **XPOS**, and **Lemma** are reproducible and expressed by their F_1 scores. **UAS** (Unlabeled Attachment Score) and **LAS** (Labelled Attachment Score) are standard evaluation metrics in dependency parsing results. These results were output by the evaluation scripts of CoNLL 2018 shared tasks (Zeman et al., 2018).

When the training and testing data are different (e.g. train/dev GSD and test CEJC, or train/dev CEJC and test GSD), tokenisation (**Tokens**) and POS tagging (**UPOS** and **XPOS**) exhibit poor performance. This is because there are differences in vocabulary and distributions of POS and DEPREL. During tokenisation, spoken utterances have significantly different delimiters compared to those observed in written sentences, as the former include fillers, disfluencies, and repairs. It is also difficult to tokenize without spaces, as required by Japanese. POS tagging presents similar challenges. Although the major POSs of the CEJC are `INTJ`, `CCONJ`, and `PRON` (e.g. first personal pronoun, second personal pronoun), the POS `INTJ` is very rare in GSD. Consequently, the assignment of `INTJ` requires training data from the CEJC. Overall, the combined training data (train/dev: CEJC+GSD) achieved the best performance for both GSD and CEJC tokenisation and tagging.

[7] https://github.com/cl-tohoku/bert-japanese/

Results of filter and reparandum detection are shown in Table 9. The simultaneous analysis model tended to be slightly more accurate than the two-step analysis model. This is thought to be an effect of learning-dependent structure analysis, as well as the simultaneous identification of fillers and reparanda. However, compared to the overall evaluation (in Table 8), the accuracy of tokenisation, POS tagging, and dependency analysis for both fillers and reparanda decreased by more than 6 points.

The dependency attachment (**UAS** and **LAS**) of the CEJC is also difficult, and even the CEJC tree length (avg. 4.3) is shorter than that of the GSD tree (avg. 23.9). GSD also encompasses punctuation in written texts, which helps determine the roots of trees and resolve long-distance dependencies. In contrast, the CEJC does not include punctuation in the transcription, making it difficult to determine the roots of trees as well as presenting challenges with respect to fillers and disfluencies.

5 Conclusions

This study introduces a novel UD Japanese resource derived from the Corpus of Everyday Japanese Conversation (CEJC), representing the first spoken language resource in the UD Japanese framework. The UD resource was built upon transcriptions of audio files from individual speakers, accompanied by two types of video recordings (standard camera and omnidirectional 360-degree camera). Whereas previous efforts have been limited in their incorporation of text-to-video alignment, this study presents a substantial treebank with video, surpassing existing UD resources in this aspect. In the future, we plan to primarily expand the annotation based on audio information; e.g., overlap markers similar to those used in the UD French Rhapsodie (Kahane et al., 2021a).

Parser evaluations were conducted to compare the performance of the parser on UD_Japanese-CEJC (spoken) and GSD (written) datasets. The findings clearly demonstrate the challenges associated with parsing spoken Japanese using a model trained on written corpora. The presence of fillers, disfluencies, and repairs significantly impacted tokenisation and POS tagging accuracy, highlighting the unique characteristics of spoken language that must be accounted for to improve parsing performance.

The UD version of CEJC is currently available

train/dev	test	**Token**	**UPOS**	**XPOS**	**Lemmas**	**UAS**	**LAS**
spaCy two-stage analysis model (eliminating gold fillers and reparandums)							
CEJC–	GSD	98.15%	84.54%	96.96%	94.38%	80.58%	71.97%
CEJC–	CEJC–	96.38%	94.45%	92.33%	86.33%	89.71%	87.54%
spaCy simultaneous analysis model (including fillers and reparandums)							
GSD	GSD	98.14%	97.04%	96.96%	94.38%	91.72%	90.84%
GSD	CEJC	81.16%	84.33%	89.32%	84.92%	80.74%	74.71%
CEJC	GSD	98.14%	84.31%	96.96%	94.38%	79.58%	70.52%
CEJC	CEJC	95.44%	93.39%	89.32%	84.92%	88.19%	84.51%
CEJC+GSD	GSD	98.14%	97.16%	96.96%	95.64%	91.49%	90.56%
CEJC+GSD	CEJC	95.55 %	93.47%	93.47%	89.32%	88.38%	86.57%

Table 8: Results of tokenisation, tagging, lemmatisation and dependency parsing using CEJC and GSD.

Category	Occurence train / dev / test	**Two-stage analysis model Token P / R / F**	**Simultaneous analysis model Token P / R / F**	**Simultaneous analysis model UPOS / UAS / LAS**
Filler	1,736 / 524 / 559	88.6% / 87.3% / 87.9%	86.9% / 90.4% / 88.6%	87.7% / 82.4% / 82.0%
Reparandum	2,122 / 741 / 793	90.5% / 86.0% / 88.2%	88.4% / 87.4% / 87.9%	87.9% / 83.7% / 83.2%

Table 9: Results of accuracy detection for fillers and reparanda analyses.

to CEJC subscribers through the dedicated download site on the CEJC platform. Additionally, the UD will be made available on the Universal Dependencies site and the UD Japanese Github repository [8][9] in a standoff format. This wider distribution will enable researchers to access and utilize the UD Japanese CEJC data for various linguistic analyses and applications. The spaCy models employed in the conducted experiments will be made publicly available in Github repository [10]. These resources will allow researchers and practitioners to utilize the models for their own natural language processing tasks and further contribute to the advancement of linguistic research in the field of Japanese spoken language processing.

Acknowledgements

This work is supported by JSPS KAKENHI 19K13195, a collaborative research project with Recruit Co. Ltd., and a NINJAL Collaborative Research Project 'Evidence-based Computational Psycholinguistics Using Annotation Data'.

[8] https://github.com/udjapanese/UD_Japanese-CEJCSUW

[9] https://github.com/udjapanese/UD_Japanese-CEJCLUW

[10] https://github.com/megagonlabs/UD_Japanese-GSD/releases/tag/nlp2023

References

Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki. 2018. Universal Dependencies version 2 for Japanese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1824–1831, Miyazaki, Japan. European Language Resources Association.

Masayuki Asahara and Yuji Matsumoto. 2003. Filler and disfluency identification based on morphological analysis and chunking. In *Proceedings of ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pages 163–166, Tokyo, Japan. ISCA.

Masayuki Asahara and Yuji Matsumoto. 2016. BCCWJ-DepPara: A syntactic annotation treebank on the 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of the 12th Workshop on Asian Language Resources*, pages 49–58, Osaka, Japan. The COLING 2016 Organizing Committee.

Anouck Braggaar and Rob van der Goot. 2021. Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58, Kyiv, Ukraine. Association for Computational Linguistics.

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, se-

mantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419.

Bernard Caron, Marine Courtin, Kim Gerdes, and Sylvain Kahane. 2019. A surface-syntactic UD treebank for Naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 13–24, Paris, France. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Yasuharu Den, Hanae Koiso, Takehiko Maruyama, Kikuo Maekawa, Katsuya Takanashi, Mika Enomoto, and Nao Yoshida. 2010. Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2103–2110, Valletta, Malta.

Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. 2008. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1019–1024, Marrakech, Morocco. European Language Resources Association.

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

Kaja Dobrovoljc. 2022. Spoken Language Treebanks in Universal Dependencies: an Overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1566–1573, Portorož, Slovenia. European Language Resources Association (ELRA).

Itsuko Fujimura, Shoju Chiba, and Mieko Ohso. 2012. Lexical and grammatical features of spoken and written Japanese in contrast: exploring a lexical profiling approach to comparing spoken and written corpora. In *Proceedings of the VIIth GSCP International Conference : Speech and Corpora*, pages 393–398, Belo Horizonte, Brazil. Firenze University Press.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Proceedings of 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)*, volume 1, pages 517–520 vol.1.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Yuriko Iseki, Keisuke Kadota, and Yasuharu Den. 2019. Characteristics of everyday conversation derived from the analysis of dialog act annotation. In *Proceedings of 2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques*, pages 1–6, Cebu, Philippines. IEEE.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021a. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021b. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria. Association for Computational Linguistics.

Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe. 2022. Design and evaluation of the Corpus of Everyday Japanese Conversation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5587–5594, Marseille, France. European Language Resources Association.

Shunsuke Kozawa, Kiyotaka Uchimoto, and Yasuharu Den. 2014. Adaptation of Long-Unit-Word analysis system to different part-of-speech tagset [in Japanese]. *Journal of Natural Language Processing*, 21(2):379–401.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, pages 1—7. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Kikuo Maekawa. 2003. Corpus of Spontaneous Japanese : Its design and evaluation. In *Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 7–12, Tokyo, Japan. ISCA.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguti, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language resources and evaluation*, 48(2):345–371.

Mitchell P Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3. *Linguistic Data Consortium, Philadelphia*, 14.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association.

Joakim Nivre and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 99–106, Ann Arbor, Michigan. Association for Computational Linguistics.

Mai Omura and Masayuki Asahara. 2018. UD-Japanese BCCWJ: Universal Dependencies annotation for the Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the Second Workshop on Universal Dependencies*, pages 117–125, Brussels, Belgium. Association for Computational Linguistics.

Mai Omura, Aya Wakasa, and Masayuki Asahara. 2021. Word delimitation issues in UD Japanese. In *Proceedings of the Fifth Workshop on Universal Dependencies*, pages 142–150, Sofia, Bulgaria. Association for Computational Linguistics.

Lilja Øvrelid, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg, and Janne Bondi Johannessen. 2018. The LIA treebank of spoken Norwegian dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. 2016. Universal dependencies for Japanese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 1651–1658, Portorož, Slovenia. European Language Resources Association.

Takaaki Tanaka and Masaaki Nagata. 2013. Constructing a practical constituent parser from a Japanese treebank with function labels. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 108–118, Seattle, Washington, USA. Association for Computational Linguistics.

Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204, Barcelona, Spain (Online). Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tak-sum Wong, Kim Gerdes, Herman Leung, and John Lee. 2017. Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 266–275, Pisa,Italy. Linköping University Electronic Press.

Adam Yaari, Jan DeWitt, Henry Hu, Bennett Stankovits, Sue Felshin, Yevgeni Berzak, Helena Aparicio, Boris Katz, Ignacio Cases, and Andrei Barbu. 2022. The Aligned Multimodal Movie Treebank: An audio, video, dependency-parse treebank. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9531–9539, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, and Akinori Ito. 2020. Construction and analysis of a multimodal chat-talk corpus for dialog systems considering interpersonal closeness. In *Proceedings*

of the Twelfth Language Resources and Evaluation Conference, pages 443–448, Marseille, France. European Language Resources Association.

Ishimoto Yuichi and Ohsuga Tomoko. 2018. Spontaneous speech resources in Japan. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC 2018) Special Speech Sessions*, pages 1–5.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech 2019)*, pages 1526–1530, Graz, Austria. ISCA.

Unravelling Indirect Answers to Wh-Questions: Corpus Construction, Analysis, and Generation

Zulipiye Yusupujiang and **Jonathan Ginzburg**
Université Paris Cité, CNRS, Laboratoire de Linguistique Formelle
zulipiye.yusupujiang@linguist.univ-paris-diderot.fr
yonatan.ginzburg@u-paris.fr

Abstract

Indirect answers, crucial in human communication, serve to maintain politeness, avoid conflicts, and align with social customs. Although there has been a substantial number of studies on recognizing and understanding indirect answers to polar questions (often known as yes/no questions), there is a dearth of such work regarding *wh*-questions. This study takes up the challenge by constructing what is, to our knowledge, the first corpus of indirect answers to *wh*-questions. We analyze and interpret indirect answers to different *wh*-questions based on our carefully compiled corpus. In addition, we conducted a pilot study on generating indirect answers to *wh*-questions by fine-tuning the pre-trained generative language model DialoGPT (Zhang et al., 2020). Our results suggest this is a task that GPT finds difficult.

1 Introduction

Indirect answers (INDs) to questions hold a distinctive position in the realm of human communication, as they provide related or implied information instead of offering the speaker's intentions or knowledge directly through an utterance's *grammatically governed content.* (i.e., *literal content*) (Ginzburg et al., 2022). Grasping the intrinsic nuances of indirect answers and accurately deducing the expected direct answer from them is essential to facilitate effective communication and information sharing between dialogue participants.

It is a natural part of human communication to produce and understand indirect answers. People use indirect speech to maintain politeness, avoid confrontations, adhere to social norms, or convey information without explicitly stating it (Searle, 1975; Brown et al., 1987). However, understanding and generating indirect answers to questions can be quite challenging for dialogue systems. To engage in human-like conversation, these systems must be able to grasp the conversational context, background information, and relationships between participants. By accurately interpreting the meaning behind an indirect answer, the system can then provide a more appropriate response, contributing to a more natural interaction.

In the field of dialogue studies, considerable attention has been given to the interpretation and generation of indirect answers to polar questions (Green and Carberry, 1994a,b, 1999; de Marneffe et al., 2009, 2010; de Marneffe and Tonhauser, 2016; Louis et al., 2020; Damgaard et al., 2021). However, there still exists a gap when it comes to the identification and interpretation of indirect answers to *wh*-questions. Studying indirect answers to *wh*-questions is a challenging task for several reasons: a). Unlike polar questions that have only *yes* or *no* (or rather the propositions they convey in context) as direct, resolving answers, *wh*-questions can have a wide range of possible direct answers. This makes it harder to interpret indirect answers to *wh*-questions; b). Compiling a corpus of indirect answers to *wh*-questions is a challenging task, since indirect answers to *wh*-questions are significantly less frequent than those of polar questions. It requires annotating a huge number of *wh*-questions within conversational context to collect a reasonable amount of *WhQ-IND* pairs for analysis and training machine learning algorithms; c). The implied meaning of indirect answers to *wh*-questions often depends heavily on the context of the conversation. It usually also involves nuanced linguistic features like sarcasm, irony, and figurative expressions which can be a challenge for humans (overhearers) to interpret, let alone for dialogue systems.

The aim of this paper is, therefore, to conduct a preliminary study by constructing what is, to our knowledge, the first corpus of indirect answers to *wh*-questions, and to investigate how direct answers are deduced from indirect answers.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 336–348
September 11–15, 2023. ©2023 Association for Computational Linguistics

This paper is structured as follows: Section 2 consists of a literature review, whereas Section 3 provides the requisite theoretical background. In Section 4, we present the data collection and annotation process. In Section 5 we propose possible information resources needed for interpreting indirect answers to *wh*-questions. Following this, in Section 6 we briefly describe a pilot study on generating indirect answers by using a pre-trained language model. Our results suggest this is a task that GPT finds difficult. The final section offers conclusions and some potential future work.

2 Related Work

Several studies exist concerning the interpretation and generation of indirect answers to polar questions: Green and Carberry (1994a,b, 1999) proposed both pragmatic and computational methods for understanding and generating indirect answers to polar questions. Specifically, they introduced a discourse-plan-based strategy for implicatures and a combined reasoning model to simulate a speaker's incentive for offering pertinent, unsolicited information. Furthermore, they designed a computational model that is capable of interpreting and generating indirect answers to polar questions in English. Their model relies on shared knowledge of discourse strategies and coherence relations to recognize and formulate a responder's discourse plan for a complete response.

Takayama et al. (2021) released the corpus *DIRECT*, which provides 71,498 indirect-direct pairs together with multi-turn dialogue history extracted from the MultiWoZ dataset, and conducted three experiments to examine the model's ability to recognize and generate indirect and direct utterances. The *DIRECT* corpus provides triples of paraphrases for each user's utterance: *original utterance, indirect utterance*, and *direct utterance*. This is the first study that offers a large-scale corpus of pragmatic annotations, which is very useful for understanding users' intentions in dialogue systems.

In another recent work, Louis et al. (2020) created and released the first large-scale English corpus of more than 34K *polar question–indirect answer* pairs, named *Circa*. That is a collection of natural responses obtained by crowd-sourcing and contains responses with yes-no meaning, as well as uncertain, middle-ground, and conditional responses. The authors also conducted experiments by fine-tuning a multiclass classifier over the BERT model (Devlin et al., 2019), and then further fine-tuned those models with polar question-answer pairs from the *Circa* corpus. They examined the performance of different models for the classification of polar question-indirect answer pairs into the following meaning categories: 1. STRICT labels: *Yes; No; Probably yes / sometimes Yes; Yes, subject to some conditions; Probably no; In the middle; neither yes nor no; I am not sure; Other; N/A.*, and 2. RELAXED labels: *Yes; No; Yes, subject to some conditions; In the middle, neither yes nor no; Other; N/A.* [1] The study evaluated various baseline models and compared the performance of the models using only questions, only answers, and both questions and answers. The results indicated that joint models (that is, models trained both with questions and answers) outperformed answer-only models. The study also highlighted the challenges of classifying uncertain or ambiguous responses and suggested that incorporating the right information for the task remains a challenge.

Taking inspiration from the research of Louis et al. (2020), Damgaard et al. (2021) studied how to understand indirect answers to polar questions. Instead of crowdsourcing, they collected polar questions and indirect answers from the transcripts of the *Friends* TV series. After manual annotations, they released the FRIENDS-QIA dataset with 5,930 *polar question–indirect answer* pairs in English, both with the majority label and with the raw annotations. They further experimented with Convolutional Neural Networks (CNNs) with different word embeddings: CNN with GloVe embeddings and CNN with BERT embeddings. Furthermore, an additional crowd layer was added to enable the model to learn from the disagreement of human annotators. As a result, CNNs trained with BERT embeddings outperformed CNNs trained with GloVe word embeddings when the model was trained both with questions and answers. Furthermore, using Convolutional Neural Networks (CNNs) to evaluate the task, the authors showed that there was still room for improvement in the interpretation of indirect answers. However, they also found encouraging improvements when explicitly modeling human disagreement in the annotations.

[1] The RELAXED labels were achieved by collapsing the more uncertain and confusing classes from the STIRICT labels: "Probably yes / sometimes Yes" → "Yes", "Probable No" → "No", and "I am not sure" → "In the middle, neither yes nor no".

3 Background

The taxonomy of the response space to questions we use is formally characterized using the KoS framework (Ginzburg, 2012) which provides a theory of dialogue context and dialogue management. The *Question-Specific* responses are the most important subgroup of the taxonomy of the response space to questions. This includes responses providing *answers* (*Direct Answers* and *Indirect Answers*), and *Dependent Questions* where the response to the original question depends on the response to the question-response to that original question. Other subgroups of the taxonomy are the *Metacommunicative* responses (*Clarification Response* and *Acknowledgement*), and the *Evasion* responses (*Motivation*, *Ignore*, *Change the topic*, and *Difficult to Provide an Answer*). Detailed descriptions of each class are presented in Appendix B.

Direct Answers are defined as those that, given a proposition: p, a question: q, p is a direct answer to q, if and only if p is *about* q, and is entailed by either the meet of q's atomic or negative atomic answer set.[2] `Indirect Answers` are distinguished from direct answers under two basic conditions: a). the indirect answer p is not a direct answer to the question q, and b). the indirect answer p, together with a *bridging proposition bridgeprop* (some shared knowledge), entails r, which is a direct answer to the question q. The formal definition of indirect answers is stated as follows:

> Given $p : Prop, q : Question, dgb : DGBType$ InDirectAns(p,q,dgb) iff ¬DirectAns(p,q) and there exist $bridgeprop, r : Prop$ such that DirectAns(r,q) and In(dgb.FACTS, bridgeprop) and $\rightarrow (p \wedge bridgeprop, r)$. (Ginzburg et al., 2022)

As reflected in the definition, the implied direct answer from the indirect answer can be inferred with the help of shared knowledge during the conversation and some domain-independent information. However, in some cases, the interpretation of indirect answers might involve reasoning about the speaker's intentions. Thus, the process of inference will be influenced by the specific perspective, knowledge, goal, or interests of the individual making the inference.

In the following section, we present our methods and processes for collecting a corpus of indirect answers to *wh*-questions.

4 Corpus Collection

We aim to collect the first publicly available corpus of indirect answers to various content questions in English dialogue. To start with, we follow the annotation guidelines for the entire response space of the questions presented in previous works by Ginzburg et al. (2019, 2022), and also updated their annotation guidelines by adding extra instructions specific to indirect answers to *wh*-questions. We annotated various *wh*-questions and their corresponding responses from four different English corpora. Namely, BNC (Burnard, 2007), Cornell-Moive corpus (Danescu-Niculescu-Mizil and Lee, 2011), COCA (The Corpus of Contemporary American English, Davies, 2010), and LLC (The London–Lund corpus of spoken English, Svartvik, Jan, 1990).

4.1 Annotations

There are several steps involved in collecting the corpus of indirect answers to *wh*-questions:

- `Step 1:` we started by investigating the collections of question-answer pairs from the BNC with the response space annotations, shared by the authors of Ginzburg et al. (2022) on the OSF platform.[3] We re-annotated those collections following our updated guidelines and then extracted the *WhQ-IND* pairs.
- `Step 2:` we searched for various *wh*-questions (involving the *wh-* words *what, why, how, which, when, where and who*) and their responses using the SCoRE [4] search engine for the BNC. Table 1 presents the search patterns used for each *wh*-question, the number of examples obtained from them, and also the number of examples we annotated for this study. During this annotation process, we only focused on adjacent pairs of *wh*-questions and their responses, uttered by two distinct interlocutors. In addition, we also eliminated utterances in which the content is unclear (for instance, cases where the main parts of the utter-

[2] For the detailed description of the definition and formalization, see Ginzburg et al. (2022); for a detailed discussion of *Aboutness*, see (Ginzburg and Sag, 2000, pp. 129–149).

[3] `https://osf.io/mq6r7/`

[4] `http://www.dcs.qmul.ac.uk/imc/ds/score/saved.html`

Search Pattern	Search Result	Annotated
^when <V??>, ?	420	98
^where <V??>, ?	1877	94
^why <V??>, ?	1328	656
^how <V??>, ?	1640	359
^what <V??>, ?	7965	318
^who <V??>, ?	1696	366
^which <?N?> <V??>,	225	149
Total	**15151**	**2040**

Table 1: Search patterns from BNC, their results, and the number of annotated examples in this study.

ance are not available and marked with *<unclear>* tag, thereby reducing understanding of the utterance's meaning). As a result, we collected 35 *wh*-question and indirect answer pairs from 2040 examples of annotated *wh*-questions.

- `Step 3:` Ginzburg et al. (2022) reported that the CornellMovie corpus has the highest percentage of indirect answers in their data set. Therefore, we also annotated dialogues from the CornellMovie corpus and collected 12 pairs of *wh*-question and indirect answer pairs.
- `Step 4:` We searched for *wh*-questions and their responses in the conversational part of the London-Lund Corpus of Spoken English (LLC) corpus. This resulted in a total of 21 *wh*-question and indirect answer pairs.
- `Step 5:` we utilized the Corpus of Contemporary American English (COCA) [5], and searched for different types of *wh*-questions using various search patterns. The details of the search patterns are provided in Appendix C. Most of the examples taken from this corpus are from the sub-corpora: Movie, TV, and Spoken. An intern who is studying for a master's degree in English linguistics, specially trained in dialogue semantics, participated in this process. He went through at least 400 examples (around 1200 examples for some *wh*-question types) for each type of question and selected examples that are potential *WhQ-ID* pairs. These examples were then checked by the first author of this paper. In the end, we obtained 390 *wh*-question and indirect answer pairs from around 5000 *wh*-questions from the COCA corpus.

4.2 Corpus Description

The annotation and re-checking processes resulted in a collection of 458 *wh*-question and indirect answer pairs. Among these, 390 examples were selected from the COCA corpus, 35 from BNC, 12 from CornellMovie, and 21 from the LLC corpus. The collected *WhQ-IND* pairs, their annotations, and the updated annotation guidelines are shared with the public on the OSF platform: https://osf.io/zuhvp/.

The number of indirect answers collected for various *wh*-questions also varies. As presented in Table 2, almost half (214 out of 458) of the collected examples are *how*-questions. Other frequent questions are *what*-questions and *why*-questions, 75 and 63 examples, respectively. In addition, we found 32, 31 and 29 examples, respectively, from *where*-questions, *when*-questions and *who*-questions. However, we only found 14 examples from *which*-questions.

***wh*-question**	**No. Indirect answers**
What	75
Why	63
How	214
Which	14
When	31
Where	32
Who	29
Total	**458**

Table 2: Distribution of indirect answers across different *wh*-questions.

Inter Annotator Agreement To evaluate the reliability of the corpus annotation, we performed an experiment to determine whether the response in each dialogue instance within our corpus qualifies as an indirect answer.

In this annotation experiment, four annotators participated: the first author (referred to as First Annotator), an English L2 speaker enrolled in a Ph.D. program in linguistics and an expert in response space annotation tasks; an intern (referred to as Second Annotator), an English L2 speaker pursuing a master's degree in English linguistics; a volunteer native English speaker (referred to as Third Annotator) who is pursuing a master's degree in English linguistics, and another volunteer (referred to as Fourth Annotator), an English L2 speaker enrolled in a Ph.D. program in English linguistics. Before starting the annotation process, all annotators familiarized themselves with the updated annotation guidelines. Additionally, they underwent several training sessions and discussed any disagreements

[5] https://www.english-corpora.org/coca/

together to ensure a shared understanding of the annotation criteria. In the end, they co-annotated 65 *WhQ-IND* pairs from the collected examples. Each of the four annotators, when marking an indirect answer, was also required to infer and supply the implied direct answer from the indirect answer.

We calculated the inter-annotator agreement score among four annotators using Fleiss's Kappa (Fleiss, 1971; Fleiss et al., 2003) and Krippendorff's Alpha (Krippendorff, 2011) methods in Python. As a result, the agreement scores among the four annotators are rather low: Fleiss's κ is -0.51, and Krippendorff's α is 0.025. This indicates substantial disagreement among the four annotators. In addition, we also calculated the inter-annotator agreement level between annotators with the average pairwise Cohen's Kappa scores (Carletta, 1996) using the *Scikit-learn* (Pedregosa et al., 2011) data mining and data analysis tool in Python with its *sklearn.metrics* package. The pairwise Cohen's κ obtained are presented in Table 3. These pairwise agreement scores (0.22 - 0.44) indicate that the agreement between the annotators ranges from fair to moderate agreement.

Annotators	**Cohen's κ**
First vs. Second	0.44
First vs. Third	0.28
First vs. Fourth	0.38
Second vs. Third	0.33
Second vs. Fourth	0.22
Third vs. Fourth	0.36

Table 3: The average pairwise Cohen's Kappa scores between annotators.

The low inter-annotator agreement scores can be attributed to the fact that annotating and interpreting indirect answers is a highly inference-based task with inherent subjectivity and pragmatic complexity. To further address this issue, 60 *wh*-question indirect answer pairs from the collected corpus were randomly selected and then annotated by both authors of the paper (both are experts in the response space classification task). In this way, our aim was to evaluate inter-annotator agreement among expert annotators. Cohen's Kappa score between the two experts is 0.60, which indicates a moderated to substantial agreement between the experts. This agreement score also corroborates the difficulty in annotating *WhQ-IND* pairs.

We hypothesize that the low levels of agreement among annotators arise because identifying indirect answers to *wh*-questions involves a high level of pragmatic complexity. In addition to relying on the annotation guidelines, annotators need to use their semantic and pragmatic knowledge and experience, as well as their subjective judgments for identifying and inferring indirect answers. These low inter-annotator agreement results are also in line with the inter-annotator results reported in Ginzburg et al. (2022), who note a sharp decline when including annotations of indirect answers to calculate annotator agreements on different sets of response types. Yusupujiang et al. (2022) also reported that automatic classification results obtained for indirect answers are pretty low: F1-scores are 0.25 and 0.07 on their full taxonomy and coarser taxonomy respectively. Therefore, the authors suggest that a targeted set of features is necessary to automatically classify indirect answers.

5 Interpreting Indirect Answers to *Wh*-questions

Wh-questions are one of the most commonly observed question types in English conversation. Stivers (2010) reported that among the 328 questions that occurred in a videotaped American English conversation 27% ($n = 90$) of the questions were *wh*-questions. She indicated that the two commonest *wh*-questions types were *what*-questions (38%) and *how*-questions (23%). Other frequent types were *why*-questions (16%) and *when*-questions (12%). *Where*- and *who*-questions only accounted for 8% and 3% of their corpus, respectively. However, the distribution of *wh*-question types can vary depending on many other factors, such as conversational context, cultural and individual communication styles, as well as the specific nature of conversations.

Fox and Thompson (2010) presented the grammatical and interactional characteristics of different responses to *wh*-questions by studying a collection of 73 examples from American English conversations. The authors identified two broader types of responses to the *wh*-questions: *phrasal* and *clausal* responses. Their study suggested that phrasal responses provided simple answers to *wh*-questions, while clausal responses, specifically, clausal Phrase-in-Clause (*PiC*) responses, often signaled trouble with the question or sequences even though they also provided answers. Furthermore, the main types of clausal responses (that

is, full-clause responses) usually did not provide answers to the question, instead, they treated an assumption in the question as problematic or provided "no-access" responses, such as *I don't know*, or *he/she/they don't know*. It is worth mentioning that, the *"treating an assumption as problematic"* function of the full-clause responses corresponds to the "Clarification Response", precisely, the "Correction" response type, while the *"no-access" responses* correspond to the "Difficult to provide an answer" response type in the response space taxonomy provided by Ginzburg et al. (2019, 2022).

5.1 Information Sources

Ginzburg et al. (2022) proposed to categorize indirect answers into two main types: *shallow* and *deep* indirect answers. Shallow indirect answers are those where the implied direct answers are inferred only based on some shallow shared knowledge and domain-independent erotetic reasoning (also known as interrogative or questioning reasoning); whereas deep indirect answers require reasoning about the speaker's intentions, beliefs, and some domain-specific knowledge. Therefore, based on their suggestions, we further divide the information that one might need to interpret indirect answers into 9 categories as follows:

Basic linguistic knowledge: this is based on significant competence in the language used (grammar, vocabulary, etc.). As in Dialogue (1), the word (*daily*) used in the indirect answer helps questioner A to infer the implied direct answer from B's indirect answer, which is *"The last time it was inspected was yesterday/today."* Thus, A is required to have a good understanding of basic English grammar and vocabulary for interpretation.

(1) A: When was the last time that line was inspected, commander?
B: It's inspected daily. [COCA Corpus]

Shared knowledge: this involves shared or communally established knowledge during conversations.

(2) *previous utterances*:I also had extraordinary hearing. During dinner, I could tune out the cacophony of chewing, slurping, chewing, cutlery scraping against plates, chewing, . . .
A: Why aren't you eating, Sheldon?
B: How can I with that horrible noise? [COCA Corpus]

From the previous utterances in Dialogue (2), one learns that Sheldon has very sensitive hearing. Therefore, the noise around Sheldon is the reason he is not eating. In contrast, in Dialogue (3), by providing the indirect answer *"Look what happened in 2018."*, Speaker B invites Speaker A to recall events that happened in 2018 to infer the direct answer to his question. Here, Speaker B believes that Speaker A shares the same communal memory as he does, and is capable of finding the requested information in this way.

(3) *previous utterances*:AXELROD: Yes. So, that lack of enthusiasm if it's Joe Biden, right, on the one side, Donald Trump on the other, I can tell you whose voters are going to be more enthusiastic.
A: Well, how do you know that? How do you know that?
B: Look what happened in 2018. [COCA Corpus]

Speaker's intentions/goals: the speaker conveys the messages indirectly by mentioning her/his goals or intentions. As shown in Dialogue (4), we can learn of Speaker B's intentions of *"[getting] married to that woman"*, so can infer the direct answer that the person that Speaker B is talking to is his girlfriend.

(4) A: Who are you talking to? Your girlfriend? I didn't know you had a girlfriend.
B: I'm probably gonna marry this one. [COCA Corpus]

Speaker's belief/interest: some indirect answers convey speakers' beliefs or interest in a subject/topic, so correctly identifying these is the way to interpret the direct answer to the original *wh*-questions.

(5) A: Man, how do you know this shit's safe?
B: These guys know what they're doing. Don't worry. They've tested it on dogs and everything. [COCA Corpus]

In Dialogue (5), Speaker B indicates her/his trust in the ability of those group of people who

invented the (*medical items or drugs*). Therefore, Speaker B's full trust in those people is the basis for her/him to (believe he) know(s) that the item invented by those people is safe.

Relationships between speakers: Indirect answers can be used between strangers to be polite and to exude more professionalism, or to avoid conflict in an employer-employee relationship. On the other hand, among close friends or family members, indirect answers might be used to make the conversation more casual based on their vast amount of shared knowledge. Thus, in Dialogue (6), Speaker B's response, *"Like you don't know."* indicates that Speaker A already knows the reason based on their relationship and shared history. However, a third party might not be able to infer Speaker B's implied direct answer because of not being in that relationship.

(6) *previous utterances*:Carl: Okay, here she is. She'll clear up this whole thing. What are you doing here?! Uh, Carl... What's goin' on? It's not what it looks like.
A: Why are you wearing that?
B: Like you don't know. [COCA Corpus]

Nuanced linguistic features: these include idioms, slang, figurative expressions etc. As in Dialogue (7), the figurative expression *"I'm right inside your head."* usually implies that she/he understands the other person's thoughts, feelings, and motivations.

(7) A: How do you know that?
B: I'm right inside your head. [COCA Corpus]

Common sense: this involves common knowledge about the world, certain social norms, customs, etc. In order to infer the implied direct answer *"I'm not very hungry now"* to the question about Speaker B's hunger level in Dialogue (8), one is required to understand what *"being flexible about eating time"* means.

(8) PREVIOUS UTTERANCES:Would you like to suggest a time for eating? Would I? Either of you
A: <laughs> how hungry are you Ken? <laughs>
B: I can I could eat now, or I could manage to wait. I'm quite flexible. [LLC Corpus]

Visual context can provide important cues for interpreting indirect answers, especially when analyzing multimodal dialogue settings. The Dialogue (9) is taken from the CornellMovie corpus, so is a dialogue in a movie scenario. Both speakers are in the same physical space and, hence, share visual context. Thus, Speaker A can identify the person requested by looking in the direction provided by Speaker B, *"At the end of the bar."*.

(9) A: Who said that?
B: At the end of the bar. [CornellMovie Corpus]

Non-verbal cues: we can utilize tone of voice, facial expressions, body language, etc. to better understand speakers' motivations and intentions. This is very useful when considering multimodal dialogues. For instance, in the constructed example of Dialogue (10), the parent can infer from the child's guilty facial expression and body behaviors that the child broke the window.

(10) *scenario*:A parent enters a room and notices a broken window. So the parent initiates the following dialogue:
A: Who broke the window?
B: (The child looks guilty and tries to avoid eye contact with the parent.) [Constructed example]

5.2 Statistical Analysis of Information Sources

To study which information sources are more frequently needed for the interpretation of indirect answers to *wh*-questions, we conducted a pilot study using the examples in our collected corpus of *WhQ-IND* pairs. The first author of this paper selected 141 examples (examples whose indirectness has been annotated with high confidence) for annotation with the 9 possible information sources presented above in Section 5.1 as a pilot study.

As indicated in Table 4, `Basic linguistic knowledge` (30.50%) and `Common Sense` (24.11%) are the two most frequent information sources used for inferring direct answers from indirect answers. The third frequently used information source is the `Nuanced linguistic features` in the indirect answers, which accounts for 14.18% of all information sources in our anno-

Information Source	How	Why	What	When	Where	Which	Who	Freq. %
Basic linguistic knowledge	11	7	13	10	0	0	2	30.50% (43)
Common sense	23	2	3	1	1	1	3	24.11% (34)
Nuanced linguistic features	14	2	2	0	0	1	1	14.18% (20)
Shared knowledge	7	4	3	0	0	0	0	9.93% (14)
Speaker's intentions/goals	5	2	1	0	0	1	4	9.22% (13)
Speaker's beliefs/interests	7	4	1	0	0	0	0	8.51% (12)
Relationships between speakers	0	1	2	0	0	0	0	2.13% (3)
Visual context	1	0	0	0	0	0	1	1.42% (2)
Non-verbal cues	0	0	0	0	0	0	0	0 %
Total	**68**	**22**	**25**	**11**	**1**	**3**	**11**	**141**

Table 4: Distribution of information sources.

tations. Furthermore, the `Shared knowledge`, `Speaker's intentions/goals`, and `Speaker's beliefs/interests` have similar distributions, which are 9.93%, 9.22%, and 8.51% respectively. Other types of information sources seem to have quite lower frequency: `Relationships between speakers` (2.13%), `Visual context` (1.42%), and `Non-verbal cues` (0%).

In addition, we can learn from Table 4 that, most of the indirect answers to *how*-questions can be interpreted based on `Common sense` and `Nuanced linguistic features`. For *what-* and *when*-questions, `Basic linguistic knowledge` seems to be used more in interpreting their indirect answers. However, due to the imbalanced number of examples for each type of *wh*-question in our current data set, our results concerning the distribution of information sources must be viewed as quite provisional.

6 Generation of INDs to *wh*-questions

As a pilot study, we fine-tuned the pre-trained response generation model DialoGPT (medium) (Zhang et al., 2020) with our collected corpus of indirect answers to *wh*-questions (458 examples), and tested the fine-tuned model's ability to generate indirect answers to *wh*-questions in a new test set.

Experimental Setup We fine-tuned our model by using Hugging Face's *"Transformer"* library. During the training, we randomly split the corpus into training and evaluation sets with a ratio of $4:1$. We set the number of training epochs to $num\_train\_epochs = 10$, with a per device training batch size of 4. The model also saves its result every $10,000$ steps, while also applying a weight decay of 0.01 to avoid overfitting. In addition, we adopted a step-wise evaluation strategy *evaluation_strategy="steps"*, to evaluate the model every 500 steps during the training phase. Furthermore, we set *load_best_model_at_end=True*, to load the model that had the best performance during the evaluation steps. Finally, the input format of the data for fine-tuning is *"[PH] Previous dialogue history + [Q] Wh-Questions + [R] indirect answers +* <|*endoftext*|>*"*.

Evaluation We tested the performance of the fine-tuned model on 20 new *wh*-questions selected from the annotated 2040 examples of BNC *wh*-questions, where the original responses to these 20 examples were direct answers. We then deleted their original direct answers and created a test set with a format, *"[PH] Previous dialogue history + [Q] Wh-Questions + [R]"*. The fine-tuned model generated responses to those new *wh*-questions, and we evaluated the performance of the model by manually determining if the model-generated responses were indirect answers. However, only one example in 20 was an indirect answer. Details of the generated responses are presented in Appendix A for reference.

7 Conclusions and Future Work

In this paper, we have addressed the challenge of interpreting indirect answers to *wh*-questions. We started by collecting indirect answers to *wh*-questions from four different English corpora (BNC, CornellMovie, COCA, and LLC), and constructed a small corpus of 458 *WhQ-IND* pairs along with pre-question utterances and post-response utterances. Building such a corpus is highly labour intensive, given the difficulty of the task of classifying responses as indirect, as we demonstrated in several inter-annotator studies, including ones involving expert annotators.

In addition, we developed a scheme of 9 possible information sources used to infer direct answers from indirect answers and found—provisionally,

given problems with the imbalanced nature of our data set— that *Basic linguistic knowledge, Common sense*, and *Nuanced linguistic knowledge* are the three most frequently used information sources for the interpretation of indirect answers to *wh*-questions. Finally, we also conducted a preliminary experiment for generating indirect answers to *wh*-questions by fine-tuning a large-scale response generation language model, DialoGPT. The results of this latter experiment are hampered by the small amount of our current data set, but also suggest that this is a tricky task for GPT.

There are several clear limitations of the current study, which future work should improve on: (1). Since the size of the collected corpus is small, there is a need to continue collecting a more balanced and larger corpus of indirect answers to *wh*-questions; (2). The proposed 9 possible information sources need to be further evaluated, related to established components of context, and tested across annotators; (3). We hope to improve the performance of our generation model by fine-tuning it with a larger corpus. Other methods, such as few-shot learning, data augmentation, and transfer learning techniques may help improve the model performance on generating indirect answers to *wh*-questions.

Acknowledgements

We acknowledge the support of a public grant overseen by the French National Research Agency (ANR) as part of the program *Investissements d'Avenir* (reference: ANR-10-LABX-0083). It contributes to the IdEx Université Paris Cité-ANR-18-IDEX-0001. In addition, we would like to thank three anonymous reviewers for their thorough and detailed feedback. Finally, we thank Boško Rajkovic, Miloš Milisavljevic, Jacob Rigal, and Alatafe Abulimiti for assisting us during the annotation process as well as for the very interesting discussions about this work.

References

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Lou Burnard, editor. 2007. *Reference guide for the British National Corpus (XML Edition)*. Oxford University Computing Services on behalf of the BNC Consortium. Acess 20.03.2017.

Jean Carletta. 1996. Assessing agreement on classification task: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Cathrine Damgaard, Paulina Toborek, Trine Eriksen, and Barbara Plank. 2021. "I'll be there for you": The One with Understanding Indirect Answers. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 1–11, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Mark Davies. 2010. The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, 25(4):447–464.

Marie-Catherine de Marneffe, Scott Grimm, and Christopher Potts. 2009. Not a Simple Yes or No: Uncertainty in Indirect Answers. In *Proceedings of the SIGDIAL 2009 Conference*, pages 136–143, London, UK. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2010. "Was It Good? It Was Provocative." Learning the Meaning of Scalar Adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 167–176, Uppsala, Sweden. Association for Computational Linguistics.

Marie-Catherine de Marneffe and Judith Tonhauser. 2016. Inferring Meaning from Indirect Answers to Polar Questions: The Contribution of the Rise-Fall-Rise Contour. In *Questions in discourse*, pages 132–163. Brill, Leiden, The Netherlands.

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

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2003. *Statistical Methods for Rates and Proportions*. John Wiley & Sons.

Barbara A Fox and Sandra A Thompson. 2010. Responses to Wh-Questions in English Conversation. *Research on Language and Social Interaction*, 43(2):133–156.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.

Jonathan Ginzburg and Ivan A. Sag. 2000. *Interrogative Investigations: the form, meaning and use of English Interrogatives*. Number 123 in CSLI Lecture Notes. CSLI Publications, Stanford: California.

Jonathan Ginzburg, Zulipiye Yusupujiang, Chuyuan Li, Kexin Ren, Aleksandra Kucharska, and Pawel Lupkowski. 2022. Characterizing the response space of questions: data and theory. *Dialogue & Discourse*, 13(2):79–132.

Jonathan Ginzburg, Zulipiye Yusupujiang, Chuyuan Li, Kexin Ren, and Paweł Łupkowski. 2019. Characterizing the Response Space of Questions: a Corpus Study for English and Polish. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 320–330, Stockholm, Sweden. Association for Computational Linguistics.

Nancy Green and Sandra Carberry. 1994a. Generating indirect answers to yes-no questions. In *Proceedings of the Seventh International Workshop on Natural Language Generation*.

Nancy Green and Sandra Carberry. 1994b. A Hybrid Reasoning Model for Indirect Answers. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 58–65, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Nancy Green and Sandra Carberry. 1999. Interpreting and generating indirect answers. *Computational Linguistics*, 25(3):389–435.

Klaus Krippendorff. 2011. Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2):93–112.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding Indirect Answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

John R Searle. 1975. Indirect speech acts. In *Speech acts*, pages 59–82. Brill.

Tanya Stivers. 2010. An overview of the question–response system in American English conversation. *Journal of Pragmatics*, 42(10):2772–2781.

Svartvik, Jan, editor. 1990. *The London–Lund corpus of spoken English : Description and research*, volume 82 of *Lund Studies in English*. Lund University Press. Book Editor.

Junya Takayama, Tomoyuki Kajiwara, and Yuki Arase. 2021. DIRECT: Direct and Indirect Responses in Conversational Text Corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1980–1989, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zulipiye Yusupujiang, Alafate Abulimiti, and Jonathan Ginzburg. 2022. Classifying the response space of questions: A machine learning approach. In *SemDial 2022-26th Workshop on the Semantics and Pragmatics of Dialogue*, pages 59–69.

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

Appendix A: Details of the responses generated as indirect answers

Generated Responses:

(1) *[PH] And she's a bit of a slag. She don't really, you know, I mean she's got a bit of a reputation. I suppose everyone has but I hear about her a lot. In school and everything. Which one was that? Pardon?*
Question: Which one was this?
Response: Pardon. I think I know who you are.

(2) *[PH] Is it a big fire? Is it? You've seen it. I can't remember.*
Question: Which room is it in?
Response: I don't know. I left the building when I was nine.

(3) *[PH] Let me put that question to the hundred women here, are you looking forward to Christmas? Button one for yes, and button two for no. And, this hundred don't share my apparent cynicism, seventy three of them are, twenty seven said no. Of those who said yes, why?*
Question: What are you looking forward to?
Response: I'm looking forward to a good time.

(4) *[PH] Any problems with these? Erm no. They they they were they were straightforward, just. Okay.*
Question: How did you do this one?
Response: I just, uh... I used a program called CamScanner.

(5) *[PH] We want a difference of nineteen somewhere don't we? That's easy. We drop the calculator everywhere. Well you can drop the calculator if you want it would be better if you pick 'em up and stick the numbers in.*
Question: How're you gonna do?
Response: Well, I'm going to use my powers for good. (Indirect Answer)

(6) *[PH] So nine hundred, it works out at nine pound odd What would two thirds of three quarters come to roughly. Just sort of thinking about it having a guess. A half.*
Question: How did you get that?
Response: <pause> I think I know. I think I know. I think I know.

(7) *[PH] What I, where I gonna go? You're not going anywhere, you're staying here Can I help you plant the? Bulbs, er probably not cos I think I'll have to do while you're at school cos daddy tells me it's gonna rain this afternoon, I'll have to do that this morning, pity that.*
Question: Why can't we do them tomorrow?
Response: I don't know, I think it's best if we wait till tomorrow.

(8) *[PH] he save, he keeps swords. He saves those. Well, oh do they save them? Mm.*
Question: Why do they save them?
Response: Well, they're not exactly cheap.

(9) *[PH] Do you want mum to come to Argos with me tomorrow morning? Why? Well she isn't, I don't think she's doing anything tomorrow morning.*
Question: Are you asking me? Why are you asking me?
Response: Why are you asking me? Why are you asking me?

(10) *[PH] Yes, Christmas eve, well. Did you have many guests at sunrise? Well we, we had about sixty erm, people that trudged up to the top of the hill including erm, one who didn't know there was a wedding, he was a runner, and then he and he had tagged the top of Arthur's Seat and was astonished to see a crowd there.*
Question: What do you wear to a mountain top wedding in the middle of winter?
Response: Well, I'm a guy so I don't really have to worry about that sort of thing.

(11) *[PH] Erm forty nine. Alright and five squared? Twenty five. Right.*
Question: What's forty nine and twenty five come to?
Response: Erm five centimeters per second.

(12) *[PH] yes It means you can pack in a lot more things in your day, you can work, you can look after your children, you can take them out, you can do a lot of things that you couldn't do otherwise. mm, mm, yeah Er from the necessity in an area with no public transport*
Question: where's that?
Response: Er, I don't know, I moved to London. And I'm not even a very good driver.

(13) *[PH] Yeah? Negative one. Now the tricky one, let's do plus four make it plus five, plus five a a takeaway a negative three. Okay?*
Question: Where do we start?
Response: Well, I'm sure we can come up with a few ideas.

(14) *[PH] Right. Twelve over three and Okay. Now the notation, do you understand the notation?*
Question: Where's the notation?
Response: I don't know, I'm just?

(15) *[PH] I would only been six or seven when the war started. It started in nineteen fourteen. First I mind about it was the first zeppelin that came over. Great big thing like a huge balloon and lived out at Ness and I mind me mother coming and getting us early and this great big thing come over.*
Question: Who was in it?
Response: Zebediah Kerman, a Russian spy that was captured during the Russian Revolution.

(16) *[PH] Yes I'm getting you out treasure. This this er thing's got stuck on the rails. he's got Yes but one of them's going to get in.*
Question: Who's going to get in first and who's going to get in later?
Response: I'm gon na get in there, but I can't get in the car.

(17) *[PH] What are you going to do? I'm, making another lot of marmalade. Again?*
Question: Who's going to eat it all?
Response: I'll make it up to you.

(18) *[PH] No, but I just want you to tell me. Science. With whom? I don't know.*
Question: When are you supposed to come up here again?
Response: Well, I'm supposed to be in Las Vegas by the time I'm eighteen.

(19) *[PH] Oh yeah. do the one that you can do. Okay, well. What what we need to do, is make sure these scripts are done before half term really, don't we.*
Question: When's half term?
Response: It starts in a few minutes.

(20) *[PH] Have you had any erm problems in the past? No. No problems at all? No.*
Question: When was the last time you saw your doctor?
Response: I haven't.

Appendix B: Taxonomy for Responses to Queries

Full-Taxonomy	Description
Direct Answer **(DA)**	the response directly offers an answer to the question.
Indirect Answer **(IND)**	the answer to the question can be indirectly inferred from this utterance.
Dependent questions **(DP)**	the answer to the original question depends on the answer to this query response.
Clarification Response **(CR)**	the speaker asks for extra information to confirm (s)he understood the question correctly, requires additional information to understand it better, or provides some information to clarify/correct misinformation from the previous utterance.
Acknowledgement **(ACK)**	the speaker acknowledges that (s)he heard the question, such as mhm, aha,... etc.
Motivation **(MOTIV)**	a query response about the motivation of asking the initial question.
Ignore **(IGNORE)**	the utterance does not relate to the question, but to the situation.
Change the topic **(CHT)**	the utterance signals that the speaker does not want to answer the question, instead (s)he changes the topic, and gives an evasive response.
Difficult to provide an answer **(DPR)**	the speaker indicates that (s)he does not know the answer, or it is difficult for her/him to provide an answer, so points at a different information source,
OTHER	utterance that does not fit in any of the categories above.

Appendix C: Details of search patterns and annotated questions from the COCA corpus

Search Pattern	Annotated Questions	Number of INDs
what * * * PUNC	What do you think?	28
what are * * *	What are you * ?	14
	What are you going to * ?	17
why are * * * PUNC	Why are you doing this?	27
	Why are you still here?	3
	Why are you following me?	6
	Why are you calling me?	1
	Why are you so nervous?	2
	Why are you so happy?	1
	Why are you wearing that?	3
	Why are you protecting him?	1
	Why aren't you eating?	1
	Why are you so calm?	1
	Why are you ignoring me?	1
	Why are you helping us?	1
	Why are you here?	1
	Why are you so late?	1
	Why are you so surprised?	1
how do * * * PUNC	How do you know that?	35
	How do you do that?	7
	How do you explain that?	25
	How do you know this?	40
	How do you figure that?	13
	How do we do that?	26
	How do you feel?	51
which one * * * PUNC	Which one do you want?	3
	Which one do you like?	5
	Which one do you think?	2
who was * * * PUNC	Who was on the phone?	1
who is * * * PUNC	Who is responsible for this?	1
who are * * * PUNC	Who are all these people?	2
	Who are you working for?	2
	Who are you looking for?	1
	Who are you voting for?	1
	Who are you talking to?	14
	Who are you talking about?	1
when was * * * PUNC	When was the last time?	23
when is * * * PUNC	When is he coming back?	2
where did * * * PUNC	Where did you get that?	21
where * * * PUNC	Where is he now?	4
Total		**390**

A New Dataset for Causality Identification in Argumentative Texts

Khalid Al-Khatib ♠ **Michael Völske** ‡♣ **Shahbaz Syed** † **Anh Le** ‡
Martin Potthast †◇ **Benno Stein** ‡

♠University of Groningen ‡Bauhaus-Universität Weimar
†Leipzig University ◇ScaDS AI ♣Artefact Germany GmbH
<khalid.alkhatib@rug.nl>

Abstract

Existing datasets for causality identification in argumentative texts have several limitations, such as the type of input text (e.g., only claims), causality type (e.g., only positive), and the linguistic patterns investigated (e.g., only verb connectives). To resolve these limitations, we build the WEBIS-CAUSALITY-23 dataset, with sophisticated inputs (all units from arguments), a balanced distribution of causality types, and a larger number of linguistic patterns denoting causality. The dataset contains 1485 examples derived by combining the two paradigms of distant supervision and uncertainty sampling to identify diverse, high-quality samples of causality relations, and annotate them in a cost-effective manner.

1 Introduction

Causality identification is a vital task in natural language processing that can contribute to different downstream applications such as question answering, fact-checking, and commonsense reasoning. The task which concerns identifying texts with causality relations, the type of relations (positive or negative), and the concepts involved in the relations, is studied in diverse domains including biomedicine (Kyriakakis et al., 2019), education (Stasaski et al., 2021), and recently computational argumentation (Al-Khatib et al., 2020).

In computational argumentation, causality identification impacts fundamental tasks such as topic-independent argument mining and building large-scale argumentation graphs (Reisert et al., 2018). Despite its importance, only a few annotated datasets for identifying causality have been built so far. Moreover, these datasets often focus only on one argument component (e.g., claim), encode bias towards the 'positive' type of causality, and/or consider a limited number of linguistic patterns that capture causality (e.g., verb connectives). As such, developing *robust* supervised learning approaches based on these datasets for causality identification becomes more laborious.

This paper aims to expand and enrich the available data for causality identification in argumentative texts written in English with WEBIS-CAUSALITY-23, a new dataset comprised of 1485 examples of more sophisticated input text (i.e., the whole argument), covers more causality patterns, and maintains a balanced distribution of causality types. To this end, we develop an approach that comprises two main steps: *distant supervision* and *uncertainty sampling*. First, we identify 10,329 candidate sentences for causality via distant supervision. Next, we employ uncertainty sampling on these candidates and manually annotate 1485 argumentative sentences (via crowdsourcing), 867 of which contain at least one causal relation. Of these, 515 sentences are further annotated as containing a positive cause-effect relation, and 536 as containing a negative one. Many sentences encode multiple relations, and involve diverse linguistic patterns (see Section 3).

We train transformer-based classifiers using our newly built dataset, and reach high effectiveness in identifying causal relations compared to several baselines. The developed resources in the paper (e.g., data and code) are made freely available.[1]

2 The WEBIS-CAUSALITY-23 Dataset

In this section, we describe our method for constructing the dataset. In particular, we first outline the distant supervision step, then, we discuss the uncertainty sampling.

2.1 Distant Supervision

Distant supervision is the process of mining suitable training examples from weakly labeled data sources using task-specific heuristics (Mintz et al.,

[1] https://github.com/webis-de/SIGDIAL-23

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 349–354
September 11–15, 2023. ©2023 Association for Computational Linguistics

2009). These examples can then be used for supervised learning of the task at hand. Here, we employ distant supervision to find argumentative sentences that are more likely to encode causality relations, without being restricted to certain topics or linguistic patterns. Specifically, we first collect pairs of concepts that are involved in a causality relation, using the corpus of Al-Khatib et al. (2020). Second, we acquire a set of argumentative texts and segment them into self-contained sentences. Lastly, we extract the argumentative sentences that contain at least one of the concept pairs.

Concept Pair Collection In this step, we utilize the corpus of Al-Khatib et al. (2020) to collect various concepts related to causality. The corpus covers 4740 claims extracted from Debatepedia – an online debate portal. Each claim is manually annotated for the presence of a causality relation (called 'effect'), the type of the relation (positive or negative), and the concepts that are involved in the found relation. For example, the claim "legalization of drugs increases drug consumption" exhibits a *positive effect relation* involving the concepts of *legalization of drugs* and *drug consumption*.

We carefully review these concepts and manually perform two filtering steps: (1) we simplify the complex concepts (e.g., from *"the state can regulate the sale"* to *"sale regulation"*), and we (2) split some concepts into multiple ones (e.g., *"crime and safety problems"* is split into *"crime"* and *"safety problems"*). Overall, we end up with 1930 unique concepts grouped into pairs, each of which consists of two concepts involved in the same relation (e.g., *"legalizing marijuana"* and *"safety"*).

Argumentative Data Acquisition and Simplification We rely on the Args.me corpus (Ajjour et al., 2019) as the source of the argumentative data. The corpus includes 387,606 arguments from various debates regarding controversial topics. The arguments are derived from four popular debate portals: Debatewise (14,353 arguments), IDebate.org (13,522 arguments), Debatepedia (21,197 arguments), and Debate.org (338,620 arguments).

To split the arguments into coherent and self-contained sentences, we use Graphene (Cetto et al., 2018), an open information extraction tool. This tool performs discourse simplification, in which an input sentence is syntactically simplified and split (if necessary) into sentences with resolved coreference and high coherence. Altogether, the arguments are segmented into 10,720,451 sentences.

Concept Pairs and Argumentative Data Matching In this step, for each sentence in the acquired argumentative data, we check whether it includes any of the concept pairs. Using full-string matches between the concepts and the sentences' tokens (after stemming with Porter Stemmer (Porter, 1980)), we obtain around 28,000 sentences that match at least one concept pair. We additionally filter them by removing duplicates, all hyperlinks, and special characters contained in the sentences. Besides, on manual inspection, we observed that matching with generic concepts such as *"individuals"* or *"corporation"* lead to noisy sentences not actually containing any causality relations and were therefore excluded. As a result, we end up with 10,329 sentences. To evaluate the filtering process, we check a random sample of 100 sentences before and after filtering. We observe an increase in the number of sentences with causality relations (from 56 to 70).

2.2 Uncertainty Sampling

Uncertainty sampling is one of the strategies employed in active learning (Settles, 2012). Given an initial classification model and a pool of unlabeled samples, the goal is to select those samples for labeling for which the classifier's confidence is lowest, i.e., the predicted class distribution is closest to uniform, and thus maximize the information gain to the model. Following this idea, we train causality identification models on the labeled samples in the Al-Khatib et al. (2020) dataset, and use the argumentative sentences acquired from the distant supervision step as the unlabeled pool. Next, based on the confidence of these models, we sample a subset of the sentences and annotate them manually via crowdsourcing.

Candidate Sentence Selection Causality identification is often comprised of three classification sub-tasks; given an input text, (1) detect whether the text contains a causality relation, (2) identify the type of causality, and (3) determine the entities or events representing the cause and effect relation.

For the first two sub-tasks, we develop several classification models using the corpus of Al-Khatib et al. (2020) containing labels for the causality relation ('effect'), and the type of relation (positive or negative). The models are based on XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2019), BERT (Devlin et al., 2019), NBSVM (Wang

and Manning, 2012), and Fasttext (Joulin et al., 2016). The implementation is done using the HuggingFace library (Wolf et al., 2020) with default settings. In particular, RoBERTa and XLNet achieve high effectiveness with F_1 scores of 0.88 and 0.91 for the first and second tasks respectively, compared to 0.81 and 0.86 achieved by the feature-rich SVM approach of Al-Khatib et al. (2020). We apply the two best-performing transformer-based models (RoBERTa and XLNet) to the 10,329 argumentative sentences obtained from the distant supervision step. Using these models' confidence scores, we distribute the sentences into nine bins: the first bin represents the highest confidence for the 'no-causality' class, and the last bin represents confidence for the 'causality' class.

We aim to find sentences that encode new causality patterns while maximizing the number of sentences with the 'negative-causality' class. Thus, our uncertainty sampling filters out the sentences with high confidence for the 'causality', 'no-causality', and 'positive-causality' classes. This results in 1937 sentences for our manual annotation.

Sentence-level Manual Annotation We conduct an annotation task for causality identification via Amazon Mechanical Turk for the 1937 sampled sentences,which requires identifying all the causality relations in a sentence. In particular, for each identified causality relation, the workers are asked to specify the causality relation's type, the concepts involved in the relation, and the sentence's phrase(s) that indicate the presence and type of the relation. The workers also have the option to point out sentences that are not comprehensible or/and include several grammatical errors. The task instructions are carefully explained using written guidelines and demonstration videos, covering various causality relations with different linguistic patterns.

We first ask three experts in computational linguistics to annotate 100 sentences, and use their feedback to refine the guideline and improve the annotation interface for the crowdsourcing task. Each sentence is annotated by three different workers. For quality control, we hire native English speakers with a task approval rate of at least 98%. We closely monitor and review the annotations, rejecting workers that perform poorly. In total, 285 workers successfully participated in our task, resulting in 1485 sentences with high-quality annotations.

	Causality	**Positive**	**Negative**	**Multiple**
Expert	0.34	0.66	0.70	0.28
Crowd	0.27	0.31	0.36	0.03

Table 1: Inter-annotator agreement (Krippendorff's alpha) for the expert and crowdsourcing annotations.

We pay a fair hourly wage for the annotators.

3 Dataset Analysis

In this section, we present both qualitative and quantitative analyses of the WEBIS-CAUSALITY-23 dataset. The qualitative analysis encompasses an examination of inter-annotator agreement, dataset statistics, and identified patterns within the dataset. On the other hand, the quantitative analysis involves leveraging the constructed dataset to develop a new causality classifier.

3.1 Qualitative Analysis

Inter-annotator Agreement The inter-annotator agreement, measured using Krippendorff's alpha and presented in Table 1, provides insights into the level of agreement among both experts and crowds. While the crowd's agreement is relatively lower compared to experts, they still achieve a reasonable level of agreement for causality and types (ranging from 0.27 to 0.36). However, the crowd tends to prioritize annotating only one relation per sentence, potentially overlooking instances with multiple relations. These findings highlight the subjective nature of the task and the intricate linguistic patterns within the sentences. It is worth noting that the majority of cases fall into the scenario where two out of the three annotators agree, which significantly helps in obtaining a reliable gold standard.

Dataset Statistics The annotations are aggregated based on majority vote, with one exception: we consider a sentence to have multiple relations as long as at least one annotator found multiple relations there. Table 2 shows statistics for our dataset: there is a high percentage of causal relations, especially of the negative type; a quarter of the sentences contain more than one relation. This demonstrates the cost-effectiveness of our construction method; we obtain a rich set of causal sentences by annotating only 1937 examples. The annotation study costs around 400 EUR.

Dataset Inspection We manually examine the dataset, exploring the causality linguistic patterns

	Expert		Crowd	
Causality				
Overall	80	100%	1324	100%
Relation	48	60%	819	62%
No Relation	32	40%	505	38%
Relation Type				
Overall	48	100%	819	100%
Positive	29	60%	486	59%
Negative	29	60%	507	62%
Multiple Relations				
Overall	48	100%	819	100%
Single	34	71%	614	75%
Multiple	14	29%	205	25%

Table 2: Sentence statistics of the WEBIS-CAUSALITY-23 dataset. Relation type percentages do not sum to 100 since sentences can have multiple relations.

X $\overset{\text{positive}}{\longmapsto}$ **A, B** Social media^{X} can fuel anxiety^{A} and depression.^{B}
X $\overset{\text{negative}}{\longmapsto}$ **A, C, D; X** $\overset{\text{positive}}{\longmapsto}$ **B** GM foods^{X} are safe for human consumption, reduce pesticide,^{A} increase yield,^{B} decrease cost,^{C} and combat global warming.^{D}
X $\overset{\text{negative}}{\longmapsto}$ **A, B, C, D; Y** $\overset{\text{positive}}{\longmapsto}$ **A, B, C, D** Marijuana^{X} can relieve certain types of pain,^{A} nausea,^{B} vomiting^{C} and other symptoms^{D} caused by such illnesses as cancer.^{Y}
X $\overset{\text{negative}}{\longmapsto}$ **Y; X** $\overset{\text{positive}}{\longmapsto}$ **<Z** $\overset{\text{negative}}{\longmapsto}$ **Y>** Genetic screening^{X} for the embryos can reduce the chance of giving birth to more than one child;^{Y} because clinics^{Z} now want to prevent this by planting one embryo at a time and they have to do this through genetic screening.

Table 3: Examples of the found patterns for causality in the set of the sentences with multiple relations.

and the structure of the sentences with multiple relations. As for the linguistic patterns, we look at the list of phrases (provided by the annotators) that indicate a causality relation, finding different causal connectives such as verbs (*"prevent, "promote"*), verb phrases (*"leads to"*), conjunctions (*"because"*), prepositional phrases (*"because of, due to"*), and clauses (*"the source of, is an addition to, can be tied to, becomes a burden for"*).

Besides, we find different patterns for causality in the sentences that contain multiple relations. Examples of these patterns are shown in Table 3. The examples exemplify diverse levels of complexity in encoding relations within different argument components. For instance, the last example demonstrates relations found in a complete argument.

3.2 Quantitative Analysis

To evaluate the impact of our constructed dataset, we employ it to develop a new classifier for causality identification. We compare the effectiveness of this classifier to another one that is developed using the corpus of Al-Khatib et al. (2020).

To build a classifier based on our new dataset, we first split the dataset into training (80%) and test (%20) sets. The split considers the topics of the sentences, placing sentences with the same topic in either the training or test sets.

For evaluation, we tackle the task of identifying whether a sentence has causality relation(s), implementing three classifiers based on the XLNet model: (C_1) this classifier is trained by the training set of Al-Khatib et al. (2020), (C_2) this classifier is trained by the training set of our new constructed dataset, and (C_3) this classifier is trained by the combination of the training sets of the new and old classifiers. We focus on causality identification because Al-Khatib et al. (2020) do not consider multiple relations, making their dataset partially incompatible for causality type identification.

We apply the three classifiers to the test set of Al-Khatib et al. (2020) (D_1), the test set of our new dataset (D_2), and both test sets combined (Table 4). In general, the classifier trained on (D_2) outperforms the baseline, and using the classifier that is trained with the combined training set (C_3) always leads to the best effectiveness, which speaks for the positive impact of our new dataset.

Classifier	D_1	D_2	D_1+D_2
C_1	0.88	0.63	0.82
C_2	0.74	0.71	0.74
C_3	**0.89**	**0.75**	**0.85**
Majority Class Baseline	0.64	0.53	0.62
(Al-Khatib et al., 2020)	0.81	-	-

Table 4: F_1 scores for causality identification. D_1 is the test set of Al-Khatib et al. (2020), D_2 is our test set.

4 Related Work

In general, causality datasets are expensive to build, scarce, small, biased towards one class, focused on only a single aspect of causality (e.g., whether a sentence has a causal relation or not), and include limited linguistic patterns (due to their sampling method, e.g., via a seed list of causal verbs). Recently, Xu et al. (2020) reviewed six publicly-available datasets. The largest, AltLex (Hidey and

McKeown, 2016), comprises nearly 45,000 sentences, but they are annotated only for the presence of causal relations, and only 10% are causal; the other five datasets are an order of magnitude smaller, and exhibit similar bias.

In addition, the EventStoryLine Corpus (Caselli and Vossen, 2017), which is frequently used in related work, comprises several thousand causal links but no annotated negative samples. Additionally, Al-Khatib et al. (2020) introduce a corpus comprising 4740 claims extracted from argumentative texts, with 36% of these claims being annotated as containing a causal relation. Given that this corpus is the only one specifically focused on argumentative texts, we utilize it in our distance supervision and uncertainty sampling techniques. Our objective is to achieve broader coverage of new causality patterns through the incorporation of this corpus.

Du et al. (2022) present a human-annotated dataset consisting of over 21,000 causal reasoning questions, each accompanied by a natural language explanation providing insight into the underlying cause of the observed causation. Due to the scarcity of multilingual datasets with reliable and consistent annotations for event causality relations, Lai et al. (2022) present a new multilingual dataset that utilizes consistent annotation guidelines for five typologically distinct languages.

Perhaps most closely related to our own work, Zuo et al. (2020) propose a distant supervision-based data augmentation framework to address the data scarcity problem in causality. Whereas their approach involves a fully automated causal event pair extraction for distant supervision, we propose a framework based on uncertainty sampling, aiming to both improve the quality, and drive down the cost of hand-labeled corpora.

5 Conclusion

In this paper, we present WEBIS-CAUSALITY-23, a new dataset for causality identification in argumentative texts that considers all argument units (claims, premises) as inputs. The 1485 argumentative sentences in the dataset comprise a balanced distribution of the positive and negative causality types and encode diverse linguistic patterns denoting causality. Initial experiments on causality identification using transformer-based classifiers demonstrate the effectiveness of our smaller yet high-quality dataset in comparison to a larger existing corpus with some limitations.

Our future plans involve utilizing our dataset to extract causality relations from diverse argumentative resources on the Web. Our main objectives are to construct a large-scale argumentation graph, enhance argument scheme classification, and improve argument explanation and simplification methods. Additionally, we intend to leverage this dataset to develop a question-answering system specifically designed to address causal questions.

References

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data Acquisition for Argument Search: The args.me corpus. In *42nd German Conference on Artificial Intelligence (KI 2019)*, pages 48–59, Berlin Heidelberg New York. Springer.

Khalid Al-Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-end argumentation knowledge graph construction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7367–7374. AAAI Press.

Tommaso Caselli and Piek Vossen. 2017. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Matthias Cetto, Christina Niklaus, André Freitas, and Siegfried Handschuh. 2018. Graphene: Semantically-linked propositions in open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2300–2311. Association for Computational Linguistics.

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

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446,

Dublin, Ireland. Association for Computational Linguistics.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Manolis Kyriakakis, Ion Androutsopoulos, Artur Saudabayev, and Joan Ginés i Ametllé. 2019. Transfer learning for causal sentence detection. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 292–297. Association for Computational Linguistics.

Viet Dac Lai, Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2022. MECI: A multilingual dataset for event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2346–2356, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program: electronic library and information system*, 14(3):130–137.

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 2018. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In *Proceedings of the 5th Workshop on Argument Mining*, pages 79–89, Brussels, Belgium. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC2̂ Workshop*.

Burr Settles. 2012. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, and Marti A. Hearst. 2021. Automatically generating cause-and-effect questions from passages. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 158–170, Online. Association for Computational Linguistics.

Sida I. Wang and Christopher D. Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*, pages 90–94. The Association for Computer Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jinghang Xu, Wanli Zuo, Shining Liang, and Xianglin Zuo. 2020. A review of dataset and labeling methods for causality extraction. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1519–1531. International Committee on Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. Knowdis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1544–1550. International Committee on Computational Linguistics.

Controllable Generation of Dialogue Acts for Dialogue Systems via Few-Shot Response Generation and Ranking

Angela Ramirez and **Kartik Aggarwal** and **Juraj Juraska**
and **Utkarsh Garg** and **Marilyn A. Walker**
University of California Santa Cruz
aramir62, kartik, mawalker@ucsc.edu

Abstract

Dialogue systems need to produce responses that realize multiple types of dialogue acts (DAs) with high semantic fidelity. In the past, natural language generators (NLGs) for dialogue were trained on large parallel corpora that map from a domain-specific DA and its semantic attributes to an output utterance. Recent work shows that pretrained language models (LLMs) offer new possibilities for controllable NLG using prompt-based learning. Here we develop a novel few-shot overgenerate-and-rank approach that achieves the controlled generation of DAs. We compare eight few-shot prompt styles that include a novel method of generating from textual pseudo-references using a textual style transfer approach. We develop six automatic ranking functions that identify outputs with both the correct DA and high semantic accuracy at generation time. We test our approach on three domains and four LLMs. To our knowledge, this is the first work on NLG for dialogue that automatically ranks outputs using both DA and attribute accuracy. For completeness, we compare our results to fine-tuned few-shot models trained with 5 to 100 instances per DA. Our results show that several prompt settings achieve perfect DA accuracy, and near perfect semantic accuracy (99.81%) and perform better than few-shot fine-tuning.

1 Introduction

Dialogue systems need to faithfully produce utterances that realize multiple types of dialogue acts (DAs), such as providing opinions, making recommendations, or requesting information. In the past, natural language generators (NLGs) for dialogue have been trained on large parallel corpora that map from a domain-specific meaning representation (MR) that specifies the desired DA and semantic attributes to an output utterance. The NLG must faithfully generate utterances that realize the style and form of the DA, and all of the specified attributes, as shown by the reference utterances in Table 1. Recent work shows that pretrained language models (LLMs) offer new possibilities for controllable NLG using prompt-based learning (PBL) (Brown et al., 2020; Radford et al., 2019; Liu et al., 2021). Here we present a novel few-shot overgenerate-and-rank approach that achieves the controlled generation of DAs.

Attributes and Values
(NAME [**Call of Duty: Advanced Warfare**], RATING [**excellent**], DEVELOPER [**Sledgehammer Games**], ESRB [**M (for Mature)**])
give_opinion
Call of Duty: Advanced Warfare must be **one of the best games** I've ever played. **Sledgehammer Games** always nail their **M-rated** games.
recommend
Since you seem to **love M-rated** games developed by **Sledgehammer Games**, I wonder if you have tried **Call of Duty: Advanced Warfare**.
inform
Developed by **Sledgehammer Games**, **Call of Duty: Advanced Warfare** is targeted at **mature audiences** and has overall **very positive ratings**.

Table 1: Sample ViGGO dialogue acts (DAs) (Juraska et al., 2019). The same attributes and values can be realized as different DAs.

Previous work on semantically-controlled NLG has focused on improving semantic accuracy (Rastogi et al.; Xu et al., 2021; Du et al., 2022; Wen et al., 2015; Kedzie and McKeown, 2020; Juraska and Walker, 2021). However, Table 1 shows how the the same set of semantic attributes can be realized by different DAs, such as *give_opinion*, *recommend* and *inform*, each of which affect the dialogue state differently (Traum and Allen, 1994).

Obviously an NLG for dialogue needs to faithfully realize the DA as well as the semantic attributes. However, previous work has neither *controlled for* nor *evaluated* DA accuracy. We speculate that this is because many NLG training sets, such as E2E, Weather, WebNLG, WikiBio, DART and ToTTo, only include *inform* DAs (Novikova

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 355–369
September 11–15, 2023. ©2023 Association for Computational Linguistics

et al., 2017b; Belz, 2008; Gardent et al., 2017; Lebret et al., 2016; Nan et al., 2021; Parikh et al., 2020). Yet NLG training sets for spoken dialogue include many types of DAs, e.g. the ViGGO corpus has 9 DAs (Juraska et al., 2019), the RNNLG corpus provides 13 DAs (Wen et al., 2015), MultiWOZ has 34 DAs (Eric et al., 2021), and Topical Chat was automatically labelled with 11 DAs (Hedayatnia et al., 2020; Mezza et al., 2018).

We present a few-shot PBL framework that overgenerates and ranks NLG outputs and achieves high accuracy for both semantic attributes and DAs. We develop high accuracy DA classifiers for three domains and use them to define 6 ranking functions that combine estimates of DA probability with measures of semantic accuracy. We also compare a combination of prompt formats, prompt sampling methods, and DA representations. Several prompt templates take the novel approach of treating DA control as a textual style transfer (TST) problem (Reif et al., 2022). For completeness, we report results for few-shot fine-tuned models trained with 5 to 100 instances per DA. Our contributions include:

- The first results showing that dialogue acts can be controlled with PBL;
- A new overgenerate-and-rank framework that automatically ranks generation outputs for DA accuracy at generation time;
- A systematic exploration of both domain-specific and general measures in ranking functions, and a comparison of their performance;
- Results showing that a ranking function that prioritizes DA correctness results in higher semantic accuracy.
- The definition of novel textual DA representations that support automatic ranking for semantic accuracy using off-the-shelf metrics such as BLEU and Beyond-BLEU;
- The systematic testing of 8 prompt formats that re-cast data-to-text generation as a text-to-text task, and an examination of their performance across 4 LLMs.

The results demonstrate large performance differences across prompt styles, but show that many prompts achieve perfect DA accuracy, and semantic accuracy as high as 99.81% with only 10 examples, while 100-shot per DA fine-tuning only achieves 97.7% semantic accuracy, and 80.6% DA accuracy.

2 Related Work

This paper applies few-shot PBL to the task of controllable generation of DAs using an overgenerate-and-rank NLG framework. The overgenerate-and-rank paradigm for NLG has primarily used two methods for ranking: (1) language model probability (Langkilde and Knight, 1998); and (2) ranking functions trained from human feedback (Rambow et al., 2001; Bangalore et al., 2000; Liu et al., 2016). We extend this framework by applying it in the context of PBL, by using DA probability in ranking, and by comparing many ranking functions, including Beyond-BLEU and BLEU baselines (Wieting et al., 2019; Papineni et al., 2002).

We know of only a few previous studies on controllable generation of DAs in the context of dialogue systems, each of which has only focused on one or two types of DAs. Obviously, tasks like question generation (QG) aim at controllable generation of questions (Harrison and Walker, 2018; Zhang et al., 2021) but research on QG is not focused on trying to control the generation of questions as opposed to other types of DAs. However, some work has focused on controlling questions in dialogue, e.g. Hazarika et al. (2021) learned a latent representation of questions from a labelled corpus and then used this as a prompt prefix to control question generation. See et al. (2019) fine-tuned a Persona Chat model and tested decoding methods that controlled question frequency, but did not guarantee a question on a particular turn. Other work has focused on dialogue acts like opinions and recommendations. For example, Oraby et al. (2019) curated opinionated utterances from user reviews that had been marked with exclamation points, and then used the exclamation points as a way to control the production of exaggerated opinions. Reed et al. (2020) used token supervision to control the production of `recommendation` as opposed to `inform` dialogue acts where `recommendation` DAs stated that a particular restaurant was the best and then justified the recommendation with attributes from the MR. Ramirez et al. (2023) used PBL with similar prompts to control the expression of Big 5 personality types (Harrison et al., 2019), rather than dialogue acts.

It is well known that data-to-text NLGs based on fine-tuned LLMs are prone to semantic errors (Ji et al., 2022; Rashkin et al., 2021), thus previous work has focused on methods for ensuring semantic

correctness. This includes automatically augmenting the training data (Xu et al., 2021; Du et al., 2022), modifying the input representation (Kedzie and McKeown, 2020; Heidari et al., 2021), using rankers or classifiers or decoding methods that identify semantically accurate or acceptable candidates (Harkous et al., 2020; Juraska and Walker, 2021; Wen et al., 2015; Shen et al., 2019; Batra et al., 2021). Previous work on few-shot PBL for semantically-controlled NLG has not attempted to control DA accuracy (Reed et al., 2022; Soltan et al., 2022), and has not used an overgenerate and rank approach, resulting in lower semantic accuracies than we report here.

Much previous work on few-shot NLG has investigated few-shot finetuning rather than few-shot PBL. Previous work on the ViGGo, TV and Laptop corpora (Xu et al., 2021; Du et al., 2022; Kedzie and McKeown, 2020; Juraska and Walker, 2021) supports direct comparison to our work, but is not few-shot, does not rank outputs or use PBL. FewShotWoz trains a model called SC-GPT on a 400K data-to-text corpus, and then tests transfer learning with only 40 or 50 fine-tuning examples (Peng et al., 2020). Other recent work develops methods for augmenting FewShotWoz using synthetic data or by self-training and shows improvements in semantic accuracy and BLEU score. The FewShotWoz corpus includes many types of DAs but none of this previous work includes an evaluation of NLG DA accuracy. Previous work on few-shot finetuning in the weather domain used 300 examples in fine tuning, and also explored different ways of textualizing the MR (Heidari et al., 2021), but did not attempt to control DAs, develop ranking functions, evaluate DA accuracy, or use instructions such as our novel definitional prompts and the templates for TST tasks. Heidari et al. (2021) achieve an 85% reconstruction accuracy, while our best prompt/LLM combinations achieve 99.44% PERF score for ViGGO, 99.57% PERF for TV and 99.47% PERF for Laptop, a similar metric to reconstruction accuracy, with only 10 examples.

3 Automatically Ranking NLG Outputs

We start by providing a mathematical formulation of our problem. When generating from a DA representation, a high-quality response should: (1) manifest the specified DA; (2) have no missing or incorrect mentions of the attributes; (3) hallucinate no additional attributes; and (4) be fluent. Thus the generated utterance y, conditioned on an input x composed of DA d and attribute values a, can be formulated as $y = f(d, a)$. The conditional likelihood of y given the MR can then be decomposed using Bayes Rule into the product of three probabilities:

$$p(y|d, a) = p(d|y, a) * p(a|y) * p(y) \quad (1)$$

The term $p(d|y, a)$ is the DA probability given the generated utterance y and the semantic attributes a. The term $p(a|y)$ represents the semantic accuracy. The term $p(y)$ is the unconditional probability of the generated utterance, which is commonly used as a measure of fluency. Below, we show how we compute estimates of these terms at generation time, and then explain their use in the ranking functions.

Dialogue Act Classifier. The term $p(d|y, a)$ requires highly accurate DA classifiers to use in automatic ranking. We fine-tuned two classifiers using pre-trained bert-base-uncased on HuggingFace. We discovered that even though the ViGGO, Laptop and TV training corpora are good size (Juraska et al., 2019; Wen et al., 2015), producing high accuracy classifiers required us to modify the training data.[1] We originally trained the ViGGO classifer with the original ViGGO training set, when we applied this classifier to the generated outputs, we noticed many cases of low confidence classification. A qualitative analysis of the data showed that many generated outputs did not actually fit into the original ViGGO ontology, which is not surprising, given that the training data for an LLM would have included many different types of DAs.

To increase the ViGGO classifier performance, we introduced an "Other" class of dialogue acts, doubly annotated another 1000 ViGGO NLG outputs by hand, and added them to the original training set. Final results are shown in Table 2.

The second classifier was trained using the complete RNNLG corpus with all 4 domains to maximize classifier domain transfer. When we tested it on the RNNLG test set, we discovered that several classes had low F1. Examination of the confusion matrix showed that the *recommend* and *inform* DAs were highly confusable, so we created a new type of DA we call “describe” by combining their

[1]We also experimented with training classifiers for MultiWoz but were unable to get high accuracies due to noise in DA labelling, which is known to be an issue with MultiWoz (Zou, 2022).

Dialogue Act	ViGGO
confirm	0.99
inform	0.98
suggest	0.91
give_opinion	0.90
recommend	0.92
request	0.94
request_attribute	0.93
request_explanation	0.99
verify_attribute	0.94
other	0.78
Weighted Average	0.97

Table 2: ViGGO DA classification F1 scores.

training sets. The final results for for the RNNLG classifiers is shown in Table 3.

Dialogue Act	Laptop	TV
compare	1.00	1.00
confirm	0.96	0.95
describe	1.00	1.00
inform all	0.86	0.92
inform count	1.00	1.00
inform no info	1.00	1.00
inform no match	0.98	0.94
inform only match	0.83	0.87
suggest	1.00	1.00
Weighted Average	0.99	0.99

Table 3: Laptop and TV DA classification F1 scores. The *describe* DA = combination of the *inform* and *recommend* DAs in the original dataset.

We provide these DA classifiers along with additional human-labelled model outputs so that other researchers can duplicate our setup.[2] The resulting classifiers achieve average F1s over .97 for all three domains.

Semantic Accuracy. Work on data-to-text NLG often computes semantic accuracy as the Slot Error Rate (SER), i.e., the percentage of slots across all outputs y that the NLG realized incorrectly, with models either carefully tuned by hand, or trained by artificially creating incorrect realizations (Wen et al., 2015; Dusek et al., 2019; Juraska et al., 2018; Reed et al., 2020; Wiseman et al., 2017; Harkous et al., 2020; Kedzie and McKeown, 2019, 2020). There is a toolkit for SER for all three domains,[3] which we use to calculate SACC:

$$\text{SACC} = 1 - \text{SER} \quad (2)$$

Because the SACC scripts are domain specific, we also create new metrics that are based on BLEU, BLEURT, Beyond-BLEU and BertScore, widely used measures of semantic accuracy and semantic preservation (Papineni et al., 2002; Wieting et al., 2019; Sellam et al., 2020; Zhang et al.; Gehrmann et al., 2021). Because these metrics require comparisons with reference utterances, which are not available at generation time, we define reference-less versions based on pseudo-references, S_{pseudo}, created from the input DAs Juraska (2022). For any MR, we create its S_{pseudo} by omitting the slot names and the DA name and then concatenating the categorical attribute values with spaces between them, and converting boolean attributes, such as HAS_MULTIPLAYER = no, into phrases using the attribute name, with a negation when needed, e.g. "no multiplayer". For example, S_{pseudo} for the MR at the top of Table 1 would be "Call of Duty: Advanced Warfare excellent Sledgehammer Games M for Mature". Pseudo-references are available at generation time, so we use them to calculate pseudo-metrics for semantic accuracy and use them in ranking. Juraska et al. (2019) shows that the *relative* differences of these pseudo-metrics distinguish errorful NLG utterances from correct ones.

Fluency. Recent work suggests that the probability P(S) of a generated output S according to an LLM is a good automatic and referenceless measure of fluency (Kann et al., 2018; Suzgun et al., 2022). We thus adopt P(S) to measure fluency, and use GPT-2 to calculate P(S).

Ranking. The ranking functions in Table 4 aim to select NLG outputs that maximize DA accuracy, semantic accuracy, and fluency. Ranking function RF1 scores each candidate according to Equation 1.

RF1: DAC * SACC * P(S)
RF2: DAC * SACC * pBLEU * P(S)
RF2$_{DA}$: DAC | SACC | pBLEU | P(S)
RF3: DAC * pBBLEU * P(S)
RF4: pBBLEU
RF5: pBLEU

Table 4: Ranking functions. DAC = probability of the correct DA using a classifier. SACC = semantic accuracy using domain-specific SACC scripts. P(S) = LM probability as a measure of fluency. pBBLEU = pseudo-Beyond-BLEU to measure semantic accuracy. pBLEU = pseudo-BLEU as a baseline.

After a qualitative analysis of the ranking outputs from RF1 on pilot data, we developed ranker RF2 and RF2$_{DA}$ in Table 4. Our analysis revealed

[2] https://github.com/aramir62/da-nlg
[3] https://github.com/jjuraska/data2text-nlg

that the SER scripts often do not detect hallucinations, but pBLEU appeared to detect some hallucinations, so we add pBLEU to RF2. Ranking function RF2$_{DA}$ prioritizes one metric at each step, as represented by | in RF2$_{DA}$, enforcing DA correctness as more important for dialogue than perfect SACC. Matching DA candidates are preferred, but if no candidates match the required DA, the DA class *other* is preferred, or otherwise, all k candidates are selected. The second step selects candidates with the highest SACC. The third step aims to remove candidates with hallucinations by choosing the highest pBLEU outputs. The final step selects outputs with the highest fluency (P(S)).

So far RF1, RF2 and RF2$_{DA}$ all use the domain-specific SACC score for measuring semantic accuracy. To define a domain-independent ranking function, we calculate the correlation of SACC with pBLEU, pBBLEU, pBERT, and pBLEURT, defined in Section 3, on sample model outputs. See Table 12 in Appendix A.2. The results show that pBBLEU (Wieting et al., 2019) has the highest correlation across all three domains with 0.52 for Viggo, 0.32 for Laptop and 0.45 for TV. We thus define RF3 by replacing SACC in RF1 with pBBLEU. We then define RF4 as pBBLEU alone, so we can compare our novel ranking functions to pBBLEU. Finally, as a baseline reflecting the fact that previous work uses BLEU as a single measure of goodness for NLG, we define R5 as pBLEU.

4 Experimental Overview

Figure 1 provides an overview of the experimental architecture. Given a set of DA representations for a domain, we sample prompt examples from the original training sets while varying the number of samples. We then textualize the DA representations in the sample to look more similar to the LLMs free-text training data. The samples are then fed through the 8 prompt formats in Table 5. We apply this method to the ViGGO, Laptop and TV domains and utilize the 6 ranking functions in Table 4.

Prompt Formats. LLMs are typically trained on far more monologic data than dialogue, and will have rarely, if at all, seen examples of data-to-text NLG (Brown et al., 2020; Raffel et al., 2020; Devlin et al., 2018). While there are LLMs trained on dialogue such as DialoGPT (Zhang et al., 2020), and semantically-controlled dialogue data such as KGPT (Chen et al., 2020), and SC-GPT (Peng et al., 2020), there are clear benefits to using a general

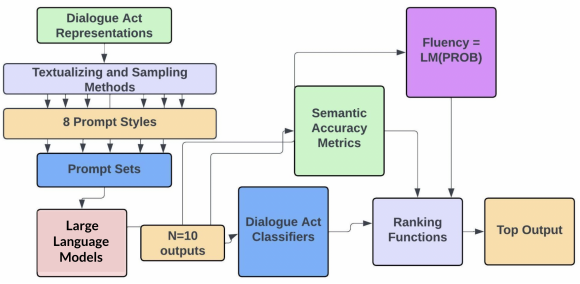


Figure 1: Experimental Architecture

LLM. Previous work also shows that without specific dialogic data, many LLMs do well on NLG for dialogue (Soltan et al., 2022). Here, we test the hypothesis that performance can be improved by using prompt formats that make the data-to-text task look more like the LLM's textual training data.

Prompt ID	Prompt Template
TST VANILLA	Here is a text: "s_{pseudo}". Here is a rewrite of the text which is a(n) d dialogue act: "r_{text}"
TST DIALOGUE	Here is a text: "s_{pseudo}". Rewrite it to be a(n) d dialogue act: "r_{text}"
TST PARAPHRASE	Here is a text: "d_r s_{pseudo}". Here is a paraphrase of the text: "r_{text}"
DEFINITIONAL	description of $< d >$: D^d. Data: $d = yes \mid sa^1 = v^1..sa^n = v^n$ Data to Text for $< d >$: r_{text}
PARAPHRASE	d_r s_{pseudo} r_{text}
DIALOGIC	d_r s_{pseudo} r_{text}
PSEUDO	d s_{pseudo} r_{text}
S2S	$d = yes \mid a^1 = v^1..a^n = v^n$ r_{text}

Table 5: Prompt IDs and templates. Instantiations of each template are given in Table 11 in the Appendix.

Table 5 shows the 8 prompt templates, with full instantiations in the Appendix in Table 11. The templates vary the representation of the DAs and their attributes. We represent the DA directly by its name d, or convert the DA to a sentence starter d_r such as "I recommend". The attributes of the DA constitute a set $a = a^1, a^2, ..., a^n$, each with a value in v where $v = v^1, v^2, ..., v^n$. The attributes can be represented directly or using a textual pseudo-reference s_{pseudo}, as described in Section 3. The reference text r_{text} then varies the representation of the DA and the attributes.

Prompts TST Vanilla, TST Dialogue, and TST Paraphrase of Table 5 treat data-to-text generation as a textual style transfer (TST) task, where each DA is a style, and the prompt provides instructions,

e.g., "Rewrite it to be a suggest dialogue act" (Reif et al., 2022; Suzgun et al., 2022). TST Vanilla and TST Dialogue represent the MR as its pseudo-reference s_{pseudo}, while TST Paraphrase prefixes the sentence starter d_r for the DA to s_{pseudo}.

We also define a Definitional prompt with definitions of the DAs, represented as D^d, based on the instructions given to crowdworkers when ViGGO was collected, inspired by previous work providing slot descriptions (Gupta et al., 2022).

The Paraphrase prompt is based on the fact that producing paraphrases is a common task. This prompt rewrites the DA as a first-person sentence starter, e.g., "I suggest" for the *suggest* DA. The Dialogue Response prompt is similar, but mimics a request and its response, with sentence starters written as requests, e.g., "can you recommend a game Worms: Reloaded Steam?" for the *recommend* DA.

To directly evaluate the benefit of instructions, we also input the pseudo-reference without instructions as a baseline (Pseudo), as well as input the commonly used S2S format which linearizes the MR as a sequence of attributes and values (Soltan et al., 2022; Wen et al., 2015; Harkous et al., 2020).

5 Results

Experimental Roadmap. We first experiment with ViGGO over all the experimental settings from Section 4 using Jurassic-1 Jumbo, a 175B auto-regressive transformer-based LLM with a different depth-width tradeoff than GPT3 (Levine et al., 2020; Lieber et al., 2021). All experiments set top P = 1, and T = 0.7 based on pilot experiments. We compare prompting to few-shot fine-tuning using 5, 25, 50 and 100 examples per DA sampled from the training data. We test the 8 prompt formats in Table 5 with 1, 5 or 10 prompt examples. Our focus is DA control, so we create a ViGGO test set with 40 instances per DA (360 total). We look-ahead to see which ranking function performs best for ViGGO and use that for the results in Table 6.

We then test the best settings from ViGGO on the Laptop and TV corpora (Wen et al., 2015) with results in Table 7. We compare ranking function performance across all domains in Table 8, and demonstrate the improved performance of our ranking functions compared to simply using BLEU. We then test for generalization with additional LLMs: we select the top three prompt settings, and test of GPT-Neo as a smaller LLM, and GPT-3 and ChatGPT as instruction-tuned LLMs, and compare them to Jurassic-1, for all three domains. These results are shown in Table 9. Table 10 then compares our best performance to recent SOTA results for both fine-tuning and few-shot fine-tuning on ViGGO, Laptop and TV. Finally we report the results of our human evaluations. We make the DA classification models, the prompts and their instantiations, and the model outputs for all experiments available.[4]

ID	N	PERF	SACC	DAC
Few-Shot Fine-Tuning Experiments				
FTune 5-per	45	38.88	85.71	54.44
FTune 25-per	225	62.22	92.19	79.72
FTune 50-per	450	71.94	96.43	79.44
FTune 100-per	900	**78.61**	**97.74**	**80.56**
Prompt Styles and Samples Experiments				
TST Vanilla	10	**85.56**	**94.73**	**100.00**
TST Dialogue	10	83.89	94.17	100.00
TST Paraphrase	10	83.90	94.20	100.00
Definition (each)	10	76.94	91.16	100.00
Definition (top)	10	82.22	93.51	100.00
Paraphrase	10	77.78	92.10	100.00
Dialogic	10	77.22	91.53	100.00
Pseudo	10	75.83	94.17	100.00
S2S	10	70.56	86.45	100.00
TST Vanilla	5	80.56	92.57	99.72
TST Dialogue	5	**83.61**	**93.88**	**100.00**
TST Paraphrase	5	80.20	92.60	99.70
Definition (each)	5	80.00	92.66	99.40
Definition (top)	5	77.22	91.25	100.00
Paraphrase	5	70.83	89.71	100.00
Dialogic	5	66.94	88.34	99.10
Pseudo	5	52.22	82.60	85.56
S2S	5	66.67	83.54	99.72
TST Vanilla	1	68.06	86.64	91.94
TST Dialogue	1	69.17	88.15	93.30
TST Paraphrase	1	**72.20**	**89.80**	93.60
Definition	1	63.89	85.32	**98.30**
Paraphrase	1	41.94	75.14	83.88
Dialogic	1	38.89	71.83	82.30

Table 6: Results after ranking via $RF2_{DA}$ for ViGGO. N = number of prompt examples. PERF = % outputs that are perfect. SACC = semantic accuracy using SACC scripts. DAC = DA accuracy using a classifier.

Few-Shot Fine-Tuning. To compare prompting to fine-tuning, we use the traditional linearized MR in the S2S format and vary the number of training examples per DA in few-shot fine-tuning from 5, to 25, to 50, to 100. The results in Rows 1-4 of Table 6 show that, as expected, increasing the number of training examples improves performance, with 100 examples per DA (900 overall) achieving a SACC of 97.74 after ranking. However, interestingly, the highest DAC performance is only 80.56, and the PERF score (both perfect DA and perfect SACC) is only 78.61. Table 13 in the Appendix shows more

[4] https://github.com/aramir62/da-nlg

detail, providing before and after ranking performance for fine-tuning. Overall, the results affirm previous findings that few-shot prompting beats few-shot fine-tuning (Le Scao and Rush, 2021).

Prompt Styles. All experiments provide examples for a single DA and then generate that DA, while varying the prompt style and the number of examples. The TST format provides N examples using one of the TST prompts in Table 5. The Definitional (each) format, for 10 prompts, provides 10 triplets of (definition, MR, text). For Definitional (top), the definition is mentioned once before all the MRs and examples, so for 1 prompt, there is no difference between *top* and *each*.

We first notice in Table 6 that the PERF score improves with the number of prompt examples, from 1 to 5 to 10 for all the prompt styles, with TST Vanilla, TST Dialogue, and TST Paraphrase, which provide the MR as text and include instructions (see Table 5) consistently performing the best overall. TST Vanilla-10 performs signicantly better than the other TST styles with 10 examples ($p < .01$), but TST Dialogue is the best for 5 examples and TST Paraphrase is the best for 1 example. The Definitional, Paraphrase and Dialogic formats all perform significantly worse than the TST formats, but interestingly the Definitional format gets the highest DAC with only 1 example perhaps showing the advantage of explicit definitions in PBL.

The Pseudo and S2S prompt styles are baselines, and only reported for the 5 and 10 example settings. Both baselines indicate the benefits of instructions. The S2S 10 performance is the worst for 10 examples, and the Pseudo performance is the worst for 5 examples. It is worth noting that the poorly performing S2S representation is commonly used in both fine-tuning and PBL (Soltan et al., 2022; Wen et al., 2015; Harkous et al., 2020).

Domain	ID	N	PERF	SACC	DAC
Laptop	TST Van.	10	80.95	95.90	100.00
TV	TST Van.	10	98.85	99.76	100.00

Table 7: Results for Laptop and TV for TST 10 using $RF2_{DA}$. N = number of examples. PERF = % outputs that are perfect. SACC = semantic accuracy using SACC scripts. DAC = DA accuracy using a classifier.

We then take the best performing prompt (TST Vanilla) and experiment with TV and Laptop. The results are shown in Table 7. $RF2_{DA}$ performs the best for both Laptop and TV so these results are ranked with $RF2_{DA}$. Interestingly, TV has the highest PERF and SACC seen so far, while Laptop also has a higher SACC than any ViGGO setting, suggesting that it is easier to achieve high performance with Laptop and TV than ViGGO.

RF	Terms	PERF	SACC	DAC	BLEU
ViGGO					
RF1	DAC, SACC, P(S)	79.17	91.82	99.72	38.41
RF2	DAC, SACC, pBLEU, P(S)	78.33	91.72	99.00	38.67
$RF2_{DA}$	DAC, SACC, pBLEU, P(S)	**85.56**	**94.73**	**100.00**	40.08
RF3	DAC, pBBLEU, P(S)	62.78	84.38	100.00	**49.87**
RF4	pBBLEU	60.55	91.63	77.78	42.82
RF5	pBLEU	44.22	81.66	75.28	40.08
TV					
RF1	DAC, SACC, P(S)	85.40	96.86	100.00	72.55
RF2	DAC, SACC, pBLEU, P(S)	88.19	97.43	100.00	72.55
$RF2_{DA}$	DAC, SACC, pBLEU, P(S)	**98.85**	**99.76**	**100.00**	60.51
RF3	DAC, pBBLEU, P(S)	73.96	93.87	100.00	**72.89**
RF4	pBBLEU	90.14	97.88	99.71	60.51
RF5	pBLEU	63.45	91.50	99.57	66.71
Laptop					
RF1	DAC, SACC, P(S)	49.25	86.70	100.00	61.24
RF2	DAC, SACC, pBLEU, P(S)	57.29	89.47	100.00	59.39
$RF2_{DA}$	DAC, SACC, pBLEU, P(S)	**80.95**	**95.90**	**100.00**	**61.36**
RF3	DAC, pBBLEU, P(S)	35.55	80.41	100.00	45.03
RF4	pBBLEU	61.79	90.97	98.88	36.32
RF5	pBLEU	42.38	84.25	97.77	**61.36**

Table 8: Ranking functions performance.

Ranking Functions. Our results show that our overgenerate-and-rank method has a huge effect on performance as compared to taking the first output from the model. Section A.3 in the Appendix provides more detail, e.g. showing for Viggo, across all the experiments, *Before Ranking* has an average SACC of 65.29% versus an *After Ranking* average of 86.82%, while DAC has an almost a 30% increase with a *Before Ranking* average of 62.11%, and an *After Ranking* average of 91.04%.

Table 8 compares the 5 ranking functions from Section 3 on all three domains for the best prompt so far: TST Vanilla 10. The differences between RF1 and RF2 (addition of pBLEU) are not significant for ViGGO, but are significant for TV (t-test, $p < 0.001$) and Laptop (t-test, $p < 0.001$), with Laptop improving from 49.24 PERF to 57.29 PERF. Note that in all domains ranking by $RF2_{DA}$ results in significantly higher performance across all metrics (t-test, $p < 0.001$): **prioritizing DA correctness results in higher SACC and higher PERF**.

Table 8 also shows that replacing SACC with pBBLEU in RF3 results in a clear drop in performance. As shown in Appendix Section A.2 pBBLEU is the best performing pseudo-metric overall, but there are clear advantages to the domain-specific SACC. Recent work explores automatic methods for training domain-specific semantic fidelity classifiers, but these methods rely on large training corpora making them difficult to apply in few-shot settings (Harkous et al., 2020; Batra et al., 2021).

The baseline RF4 with only the pBBLEU term performs surprisingly well in SACC across all three domains, suggesting that it might be worth examining further combinations of BBLEU with DAC.

MODEL	PROMPT	PERF	SACC	DAC	BLEU
ViGGO					
ChatGPT	TST 10	98.89	95.58	99.44	45.05
ChatGPT	TST 5	94.72	99.34	96.67	40.88
ChatGPT	Def 10	98.89	**100.00**	100.00	42.40
ChatGPT VO	Def 10	95.28	99.85	95.83	14.79
GPT 3	TST 10	95.00	98.49	98.33	40.26
GPT3	TST 5	95.28	98.31	98.89	**54.11**
GPT3	Def 10	**99.44**	99.81	**100.00**	42.75
GPT3 VO	Def 10	95.28	99.83	95.55	9.55
Jurassic	TST 10	85.56	94.70	100.00	40.08
Jurassic	TST 5	83.61	93.88	100.00	32.54
Jurassic	Def 10	82.22	93.51	100.00	15.77
GPT NEO 1.3B	TST 10	17.78	85.32	35.56	25.25
GPT NEO 1.3B	TST 5 dial	64.17	86.74	94.72	43.47
GPT NEO 1.3B	Def 10	35.56	78.27	81.94	15.44
TV					
ChatGPT	TST 10	98.00	99.57	99.93	45.98
ChatGPT	TST 5	91.23	98.14	100.00	38.22
ChatGPT	Def 10	98.00	99.30	99.64	50.97
GPT 3	TST 10	**99.57**	99.91	**100.00**	57.92
GPT3	TST 5	99.07	99.81	100.00	71.80
GPT3	Def 10	99.22	**99.94**	100.00	73.81
Jurassic	TST 10	98.85	99.76	100.00	60.51
Jurassic	TST 5	91.80	98.26	100.00	**74.73**
Jurassic	Def 10	95.01	98.94	100.00	73.66
GPT NEO 1.3B	TST 10	83.15	96.37	100.00	66.28
GPT NEO 1.3B	TST 5 dial	50.78	93.15	73.93	31.95
GPT NEO 1.3B	Def 10	15.74	78.61	65.88	19.29
Laptop					
ChatGPT	TST 10	**97.08**	99.47	99.58	41.45
ChatGPT	TST 5	85.95	97.19	99.43	23.36
ChatGPT	Def 10	67.54	90.37	99.92	36.00
GPT 3	TST 10	84.79	**99.91**	**100.00**	33.20
GPT3	TST 5	94.79	97.14	100.00	32.41
GPT3	Def 10	81.45	92.54	100.00	**85.40**
Jurassic	TST 10	80.95	95.90	100.00	61.36
Jurassic	TST 5	81.55	96.10	99.81	12.94
Jurassic	Def 10	55.98	45.60	100.00	29.12
GPT NEO 1.3B	TST 10	68.89	92.66	100.00	46.21
GPT NEO 1.3B	TST 5 dial	71.89	93.55	100.00	19.49
GPT NEO 1.3B	Def 10	1.33	43.73	99.96	14.59

Table 9: Experiments with additional LLMs, with the top three prompt settings, for ViGGO, Laptop and TV, using the $RF2_{DA}$ ranking function. We also tested here with the original ViGGO test set, with ChatGPT Def 10 and GPT-3 Def 10, with results shown in cyan, to facilitate comparison with previous work.

Finally, the pBLEU baseline of RF5 reinforces work emphasizing the inadequacies of BLEU as a metric for NLG (Belz, 2008; Liu et al., 2016; Novikova et al., 2017a). We report BLEU for comparison with related work, but Table 8 clearly shows that the highest BLEU score doesn't correspond to the best PERF or SACC, and that even ranking with pBLEU (RF5) doesn't maximize BLEU. RF5 gets the lowest PERF, SACC and DAC scores for ViGGO and TV, and $RF2_DA$ achieves the same BLEU score, with much higher PERF, SACC and DAC for both ViGGO and Laptop.

Experiments with other LLMs. We also compare our results with Jurassic to other LLMs. We select the three best prompt settings, namely TST 10, TST 5, and Definitional Top 10, and experiment with ChatGPT and GPT-3 as large instruction-based models and GPT-Neo 1.3 as a small model.

Table 9 presents the results. Our primary metric is PERF with best PERF shown in bold. Note in the table that the highest PERF score does not necessarily correspond with the highest SACC or highest BLEU. Interestingly, GPT-3 performs slightly better than ChatGPT for both ViGGO and TV while ChatGPT performs best for Laptop. Both ChatGPT and GPT-3 perform significantly better than Jurassic across all three domains. Table 9 shows that the Definitional prompt performs better than TST 10 with both ChatGPT and GPT-3 for Viggo, while TST 10 for TV was comparable to Definitional and performs the best for Laptop in terms of PERF. We add results here for the original ViGGO test set shown in cyan, which has a skewed distribution of DAs with more long Inform DAs, and which appears to be more challenging for DAC but not SACC. Finally, we see much worse performance with GPT Neo, reinforcing results suggesting a model size threshhold for PBL (Wei et al.).

Comparison with SOTA. Table 10 compares our best results with recent work on the VIGGO, Laptop and TV corpora (Xu et al., 2021; Du et al., 2022; Juraska and Walker, 2021; Kedzie and McKeown, 2020; Harkous et al., 2020; Peng et al., 2020). The related work either used fine-tuning or few-shot fine-tuning, rather than PBL. JW21, DT and K-McK are based on fine-tuning. SC-GPT, AUGNLG and ST-SA are all based on FEWSHOTWOZ. In each case, we take the results exactly as reported in the related work. These results are indicative only as e.g. FEWSHOTWOZ does not use the original RNN-NLG test set for Laptop and TV, which we use here. We created our own ViGGO test set to have equal numbers of each DA, but the original test set has many more long *inform* DAs.

Human Evaluation. Given the almost perfect performance reported in Table 9, we conducted a human evaluation to check whether the outputs were indeed perfect (the right DA and the correct semantics), and whether there were any hallucinations. Two expert annotators hand-labelled 100 outputs from ChatGPT with TST-10 Vanilla prompts. Amazingly, neither annotator found any outputs that weren't perfect and neither did they find any hallucinations. They agreed 100% on the results, resulting in a Cohen's Kappa of 1.0.

Model	Laptop		TV		ViGGO	
	BLEU↑	ERR↓	BLEU↑	ERR↓	BLEU↑	ERR↓
Ours	33.20	**0.08**	**73.81**	**0.06**	14.79	**0.15**
JW21	–	–	–	–	**53.60**	0.46
DT					**53.60**	1.68
K-McK	–	–	–	–	48.50	0.46
SC-GPT	32.73	3.39	32.95	3.38	–	–
AUGNLG-SC	34.32	2.83	34.99	5.53	–	–
ST-SA	**35.42**	2.04	36.39	1.63	–	–

Table 10: Ours = Our best model for each domain from Table 9 compared to recent SOTA results. Our VIGGO result is for the ViGGO ORIGINAL test set. JW21 = SeaGuide (Juraska and Walker, 2021). DT = Data Tuner (Harkous et al., 2020). K-McK = (Kedzie and McKeown, 2020). SC-GPT = (Peng et al., 2020). AugNLG = (Xu et al., 2021). ST-SA = (Du et al., 2022). We convert SACC to SER, which other work calls ERR, and report BLEU, and ERR as in that other work. Note that we use our best SACC score from Table 9 to select the row to include here, but this doesn't necessarily correspond to the best BLEU score or the best PERF score.

We also test whether our addition of pBLEU to RF2 has an effect on hallucinations, by testing in general whether pBLEU helps identify hallucinations. We annotate hallucinations for ViGGO, by having 3 annotators label all 360 outputs for each ranking function (6*360) shown in Table 8. The number of hallucinations for RF1 was 34, RF2 was 19, RF3 was 26, RF4 was 40 and RF5 was 14. We compared the mean number of hallucinations of ranking functions with pBLEU, namely RF2, $RF2_{DA}$, and RF5 to those without, namely RF1, RF3 and RF4. We find that the mean number of hallucinations of those with pBLEU is 31.67, while the mean number of those without is 19.67. This difference seems large, but the sample size is small and therefore it's not significant (t = 1.82, p = .14)

6 Conclusion and Future Work

Here we apply an overgenerate-and-rank NLG approach and and provide the first experiments using automatic ranking functions that optimize both DA and semantic accuracy in few-shot prompt-based NLG. We test and compare a combination of prompt formats, sampling methods, and DA representations. We test prompts used for textual style transfer (TST) by treating DAs as styles to be controlled. We also create novel prompts that provide definitions of DAs, For completeness, we fine-tune few-shot models and compare them with the few-shot results. The results show that several prompting styles achieve perfect DA accuracy, and that few-shot methods can achieve semantic accuracy as high as 99.81% with the right ranking function, while 100-shot fine-tuning achieves 97.7%, and performs much worse on DA accuracy (80.6%).

Our contributions include systematic experimentation with different ways of textualizing MRs, providing instructions to the LLM, and ranking outputs. Our results also show that formulating the data-to-text task as textual style transfer using pseudo-references yields the highest performance. We achieve SOTA semantic accuracy with only 10 prompt examples with our best prompt styles, and achieve the surprising results that a ranking function that prioritizes DA correctness results in higher semantic accuracy.

Limitations and Risks One limitation arises from the challenges of prompt-engineering: it is impossible to tell whether another prompt format could perform better, e.g. with smaller LLMs like GPT-Neo, where we get poor comparative results. Another limitation is the need for a high-accuracy DA classifier that works well on out-of-domain model outputs. We address this limitation by releasing our classifiers. Another possible limitation is the use of the overgenerate and rank approach in real-time. In future work we plan to use the high quality (ranked) generated data, to fine-tune a smaller real-time language model, without the need for overgeneration. Another limitation arises from the comparison to few-shot fine-tuning – there are many ways to fine tune and many representations of the MRs, so it is possible that some other method of fine-tuning would lead to better fine-tuning results (Liu et al., 2022). Our main goal here was to show that with a small-number of examples, using reasonable assumptions, few-shot fine-tuning performs worse than PBL.

A potential risk of using LLMs is the possibility of disinformation, often called hallucinations. Control of hallucinations is an active area of research. One of the challenges is that it is very difficult to automatically identify them. Here we experiment with ranking functions for better control of hallucinations, hand-label hallucinations and characterize them. Another potential risk of our work is that some of our dialogue acts like recommend and suggest could be used, in an application context, to persuade a user to buy something. In this context, it is even more important to ensure that the system is not providing false information to users.

References

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 1–8.

Soumya Batra, Shashank Jain, Peyman Heidari, Ankit Arun, Catharine Youngs, Xintong Li, Pinar Donmez, Shawn Mei, Shiunzu Kuo, Vikas Bhardwaj, Anuj Kumar, and Michael White. 2021. Building adaptive acceptability classifiers for neural NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 682–697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anja Belz. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14(4):431–455.

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

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. Kgpt: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Wanyu Du, Hanjie Chen, and Yangfeng Ji. 2022. Self-training with two-phase self-augmentation for few-shot dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2770–2784.

Ondrej Dusek, David M Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *INLG*.

Mihail Eric, Nicole Chartier, Behnam Hedayatnia, Karthik Gopalakrishnan, Pankaj Rajan, Yang Liu, and Dilek Hakkani-Tur. 2021. Multi-sentence knowledge selection in open-domain dialogue. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 76–86, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120. Association for Computational Linguistics.

Raghav Gupta, Harrison Lee, Jeffrey Zhao, Abhinav Rastogi, Yuan Cao, and Yonghui Wu. 2022. Show, don't tell: Demonstrations outperform descriptions for schema-guided task-oriented dialogue.

Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2410–2424, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn Walker. 2019. Maximizing stylistic control and semantic accuracy in nlg: Personality variation and discourse contrast. *INLG Workshop on Discourse Structure in NLG 2019*, page 1.

Vrindavan Harrison and Marilyn Walker. 2018. Neural generation of diverse questions using answer focus, contextual and linguistic features. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 296–306.

Devamanyu Hazarika, Mahdi Namazifar, and Dilek Hakkani-Tür. 2021. Zero-shot controlled generation with encoder-decoder transformers. *arXiv preprint arXiv:2106.06411*.

Behnam Hedayatnia, Seokhwan Kim, Yang Liu, Karthik Gopalakrishnan, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialogue systems. *arXiv preprint arXiv:2005.12529*.

Peyman Heidari, Arash Einolghozati, Shashank Jain, Soumya Batra, Lee Callender, Ankit Arun, Shawn Mei, Sonal Gupta, Pinar Donmez, Vikas Bhardwaj, Anuj Kumar, and Michael White. 2021. Getting to production with few-shot natural language generation models. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 66–76, Singapore and Online. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *CoRR*, abs/2202.03629.

Juraj Juraska. 2022. *Diversifying Language Generated by Deep Learning Models in Dialogue Systems*. Ph.D. thesis, UC Santa Cruz.

Juraj Juraska, Kevin K Bowden, and Marilyn Walker. 2019. ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*.

Juraj Juraska, Panagiotis Karagiannis, Kevin Bowden, and Marilyn Walker. 2018. A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 152–162.

Juraj Juraska and Marilyn Walker. 2021. Attention is indeed all you need: Semantically attention-guided decoding for data-to-text NLG. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 416–431, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323.

Chris Kedzie and Kathleen McKeown. 2020. Controllable meaning representation to text generation: Linearization and data augmentation strategies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5160–5185.

Chris Kedzie and Kathleen R McKeown. 2019. A good sample is hard to find: Noise injection sampling and self-training for neural language generation models. In *INLG*.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213.

Yoav Levine, Noam Wies, Or Sharir, Hofit Bata, and Amnon Shashua. 2020. The depth-to-width interplay in self-attention. *arXiv preprint arXiv:2006.12467*.

Opher Lieber, Barak Lenz Sharir, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. *Technical report, AI21 Labs*.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Stefano Mezza, Alessandra Cervone, Evgeny Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi. 2018. Iso-standard domain-independent dialogue act tagging for conversational agents. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3539–3551.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017a. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017b. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

Shereen Oraby, Vrindavan Harrison, Abteen Ebrahimi, and Marilyn Walker. 2019. Curate and generate: A corpus and method for joint control of semantics and style in neural nlg. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5938–5951.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 172–182, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Owen Rambow, Monica Rogati, and Marilyn Walker. 2001. Evaluating a trainable sentence planner for a spoken dialogue system. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 434–441.

Angela Ramirez, Mamon Alsalihy, Kartik Aggarwal, Cecilia Li, Liren Wu, and Marilyn Walker. 2023. Controlling personality style in dialogue with zero-shot prompt-based learning. ArXiv preprint arXiv:2302.03848.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.

Lena Reed, Vrindavan Harrison, Shereen Oraby, Dilek Hakkani-Tur, and Marilyn Walker. 2020. Learning from mistakes: Combining ontologies via self-training for dialogue generation. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2020)*.

Lena Reed, Cecilia Li, Angela Ramirez, Liren Wu, and Marilyn Walker. 2022. Jurassic is (almost) all you need: Few-shot meaning-to-text generation for open-domain dialogue. In *Conversational AI for Natural Human-Centric Interaction: 12th International Workshop on Spoken Dialogue System Technology, IWSDS 2021, Singapore*, pages 99–119. Springer.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically informative text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4060–4067, Minneapolis, Minnesota. Association for Computational Linguistics.

Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, et al. 2022. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models.

David Traum and James Allen. 1994. Discourse obligations in dialogue processing. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 1–8.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263.

Xinnuo Xu, Guoyin Wang, Young-Bum Kim, and Sungjin Lee. 2021. Augnlg: Few-shot natural language generation using self-trained data augmentation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1183–1195.

Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2021. A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, 40(1):1–43.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Deyan Zou. 2022. Multi-dimensional consideration of cognitive effort in translation and interpreting process studies. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 416–426, Orlando, USA. Association for Machine Translation in the Americas.

A Appendix

A.1 Full Prompt Descriptions and Examples

Table 11 shows a sample instantiation for each prompt type and template. When this paper is accepted, we will provide all the prompt files and instantiated prompts for all experiments in our github: https://github.com/aramir62/da-nlg.

A.2 Semantic Accuracy Pseudo Metrics

We estimate the goodness of the pseudo versions of BLEU, Beyond-BLEU, BERT and BLEURT by examining their correlations with the domain-specific SACC scores on a sample of model outputs from our experiments, as shown in Table 12. The correlations show that the pseudo version of Beyond-BLEU (Wieting et al., 2019) – pBBLEU – performs the best across all three domains. Interestingly, pBLEU, despite BLEU's popularity, performs the worst.

A.3 Before & After Ranking

Our results show that ranking by any ranking function significantly and greatly improves performance, with the greatest performance improvements arising from the RF2$_{DA}$ ranking function for all three domains. We calculate *Before Ranking* by averaging all metrics over the entire set of test outputs (test set size X 10 outputs into ranking). When taking averages across all experiments (per, fine-tuned, and specific), average SACC and DAC are significantly higher after ranking.

Table 13 provides more detail on how the ranking affects the results for few-shot fine-tuning. Comparing Row 1 to Row 4 shows that ranking improves the performance of SACC for 5-shot fine-tuning (85.71) to perform almost as well as 100-shot fine-tuning before ranking (88.71). Ranking also improves the performance of DAC for 100-shot fine-tuning from 57% to 80.56%, a huge improvement.

Table 14 shows more detail for Viggo across all the experimental settings. *Before Ranking* has an average of 65.29% versus *After Ranking* with an average of 86.82% for SACC. DAC has an almost a 30% increase where *Before Ranking* has an average of 62.11%, and *After Ranking* has an average of 91.04%. Table 15 shows the effect of ranking for TV and Laptop, illustrating a similarly large performance improvement due to ranking.

Prompt ID	Example
TST VANILLA	Here is a text: "Worms: Reloaded Steam". Rewrite of the text, which is a suggest dialogue act: "I bet you like it when you can play games on Steam, like Worms: Reloaded, right?"
TST DIALOGUE	Here is a text: "Worms: Reloaded Steam". Rewrite it to be a suggest dialogue act: "I bet you like it when you can play games on Steam, like Worms: Reloaded, right?"
TST PARAPHRASE	Here is a text: "I suggest Worms: Reloaded Steam". Paraphrase of the text: "I bet you like it when you can play games on Steam, like Worms: Reloaded, right?"
DEFINITIONAL	Description of $< suggest >$: A question asking if your friend has any experience with a certain type (based on data) of video games. Use the name of the game in data with 'such as', 'like', etc. The response should consist of a single yes/no question. Generate diverse responses. Data: suggest = yes | name = Worms: Reloaded | available_on_steam = yes. Data to Text for $< suggest >$: I bet you like it when you can play games on Steam, like Worms: Reloaded, right?
PARAPHRASE	I suggest a game Worms: Reloaded Steam. I bet you like it when you can play games on Steam, like Worms: Reloaded, right?
DIALOGIC	Can you suggest a game Worms: Reloaded Steam? I bet you like it when you can play games on Steam, like Worms: Reloaded, right?
PSEUDO	Suggest Worms: Reloaded Steam. I bet you like it when you can play games on Steam, like Worms: Reloaded, right?
S2S	suggest = yes | name = Worms: Reloaded | available_on_steam = yes. I bet you like it when you can play games on Steam, like Worms: Reloaded, right?

Table 11: Prompt IDs and Instantiation of each Prompt Template Type

Measure	ViGGO	Laptop	TV
pBLEU	0.08	-0.12	0.05
pBBLEU	0.52	0.32	0.45
pBLEURT	0.38	0.17	0.26
pBERT precision	0.33	0.14	0.36
pBERT recall	0.03	-0.06	0.14
pBERT F1	0.20	0.04	0.26

Table 12: Pearson correlation between SACC and common semantic preservation measures when applied to pseudo-references. All correlations are statistically significant at p < 0.001 .

N	SACC		Perf		DAC	
	Before	After	Before	After	Before	After
5	65.57	85.71	9.10	38.88	21.10	54.44
25	76.01	92.19	16.39	62.22	31.10	79.72
50	86.70	96.43	29.10	71.94	42.00	79.44
100	88.71	97.74	40	78.61	57.00	80.56

Table 13: Few-shot fine-tuning performance with increasing training examples per DA - before and after ranking. DAC = DA accuracy.

Format	N	Perfect		SACC		DAC	
		Before	After	Before	After	Before	After
TST Vanilla	10	37.2	**85.6**	76	**94.7**	84.3	**100**
TST Dialogue	10	**39.5**	83.9	**76.7**	94.2	84.7	**100**
S2S	10	32.0	70.6	68.3	86.5	85	100
Pseudo	10	32	75.8	70.3	94.2	84.5	100
Definitional (each)	10	37.2	76.9	73.4	91.2	88.3	100
Definitional (Top)	10	38.2	82.2	72.3	93.5	**88.8**	100
TST Vanilla	5	38.7	**83.6**	76.8	92.6	76.9	98.7
TST Dialogue	5	**40.7**	**83.6**	**76.9**	**93.9**	79.1	100
S2S	5	34.1	66.7	65.5	83.5	77.9	98.7
Pseudo	5	14.7	52.2	47.5	82.6	47.2	88.6
Definitional (each)	5	40.2	80.0	75.1	92.7	81.9	99.4
Definitional (Top)	5	38.4	77.2	74	91.3	**82**	**100**
TST Vanilla	1	25.6	**69.2**	**69.3**	**88.2**	58	92
TST Dialogue	1	25.5	**69.2**	68.2	**88.2**	62.3	93.3
Definitional	1	**25.7**	63.9	67	85.3	**66.2**	**98.3**

Table 14: Results Before and After Ranking

Format	N	SACC		Perf		DAC	
		Before	**After**	**Before**	**After**	**Before**	**After**
TV	10	92.59	99.76	65.30	98.85	95.90	100
Laptop	10	80.73	95.90	36.35	80.95	99.71	100

Table 15: Laptop and TV Before and After ranking. DAC = DA Accuracy.

Reference Resolution and New Entities in Exploratory Data Visualization: From Controlled to Unconstrained Interactions with a Conversational Assistant

Abari Bhattacharya[*1], **Abhinav Kumar**[*1], **Barbara Di Eugenio**[1], **Roderick Tabalba**[2], **Jillian Aurisano**[3], **Veronica Grosso**[1], **Andrew Johnson**[1], **Jason Leigh**[2], and **Moira Zellner**[4]

[1]University of Illinois Chicago
{abhatt62, akumar34, bdieugen, vgross3, ajohnson}@uic.edu
[2]University of Hawaii at Manoa{tabalbar, leighj}@hawaii.edu
[3]University of Cincinnatijillian.aurisano@uc.edu
[4]Northeastern University m.zellner@northeastern.edu

Abstract

In the context of data visualization, as in other grounded settings, referents are created by the task the agents engage in and are salient because they belong to the shared physical setting. Our focus is on resolving references to visualizations on large displays; crucially, reference resolution is directly involved in the process of creating new entities, namely new visualizations. First, we developed a reference resolution model for a conversational assistant. We trained the assistant on controlled dialogues for data visualizations involving a single user. Second, we ported the conversational assistant including its reference resolution model to a different domain, supporting two users collaborating on a data exploration task. We explore how the new setting affects reference detection and resolution; we compare the performance in the controlled vs unconstrained setting, and discuss the general lessons that we draw from this adaptation.

1 Introduction

Conversation is understood in context. When the world, whether real or simulated, can change because of the user's actions, new entities are created by the processes that change the world itself: then, reference resolution, which links what the user refers to with objects in the world, is crucial for a dialogue system to effectively respond to the user, including by creating new entities.

Our overall research program aims to develop and deploy flexible conversational assistants to support users, whether causal or professional, and whether alone or in teams, explore data via visualizations on large screen displays - large screen displays better support exploration and collaboration (Andrews et al., 2011; Rupprecht et al., 2019; Lischke et al., 2020). In this paper, we focus on new entity establishment via reference in such contexts. We start from the corpus *Chicago-Crime-Vis* we collected a few years back (Kumar et al., 2016, 2017) in which a user exploring crime data in Chicago interacts with a Visualization Expert (VE) whom they know to be a person generating visualizations on the screen remotely from a separate room. On the basis of *Chicago-Crime-Vis*, we designed and developed a version of our assistant which was called *Articulate2* (Aurisano et al., 2016; Kumar et al., 2020) [1]. We will report the performance of *Articulate2* on reference resolution, and especially reference establishment, with respect to the transcribed and annotated *Chicago-Crime-Vis* corpus, evaluated in an offline manner. The second part of our paper discusses the challenges that arose when we ported *Articulate2* to a new setting: two collaborators work together to assess COVID policies given geographic and demographic features of the data, and interact exclusively with the deployed *Articulate+* (see Figure 1). We will illustrate the many issues which degrade performance, from speech processing errors, to the adaptation of models to new domains, to the inherently more complex setting in which the assistant is now behaving like an overhearer of somebody else's conversations. For clarity, we will refer to *Articulate2* in the city crime domain as *Art-City-Asst*, and to *Articulate+* in the COVID domain, as *Art-COVID-Asst*.

A disclaimer before we proceed: the purpose

[*]Co-first authors

[1]The first interface we developed in this space was called Articulate (Sun et al., 2010).

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 370–380
September 11–15, 2023. ©2023 Association for Computational Linguistics

of this work was to adapt a previously developed conversational assistant and to evaluate it in a more unconstrained setting. We do not believe in chasing after the latest shiny approach, including ChatGPT[2], and undertake a potentially infinite loop of changes which would never bring us to real user studies. Additionally, we strongly believe in ecologically valid data, such as our *Chicago-Crime-Vis* data. This data is by nature small, in fact tiny as compared to most current datasets. We will return to these issues in the Conclusions.

Figure 1: User setting for COVID data exploration, with two collaborators

2 Related Work

2.1 Conversational assistants for data visualization

Earlier work on conversational assistants for data visualization include (Cox et al., 2001), which established the benefits of using NL to generate visualizations for exploratory data analysis. In the ensuing 20 years, several such systems emerged in this area, see (Shen et al., 2023) for a systematic survey: e.g., DataTone (Gao et al., 2015a), FlowSense (Yu and Silva, 2020), Eviza (Setlur et al., 2016a) and DT2VIZ (Jiang et al., 2021). Lately, Large Language Models (LLMs) have started being integrated into visualization tools (e.g., PandasAI from OpenAI[3]), but not as part of a conversational assistant that keeps track of dialogue history.

Our previous work - the Articulate assistant series. Our research program started more than 10 years ago with *Articulate*, one of the first conversational assistants for creating data visualizations (Sun et al., 2010), as also noted by (Shen et al., 2023). The first *Articulate* would only respond to individual commands, but even so, users were 12 times faster when using *Articulate* to generate a chart in comparison to a spreadsheet program (Microsoft Excel). Still, the commands that Articulate would answer to were not grounded in actual human data; hence, we collected the *Chicago-Crime-Vis* corpus (Aurisano et al., 2015; Kumar et al., 2016, 2017) that informed a new prototype, *Articulate2*, a multimodal system that could support speech commands and gestures to facilitate data exploration tasks (Kumar et al., 2020; Kumar, 2022); and whose reference resolution component we are discussing in this paper. Subsequently, we ported *Articulate2* to the COVID domain, dubbed it *Articulate+* and developed two versions of the NLI: *Articulate+-PE* and *Articulate+-DM*. *Articulate+-PE* (Tabalba et al., 2023, 2022), was developed independently (from scratch), and works by identifying database properties or attributes mentioned directly or indirectly in the utterances. To identify the chart types given the utterance, it uses a Chart Classifier Neural Network trained on a small dataset of utterances from a preliminary user study using NLP.js library [4]. However it lacks dialogue management as well as reference resolution. The other version, *Articulate+-DM*, is *Articulate2* ported to the COVID domain. To reiterate then, in this paper we discuss the evaluation of *Articulate2*, in its incarnation as *Art-City-Asst* evaluated offline on the *Chicago-Crime-Vis* data (Section 5), and in its second incarnation in the COVID domain as *Art-COVID-Asst* evaluated in an actual user study (Section 6).

2.2 Co-Reference Resolution

This field is as old and as vast as NLP; here we focus on its applications to visualization, which are hindered by several limitations: e.g., only referents to objects within the current visualization are handled (Sun et al., 2010; Gao et al., 2015b; Narechania et al., 2020), or only referents for follow-up queries on a current visualization are tracked (Reithinger et al., 2005; Setlur et al., 2016b; Hoque et al., 2017; Srinivasan and Stasko, 2017). As (Shen et al., 2023) concludes, "existing [approaches] mostly leverage NLP toolkits to perform co-reference resolution. Although useful, they lack detailed modeling of visualization elements" or, we would add, of what has transpired earlier in the

[2]https://openai.com/product/chatgpt

[3]https://www.kdnuggets.com/2023/05/pandas-ai-generative-ai-python-library.html

[4]https://github.com/axa-group/nlp.js/

dialogue. In contrast to this, we focus on reference resolution within an environment in which visualizations are dynamically added to and removed from the screen, and can subsequently be referred to. This requires *accommodating context change*, a notion first introduced by (Webber and Baldwin, 1992) in their discussion of new entities that are the results of physical processes as in cooking (e.g., *the dough* resulting from *mixing flour, butter and water*). In the 30 years since, not much work has been done on how to accommodate the creation of new entities[5] (see (Wilson et al., 2016) for documents and (Li and Boyer, 2016) for tutoring dialogues about programming), and none in the visualization domain. Note we do not focus on multimodal reference resolution, another vast area (Navarretta, 2011; Qu and Chai, 2008; Eisenstein and Davis, 2006; Prasov and Chai, 2008; Iida et al., 2011; Kim et al., 2017; Sluÿters et al., 2022), even if we will briefly touch on deictic gestures in Section 3.

3 Controlled Dataset: Chicago-Crime-Vis

Our *Chicago-Crime-Vis* corpus comprises multimodal interaction for 16 subjects that explored public crime data in our city to better deploy police officers.[6] As noted, they spoke with a human VE who remotely created visualizations on a large screen, was not visible and did not speak back. The corpus contains 3.2K utterances. Since the user was encouraged to reason out loud about the patterns discovered via visualization, conversational turns often start with *think aloud*, followed by what we call an *actionable request* (AR) for the VE.

Using ANVIL (Kipp, 2001, 2014), we annotated 449 CARs (*contextual actionable requests*), covering 1545 utterances: a CAR consists of *setup*, i.e. think aloud prior to the AR (up to and including utterances that mention data attributes, if any); the AR; and the *conclusion*, the think aloud subsequent to the AR (also based on data-attribute mentions). While each AR is just one utterance, each of set-up and conclusion may include more than one —on average, 1.8 and 2 respectively. (See Table 1, *Chicago-Crime-Vis*(H) column for the distribution of set-ups and ARs annotated in the dataset). Figure 2 shows two CARs from our corpus, which we will use as our running example.

	Chicago-Crime-Vis (H)	COVID (A)	COVID (T)
Set-up	218	73	149
AR	449	1296	2563

Table 1: Total count of Set-ups and ARs in the 3 user studies —H: Human; A: Automatic; T: Transcript

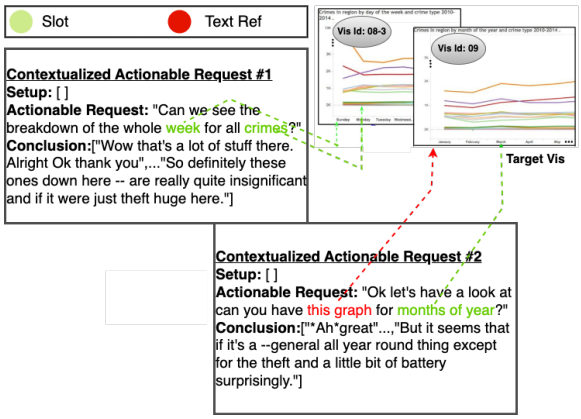


Figure 2: Excerpt comprising two CARs;references shown in *red* and slot fillers in *green*. In CAR #1 visualization *"*$08-3$*"* is specified via temporal axis DAY associated with slot filler *"week"* and similarly CRIME for *"crimes"*. CAR #2 creates *"*09*"* substituting temporal axis DAY in *"*$08-3$*"* with MONTH, associated with slot filler *"month of year"*. The identifiers are internal to the system but not visible to the users.

Each AR is annotated for user intent with one of 8 Dialogue Acts (DA) labels (with excellent intercoder agreement on the 8-way annotation, $k = 0.74$), including: WINMGMT for window management operations, e.g., closing, or minimizing; CREATEVIS for creating a new visualization from scratch; MODIFYVIS for creating a new visualization based on an existing one. The transcribed corpus is publicly available[7], and so is an augmented dataset built to alleviate data scarcity, comprising a 10-fold increase to 160 subjects covering approximately 15K utterances obtained via delexicalization and paraphrasing.

Referring Expression Annotation. We annotated both *text* (NPs) and *gestural* references to visualizations. Hand gestures were coded with various labels (e.g., the kind of gesture, the objects pointed to on the screen, and so on); approximately

[5]Work in formal pragmatics that models extra-linguistic context exists - e.g. see (Stojnic et al., 2013; Hunter, 2014), but as far as we know, it has not been used to model references in actual physical contexts.

[6]We acknowledge that this task may be fraught in the era of Black Lives Matter in the United States. This data was collected prior to 2020, when the current awakening as concerns policing and racism surfaced to public consciousness.

[7]https://github.com/uic-nlp-lab/Chicago-crime-vis-corpus

Category	Setup	AR
Overall	19	109
Single Referents	18	86
Single Targets	14	66

Table 2: *Chicago-Crime-Vis* text reference distribution

a third were identified as referential when they co-occur with text references. We labeled a total of 294 references in the 449 CAR's, of which 176 textual, and 118 gesture. We obtained an excellent intercoder agreement of $\kappa = 0.85$ with 2 judges on the full interaction from one subject. Given lack of space, and because in our unconstrained setting gestures were not addressed, we will not discuss gestures further. Table 2 shows the text reference distribution where within the 176 text references (of which 19 appear in set-up, 109 in AR, and 58 in conclusions). We also annotated 680 phrases as slot fillers corresponding to data attributes (i.e., *slots*) in our knowledge ontology (KO). The KO was semi-automatically constructed via external sources such as our city portal, augmented with synsets extracted from Wordnet[8] and Babelnet [9]; it comprises 3.5K total terms categorized into 11 parent types such as CRIME TYPE, NEIGHBORHOOD, TIME etc, of which about half are common nouns and about half proper nouns pertaining to Chicago.

4 Co-Reference: Detection, Resolution, and New Entity Establishment

We briefly discuss the NLP engine (in the context of the full conversational assistant, see Figure 3), focusing on its reference resolution component - full details on the NLP engine can be found in (Kumar et al., 2020; Kumar, 2022). The NLU pipeline relies on an information state architecture with dialogue state tracking. After speech recognition (please see below for further discussion), traditional parsing and semantic role labeling are performed, and then a semantic frame is computed (see below). The dialogue management module is responsible for: classifying the intent of the user as one of the 8 DAs mentioned in Section 3; performing reference resolution; and updating and maintaining the dialogue history (DH). The NLP engine transforms the user request (when appropriate) into an SQL query; and in a visualization specification that is passed to Vega-Lite2[10], a separate visualization interface software, to create a visualization of the data returned by the SQL query and add it to the display.

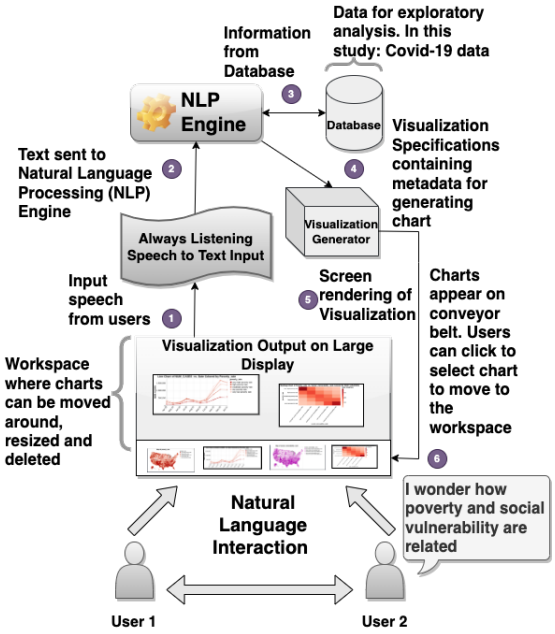


Figure 3: The Conversational Assistant—in its COVID incarnation, with two collaborators. The annotated arrows denote the workflow of the architecture, the numbers signify the order of events when the users interact with the conversational assistant.

4.1 Semantic Frame Construction

Each time a visualization is mentioned in the dialogue (whether it refers to a previous one or not) our model looks for slots in the request to form its semantic frame. We find phrases that are in close proximity in the embedding vector space to terms in the KO, by using a domain targeted word embedding model (WEM)[11]. Subsequently the candidate words are pruned based on linguistic patterns using the SpaCy[12] dependency parse of the entire utterance to form the final list of slot fillers. For example in the AR in CAR #2 in Figure 2, the prepositional phrase *"for months of year"* contains *"month"* and *"year"*, both of which are known as temporal slots in KO. Here, the terms are merged to form *"months of year"*, and mapped to the parent slot MONTH - see User Action (1) in Figure 4.

[8]https://wordnet.princeton.edu/
[9]https://babelnet.org/
[10]https://vega.github.io/vega-lite/
[11]100-dimensional continuous bag-of-word model trained on 5GB of online articles and wikipedia pages related to crime.
[12]http://spacy.io

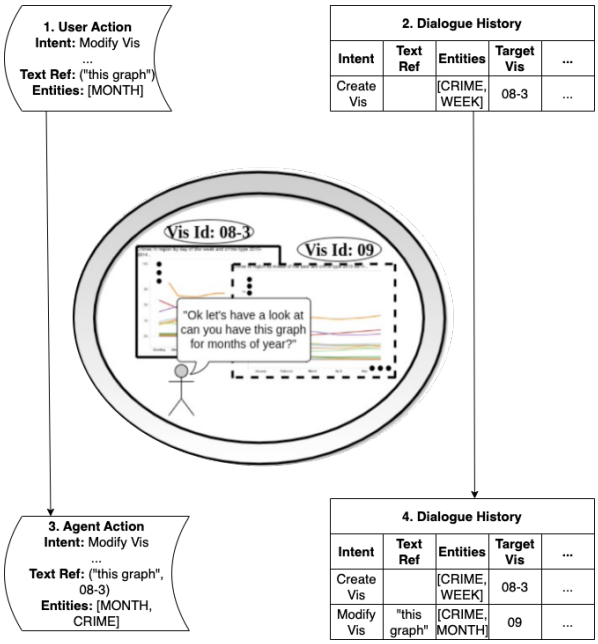


Figure 4: The user (inside the circle) currently has visualization $''08-3''$ on the screen and is asking to construct a new visualization $''09''$(in dashed lines since it is being built - visualization identifiers are internal to the system but not visible to the users). Reference resolution operates in four stages. NLU creates user action (1); the DM uses DH (2) to create agent action (3); finally the state tracker updates DH (4).

4.2 Dialogue Manager (DM)

The DM executes a dialogue policy which aside from making back-end decisions such as forming an SQL query for data retrieval, also seeks to populate unknown frame attribute values - semantic frames are constructed in response to either a CREATEVIS or a MODIFYVIS DA, and in the case of MODIFYVIS, reference resolution may be used to fill some of those unknown values. When the semantic frame is complete, the state tracker adds it as a new entry to the DH while the system also outputs a json object (which we call a *visualization specification*) that instructs Vega-Lite2 to accordingly update the screen.

For example, in Figure 4, in CAR 1, AR #1 "Can we see the breakdown of the whole week for all crimes"? has resulted in updating both the DH and the screen with a new linechart (the new visualization $''08-3''$). After AR #1, the DH contains a single entry for $''08-3''$ and its specifications in a frame-slot format, including: the user intent (CREATEVIS), the type of plot, and its semantic frame in terms of attributes that were mentioned (*crime, week*) - see Dialogue History 2 in Figure 4. Note that IDs like $''08-3''$ are for internal reference, and not shown to the user, but are included in Figures 2 and 4 for ease of exposition.

When AR #2 is processed and a MODIFYVIS DA is recognized, a new frame is created (see Agent Action 3 in Figure 4); while user intent (MODIFYVIS) and some slots (MONTH) are filled, others are left empty either because of underspecification by the user (e.g., axes labels, plot type, and so on) or they require additional processing by the DM; in this particular case, the previous visualization $''08-3''$ will be found as the referent for *this graph* and both CRIME will be added as an additional slot, and the plot type will be inferred to be line chart (see below) - see Dialogue History 4 in Figure 4.

Next, we describe reference detection and resolution.

4.2.1 Reference Detection

We trained a sequence tagging model to detect text references (DTR). The model predicts tags using the standard IOB2 format (i.e., *"B-REF"/"I-REF"/"O-REF"* for beginning of / inside / outside text reference respectively). We trained a simple CRF model that uses POS tags as features, and two baseline models, BiLSTM-CRF and BERT-CRF. Further, to remedy data insufficiency - there are only 176 text references appearing across 449 CARs in the corpus, we investigated Sequential Transfer Task Learners and Multi Task Learners, in both cases, as applied to BiLSTM-CRF and BERT-CRF. As transfer or additional task, we use a NER task based on our augmented dataset, which is also automatically labelled for 23 NER tags, based on the B/I/O scheme: the *"B"* and *"I"* tag for each of the 11 parent slots in the KO (e.g., B-visualization, I-visualization) plus *"O"* tag (the slot names are known because they are manually labelled in the 449 CARs and delexicalization maintains their type).

4.2.2 Reference Resolution

To understand to which visualization the current referring expression refers, we use heuristics based on recency and similarity. The slot fillers from the frame of the current referring expression and from the candidate visualizations in the DH are transformed into *visualization vectors*, ie, they are projected onto an embedding space along 11 dimensions, corresponding to the 11 slots in the KO, using the WEM mentioned earlier. Before compar-

ing the two visualization vectors, a recency factor is applied. If n represents the total entries in the DH, then the visualization vectors of the most recent $\frac{n}{2}$ entries in the DH are associated with a multiplicative factor of 1.0 signifying that they are equally preferred. The latter $\frac{n}{2}$ entries in the DH however are associated with a linear decrease by a factor of $\frac{1}{n}$. Finally, cosine similarity is used to score each visualization in the DH relative to the referring expression and the visualization with the highest score is selected, as long as it exceeds a cut-off of 0.40 (established empirically).

For example in Figure 4, the DH contains only an entry for $''08-3''$ (other earlier visualizations must have been closed and are not relevant any more). Since the cosine similarity score between $''08-3''$ and the current semantic structure exceeds 0.4, $''08-3''$ is chosen as the referent for *this graph*.

4.2.3 New Entity Establishment

Once the referent of the specific referring expression has been established, a new visualization ($''09''$) is constructed using the referent's frame representation to infer missing information ($''08-03''$). Explicit information in the current request is used to replace identical slots: e.g. MONTH, which was used to resolve the referring expression via WE embedding and cosine similarity among semantic structures, replaces WEEK as the temporal axis in $''09''$. Information that is unspecified in the request but present in the referred-to visualization is imported to establish the new visualization; in this particular case, CRIME is added to the slot list because $''08-3''$ of the previous request includes it. Finally, to generate the new visualization corresponding to a referring expression, the chart type (heat map, line chart, or bar graph) also needs to be inferred; it is simply copied from the referent, resulting in the new linechart $''09''$ being added to the screen, and the updated entry being added to DH ($\#4$ in Figure 4).

5 Constrained Evaluation on Chicago-Crime-Vis

The results we present now were obtained by manually evaluating the pipeline, which was run on the transcribed *Chicago-Crime-Vis* data in an offline manner: hence, we did not have to contend with speech errors, or with error propagation, since for every utterance, the DH up to that point was reset to a correct state if necessary. Currently, our model focuses on references occurring in *setup* and *AR* for detection, and in AR only for evaluation of semantic frame correctness. Additionally, we focus on single referents and single targets: e.g. in *"Can you bring up the graph behind the River North one?"* the user refers to two visualizations; whereas *"well I would like to see battery by day of week, battery by month, and battery by year."* results in 3 new corresponding visualizations. However, our model only adds one of these visualizations to the dialogue history (DH) as part of the evaluation. Table 2 presents text reference counts only for setup and ARs (hence, excluding 58 references in conclusion). Single referents account for about 94.7% of references in setup and for about 80% of those in ARs. Finally, when filtering on single targets, we are left with the 80 text references (last row in table) on which we will focus.

5.1 Detection

Notwithstanding the lack of training data, the CRF performed the best, achieving a 61.2% F1 on the B-REF, I-REF, O-REF task. This is statistically significantly better than any other models (the next best is Multitask BERT-CRF with F1= 43.5%). Hence, the CRF model is used in the subsequent steps in the pipeline. The five-fold cross validation accuracy of this CRF model on the Chicago-Crime-Vis data is shown in Table 3.

5.2 Resolution

Accuracy on resolving text references for varying *WINDOW* sizes is shown in Table 4. If one only takes into account the visualization introduced by the preceding AR (recall that we currently don't deal with multiple references), accuracy is 85.3% for set-up and 74.4% for AR. Interestingly, in the *Chicago-Crime-Vis* corpus, users also refer to the most recent visualization over 75% of the time. However, when we provide unlimited window size (∞ means all referent visualization candidates are eligible), resolution of references in ARs decreases; this suggests our linear decay function may need further tuning to better model the user preference behavior.

5.3 Semantic Frame Accuracy

We report the performance of semantic structure construction as concerns CREATEVIS and MODIFYVIS AR's. Our model achieved a slot accuracy metric (Takanobu et al., 2020) of 66.2% for semantic slots: this concerns the specification of the slots

	Chicago-Crime-Vis	COVID (A)	COVID (T)
Set-up	60.0	50.0	33.3
AR	55.0	25.0	45.8

Table 3: Evaluation of reference detection model. Chicago-Crime-Vis: five-fold cross validation accuracy calculated on Single Targets of Table 2; COVID (A): Accuracy in real-time user study; COVID (T): Accuracy on correct transcripts of real-time user study. COVID (A) and COVID (T) evaluated on a significant sample size

	Setup Window		AR Window	
	1	∞	1	∞
Chicago-Crime-Vis	85.3	85.3	74.4	68.3
COVID (T)	-	-	36.3	54.0

Table 4: Resolution accuracy for varying window sizes. COVID (T) evaluated on a significant sample size

of the *Visualization Frame(VH)* in the DH, and includes slots that were explicit in the utterance, and those that were inferred. Given the example in Figure 2, for "$08-3$" the two slot values are "crime" and "week", and for "09" "month" (explicit) and "crime", inferred via reference resolution. Table 5 reports the number of VFs for which a certain percentage of slots has been correctly recognized, by quartile. The 100% quartile is equivalent to the *Joint Goal Accuracy (JGA)* metric used in some of the Dialogue State Tracking challenges, which *compares the predicted dialog states to the ground truth at each dialog turn, and the output is considered correct if and only if all the predicted values exactly match the ground truth* (Takanobu et al., 2020). For the *Chicago-Crime-Vis*, these were manually annotated when annotating for references, and the results are computed by evaluating the resolution pipeline turn by turn, with the gold-standard DH up to the previous turn: in 131 of those (55%), all slots were correctly recognized; in 83% of these VFs, at least 75% of the slots were correct; only in 17 (7%) of these 238 VFs, no slots were correctly recognized. Beyond Joint Goal Accuracy, we report partial accuracy to provide a more nuanced analysis of the assistant's performance, which cannot be simply measured in a binary "Correct/Incorrect" fashion: in an dialogue based application for data exploration like ours, a partially recognized vizualization frame can generate charts which may help the users move forward. Papers exploring similar views are Selfridge et al. (2011) and Schlangen et al. (2009), where partial speech recognition and reference resolution were found to be beneficial for dialogue systems that react satisfactorily to the user.

	0%	25%	50%	75%	100% (JGA)	Total VF
Chicago-Crime-Vis	17	5	19	66	131	238
COVID (A)	22	1	25	8	66	122
COVID (T)	23	4	25	15	75	142

Table 5: Distribution of *Visualization Frames* wrt % correct slots. COVID (A) and COVID (T) evaluated on a *significant sample size*

6 Unconstrained setting: User studies in a COVID domain

A realistic evaluation of the NLP Architecture was conducted through user studies: pairs of participants interact with the conversational assistant (*Art-COVID-Asst*) to perform two open-ended exploratory data analysis tasks, concerning which factors may affect COVID mitigation strategies, such as access to doctors or elderly population. Overall, 15 groups of 2 participants, performed the two tasks in a specified sequence, within a time limit of 25 minutes per task. The participants, aged $18+$, were recruited from UIC and were mostly graduate students. With their consent, we audio and video recorded them, and collected logs generated by the back-end code of *Art-COVID-Asst* for analysis purpose. As shown in Figure 1, they are sitting and wearing a mike; also, each has a mouse with which they are able to reposition and click the visualizations on the screen. We encouraged the users to freely interact with each other and with *Art-COVID-Asst*, and we did not provide specific instructions about the tasks, the interface, or the collaboration. The system is designed to "always listen" to the participants, whether or not they are addressing the assistant directly. This is implemented using the Web Speech API[13].

It was relatively simple to port *Art-City-Asst* to *Art-COVID-Asst* (Figure 3 shows the architecture) and mostly required to update the KO. For the COVID data, we identify 13 semantic slots like "COVID vulnerability rank", "Access to doctors", "Diabetes risk", "Uninsured rate" etc. and the possible values for these slots. As earlier, we enlarged the KO with synonyms for each slot and their values by using Wordnet and Babelnet to generate

[13] https://wicg.github.io/speech-api/

these synonyms. The generated KO has a vocabulary of 710 terms. This, as we describe in Section 4 forms the backbone of semantic slot filling and new entity establishment. We keep the same Dialogue Manager as before and use the best Reference Detection model built using the *Chicago-Crime-Vis* corpus, namely, the CRF model. The Reference resolution algorithm also remains the same. Finally for screen rendering of the generated charts, the relevant data obtained from the database is converted to Vega-Lite grammar.

6.1 Findings of the User Study

To evaluate the reference detection and resolution pipeline in this setting, in principle we only need the log of the interactions to assess real-time performance wrt the utterances from the conversations of the participants. However, after we realized that speech recognition errors were a major bottleneck in the real-time study, we conducted additional experiments on the transcripts. These are generated using the Whisper speech recognition model [14] followed by light manual inspection. The corrected transcripts are then fed to the back-end code of the conversational assistant and new logs are generated. We name this version of the user study data as COVID (T) (for *Transcript*), while the real-time logs are named COVID (A) (for *Automatic*).

Since, as we noted earlier, reference detection applies to set-up and ARs, Table 1 shows the distribution of setups and requests in these two versions along with those from the Chicago-Crime-Vis corpus. An important difference is that set-ups and ARs for *Chicago-Crime-Vis* were manually annotated, whereas these are the results of automatic recognition for the COVID study (whether A or T). The table shows that there are many more set-ups in the *Chicago-Crime-Vis* data; this difference is significant, as confirmed by $\chi^2 = 489.9511, p < 0.00001$ (with Bonferroni correction). There may be various reasons for this, one being that the classifiers that recognize setup and ARs were trained on the augmented *Chicago-Crime-Vis* corpus and perform worse here to start with. However, it is also possible that in fact, think aloud that feels natural when somebody is by themselves is not in a collaborative situation: a set-up by definition doesn't talk about a data attribute, but we surmise that the two collaborators are more focused on data attributes than on thinking aloud, precisely because they are interacting with another person.

For the purpose of the evaluation, we need to manually verify the results returned by the reference pipeline. Given the size of the data, we obtain two samples, one from COVID (A) (# utterances: 3096) and one from COVID (T) (# utterances: 8440). A significant sample size is computed for both with 95% confidence interval and 5% margin of error. This results in a random sample of 340 (11%) utterances for COVID (A), and of 370 (4.38%) utterances for COVID (T). Subsequently, we use COVID (A) and COVID (T) to refer to these samples of the respective groups, not to the whole group; all evaluation and analysis are done on these samples only.

6.1.1 Reference Detection

Table 3 shows the accuracy of the detected references in Set-up and Request utterances of COVID (A) and COVID (T). As expected, the performance degrades in a real-time user study scenario. Unlike the controlled study setting with one participant, when two people collaborate for an exploratory task, three things happen. First they talk to each other; next, they make requests to the system and finally they draw conclusions. These make reference detection in utterances extremely complex. In the case of COVID (A), we also attribute the lack of accuracy to speech-recognition errors.

6.1.2 Reference Resolution

We limit the evaluation of the reference resolution pipeline to COVID(T) as there were no references resolved during the actual study—DAs of around 44% of those utterances with detected references were misclassified (note that useful visualizations may have been created all the same in response to those specific utterances, but not because a reference resolution was resolved). After conducting a thorough manual inspection of the issue we find the speech recognition errors to be the major roadblock yet again. However using the corrected transcript (COVID (T)) we get a comparatively better performance as shown in Table 4. Since in this study setting, only ARs where references are detected are resolved, we limit our evaluation to ARs only. Contrary to the constrained *Chicago-Crime-Vis* setting, where considering only the previous AR was the better strategy, here limiting window size to 1 results in lower accuracy. We observe that in a more real scenario, especially when two peo-

[14] https://github.com/openai/whisper - it became available in September 2022, after our conversational assistant was developed and hence could not be used for the user study.

ple are involved in the conversation, there are more relevant entries in the dialogue history. This may also be due to the nature of the interaction with the large screen: in *Chicago-Crime-Vis*, the user was standing in front of the large display, and often fairly close so that they would in fact mostly focus on only a portion of the display; in the COVID study, the two collaborators were sitting at about 6 ft from the screen (see Figure 1), and hence all visualizations on the screen are more readily available to them.

6.1.3 Semantic Frame Accuracy

For the user study settings of *Art-COVID-Asst*, VFs were recognized for utterances having DAs CREATEVIS and MODIFYVIS. Similar to what we observed in the controlled setting of Chicago-Crime-Vis (as described in Section 5.3) in Table 5, more than 50% VFs had all their semantic slots recognized as fully correct in the unconstrained settings with *Art-COVID-Asst*. In fact, we see comparable performances of COVID (A) and COVID (T) across all quartiles. This shows that irrespective of the problematic performance of the speech-to-text algorithm, more than 60% VFs had 75% or more slots correctly filled and more than 80% VFs had at least 50% slots correctly identified. This also explains the reasonable success of the user study that we observed despite the subpar performance of the speech-to-text algorithm. This is attested by questionnaires the users filled. On a 5 point Likert scale, mean scores of 4 and 3 were respectively obtained for usefulness of the charts generated, and for ease of command system use.

7 Conclusions and Future Work

We have presented a reference resolution model for conversational assistants that help user in exploratory data visualization. In particular, the model resolves visualization references in the context of the current interaction, crucially tracking visualizations constantly being added to the screen. The model is central to the creation of new visualizations: visualization features encoded in the DH as slot values, help the model know how to refer to a visualization later on. We have also shown how the initial assistant, *Art-City-Asst*, was ported to a completely different domain. We presented the evaluation of the reference pipeline in both settings, the constrained *Chicago-Crime-Vis* and the "wild" COVID setting, in which two collaborators were exploring COVID data. We are fully aware that our results are not compared to an external baseline, but we contend that evaluations in grounded settings are important, and do not require creating some artificial baseline or evaluating the pipeline on existing reference resolution datasets.

Not surprisingly, the user evaluation brought several issues to the fore. First, we discovered that the speech API that we had chosen did not work very well (it would have been impossible to change it during the user study even if we had noticed it). Whereas this is unfortunate, we were able to obtain correct transcripts and run a second evaluation. Second, the nature of the interaction and the setting affected the conversations and the results: for example, we found many fewer set-ups in the COVID data, but on the other hand, more references to referents further back in the conversation.

Potential extensions for future work include ways to better model user behavior for referring to more distant visualizations and using sophisticated machine learning approaches in our resolution algorithm to take advantage of the rich visualization feature space in our case. Additionally, in the COVID user study users don't use hand gestures to interact with the screen, however they do use their mouses to click and reposition visualizations, hence bringing multimodality to the fore; not to mention gaze that can be approximated with head movement tracking, that another researcher in the group is investigating (see the instrumented caps in Figure 1).

Finally, as we had mentioned in the introduction, our goal was to evaluate our assistant in a realistic user study, and not jump into experiments with Large Language Models. However, we have started experiments in that respect, both as concerns the specific modules in our pipeline (for example, the embeddings of the semantic slots) and the system as a whole. So far, we have noticed that while ChatGPT (released exactly after we finished the COVID user study) is able to generate charts in response to specific language instructions, if appropriately connected to visualization software, it is not able to resolve referring expressions, i.e., to create a new visualization whose specification is partly derived from the referent. But this will be the topic of a future paper.

Acknowledgments

This work is supported by awards 2007257 and 2008986 from the National Science Foundation.

References

Christopher Andrews, Alex Endert, Beth Yost, and Chris North. 2011. Information visualization on large, high-resolution displays: Issues, challenges, and opportunities. *Information Visualization*, 10(4):341–355.

Jillian Aurisano, Abhinav Kumar, Alberto Gonzales, Khairi Reda, Jason Leigh, Barbara Di Eugenio, and Andrew Johnson. 2015. "Show me data": Observational study of a conversational interface in visual data exploration. In *IEEE VIS*, volume 15, pages 1–2.

Jillian Aurisano, Abhinav Kumar, Alberto Gonzalez, Jason Leigh, Barbara Di Eugenio, and Andrew Johnson. 2016. Articulate2: Toward a conversational interface for visual data exploration. In *IEEE Visualization*, volume 8.

Kenneth C. Cox, Rebecca E. Grinter, Stacie Hibino, Lalita Jategaonkar Jagadeesan, and David Mantilla. 2001. A multi-modal natural language interface to an information visualization environment. *International Journal of Speech Technology*, 4:297–314.

Jacob Eisenstein and Randall Davis. 2006. Gesture improves coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*.

Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. 2015a. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software amp; Technology*, UIST '15, page 489–500, New York, NY, USA. Association for Computing Machinery.

Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G Karahalios. 2015b. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 489–500. ACM.

Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. 2017. Applying pragmatics principles for interaction with visual analytics. *IEEE transactions on visualization and computer graphics*, 24(1):309–318.

Julie Hunter. 2014. Structured contexts and anaphoric dependencies. *Philosophical Studies*, 168:35–58.

Ryu Iida, Masaaki Yasuhara, and Takenobu Tokunaga. 2011. Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 84–92.

Qi Jiang, Guodao Sun, Yue Dong, and Ronghua Liang. 2021. Dt2vis: A focus+ context answer generation system to facilitate visual exploration of tabular data. *IEEE Computer Graphics and Applications*, 41(5):45–56.

Hansol Kim, Kun Ha Suh, and Eui Chul Lee. 2017. Multi-modal user interface combining eye tracking and hand gesture recognition. *Journal on Multimodal User Interfaces*, 11(3):241–250.

Michael Kipp. 2001. Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*.

Michael Kipp. 2014. Anvil: The video annotation research tool. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook of Corpus Phonology*, pages 420–436. Oxford University Press.

Abhinav Kumar. 2022. *Towards a Context-Aware Intelligent Assistant for Multimodal Exploratory Visualization Dialogue*. Ph.D. thesis, University of Illinois Chicago.

Abhinav Kumar, Jillian Aurisano, Barbara Di Eugenio, Andrew Johnson, Alberto Gonzalez, and Jason Leigh. 2016. Towards a dialogue system that supports rich visualizations of data. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 304–309, Los Angeles. Association for Computational Linguistics.

Abhinav Kumar, Jillian Aurisano, Barbara Di Eugenio, and Andrew E Johnson. 2020. Intelligent assistant for exploring data visualizations. In *FLAIRS Conference*, pages 538–543.

Abhinav Kumar, Barbara Di Eugenio, Jillian Aurisano, Andrew Johnson, Abeer Alsaiari, Nigel Flowers, Alberto Gonzalez, and Jason Leigh. 2017. Towards multimodal coreference resolution for exploratory data visualization dialogue: Context-based annotation and gesture identification. In *The 21st Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2017–SaarDial)(August 2017)*, volume 48.

Xiaolong Li and Kristy Boyer. 2016. Reference resolution in situated dialogue with learned semantics. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 329–338.

Lars Lischke, Lena Janietz, Anna Beham, Hartmut Bohnacker, Ulrich Schendzielorz, Albrecht Schmidt, and Paweł W Woźniak. 2020. Challenges in designing interfaces for large displays: the practitioners' point of view. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pages 1–6.

Arpit Narechania, Arjun Srinivasan, and John Stasko. 2020. Nl4dv: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):369–379.

Costanza Navarretta. 2011. Anaphora and gestures in multimodal communication. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), Faro, Portugal, Edicoes Colibri*, pages 171–181. Citeseer.

Zahar Prasov and Joyce Y Chai. 2008. What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 20–29. ACM.

Shaolin Qu and Joyce Y Chai. 2008. Beyond attention: the role of deictic gesture in intention recognition in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 237–246. ACM.

Norbert Reithinger, Dirk Fedeler, Ashwani Kumar, Christoph Lauer, Elsa Pecourt, and Laurent Romary. 2005. Miamm—a multimodal dialogue system using haptics. In *Advances in Natural Multimodal Dialogue Systems*, pages 307–332. Springer.

Franca Rupprecht, Carol Naranjo, Achim Ebert, Joseph Olakumni, and Bernd Hamann. 2019. When bigger is simply better after all: Natural and multi-modal interaction with large displays using a smartwatch. In *Proceedings of the Twelfth International Conference on Advances in Computer-Human Interactions (ACHI 2019)*.

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental reference resolution: The task, metrics for evaluation, and a Bayesian filtering model that is sensitive to disfluencies. In *Proceedings of the SIGDIAL 2009 Conference*, pages 30–37, London, UK. Association for Computational Linguistics.

Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams. 2011. Stability and accuracy in incremental speech recognition. In *Proceedings of the SIGDIAL 2011 Conference*, pages 110–119, Portland, Oregon. Association for Computational Linguistics.

Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. 2016a. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, page 365–377, New York, NY, USA. Association for Computing Machinery.

Vidya Setlur, Sarah E Battersby, Melanie Tory, Rich Gossweiler, and Angel X Chang. 2016b. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 365–377. ACM.

Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang. 2023. Towards natural language interfaces for data visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 29(6):3121–3144.

Arthur Sluÿters, Quentin Sellier, Jean Vanderdonckt, Vik Parthiban, and Pattie Maes. 2022. Consistent, continuous, and customizable mid-air gesture interaction for browsing multimedia objects on large displays. *International Journal of Human–Computer Interaction*, pages 1–32.

Arjun Srinivasan and John Stasko. 2017. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. *IEEE transactions on visualization and computer graphics*, 24(1):511–521.

Una Stojnic, Matthew Stone, and Ernie Lepore. 2013. Deixis (even without pointing). *Philosophical Perspectives*, 27(1):502–525.

Yiwen Sun, Jason Leigh, Andrew Johnson, and Sangyoon Lee. 2010. Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In *International Symposium on Smart Graphics*, pages 184–195. Springer.

Roderick S. Tabalba, Nurit Kirshenbaum, Jason Leigh, Abari Bhattacharya, Veronica Grosso, Barbara Di Eugenio, Andrew E. Johnson, and Moira Zellner. 2023. An investigation into an always listening interface to support data exploration. *Proceedings of the 28th International Conference on Intelligent User Interfaces*.

Roderick S. Tabalba, Nurit Kirshenbaum, Jason Leigh, Abari Bhattacharya, Andrew E. Johnson, Veronica Grosso, Barbara Maria Di Eugenio, and Moira Zellner. 2022. Articulate+ : An always-listening natural language interface for creating data visualizations. *Proceedings of the 4th Conference on Conversational User Interfaces*.

Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. 2020. Is your goal-oriented dialog model performing really well? empirical analysis of system-wise evaluation. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 297–310, 1st virtual meeting. Association for Computational Linguistics.

Bonnie Lynn Webber and Breck Baldwin. 1992. Accommodating context change. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 96–103, Newark, Delaware, USA. Association for Computational Linguistics.

Shomir Wilson, Alan W Black, and Jon Oberlander. 2016. This table is different: A wordnet-based approach to identifying references to document entities. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 432–440.

B. Yu and C. T. Silva. 2020. Flowsense: A natural language interface for visual data exploration within a dataflow system. *IEEE Transactions on Visualization amp; Computer Graphics*, 26(01):1–11.

CONVERSER: Few-Shot Conversational Dense Retrieval with Synthetic Data Generation

Chao-Wei Huang Chen-Yu Hsu Tsu-Yuan Hsu Chen-An Li Yun-Nung Chen
National Taiwan University, Taipei, Taiwan
f07922069@csie.ntu.edu.tw y.v.chen@ieee.org

Abstract

Conversational search provides a natural interface for information retrieval (IR). Recent approaches have demonstrated promising results in applying dense retrieval to conversational IR. However, training dense retrievers requires large amounts of in-domain paired data. This hinders the development of conversational dense retrievers, as abundant in-domain conversations are expensive to collect. In this paper, we propose CONVERSER, a framework for training conversational dense retrievers with at most 6 examples of in-domain dialogues. Specifically, we utilize the in-context learning capability of large language models to generate conversational queries given a passage in the retrieval corpus. Experimental results on conversational retrieval benchmarks OR-QuAC and TREC CAsT 19 show that the proposed CONVERSER achieves comparable performance to fully-supervised models, demonstrating the effectiveness of our proposed framework in few-shot conversational dense retrieval.[1]

1 Introduction

Conversational information retrieval (CIR) has been an important area of research in recent years, aiming to retrieve relevant information from a large corpus of text in a conversational format. It has gained considerable interest due to its potential to deliver information in a natural format in response to a user's queries. Unlike traditional IR, CIR poses distinctive challenges, including its multi-turn and context-dependent nature, which require more nuanced approaches (Yu et al., 2021; Fang et al., 2022).

Dense retrieval methods have demonstrated their ability to understand the semantics of complex user queries and shown promising performance on open-domain retrieval (Karpukhin et al., 2020). One of the major obstacles to conversational dense retrieval is the scarcity of training data, given the high cost and extensive time to collect high-quality information-seeking conversations (Adlakha et al., 2022). Previous work has explored various approaches to address this issue (Dai et al., 2022; Kim et al., 2022). However, most methods still rely on the assumption that a large amount of in-domain data is present and build data augmentation models upon it.

In this paper, we aim to develop a few-shot conversational dense retrieval model that can effectively retrieve relevant passages based on a small number of in-domain dialogues. To achieve this, we leverage the in-context learning capability of large language models (LLMs) to generate synthetic passage-dialogue pairs with few-shot demonstrations. Specifically, in-domain passages are sampled from the retrieval corpus, and dialogues are synthesized by asking LLMs to generate a series of queries based on a few examples. We also employ a self-consistency filtering mechanism to automatically discard inconsistent generated queries, ensuring the accuracy and reliability of the generations.

We conduct experiments on two benchmark datasets, including OR-QuAC (Qu et al., 2020) and TREC CAsT 19 (Dalton et al., 2019). The experimental results demonstrate that our proposed framework, CONVERSER, performs comparably to fully-supervised models that are trained on *thousands* of annotated dialogues while using only 6 examples at most. Furthermore, analyses show that CONVERSER rivals other data augmentation methods that utilize full in-domain datasets, demonstrating its effectiveness.

2 Related Work

Conversational Dense Retrieval Conversational dense retrieval poses a unique challenge in that the questions are context-dependent. Prior works have explored various modeling techniques for conver-

[1]All source code and generated datasets are available: https://github.com/MiuLab/CONVERSER

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 381–387
September 11–15, 2023. ©2023 Association for Computational Linguistics

sational history to address this challenge (Huang et al., 2018; Choi et al., 2018; Yeh and Chen, 2019; Chiang et al., 2020). However, these works only examined the modeling ability for conversational question answering (CQA), where the relevant passages are provided.

More recently, Qu et al. (2020) proposed OR-ConvQA, which extends CQA to the open-domain setting where a retrieval module is required. ConvDR (Yu et al., 2021) utilizes an ad-hoc dense retriever and manually rewritten context-independent queries for training few-shot retrievers and rerankers, while our method does not require an ad-hoc model and additional annotation. Others have explored various methods for encoding conversational queries (Li et al., 2021; Fang et al., 2022; Wu et al., 2022; Liang et al., 2022), which are orthogonal to our work.

2.1 Synthetic Data Generation for Dense Retrieval

Due to the data-hungry nature of dense retrievers, synthetic data generation for dense retrieval has drawn considerable interest.

Previous works have worked on generating information-seeking conversations via transforming documents (Dai et al., 2022; Kim et al., 2022) or web search sessions (Mao et al., 2022). However, these methods all require training query generators with conversational data, which does not mitigate the data scarcity issue. Our method requires only 6 in-domain dialogues with their relevant passages and demonstrates comparable performance to models trained on thousands of manually annotated dialogues.

InPars (Bonifacio et al., 2022) and Promptagator (Dai et al., 2023) are the most closely related works to our method. They both proposed to generate synthetic queries with LLMs from few-shot examples, which achieved comparable performance to supervised methods in dense retrieval. Inspired by these works, our method further extends few-shot query generation to the conversational setting. We propose novel techniques for generating conversational queries and show that they are crucial to handle the unique challenges of conversational dense retrieval.

3 Proposed Method: CONVERSER

We propose few-shot conversational dense retrieval with synthetic data generation, CONVERSER, which aims to generate synthetic conversational queries given few examples. More formally, given a conversational retrieval task T, its retrieval corpus $\mathcal{P}_T$, and k examples, we aim to generate synthetic conversational query-passage pairs $\{\hat{C}_1, \cdots, \hat{C}_n\}$ for training dense retrievers.

3.1 Few-Shot Conversational Query Generation

The core of our method is *few-shot query generation*. We leverage the in-context learning ability of LLMs (Brown et al., 2020) to generate conversational queries. Specifically, we start with k examples $\{C_1, C_2, \cdots, C_k\}$, where each C_i is a conversation represented as a series of query-passage pairs, $(q_i^1, p_i^1), \cdots, (q_i^{n_i}, p_i^{n_i})$, with n_i denoting the length of C_i. Using these examples, we construct the following template $\mathcal{T}$ as a few-shot demonstration for LLMs:

$$\left[(p_1^{n_1}, q_1^1, \cdots, q_1^{n_1}), \cdots, (p_k^{n_k}, q_k^1, \cdots, q_k^{n_k})\right]$$

Note that we always choose the relevant passage that corresponds to the last query in the examplar, indicating that the last query $q_i^{n_i}$ is generated given $p_i^{n_i}$ and previous queries $q_i^1, \cdots, q_i^{n_i-1}$.

The generation process for a synthetic conversation starts with randomly sampling a passage $\hat{p}$ from the retrieval corpus, i.e., $\hat{p} \sim \mathcal{P}_T$. We concatenate the template and the sampled passage to form an input text sequence $[\mathcal{T}, \hat{p}]$. An LLM is employed for generating synthetic queries. It is expected to generate the first query $\hat{q}_1$ that is relevant to $\hat{p}$ based on the provided examples. We then append $\hat{q}_1$ to the input sequence, forming the input sequence for generating the next query $\hat{q}_2$, and so forth. We sequentially perform the generations for a conversation until a predefined number of turns is reached.

3.2 Two-Stage Generation

One unique characteristic of conversational queries is that the queries are *context-dependent* (Choi et al., 2018) except for the first query, which should be a self-contained query without any ambiguity. To address this difference, we propose to split the generations into two-stage: first query generation and follow-up query generation. When generating the first query for each conversation, we use an alternative template $\mathcal{T}_1 = \left[p_1^1, q_1^1, \cdots, p_k^1, q_k^1\right]$, which contains only the first queries and their relevant passages of the examples. We then replace

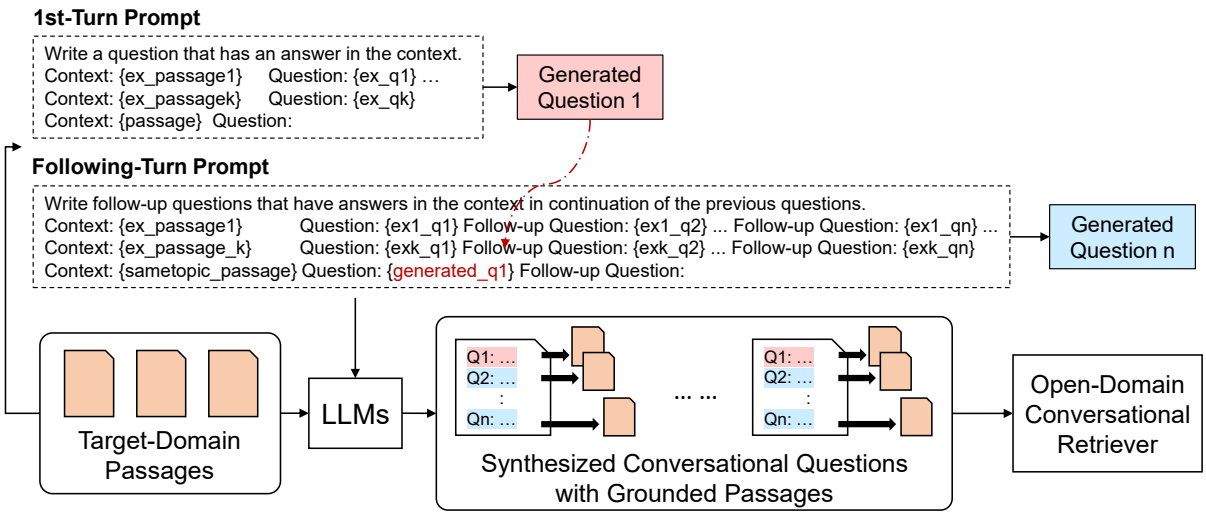


Figure 1: Illustration of our proposed framework.

$\mathcal{T}_1$ with $\mathcal{T}$ for generating all the follow-up queries. In practice, we found that this two-stage approach reduces the number of generated first queries that are not self-contained and thus ambiguous.

3.3 Passage Switching

In a conversation, relevant passages may vary for different queries. To this end, we incorporate passage switching into the generation process. We randomly replace the current passage $\hat{p}$ with a related passage $\hat{p}'$ in each turn with a probability p_{ps}. The LLM is expected to generate queries based on the new passage.

3.4 Consistency Filtering

The generation process sometimes generates queries that are nonsensical, degenerated, ambiguous, or not grounded by the given passage. We adopt a filtering mechanism via ensuring *round-trip consistency* (Alberti et al., 2019). We follow the procedure in Dai et al. (2023), where an initial retriever is trained on all synthetic query-passage pairs. For each synthetic pair $(\hat{q}, \hat{p})$, we use the initial retriever to retrieve the most relevant passages for $\hat{q}$ from $\mathcal{P}_T$. We keep the pair $(\hat{q}, \hat{p})$ only if $\hat{p}$ is in the top-k retrieved passages.

4 Experiments

To evaluate if our generated conversational questions can help train a conversational retriever, we conduct experiments on a conversational question answering dataset, OR-QuAC (Qu et al., 2020), and a conversational search benchmark, TREC CAsT-19 (Dalton et al., 2019).

4.1 Experimental Setup

We describe our experimental setup in the section. Additional details can be found in Appendix A.

Few-Shot Examples We manually select 6 examples for OR-QuAC and 5 examples for CAsT-19 and use the same set of examples in all experiments. Due to resource constraints, we use the remaining 15 conversations for evaluating on CAsT-19 without performing 5-fold cross-validation.

Generation We employ LLaMA-13B (Touvron et al., 2023) as our pretrained LLM, which is not instruction-tuned and is open to the research community. We use nucleus sampling (Holtzman et al., 2020) for decoding and set top_p $= 0.95$, temperature $= 0.75$. We generate 427k turns (61k conversations) for OR-QuAC and 230k turns (32k conversations) for An example of generation results can be found in Section 5.

Retrieval Corpus We generate synthetic conversations based on the retrieval corpus for each task respectively. For OR-QuAC, we use the provided 11M passages from English Wikipedia. For TREC CAsT-19, we use the official passage collection, which consists of 8M webpage passages from MS-MARCO (Bajaj et al., 2016) and 30M Wikipedia passages from TREC-CAR (Dietz et al., 2017).

Model Details We follow the procedures from DPR (Karpukhin et al., 2020) to train our retrievers and use BERT-base as the pretrained model. We concatenate all previous queries and the current query as the input to the retriever. Additional details can be found in Appendix A.

Method	OR-QuAC			CAsT-19	
	MRR@5	R@5	MAP@10	MRR	NDCG@3
Supervised OR-ConvQA (Qu et al., 2020)	22.5	31.4	-	-	-
Supervised DPR	50.5	64.7	49.7	29.4	19.1
Few-Shot CONVERSER (Ours)	49.6	63.4	48.7	35.8	21.4

Table 1: Evaluation results (%). We report the result of OR-ConvQA from the original paper.

Method	OR-QuAC	
	MRR@5	R@5
OR-QuAC	50.5	64.7
WikiDialog (31k)	44.6	58.2
CONVERSER (31k)	46.8	61.5
- Two-Stage	45.1	59.9
- Consistency Filtering	45.2	59.8
- Passage Switching	45.6	58.1
- Only 1-Shot	42.1	55.2

Table 2: Results of ablation study. We use the identical training procedure and training data size for each experiment to make them comparable.

Baseline Systems

- **OR-ConvQA**: A supervised dense retriever trained on OR-QuAC (Qu et al., 2020).
- **DPR**: We train a DPR model (Karpukhin et al., 2020) on the training set of OR-QuAC for a fair comparison.

4.2 Main Results

Table 1 shows the experimental results. Note that both ConvDR and WikiDialog utilized multiple additional datasets and techniques, which are complementary to our method. On the OR-QuAC dataset, our proposed CONVERSER outperforms the supervised baseline OR-ConvQA by a large margin and performs comparably to the supervised DPR trained on OR-QuAC. This result demonstrates the effectiveness of our few-shot generation strategy, as our model trained on a synthetic dataset based on only 6 annotated examples can rival the performance of supervised DPR, which is trained on 4000 annotated dialogues.

On CAsT-19, CONVERSER outperforms supervised DPR, which is trained on OR-QuAC. This shows that our task-specific generation strategy can effectively synthesize conversational queries on a new task given a few examples of the new task. Our proposed method provides better adaptability without requiring another supervised dataset as done in conventional transfer learning.

4.3 Ablation and Comparative Study

We conduct an ablation study on different settings of our proposed method, where we remove one component at a time to validate its effectiveness. We also compare our method with two datasets: OR-QuAC and WikiDialog (Dai et al., 2022). To ensure the results are comparable, we limit the size of every dataset to 31k turns, which is the same as the training set of OR-QuAC. The training process and hyperparameters are also identical for all datasets. For WikiDialog, we subsample the original WikiDialog dataset and use it to fine-tune a retriever, without further fine-tuning on OR-QuAC. The results are shown in Table 2.

Given the same number of synthesized turns, our CONVERSER outperforms WikiDialog, which requires supervised conversational datasets for training a query generator. This result validates the effectiveness of our proposed few-shot generation method. The ablation study demonstrates that all of our proposed components contribute to the improvement.

4.4 Effect of Generated Data Size

We explore the impact of the generated data size on the performance, where we conduct a series of experiments, systematically varying the number of generated turns used for training presented in Figure 2. It clearly illustrates that as the number of turns increases, the system's performance improves significantly. This finding highlights the crucial role of conversational data in enhancing the effectiveness of our model.

5 Qualitative Study

We present a generated example in Table 3 to perform qualitative analysis. WikiDialog is capable of generating follow-up questions. However, it often

Generated Question Example	
Title	Palazzo Primoli
Passage	Palazzo Primoli Palazzo Primoli is a Palazzo in Rome, Italy. It is owned by the city of Rome and houses several museums and collections. The palazzo was built in the seventeenth century. In 1901 Count Giuseppe Primoli (1851–1927) became its sole owner. He extended and partly modernised the palazzo with a new facade and entrance between 1901 and 1911. The Count's maternal grandparents were Charles Lucien Bonaparte and Zénaïde Bonaparte, and the Count brought together a collection of objects (now the Museo Napoleonico), documenting the relationship between Rome and the Bonaparte family. He also was an avid photographer. In 1927 Giuseppe Primoli donated the palazzo and his collections to the municipality of Rome. The Museo Napoleonico is located on the palazzo's ground floor, and the third floor is occupied by the Museo Mario Praz, the former residence of Mario Praz. Also located in the palazzo are the Count's library and photographic archive.
WikiDialog	Q1: What is known about Palazzo Primoli? Q2: Are there any other interesting aspects about this article? Q3: When was it built? Q4: Who were the previous owners of the palazzo? Q5: What was his role in the palazzo? Q6: What is the most interesting aspect of this article?
CONVERSER	Q1: What is the name of the building that houses the Museo Napoleonico? Q2: What is the Palazzo Primoli best known for? Q3: What year was it donated? Q4: What famous collection was donated? Q5: What other collections are housed in this museum? Q6: Who was the Count's maternal grandparents?
CONVERSER only 1 example	Q1: What is Palazzo Primoli in regards to Rome? Q2: What is the Palazzo Primoli's collection of objects? Q3: What is the Museo Napoleonico? Q4: What was the Count's library? Q5: What was the Count's photographic archive?

Table 3: A qualitative example from WikiDialog, CONVERSER, and CONVERSER with only 1 example.

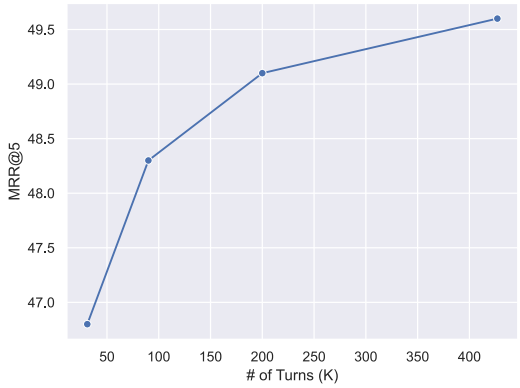


Figure 2: MRR@5 with regard to different number of generated turns on OR-QuAC.

generates generic queries, such as *Are there any other interesting aspects about this article*. On the other hand, CONVERSER with only 1 example suffers from a lack of diversity. Due to limited demonstrations, it generates queries that are very similar to the only example it is given. Our proposed CONVERSER can generate a context-independent first question and follow-up questions, demonstrating its effectiveness.

6 Conclusion

This paper introduces CONVERSER, a synthetic data generation method for training few-shot conversational dense retrievers. We leverage the in-context learning capability of LLMs and propose techniques that are designed for generating conversational queries. Experimental results demonstrate that our proposed CONVERSER achieves comparable performance to fully-supervised models while only requiring 6 annotated examples. Further analyses demonstrate that our method outperforms a fully-supervised data augmentation method. Future work could explore instruction-following LLMs, better filtering mechanisms, and synthesizing specialized data for conversational dense retrieval, such as query rewrites.

Acknowledgements

We thank the reviewers for their insightful comments. This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grants 111-2222-E-002-013-MY3, 111-2628-E-002-016, and 112-2223-E-002-012-MY5 and Google.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2387–2392, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ting-Rui Chiang, Hao-Tong Ye, and Yun-Nung Chen. 2020. An empirical study of content understanding in conversational question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7578–7585.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4558–4586. PMLR.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. Cast 2019: The conversational assistance track overview. In *TREC*.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. Trec complex answer retrieval overview. In *TREC*.

Hung-Chieh Fang, Kuo-Han Hung, Chen-Wei Huang, and Yun-Nung Chen. 2022. Open-domain conversational question answering with historical answers. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 319–326, Online only. Association for Computational Linguistics.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2018. Flowqa: Grasping flow in history for conversational machine comprehension. In *International Conference on Learning Representations*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022. Generating information-seeking conversations from unlabeled documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2362–2378, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yongqi Li, Wenjie Li, and Liqiang Nie. 2021. A graph-guided multi-round retrieval method for conversational open-domain question answering. *arXiv preprint arXiv:2104.08443*.

Tingting Liang, Yixuan Jiang, Congying Xia, Ziqiang Zhao, Yuyu Yin, and Philip S Yu. 2022. Multifaceted improvements for conversational open-domain question answering. *arXiv preprint arXiv:2204.00266*.

Kelong Mao, Zhicheng Dou, Hongjin Qian, Fengran Mo, Xiaohua Cheng, and Zhao Cao. 2022. ConvTrans: Transforming web search sessions for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2935–2946, Abu

Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. 2022. CONQRR: Conversational query rewriting for retrieval with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10000–10014, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yi-Ting Yeh and Yun-Nung Chen. 2019. FlowDelta: Modeling flow information gain in reasoning for conversational machine comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 86–90, Hong Kong, China. Association for Computational Linguistics.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 829–838.

A Implementation Details

Generation Text generation with language models often results in degeneration, i.e., repeating the same text sequence. Hence, we heuristically filter out degenerated generations. Initially, we examined the generation quality of LLaMA-7B. However, it showed an increased amount of degeneration and queries of lower quality. We have also tried several open-source instruction-tuned LLMs. To our surprise, these models failed to generate conversational queries given instructions, with or without few-shot examples. Using instruction-tuned LLMs for conversational query generation could be a direction for future exploration. Generations are conducted on 2 NVIDIA V100 GPUs. Generating one conversation takes roughly 10 seconds on a single GPU.

Training Details All retrievers are trained with a batch size of 64 queries. We use in-batch negatives as it is found to be important (Karpukhin et al., 2020). We train all retrievers for 10 epochs with a learning rate of 2e-5. To reduce GPU memory consumption, we use the DPR implementation with gradient cache (Gao et al., 2021), enabling larger batch size. The training process is done on 4 NVIDIA 2080Ti GPUs.

Evaluation Details We evaluate the models on the test sets of the evaluation datasets. There are 20 conversations for evaluation in CAsT-19. Previous work has conducted 5-fold cross-validation to address the lack of training in CAsT-19. However, due to resource constraints, we could not run generations for 5 different sets of examples. Hence, we manually select 5 conversations for building the few-shot examples and use the remaining 15 conversations for evaluation.

We report the most commonly-used evaluation metrics on each dataset: **MRR@5**, **R@5**, and **MAP@10** for OR-QuAC, and **MRR** and **NDCG@3** for CAsT-19.

Speaker Role Identification in Call Centre Dialogues: Leveraging Opening Sentences and Large Language Models

Minh-Quoc Nghiem, Nichola Roberts, Dmitry Sityaev
Connex One Limited
Bauhaus, 27 Quay St, Manchester M3 3GY, United Kingdom
{minh-quoc.nghiem,nichola.roberts,
dmitry.sityaev}@connexone.co.uk

Abstract

This paper addresses the task of speaker role identification in call centre dialogues, focusing on distinguishing between the customer and the agent. We propose a text-based approach that utilises the identification of the agent's opening sentence as a key feature for role classification. The opening sentence is identified using a model trained through active learning. By combining this information with a large language model, we accurately classify the speaker roles. The proposed approach is evaluated on a dataset of call centre dialogues and achieves 93.61% accuracy. This work contributes to the field by providing an effective solution for speaker role identification in call centre settings, with potential applications in interaction analysis and information retrieval.

1 Introduction

Speaker role identification is a fundamental process that involves recognizing different speaker roles in a conversation. Its significance has grown in various settings, such as call centres, where distinguishing between agents and customers in a call transcript is critical. Speaker role identification has numerous uses, including interaction and dialogue analysis, summarisation, and information retrieval (Lavalley et al., 2010; Jahangir et al., 2021). This paper concentrates exclusively on speaker role identification in call centre conversations. Table 1 demonstrates an example of the input and output of this process.

Sample input dialogue:	
Person_01	hello
Person_02	hello good morning is that [NAME]
Person_01	if you hang on a sec while i just get him
Person_02	sorry
Person_01	who's calling
Person_02	it's [NAME] calling you from [ORG] it's for an application
Person_03	oh hi there hi
Person_02	hi [NAME] it's just for an application
Person_03	yes this is [NAME] yes ...

Table 1: An example input of identifying speaker roles. The output should indicate that Person_01 and Person_03 are the Customers, and Person_02 is the Agent

In our application, speaker role identification takes place after speaker diarisation, where speaker turn information has been added to the transcripts. Within a call centre dialogue, two specific roles are present: the customer and the agent. While a typical call involves a single customer and a single agent, it is common for calls to involve more than one customer (as in Table 1), or more than one agent (such as when an agent transfers a customer to another agent). Identifying speaker roles is a challenging task due to various factors, such as transcription errors, interruptions, repetitions, multi-party conversations, and diverse topics.

Numerous studies have been undertaken to address the issue of speaker role identification. These efforts involve utilising text-based features (Barzilay et al., 2000; Liu, 2006; Wang et al., 2011; Sapru and Valente, 2012; Flemotomos et al., 2019), or employing multimodal approaches that integrate both text and audio features (Rouvier et al., 2015; Bellagha and Zrigui, 2020; Guo et al., 2023). In both cases, the goal is to classify each speaker in a conversation into a predefined role category. This classification task is typically accomplished using machine learning algorithms that are trained on labelled datasets, which consist of conversations where each speaker is annotated with their corresponding role category. This paper focuses on the text-based approach and formulate the task as a binary classification problem, with the categories being "customer" and "agent".

In call centre dialogues, distinguishing between

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 388–392
September 11–15, 2023. ©2023 Association for Computational Linguistics

the agent and the customer can be achieved by exploiting the language differences between them. The call centre agent typically starts the conversation by introducing themselves as a representative of their company or organization, which is referred to as the "opening sentence". We propose utilising the identification results of the opening sentence to identify the speaker roles. By combining this information with a large language model, we can accurately classify the speaker roles in call centres.

This paper makes the following two key contributions:

- We propose a model for predicting the opening sentence used by call centre agents, and provide details on how to efficiently construct the training data for this task using active learning.
- We introduce a practical approach for identifying the speaker roles in call centre dialogues by combining the opening sentence identification with a large language model.

The remainder of this paper is organised as follows. Section 2 provides a brief overview of the related work. Section 3 presents details of our methodology. Section 4 describes the experimental results and discussions. Section 5 concludes the paper and points to avenues for future work.

2 Related Work

Text-based speaker role identification often takes the form of a text classification task, aiming to categorise each speaker in a conversation into predefined role categories. Traditionally, text classification has been accomplished using machine learning algorithms trained on labelled datasets. However, with the advent of the Transformer neural network (Vaswani et al., 2017), many studies have adopted pre-trained large language models for text classification (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019). Fine-tuning these pre-trained models still requires a certain amount of labelled data. Active learning provides a means to quickly build labelled data by involving the model in the data labelling process (Settles, 2010).

On the other hand, zero-shot text classification is an approach that requires no labelled data at all (Pourpanah et al., 2022). In this method, a model is trained on a set of existing labelled examples and can subsequently classify new examples from previously unseen classes. This offers the advantage of categorising text into arbitrary categories without the requirement of data preprocessing and training. BART (Lewis et al., 2020), BLOOM (Muennighoff et al., 2022), and FLAN-T5 (Wei et al., 2021) are notable pre-trained large language models available for research purposes, offering the ability to perform zero-shot learning.

3 Method

3.1 Opening Sentence Identification

Our approach for identifying the opening sentence involves using active learning methods to acquire the necessary labelled data and constructing a classifier by fine-tuning a pre-trained large language model with the labelled data.

3.1.1 Data Preparation

An active learning approach was employed to create labelled data for opening sentence identification using a dataset of 437,135 utterances extracted from 67,719 dialogues from different call centre domains. The initial seed set of 100 samples was manually annotated using keyword searches with phrases like "calling from" and "speaking to". The identified key phrases were combined with negative examples to form a seed set. Following that, the seed set was used to train SVM classifiers, utilising two distinct embedding strategies: BERT (Bidirectional Encoder Representations from Transformers) sentence embedding and TF-IDF. This selection was primarily made to facilitate rapid training/retraining of the classifiers during the labelling process.

The classifiers are used to classify each unlabelled sample, and based on the confidence scores, human annotators decide which samples to label using a combination of two sampling strategies: Expected model change and Query-by-Diversity. Given the dataset's substantial class imbalance with only a few positive samples, the focus was on labelling positive samples. This approach aimed to identify the opening sentences that were most likely to have a significant impact on improving the current model. However, Query-by-Diversity sampling (Kee et al., 2018) was also employed to ensure a diverse range of opening sentences was identified. The classifiers underwent retraining either after labelling every 100 samples or when no samples had a score exceeding a threshold (0.7 in our specific case).

3.1.2 Classification

The system employed to identify the opening sentence comprises three key components: an input layer, a BERT model, and a classification layer. In this process, the input layer receives an utterance from the dialogue, and the input representation is generated by incorporating the corresponding token, segment, and position embeddings. The procedure adheres to the recommendations outlined in the work by Devlin et al. (2019). A fully connected neural network, positioned on top of the BERT output, functions as the classification layer to determine whether the utterance is an opening sentence or not. During the training phase, the BERT layer is initialised with pre-trained parameters, and all parameters are then fine-tuned using labelled data from the data preparation step.

3.2 Speaker Role Identification

The FLAN-T5 model is used as the baseline, using a zero-shot prompting approach. Although other models could be utilised, our experiments reveal that the FLAN-T5 yields the most favourable outcomes. The prompt provided to the model is

```
{utterances} (from a speaker)
Based on the utterances above,
{speaker} is
OPTIONS
- an agent from a call centre
- a customer
```

By inputting the utterances of each speaker, the model is able to assign them a role, either "customer" or "agent". This process is repeated for all speakers in the dialogue. Additionally, experiments were conducted using the entire conversation as input (dialogue as context), and results for both approaches are reported (the first 2 rows in Table 3).

To identify the role of speakers in a dialogue, we use a combination of the FLAN-T5 model and the opening sentence identification approach. First, we identify the opening sentences of the dialogue and designate their speakers as "agents". If a speaker does not have an opening sentence, they are labelled as "customers". However, in cases where there are no agents (i.e., no opening sentences detected) or no customers (i.e., all speakers have an opening sentence), we rely on the FLAN-T5 model to assign speaker roles. By combining the strengths of both approaches, we can improve the accuracy and reliability of speaker role identification.

4 Results and Discussion

4.1 Evaluation Dataset

We use the dataset described in 3.1.1 as the evaluation dataset. The conversations are typical of those encountered in call centre scenarios, e.g. buying mobile phones, insurance, foods, etc. A total of 867 opening utterances were labelled as positive examples, indicating they were opening sentences, while 1,982 utterances were labelled as negative examples, representing non-opening sentences, using the active learning approach. A subset of 321 dialogues from seven domains was selected for speaker role identification, which includes speaker diarisation information.

4.2 Opening Sentence Identification

To ensure balanced representation of positive and negative samples, we divided the opening sentence identification data into a train set and a test set, following an 80-20 split while maintaining an equal ratio of positive and negative samples between the two sets. Since the data was generated through active learning, there is a potential bias due to the deliberate selection of samples for labelling. To address this, we generated an additional test set by randomly selecting 100 dialogues and manually assigning labels to them. We presented the results obtained from the SVM classifiers as well as the classification performance using BERT (bert-base-uncased and bert-large-uncased). We trained the BERT classifier model for 3 epochs.

Table 2: The accuracy, precision, recall and F1 scores of different classifiers on opening sentence identification

Method	Acc	Pre	Rec	F1
Test Set				
SVM-TF-IDF	91.78	86.88	80.35	83.48
SVM-BERT	94.92	**92.12**	87.86	89.94
BERT base	**96.41**	92.09	94.22	**93.14**
BERT large	95.37	86.60	**97.11**	91.55
Additional Test Set				
SVM-TF-IDF	99.88	**98.91**	86.67	92.39
SVM-BERT	99.85	91.43	91.43	91.43
BERT base	99.05	94.23	93.33	**93.78**
BERT large	**99.89**	90.27	**97.14**	93.58

The SVM-TF-IDF method achieved an accuracy of 91.78%, highlighting its proficiency in accurately identifying opening sentences. In contrast, the SVM-BERT approach outperformed the SVM-TF-IDF method with an accuracy of 94.92%. This

improvement can be attributed to the utilisation of BERT embeddings, which incorporate the semantic meaning and contextual information of words. However, the SVM-BERT approach only utilises the last layer of BERT for embedding, resulting in slightly lower performance compared to other BERT models.

Among the evaluated methods, the BERT base model achieved the highest accuracy of 96.41%. This demonstrates the effectiveness of leveraging pre-trained language models like BERT for opening sentence identification. Although the BERT large model achieved a slightly lower accuracy of 95.37% compared to BERT base, it excelled in recall with a score of 97.11%. This indicates its strength in correctly identifying positive samples, albeit with a slightly lower precision compared to BERT base. Furthermore, the precision, recall, and F1 scores are notably high, highlighting a well-balanced trade-off in accurately identifying both positive and negative samples. The results obtained from the additional test set further validate this observation.

4.3 Speaker Role Identification

For speaker role identification, a subset of 321 dialogues from seven domains was selected. The evaluation focused on measuring the accuracy of two approaches: FLAN-T5 and the combined use of opening sentence identification and FLAN-T5. Two FLAN-T5 models were employed in the evaluation: FLAN-T5-Large and FLAN-T5-XL. The results obtained from these evaluations are presented in Table 3.

Table 3: Accuracy of different approaches on speaker role identification

Method	Acc
FLAN-T5-Large dialogue as context	70.17
FLAN-T5-Large utterances	81.25
FLAN-T5-XL utterances	86.36
Using Opening Sentence	89.49
Opening Sentence + FLAN-T5-XL	**93.61**

FLAN-T5-Large (770M parameters), when considering the whole dialogue as context, achieved an accuracy of 70.17%. However, when using utterances which belong to a specific speaker as context, FLAN-T5-Large demonstrated improved performance with an accuracy of 81.25%. This approach outperformed the dialogue-level context approach, highlighting the benefits of considering individual utterances. The FLAN-T5-XL variant (3B parameters) achieved an accuracy of 86.36%, surpassing the previous approaches. This improvement can be attributed to its larger model configuration, which enhances its ability to capture complex patterns and representations.

The utilisation of the opening sentence identification approach resulted in an accuracy of 89.49%. This method leverages labelled data, providing an advantage over FLAN-T5, which is a zero-shot approach. Combining the opening sentence identification with FLAN-T5-XL yielded the highest accuracy of 93.61%. This combination proves to be the most effective for accurate identification.

Classification errors were further analysed, and the following primary causes were identified: (1) Inaccurate speaker identifiers in the input data, particularly due to speech diarisation errors. (2) Complex contextual scenarios that pose challenges even for human understanding. (3) Instances where agents engage in conversations with each other, making it difficult to distinguish their roles. (4) Situations involving business numbers being contacted, which often share the same opening sentence pattern and are prone to misidentification as agents.

5 Conclusion

This paper proposes a text-based approach for speaker role identification in call centre dialogues. By combining the identification of the agent's opening sentence with a large language model, our approach achieves high accuracy in classifying speaker roles. This has practical implications for call centre applications, enabling improved customer-agent interaction analysis and call pattern analysis.

The use of active learning allows for efficient construction of the training dataset for opening sentence identification. Integrating this information into the classification process significantly improves the accuracy of speaker role identification.

Future work can explore enhancements to the system, such as incorporating additional contextual features and exploring multimodal approaches. Evaluating the approach on larger and more diverse datasets would also provide a better understanding of its generalisability.

References

Regina Barzilay, Michael Collins, Julia Hirschberg, and Steve Whittaker. 2000. The rules behind roles: Identifying speaker role in radio broadcasts. In *AAAI/IAAI*, pages 679–684.

Mohamed Lazhar Bellagha and Mounir Zrigui. 2020. Speaker naming in tv programs based on speaker role recognition. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE.

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

Nikolaos Flemotomos, Panayiotis Georgiou, and Shrikanth Narayanan. 2019. Linguistically aided speaker diarization using speaker role information. *arXiv preprint arXiv:1911.07994*.

Dongyue Guo, Jianwei Zhang, Bo Yang, and Yi Lin. 2023. A comparative study of speaker role identification in air traffic communication using deep learning approaches. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–17.

Rashid Jahangir, Ying Wah Teh, Henry Friday Nweke, Ghulam Mujtaba, Mohammed Ali Al-Garadi, and Ihsan Ali. 2021. Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. *Expert Systems with Applications*, 171:114591.

Seho Kee, Enrique Del Castillo, and George Runger. 2018. Query-by-committee improvement with diversity and density in batch active learning. *Information Sciences*, 454:401–418.

Rémi Lavalley, Chloé Clavel, Patrice Bellot, and Marc El-Beze. 2010. Combining text categorization and dialog modeling for speaker role identification on call center conversations. In *Eleventh Annual Conference of the International Speech Communication Association*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yang Liu. 2006. Initial study on automatic identification of speaker role in broadcast news speech. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 81–84.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. 2022. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*.

Michael Rouvier, Sebastien Delecraz, Benoit Favre, Meriem Bendris, and Frederic Bechet. 2015. Multimodal embedding fusion for robust speaker role recognition in video broadcast. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 383–389. IEEE.

Ashtosh Sapru and Fabio Valente. 2012. Automatic speaker role labeling in ami meetings: recognition of formal and social roles. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5057–5060. IEEE.

Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison*, 52.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wen Wang, Sibel Yaman, Kristin Precoda, and Colleen Richey. 2011. Automatic identification of speaker role and agreement/disagreement in broadcast conversation. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5556–5559. IEEE.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Synthesising Personality with Neural Speech Synthesis

Shilin Gao *
The University of Edinburgh
Cambridge University Press & Assessment
shilin.gao@cambridge.org

Matthew P. Aylett
CereProc Ltd.
Heriot-Watt University
matthewaylett@gmail.com

David A. Braude †
CereProc Ltd.
Sanas.ai
david@sanas.ai

Catherine Lai
The University of Edinburgh
c.lai@ed.ac.uk

Abstract

Matching the personality of conversational agents to the personality of the user can significantly improve the user experience, with many successful examples in text-based chatbots. It is also important for a voice-based system to be able to alter the personality of the speech as perceived by the users. In this pilot study, fifteen voices were rated using Big Five personality traits. Five content-neutral sentences were chosen for the listening tests. The audio data, together with two rated traits (Extroversion and Agreeableness), were used to train a neural speech synthesiser based on one male and one female voices. The effect of altering the personality trait features was evaluated by a second listening test. Both perceived extroversion and agreeableness in the synthetic voices were affected significantly. The controllable range was limited due to a lack of variance in the source audio data. The perceived personality traits correlated with each other and with the naturalness of the speech.

1 Introduction

The law of attraction in human-robot interaction means users prefer social robots with similar personality traits to themselves (Park et al., 2012). Previous work has shown that it is possible to design a text-based chatbot with a pre-defined personality (Ahmad et al., 2020; Ruane et al., 2021), and matching the personality of the agent to the personality of the user can significantly improve the user experience (Smestad and Volden, 2019; Fernau et al., 2022). Personality in voice-based conversational agent is much less investigated, but the effect is no less significant. People attribute traits to others in less than a second after hearing them in video and/or audio recordings (Reeves and Nass, 1996; Uleman et al., 2008). The same effect extends to machines that display human-like features including embodied conversational agents (Nass and Brave, 2005). The perceived personality from speech is consistent across listeners (McAleer et al., 2014). This opens the possibility of generating synthetic voices that encourage users to attribute pre-defined traits to the artificial intelligence conversational agents they interact with.

Previous work (Aylett et al., 2017) has shown that personality can be manipulated with a speech synthesis system. The effect is restrained by the system used: unit selection is heavily constrained by the corpus recorded (though there have been advances in addressing this (Buchanan et al., 2018)), whilst HMM-based Speech Synthesis (HTS) is constrained by perceived naturalness. Neural speech synthesis systems such as Wavenet (Oord et al., 2016) and Tacotron (Wang et al., 2017a) has shown an improved ability to generate natural sounding output. This has led to advancement in expressive speech synthesis (Wang et al., 2017b, 2018; Zhang et al., 2019). However the focus is on manipulating the style of single utterances and is different from synthesising a voice with a consistent personality. Recent work (Shiramizu et al., 2022) achieved altering the social perception of synthetic speech by controlling single speech-based features such as pitch. It is interesting to see the effect of using neural speech synthesis system to manipulate the perceived personality traits of the output voice.

In this work the use of Big Five scores is ex-

*This author is currently affiliated with Cambridge University Press & Assessment. Research was conducted while studying at The University of Edinburgh.

†This author is currently affiliated with Sanas.ai. Research was conducted while working at CereProc Ltd.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 393–399
September 11–15, 2023. ©2023 Association for Computational Linguistics

plored for directly controlling the perceived personality of the synthetic speech. Big Five, or OCEAN model (John et al., 1999), is widely used the domain of human-computer interaction (Vinciarelli and Mohammadi, 2014). A condensed version (Rammstedt and John, 2007) that reduces the original 44 statements to ten while preserving a high level of accuracy was used.

2 Experiments

2.1 Big Five Rating of Source Voices

Our dataset comprised of 15 English native speaker voices taken from CereProc's voice bank. The voices varied by accent and gender, see Table 1.

Gender	Received Pronunciation	Scottish	Irish English	Total
Male	5	2	0	7
Female	5	2	1	8
Total	10	4	1	15

Table 1: Accent and gender distribution

For the listening tests, five news sentences were chosen for their content being emotionally neutral but can be read with different personalities (see Appendix A Table 2). 28 English native listeners were recruited from Amazon Mechanical Turk (AMT) to rate the Big Five personality traits of each source voice. A web-based listening test was used to measure Big Five based on ten personality questions (Rammstedt and John, 2007) with an additional naturalness question using a 5-point Likert scale. Two slide bars were used to measure perceived age (10-70), and perceived gender (0-1, from woman to man). The system displayed the audio transcript and allowed participants to play the audio stimuli repeated times. A screen shot of the listening test page is in shown in Figure 2. Each participant listened to a subset of 5 speakers and for each of those speakers they listened to 5 audio examples. The audio order was randomised for each listener and each audio example was rated by nine or ten listeners.

Results were averaged by voice to give an overall personality score for that voice and are shown in Figure 1. Extroversion and agreeableness were chosen as the two personality traits to control as they showed the most variation.

Figure 3 shows the spread of the voices in the 1-5 Likert scale across both traits. The variation

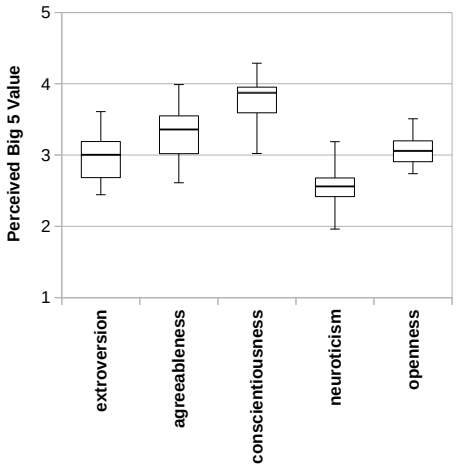


Figure 1: Box Plot of Big Five personalities averaged by voice.

across the two traits in the database is between 2 and 4. This is expected as voice talents are often chosen on similar criteria, and the recording process for speech synthesis tends to avoid high energy emotional content which puts an artificial limit on the possible perceived personality variation within the voice. There is a positive correlation between the two traits (Pearson $r = 0.664$, $df = 13$, $p < 0.05$). The *r-squared* value is relatively low (0.441), meaning that although there is a significant positive correlation, it might not be linear or the data might not be enough to make an accurate prediction. Theoretically, the Big Five model is based on factor analysis which aims at producing independent dimensions (John et al., 1999), however, this is for *actual* personality and may not translate to independence in *perceived* personality.

2.2 Building the Multiple Speaker Synthesis Voice

We used CereProc's Deep Neural Network (DNN) speech synthesis system CereWave to build a multi-speaker voice. CereWave uses a recurrent neural network architecture to firstly produce prosody targets, and then produce an intermediate acoustic feature set. After predicting the acoustic features, it uses a custom neural vocoder to produce the final output waveforms. Its inputs include phonetic, linguistic, language, accent and speaker features, in which speaker features include age and gender. For this experiment, the personality dimensions chosen at the first stage (extroversion and agreeableness) are appended to the above features in the format of an average voice score on a 5-point Likert scale. Due to the time constraints of this research and

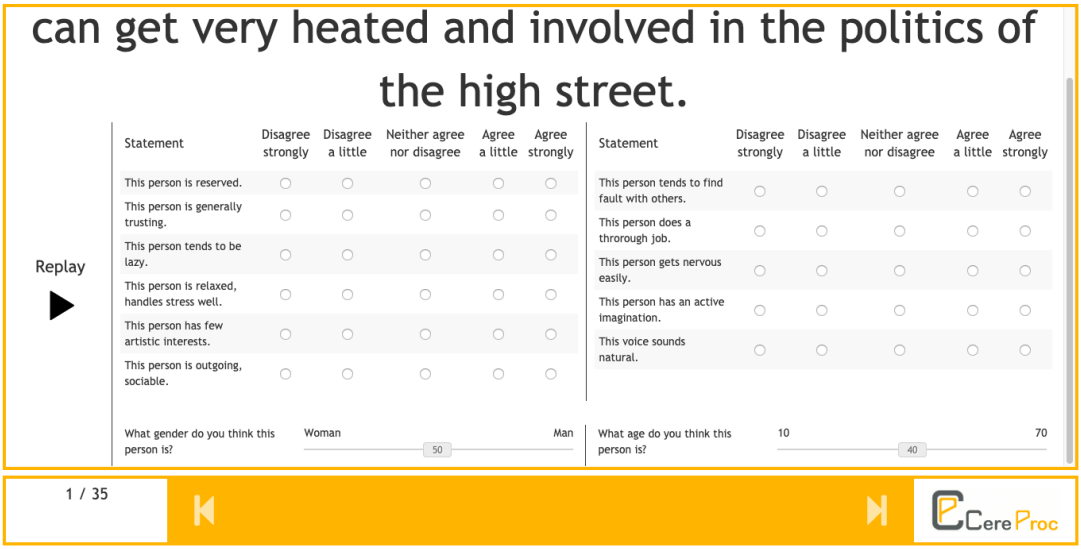


Figure 2: Screen shot of web based listening test used to evaluate Big Five.

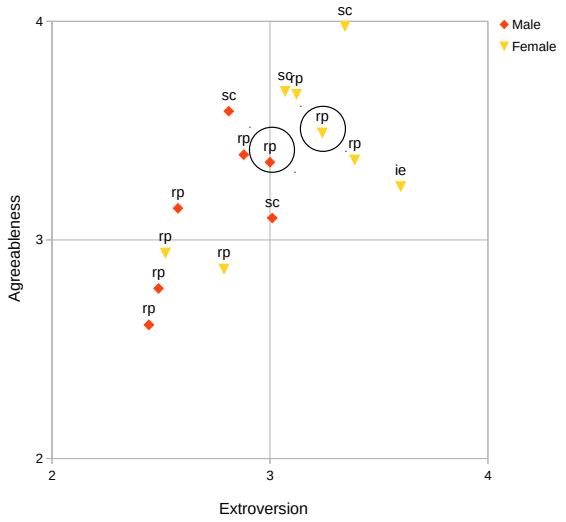


Figure 3: Distribution of average perceived extroversion and agreeableness by voice, target voices circled.

its nature of pilot study, we limited the data to a total of 1,000 utterances from the neutral speaking style data (totalling approximately 2 hours of data), which puts an limitation on the naturalness.

When synthesising from an average voice, an original speaker specification can be used to generate synthesis sounding like that speaker. Two voices, one male and one female, close to the global mean for all voices in terms of extroversion and agreeableness, were chosen to synthesise stimuli (Male voice: mean extroversion 3.0, mean agreeableness 3.4; Female voice: mean extroversion 3.2, mean agreeableness 3.5). In addition, natural recordings for each of these speakers were used as a high naturalness anchor, and synthesis using a previous generation DNN system were used as a low naturalness anchor. Five utterances were synthesised for all synthesis conditions.

2.3 Evaluating the Synthesis of Agreeableness and Extroversion

A second AMT listening test was carried out using the same interface and methodology described in section 2.1 with 18 participants. It is expected that synthesised voices' personality would not match the reference speakers exactly but should be similar. This was the case for the male voice but the synthesis process reduced both the perceived extroversion and agreeableness of the female voice (Male voice: mean extroversion 3.0, mean agreeableness 3.4; Female voice: mean extroversion 2.9, mean agreeableness 3.1).

Results were averaged over the 10 utterances (5 spoken by two voices) and a by-materials repeated measures MANOVA was carried out with perceived extroversion and agreeableness as the dependent variable. Target extroversion (tgt-e: low/high) and nested target agreeableness (tgt-a: low/high) were within-materials factors, with base synthesis voice (gender: male/female) as a between-materials factor. Both target factors were significant in a multivariate test (Wilks Lambda: tgt-a ($F(2, 7)=21.258$, $p=0.001$), tgt-e ($F(2, 7)=11.422$, $p<0.01$)), gender did not have a significant effect. Univariate tests with a Greenhouse-Geisser correction (sphericity not assumed) showed that target extroversion significantly affected perceived extroversion (tgt-e $F(1, 8)=24.981$, $p=0.001$) but not per-

ceived agreeableness, whereas target agreeableness significantly affected both perceived agreeability (tgt-e F(1, 8)=47.399, p<0.001) and extroversion (tgt-e (F(1, 8)=34.561, p<0.001)).

In terms of the adjusted means by target groups, agreeableness has the desired effect on perceived agreeableness (tgt-a low: mean 2.922, Standard Error (SE) 0.048; high: mean 3.206, SE 0.039), but also significantly affected perceived extroversion (tgt-a low: mean 2.639, SE 0.055; high: 3.156, SE 0.06). Extroversion had the opposite affect on perceived extroversion as the higher target actually reduced perceived extroversion (tgt-e low: mean 3.019, SE 0.053; high: 2.775, SE 0.034).

The effect of trait targeting on speech rate, pitch and amplitude is also evaluated using Pearson's correlation analysis. Only speech rate had a significant effect (extroversion/words-per-second: $r(40)=0.29$, $p<0.05$), agreeableness/words-per-second: $r(40)=0.23$, $p<0.005$).

Figure 4 shows the average extroversion/ agreeableness by synthesis type. The manipulation targets are: '+e+a' to be positioned at 4,4; '+e-a' at 4,2; '-e+a' at 2,4; and '-e-a' at 2,2. It is shown that the perceived variation is much lower than this (between 2.5 and 3.5), and the spread does not form the pattern expected above.

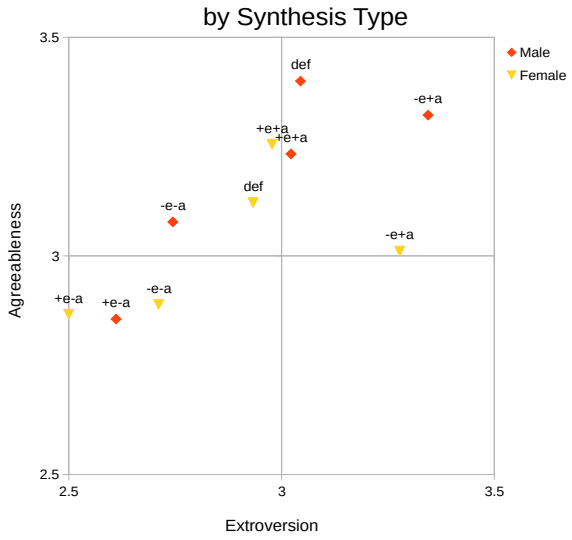


Figure 4: Distribution of average perceived extroversion and agreeableness for different synthesis types. 'def': synthesis with no personality modelling. '+': high(4), '-': low(2). 'e': extroversion, 'a': agreeableness.

2.3.1 Effect on naturalness

A univariate repeated measures ANOVA with a Greenhouse-Geisser correction (sphericity could not be assumed) was carried out to explore the effect of trait targeting on perceived naturalness. Naturalness, initially recorded on a 1-5 Likert scale, was averaged by utterance for each synthesis type and used as the dependent variable. The model matched that used in the previous MANOVA. Target extroversion (tgt-e: low/high) and nested target agreeableness (tgt-a: low/high) were within-materials factors, with base synthesis voice (gender: male/female) as a between-materials factor. Target agreeableness was significant ($F(1,8)=39.784$, $p<0.001$) where a high target increased perceived naturalness (tgt-a low: mean 2.8, SE 0.073; high: mean 3.339, SE 0.052). There was also a significant effect for an interaction between voice and target extroversion ($F(1,8)=5.967$, $p<0.05$). This effect was caused by high target extroversion increasing perceived naturalness for the female voice (tgt-e*gender low: mean 2.289, SE 0.091; high: mean 3.178, SE 0.078) and reducing naturalness for the male voice (tgt-e*gender low: mean 3.144, SE 0.091; high: mean 2.967, SE 0.078).

Values for perceived extroversion, agreeableness and naturalness were averaged across subjects for each of the utterances in all four conditions (tgt-e: low/high, tgt-a: low/high) and for both male and female voices (40 data points in total). A Pearson correlation showed a significant positive correlation between perceived extroversion, perceived agreeableness and perceived naturalness. (extroversion/agreeableness: $r(40)=0.507$, $p=0.001$, extroversion/naturalness: $r(40)=0.641$, $p<0.001$, agreeableness/naturalness $r(40)=0.512$, $p=0.001$).

3 Discussion

This pilot study shows that using the personality traits to control the perceived personality of a synthetic voice is feasible with a modern DNN / neural vocoder system. Readers are invited to listen to sample natural and synthetic speech from https://cereproc.s3-eu-west-1.amazonaws.com/samples/shilin2019/index.html. Changing input features and manipulating the target for agreeableness both alter the perceived personalities in the expected direction. However, the range in agreeableness that can be controlled, as well as the lack of a similar result for extroversion, show that controlling perceived personality is a far from simple process.

Two limitations have compromised the results of the study: 1) The corpus used as a basis for

this experiment was comprised of voices originally selected for being extrovert and agreeable, which can be seen from Figure 1 and Figure 3. With a machine learning approach this means when targets are set within outlying regions the system has to extrapolate the results which leads to unnatural results as they are not based on actual observations. This is shown for agreeableness where lower target scores (unseen in the data) generate stimuli rated lower for naturalness. In future work it will be important to source a corpus with a much wider variation in perceived Big Five personality traits. 2) The interaction between traits and naturalness appear to complicate perceived trait scores. In previous work, using actual vocal change in the data, or changing synthesis style, appeared to change Big Five without correlating with naturalness variation (Aylett et al., 2017). This work, however, shows a strong correlation between perceived agreeableness and perceived extroversion and naturalness. Such collinearity means it is difficult to produce stable results. The confounding effect is possibly intensified by using an average voice built with a limited amount of source data.

4 Conclusion and future work

To summarise our findings: 1) The prototype system showed a Big Five trait could be learned and controlled, though control may be limited in the controllable range. 2) Naturalness can interact with personality traits and ensuring the underlying average voice is as natural as possible is an important consideration. 3) Correlations across traits may interfere with final results.

The next steps would be to repeat the annotation and training with a dataset that contains a wide variety of speakers such as VCTK (Yamagishi et al., 2019), and apply the synthetic voice in a multi-turn voice-based conversational agent set-up. Spontaneous speech corpus rather than fluence read speech corpus can also be used to build synthetic voice with distinctive perceived personality (Gustafson et al., 2021). Methods of including personality features that are more sophisticated than concatenation on the input features can be explored, both in terms of architecture and training approaches (Gibiansky et al., 2017).

Further experiments can be using personality synthesis in speech together with text-based personality generation. This work suggests the possibility of making a chatbot speak in a voice with 1) pre-defined personality based on the generated text, which can be matching or mismatching, and 2) adaptive personality based on the personality of the user, as such adaptation is shown possible in text-based chatbots (Fernau et al., 2022). A multi-turn conversational set-up can be used to experiment the consistency of synthetised personality. The perception and impact of synthesised personality in different cultural context can also be explored in various user studies.

5 Acknowledgements

This work was supported by the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No 780890 (Grassroot Wavelengths). Research was conducted while Shilin Gao was studying at The University of Edinburgh. This author is currently working at Cambridge University Press & Assessment.

References

Rangina Ahmad, Dominik Siemon, and Susanne Robra-Bissantz. 2020. Extrabot vs introbot: The influence of linguistic cues on communication satisfaction. In *AMCIS*.

Matthew P Aylett, Alessandro Vinciarelli, and Mirjam Wester. 2017. Speech synthesis for the generation of artificial personality. *IEEE Transactions on Affective Computing*.

Christopher G. Buchanan, Matthew P. Aylett, and David A. Braude. 2018. Adding personality to neutral speech synthesis voices. In *SPECOM*.

Daniel Fernau, Stefan Hillmann, Nils Feldhus, Tim Polzehl, and Sebastian Möller. 2022. Towards personality-aware chatbots. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 135–145.

Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep voice 2: Multi-speaker neural text-to-speech. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2962–2970. Curran Associates, Inc.

Joakim Gustafson, Jonas Beskow, and Éva Székely. 2021. Personality in the mix-investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis. *Proc. 11th ISCA SSW*, pages 48–53.

Oliver P John, Sanjay Srivastava, et al. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138.

Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How do you say 'hello'? personality impressions from brief novel voices. *PloS one*, 9(3).

C. Nass and S. Brave. 2005. *Wired for speech: How voice activates and advances the Human-Computer relationship*. The MIT Press.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Eunil Park, Dallae Jin, and Angel P del Pobil. 2012. The law of attraction in human-robot interaction. *International Journal of Advanced Robotic Systems*, 9(2):35.

Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212.

B. Reeves and C. Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Elayne Ruane, Sinead Farrell, and Anthony Ventresque. 2021. User perception of text-based chatbot personality. In *Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event, November 23–24, 2020, Revised Selected Papers 4*, pages 32–47. Springer.

Victor Kenji M Shiramizu, Anthony J Lee, Daria Altenburg, David R Feinberg, and Benedict C Jones. 2022. The role of valence, dominance, and pitch in perceptions of artificial intelligence (ai) conversational agents' voices. *Scientific Reports*, 12(1):22479.

Tuva Lunde Smestad and Frode Volden. 2019. Chatbot personalities matters: improving the user experience of chatbot interfaces. In *Internet Science: INSCI 2018 International Workshops, St. Petersburg, Russia, October 24–26, 2018, Revised Selected Papers 5*, pages 170–181. Springer.

James S Uleman, S Adil Saribay, and Celia M Gonzalez. 2008. Spontaneous inferences, implicit impressions, and implicit theories. *Annu. Rev. Psychol.*, 59:329–360.

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017a. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Yuxuan Wang, RJ Skerry-Ryan, Ying Xiao, Daisy Stanton, Joel Shor, Eric Battenberg, Rob Clark, and Rif A Saurous. 2017b. Uncovering latent style factors for expressive speech synthesis. *arXiv preprint arXiv:1711.00520*.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*.

Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE.

A Appendix A: Sentences used in the listening tests

Sentence ID	Sentence
180	He also defended the company's policy of releasing new services and tools to users before they were finished products.
189	No charges were made, but two men have been thrown off the programme.
205	After a gruelling ten minute phone interview the reporter had a new job.
216	There is controversy around these findings: some people have tried to replicate them, although not using exactly the same methods, and got different results.
259	Even as voters drift away from party politics, they can get very heated and involved in the politics of the high street.

Table 2: Selected sentences for listening tests

Prompting, Retrieval, Training: An exploration of different approaches for task-oriented dialogue generation

Gonçalo Raposo **Luísa Coheur** **Bruno Martins**
INESC-ID, Instituto Superior Técnico, Universidade de Lisboa
{goncalo.cascalho.raposo, luisa.coheur, bruno.g.martins}@tecnico.ulisboa.pt

Abstract

Task-oriented dialogue systems need to generate appropriate responses to help fulfill users' requests. This paper explores different strategies, namely prompting, retrieval, and fine-tuning, for task-oriented dialogue generation. Through a systematic evaluation, we aim to provide valuable insights and guidelines for researchers and practitioners working on developing efficient and effective dialogue systems for real-world applications. Evaluation is performed on the MultiWOZ and Taskmaster-2 datasets, and we test various versions of FLAN-T5, GPT-3.5, and GPT-4 models. Costs associated with running these models are analyzed, and dialogue evaluation is briefly discussed. Our findings suggest that when testing data differs from the training data, fine-tuning may decrease performance, favoring a combination of a more general language model and a prompting mechanism based on retrieved examples.

1 Introduction

Task-oriented dialogue systems need to generate appropriate responses to help fulfill users' requests. Recent advancements in Natural Language Processing (NLP) have produced a shift towards leveraging large pretrained language models to tackle the generation challenge (Zhang et al., 2020). By prompting these models with a few examples, their performance has been shown to surpass traditional approaches, eliminating the need for extensive model training (Brown et al., 2020; Zhang et al., 2022).

In this paper, we explore different approaches for task-oriented dialogue generation, namely through the use of prompting, retrieval mechanisms, and fine-tuning. We investigate the best strategies to leverage these approaches, considering the integration of past conversation information, the selection of appropriate retrieval methods, and the assessment of the benefits of fine-tuning (Roller et al., 2021; Izacard et al., 2022; Peng et al., 2022).

During our investigation, we assessed various state-of-the-art instruction-based models, including different size versions of FLAN-T5 (Chung et al., 2022), GPT-3.5, and GPT-4, provided by OpenAI (Ouyang et al., 2022; OpenAI, 2023). These models, known for their impressive language generation capabilities, serve as the foundation for our experiments, through which we tested different strategies. We evaluate the performance of these models on widely used benchmark datasets, namely MultiWOZ and Taskmaster-2, which offer diverse and challenging dialogue scenarios (Zang et al., 2020; Byrne et al., 2019). Additionally, we analyze the computational costs associated with running the models, considering the trade-off between performance and resource requirements. Moreover, we discuss dialogue system evaluation, addressing the metrics and criteria that best capture the quality and effectiveness of task-oriented dialogue generation (Sellam et al., 2020; Nekvinda and Dušek, 2021).

The main contributions of this paper are[1]:

- Investigate different approaches for task-oriented dialogue generation, including prompting, use of retrieval mechanisms, and fine-tuning.
- Advocate for the combination of a large pretrained language model with the proposed retrieval mechanism when the testing data significantly deviates from the training data, showcasing its effectiveness and cost-efficiency.
- Examine the positioning of GPT-3.5 and GPT-4 models, comparing them with both pretrained and fine-tuned models, to understand their performance characteristics, advantages, and costs.

2 Related work

Task-oriented dialogue generation has garnered significant attention, leading to a wide range of research efforts. Recent studies have focused on the

[1]We make all of our code available online at https://github.com/gonced8/dialogue-retrieval

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 400–412
September 11–15, 2023. ©2023 Association for Computational Linguistics

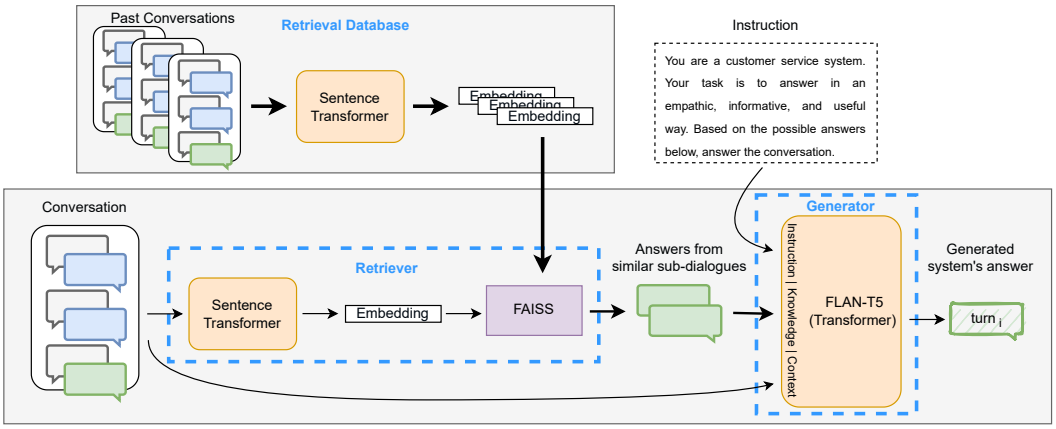


Figure 1: Our main approach for answer generation enhanced with possible answers retrieved from past conversations. During inference, our system starts by retrieving the possible answers and then includes them in the prompt given to the FLAN-T5 model, used to generate the system's response in the context of a dialogue.

use of large pretrained language models for dialogue systems. Radford et al. (2018) introduced GPT, a Transformer (Vaswani et al., 2017) using generative pretraining, which demonstrated impressive performance in various NLP tasks. Subsequent research explored the benefits of fine-tuning pretrained models specifically for dialogue generation tasks. For instance, Lin et al. (2020) proposed MinTL, a system that fine-tuned a pretrained model on task-oriented data and established new state-of-the-art results. Similarly, Thoppilan et al. (2022) employed fine-tuning on a larger pretrained model of approximately 37 B parameters and used around 1.56 T words of public dialogue data and web text, improving in all metrics.

Prompting has emerged as a valuable technique for improving the performance of pretrained language models. It involves providing specific examples or instructions as input to guide the generation process. Brown et al. (2020) demonstrated the effectiveness of prompts when using language models to generate coherent and contextually appropriate responses. A recent work by Gupta et al. (2022) addresses prompting in the context of dialogue systems, showing how instruction tuning may benefit certain test tasks.

Retrieval-enhanced methods have also been extensively explored in dialogue systems. Yang et al. (2019) integrated text retrieval and text generation models to build a hybrid conversational system that outperformed retrieval-based and generation-based approaches. In addition, several studies have also incorporated retrieval mechanisms in combination with generative models to enhance dialogue system performance (Roller et al., 2021; Shuster et al., 2022; Thoppilan et al., 2022).

While the aforementioned studies have made substantial contributions to the field, this paper aims to expand upon the existing literature by thoroughly investigating the integration of prompting, retrieval mechanisms, and fine-tuning in task-oriented dialogue generation. Specifically, we explore the efficacy of these approaches and analyze their impact on system performance, considering both the quality of generated responses and the computational costs associated with running the models. Furthermore, as far as we know, we are the first work employing the GPT-3.5 and GPT-4 models for the MultiWOZ and Taskmaster-2 datasets, establishing baselines for each.

3 Method

In our main approach, we propose to use a dense retrieval model that, given a dialogue, will retrieve other similar dialogues. We then use their answers to generate a new answer using a Transformer. Figure 1 illustrates how our system can be used for inference, depicting its components.

3.1 Dense retrieval of dialogue answers

We use dense retrieval (Karpukhin et al., 2020; Gao et al., 2023) to obtain relevant responses given a conversation context. Since the task of retrieving responses for dialogues is not necessarily equivalent to document or passage retrieval (Penha and Hauff, 2023), we considered two possible approaches: (1) Encode the current conversation context and compare it to a database of encoded past contexts. The returned relevant responses will correspond to the turns immediately after each of the indexed contexts; (2) Encode the current conversation context and query a database of encoded past

responses. The returned relevant responses will be those whose embeddings are the most similar to the query/context embedding.

The library Sentence-Transformers (Reimers and Gurevych, 2019) provides models already pretrained for tasks like text clustering or semantic search, that can be used to perform the described response retrieval. In particular, it provides Transformer-based encoders that can be used to compute text embeddings, and then compare the embeddings with a similarity function (e.g., cosine-similarity or dot product).

To implement the two approaches described, we considered two of the top pretrained models provided by Sentence-Transformers: `all-mpnet-base-v2` and `multi-qa-mpnet-base-dot-v1`. These models are both fine-tuned versions of the pretrained MPNet model (Song et al., 2020) using a contrastive loss. In particular, `all-mpnet-base-v2` was fine-tuned to be used for information retrieval, clustering, or sentence similarity tasks, making it more appropriate for our first approach. On the other hand, `multi-qa-mpnet-base-dot-v1` was fine-tuned for semantic search and it is intended to be used to pair queries/questions with relevant text paragraphs. Thus, we used it for our second approach, given the size difference between contexts and responses. For both models, we use dot product as our similarity function.

Using a conversation context as a query differs significantly from relevant passage retrieval. Some studies perform question rewriting to circumvent this issue and use a rewritten context-independent version of the last turn as the query (Raposo et al., 2022). Since question rewriting may also require additional training, we simply fine-tuned the retrieval encoder for conversational text.

We specifically used weakly supervised learning to train our encoder. Starting from an unlabeled dataset of conversations, we made sets of queries (conversation contexts) and documents (either 1. conversation contexts or 2. conversation responses). Then, given a random batch of query embeddings, we compute the similarity with the document embeddings. With the option 2., it makes sense to match the context to the corresponding response. However, with option 1., we match the context with a different context from that batch based on the similarity between responses (measured using ROUGE). We train the encoder using a cross-entropy loss (Wang et al., 2020):

$$\mathrm{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(Q_{N\times d}\cdot D_{N\times d}^{T}\right)_{i,k}}{\sum_{j=1}^{N}\exp\left(Q_{N\times d}\cdot D_{N\times d}^{T}\right)_{i,j}} \quad (1)$$

where $Q_{N\times d}$ is a matrix composed by N queries embeddings of size d, $D_{N\times d}$ a matrix composed by the corresponding N document embeddings of size d. The index k will correspond to the target document. The similarity computation uses the dot product, and no temperature parameter is applied.

3.2 Answer generation

We use the pretrained Transformer named FLAN-T5 (Chung et al., 2022), which is an enhanced version of the T5 model (Raffel et al., 2020) that was fine-tuned using instructions and is reported to achieve strong few-shot performance.

3.2.1 Generation-only approach

We start by evaluating FLAN-T5 in a zero-shot setting, using no examples of possible answers. In practice, our approach consisted in giving the model the following prompt:

```
You are a customer service system. Your
task is to answer in an empathic,
informative, and useful way. Answer the
conversation.
Conversation:
{conversation context}
```

This prompt is followed by the conversation context and the model generates the response.

3.2.2 Generating based on past answers

To incorporate the information from the retrieved past answers, we simply concatenate them in the input that is given to the generation model. This approach is similar to the work by Ram et al. (2023) and its main benefits are its simplicity and versatility, which allow it to be implemented with any generative model. Thus, FLAN-T5 is used in a few-shot setting with the following prompt:

```
You are a customer service system. Your
task is to answer in an empathic,
informative, and useful way. Based on
the possible answers below, answer the
conversation.
Possible answers:
{possible answers}
Conversation:
{conversation context}
```

3.2.3 Fine-tuning for answer generation

We described how we used retrieved past answers as examples for our generation model, which related works have shown to improve performance. In addition, we also study how fine-tuning the same model affects the achieved performance. Using the same prompts mentioned above, we train our models in both scenarios: with and without retrieval. During training and evaluation, we are careful to avoid data leakage in the retrieved answers (e.g., we index the training dataset, and we do not retrieve responses from the same conversation).

3.2.4 Open-AI models

Given the recent popularity and impressive performance of OpenAI's large language models – ChatGPT and GPT-4 – we also performed some experiments using their API. Similarly to the FLAN-T5 model, these chat-based models were also fine-tuned in an instruction-following setting but using Reinforcement Learning with Human Feedback (RLHF) for optimization (Ouyang et al., 2022; OpenAI, 2023). For reproducibility, we reused the same prompts from FLAN-T5 and evaluated both the zero-shot and few-shot settings.

4 Experimental setup

Broadly, our experiments consisted in testing different generation models on the task of answer generation in task-oriented dialogues. In some cases, this also involved the use of information retrieval mechanisms or fine-tuning models.

4.1 Implementation details

Regarding dense retrieval, we use the models from Sentence Transformers to compute the embeddings, together with FAISS (Johnson et al., 2019) to index and search them. When training the retrieval modules, we used a batch size of 64 samples and the AdamW optimizer (Loshchilov and Hutter, 2019).

For FLAN-T5, we use the checkpoints available on Hugging Face (Wolf et al., 2020). In particular, we use the small, large, and XL versions. To test and train the models, we use the Transformers library from Hugging Face along with the PyTorch framework (Paszke et al., 2019). We use the AdamW optimizer and train our models for a maximum of 20 epochs with patience of 5 steps. The batch size varied for each model due to limitations on GPU memory, but the effective batch size was kept at 64 samples. All our local models were trained and tested using a `NVIDIA Quadro RTX 6000` GPU with 24 GB of memory. As for the OpenAI models, we use their API through the provided Python package, keeping the default settings.

4.2 Task-oriented datasets

Starting from a task-oriented dataset, we extract a dataset consisting of sub-dialogues. Based on Nekvinda and Dušek (2022), we chose to use a maximum of 6 turns for each sub-dialogue, which seemed like a good compromise between providing enough context but not too long. The extracted sub-dialogues can be obtained by sliding a window of size 6 turns over the original dialogue, with a stride of 2 turns to always end in a system's turn. Depending on the speaker, we prepend each turn with "User: " or "System: ".

We apply this technique to the MultiWOZ 2.2 and Taskmaster-2 task-oriented datasets. As Taskmaster-2 has not already predefined dataset splits, we randomly select 1k dialogues for both validation and testing, ensuring a balanced distribution across domains. Table 1 shows a summary of the sizes of the obtained datasets of sub-dialogues.

Table 1: Number of samples for each dataset split after applying the preprocessing that consists of splitting each dialogue into multiple sub-dialogues.

Dataset	Train	Validation	Test
MultiWOZ 2.2	56776	7374	7372
Taskmaster-2	120892	7997	8038

4.3 Automatic evaluation metrics

To measure the performance of our models, we compare the returned answers to the ground truth answers. In particular, we use automatic metrics based on lexical similarity (i.e., BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004)) and on semantics similarity (i.e., BERTScore (Zhang* et al., 2020) and BLEURT (Sellam et al., 2020)). Additionally, we score the quality of the generated answers using QualityAdapt, a reference-free metric that achieves state-of-the-art performance on overall dialogue quality estimation through adapter fusion (Mendonca et al., 2022).

5 Results

5.1 Retrieval-only responses

The first approach to obtain the dialogue response that we evaluated consists of using a retrieval-only model. Given a conversation context as a query, its

Table 2: Performance of using only a retrieval model to return the response. Two pretrained models are compared to their fine-tuned versions on MultiWOZ. The models differ in how they perform retrieval: indexing the contexts and returning the next response, against indexing the responses.

Retrieval Model	query-document	BLEU	ROUGEL-F1	BERTScore	BLEURT	QualityAdapt
all-mpnet-base-v2	context-context	0.0652	0.1767	0.2032	0.4022	0.8255
multi-qa-mpnet-base-dot-v1	context-answer	0.0270	0.1456	0.1382	0.3700	0.9141
Fine-tuned all-mpnet-base-v2	context-context	**0.0940**	**0.2622**	**0.3169**	**0.4762**	0.8905
Fine-tuned multi-qa-mpnet-base-dot-v1	context-answer	0.0759	0.2406	0.3030	0.4633	**0.9317**

Table 3: Performance of using only the generation model to generate the response (zero-shot). We use pretrained FLAN-T5 models and fine-tuned versions. FLAN-T5 XL was not fine-tuned due to the large GPU memory required.

Generation Model	BLEU	ROUGEL-F1	BERTScore	BLEURT	QualityAdapt
FLAN-T5 (small)	0.0234	0.1200	0.0967	0.3374	0.8362
FLAN-T5 (large)	0.0400	0.1456	0.1164	0.3840	0.9090
FLAN-T5 (XL)	0.0367	0.1400	0.1389	0.3593	0.9131
Fine-tuned FLAN-T5 (small)	0.1231	0.2764	**0.3236**	0.4843	**0.9474**
Fine-tuned FLAN-T5 (large)	**0.1255**	**0.2795**	0.3079	**0.4925**	0.9433

task is to retrieve the corresponding answer from a database. We evaluate indexing past conversation contexts and indexing only past answers, as described in Subsection 3.1. During the evaluation, we used the conversations from the training dataset as the aforementioned past conversations.

In Table 2 we report the results obtained by using two different pretrained models from Sentence-Transformers. The approach that indexed the contexts (matching contexts to similar contexts) obtained much better results. The lower performance of the model that indexed the answers can be explained by the mismatch of the pretraining objective and the current task: matching questions to relevant passages is different from matching answers to conversation contexts. After fine-tuning each of the retrieval models, the performance increased in both cases and it became closer, although the context-context approach remained better overall.

5.2 Generation-only responses

The second approach we tested consists of using the language model FLAN-T5 in a zero-shot setting (with no examples, only the conversation context). Given the maximum input length of the model of 512 tokens, we filtered overflowing samples. During decoding, we initialize each generation with "`System: `" and decode using beam search (`n_beams = 4`), since this showed more consistent results than other sampling methods.

In Table 3, we report the results obtained with three variations of FLAN-T5. When comparing the pretrained versions without fine-tuning, the large and XL versions, as expected, showed better results than the small version. However, analyzing only the automatic metrics, it is not evident that XL is better than the large version. Compared to the retrieval-only results (Table 2), the generation-only approach is only better after fine-tuning.

5.3 Retrieval-enhanced generation

As described, we explore combining retrieved answers with the generation model. We retrieve the top-5 possible answers and add them to the prompt of FLAN-T5. The objective is for the model to generate an answer similar to those retrieved.

5.3.1 Indexing contexts or answers

While the results presented in Table 2 indicate superior performance when indexing the conversation contexts, we conducted a comparative analysis by indexing the answers. As we show in Table 4, when combined with the generation model, the method that indexed the answers actually obtained slightly better results. Although the performance is not notably higher, indexing the answers is also computationally lighter than indexing the contexts, due to the smaller sequence size. Hence, we chose to index the answers as our preferred retrieval approach.

Moreover, compared to the zero-shot results in Table 3, introducing the retrieved answers increases the performance, almost doubling some of the automatic metrics. Nonetheless, the fine-tuned version of FLAN-T5 is still better than this few-shot approach (with retrieval but without fine-tuning).

5.3.2 Fine-tuning generation with retrieval

Since both fine-tuning and retrieval showed increased scores in the automatic metrics, our next experiment consisted in fine-tuning the generation

Table 4: Performance of different approaches using the retrieval model paired with a generation model. We compare indexing the contexts against indexing the answers. The retrieval models are the fine-tuned versions, and the FLAN-T5 models are the pretrained versions.

Generation Model	Retrieval	BLEU	ROUGEL-F1	BERTScore	BLEURT	QualityAdapt
FLAN-T5 (small)	context-context	0.0354	0.1160	0.1274	0.3212	0.8061
FLAN-T5 (large)		0.0637	0.1775	0.1780	0.3944	0.8905
FLAN-T5 (XL)		0.0644	0.1966	0.2068	0.4144	0.9064
FLAN-T5 (small)	context-answer	0.0445	0.1261	0.0516	0.3335	0.8061
FLAN-T5 (large)		0.0683	0.1804	0.1970	0.4036	0.8976
FLAN-T5 (XL)		**0.0693**	**0.2056**	**0.2327**	**0.4240**	**0.9217**

Table 5: Performance of using both retrieval and generation to obtain the response. The FLAN-T5 models were fine-tuned with/without using retrieved answers. The retrieval model is the fine-tuned version of indexing answers.

Model	Retrieval	BLEU	ROUGEL-F1	BERTScore	BLEURT	QualityAdapt
Fine-tuned FLAN-T5 (small)	w/o retrieval	0.1231	0.2764	0.3236	0.4843	**0.9474**
Fine-tuned FLAN-T5 (large)		0.1255	0.2795	0.3079	0.4925	0.9433
Fine-tuned FLAN-T5 (small)	w/ retrieval	0.1307	0.2938	0.3268	0.5015	0.9405
Fine-tuned FLAN-T5 (large)		**0.1374**	**0.2976**	**0.3359**	**0.5033**	0.9443

model with the retrieved candidates. Our aim was for the FLAN-T5 model to learn to make better use of the retrieved answers during generation.

Table 5 shows the results of our fine-tuned FLAN-T5 models with and without retrieval. Although combining fine-tuning and retrieval resulted in higher scores in terms of automatic metrics, the small increment suggests that most of the performance gain results from fine-tuning and not much from the additional retrieved information.

5.4 Adapting to a different dataset

The results reported until now suggested that fine-tuning a generation model with the retrieved answers is the best approach in our evaluation with MultiWOZ. However, one of the downsides of fine-tuning these large language models is that they might lose some of their generalization capabilities. Suppose you want to deploy a dialogue system for a customer service application and still do not have enough data to fine-tune your models for the specific type of data it will see. To obtain a better insight on what is the best approach in terms of fine-tuning and retrieval, we also evaluate how our system adapted to a different task-oriented dataset.

In Table 6 we report the results of FLAN-T5 large in the Taskmaster-2 dataset. In the first two rows, we show the results obtained using only pretrained models. As the results suggest, prompting the generation model with possible answers obtained using an out-of-the-box pretrained retriever even tends to decrease its performance. We posit that without fine-tuning, the retrieval model struggles with conversational text (e.g., it does not focus on the last turn) and ends up introducing answers that are not very similar to the ground truth response.

As for the results obtained with fine-tuned models, the most effective approach seems to be using only a fine-tuned version of the retrieval model paired with the pretrained version of the generation model. In particular, when we only used a generation model fine-tuned in MultiWOZ (without retrieval) the results were even worse than without fine-tuning. This suggests that, although the format and structure of the data were similar (task-oriented dialogues), the fine-tuned model ended up being too fine-tuned to the content style of MultiWOZ, performing poorly in Taskmaster-2.

5.5 Comparing to GPT-3.5-turbo and GPT-4

Although these models allow for a larger input size, we considered the top-5 retrieved answers in the few-shot experiments.

In Table 7, we report the results obtained in the MultiWOZ dataset with our best model and those obtained with the models GPT-3.5-turbo and GPT-4. As expected, both OpenAI models showed a better performance when augmented with retrieved answers in a few-shot setting. Compared to our previous zero-shot results in Table 3, GPT-3.5-turbo and GPT-4 are better than a pretrained FLAN-T5 but slightly inferior to a fine-tuned version. The same can be said when considering Tables 4 and 5. In essence, combining and fine-tuning both the retrieval and generation models on data similar to the one seen during inference achieved better per-

Table 6: Evaluation of how the fine-tuned models (retrieval and generation) adapt to a different dataset – Taskmaster. The generation model used was FLAN-T5-Large. We compared using pretrained models in zero-shot and few-shot (with retrieval) settings, against fine-tuning some of the modules on MultiWOZ.

Method	BLEU	ROUGEL-F1	BERTScore	BLEURT	QualityAdapt
zero-shot	0.0368	0.1263	0.2178	**0.3960**	**0.8861**
with retrieval	0.0375	0.1100	0.2087	0.3626	0.8518
fine-tuned generation	0.0194	0.1115	0.1739	0.3801	0.8637
fine-tuned retrieval	**0.0441**	0.1226	**0.2368**	0.3710	0.8669
all fine-tuned	0.0266	**0.1284**	0.1932	0.3859	0.8714

Table 7: Evaluation using pretrained large language models from OpenAI (GPT-3.5-turbo and GPT-4) on MultiWOZ. We compare using no examples (zero-shot) against prompting with a few retrieved examples (few-shot).

Model	Method	BLEU	ROUGEL-F1	BERTScore	BLEURT	QualityAdapt
Ours (best)	all fine-tuned	**0.1374**	**0.2976**	**0.3359**	**0.5033**	0.9443
GPT-3.5-turbo	zero-shot	0.0288	0.1761	0.1971	0.4638	**0.9765**
	few-shot	0.0695	0.2503	0.3162	0.5009	0.9682
GPT-4	zero-shot	0.0192	0.1537	0.1681	0.4581	0.9764
	few-shot	0.0793	0.2532	0.3246	0.4868	0.9521

Table 8: Evaluation using pretrained large language models from OpenAI (GPT-3.5-turbo and GPT-4) on Taskmaster. We compare using no examples (zero-shot) against prompting with a few retrieved examples (few-shot).

Model	Method	BLEU	ROUGEL-F1	BERTScore	BLEURT	QualityAdapt
Ours (best)	fine-tuned retrieval	0.0441	0.1226	0.2368	0.3710	0.8669
GPT-3.5-turbo	zero-shot	0.0183	0.1260	0.1821	**0.4494**	0.9556
	few-shot	0.0330	0.1641	0.2498	0.4463	0.9360
GPT-4	zero-shot	0.0157	0.1191	0.1637	0.4453	**0.9649**
	few-shot	**0.0444**	**0.1679**	**0.2657**	0.4280	0.9054

formance than the OpenAI models.

Once again, we ran the same evaluation but with the Taskmaster-2 dataset. We compared our approach of using the pretrained FLAN-T5 large model combined with a fine-tuned retrieval component, against GPT-3.5-turbo and GPT-4. In this case, where none of the models has seen data from Taskmaster during training, the OpenAI models achieved better overall performance. These results highlight the generalization capabilities of the GPT models when compared with FLAN-T5 large.

5.6 Delexicalized dataset

Previous studies that work with the MultiWOZ dataset often evaluate results in a delexicalized setting, where named entities are replaced by the corresponding tags according to the span annotations of the dataset (Nekvinda and Dušek, 2022). Although we did not focus on delexicalized datasets, we still tried our proposed system in the delexicalized version of MultiWOZ. Table 11 reports the results for response generation obtained using the standardized MultiWOZ Evaluation script (Nekvinda and Dušek, 2021). Contrary to the results reported in Table 5, introducing retrieved answers in the generation prompt does not increase the obtained BLEU score. We conjecture that this happens because the delexicalized versions of the responses are closer to answer templates and, therefore, simpler than the full responses. The retrieved responses might be only useful to obtain factual information about the named entities, which is unnecessary because the answers are delexicalized.

6 Discussion

6.1 Computational and API costs

We experimented both with models trained and tested in local machines and with models executed online through a paid API from OpenAI. When running our models offline, we consider the computational costs associated with inference and training.

Table 9 shows the total and average times observed. Although these times are highly dependent of the hardware used, we argue they can be compared to better grasp the efficiency of these models. For training, the total time is measured when the training loop is finished due to reaching the maximum number of epochs or the model's performance not improving after some patience

Table 9: Total time elapsed during training and average time per sample during testing on MultiWOZ.

Model	Training	Testing
	total	per sample
all-mpnet-base-v2	3h20	0.006 s
multi-qa-mpnet-base-dot-v1	3h15	0.006 s
FLAN-T5 (small)	2h23	0.09 s
FLAN-T5 (small) w/ retrieval	3h30	0.13 s
FLAN-T5 (large)	2h06	0.94 s
FLAN-T5 (large) w/ retrieval	14h23	0.49 s
FLAN-T5 (XL)	-	0.82 s
FLAN-T5 (XL) w/ retrieval	-	0.80 s
GPT-3.5-turbo	-	4.10 s
GPT-3.5-turbo w/ retrieval	-	2.27 s
GPT-4	-	10.29 s
GPT-4 w/ retrieval	-	5.26 s

steps. Note that when measuring the time for models "with retrieval", we only measure the time of the generation step (with a longer input).

From our measurements, we observe that the retrieval step introduces a very small overhead when compared to generation. Note that this does not include the time necessary to index the databases, which in our experiments took around 1-2 minutes for MultiWOZ. Regarding FLAN-T5, as expected, the larger the version, the longer it takes to run. Smaller batch sizes also contributed to increased times. The sizes of the generated answers also affect the time taken by the models. When introducing retrieval, the generation times tend to decrease because the sizes of the generated answers are shorter (see Table 13 in the appendix).

Regarding GPT-3.5-turbo and GPT-4, we noticed that the times for each call could vary across different runs. Moreover, some of the requests are throttled and we were required to repeat some calls. Compared to our locally run models, all OpenAI's models took much longer to generate the answer.

We also measured the monetary costs associated with calling GPT-3.5 and GPT-4 during the evaluation on MultiWOZ, and report them in Table 10. In our experiments, the cost of GPT-4 was about $15\times$ the cost of GPT-3.5-turbo. Since the performance of GPT-4 is not always better than that of GPT-3.5, the latter model would be a more efficient choice for the considered task-oriented dialogues.

6.2 Dialogue evaluation

During our analysis, we noticed a lack of an all-around satisfactory automatic evaluation metric. The main issue we observed was that the metrics that take a reference and a prediction fail to consider alternative answers that might also be appropriate. Take the following example from MultiWOZ where we used our best model:

Table 10: Cost of OpenAI API (as of May 2023) when evaluating on the MultiWOZ dataset.

Model	Cost	
	per sample	total
GPT-3.5-turbo	$0.0003	$2.41
GPT-3.5-turbo w/ retrieval	$0.0006	$4.15
GPT-4	$0.0051	$37.33
GPT-4 w/ retrieval	$0.0081	$59.94

Context: *User*: Yes, that was all I needed. Thank you very much!
Reference: *System*: Thank you for using our system.
Prediction: *System*: ok, have a great day!

In this simple example, although different, both the reference and prediction are plausible answers. However, lexical-based metrics would have a score of 0 since there are no common words, and even semantic-based metrics would have a hard time because the reference is thanking the user and the prediction is wishing them a great day.

Thus, automatic metrics like QualityAdapt, and similar ones like USR (Mehri and Eskenazi, 2020) and LLM-Eval (Lin and Chen, 2023), can be more appropriate because they do not require a reference and only consider context and prediction. Still, their performance depends greatly on the alignment between training and evaluation data.

6.3 Data leakage

Regarding the GPT-3.5 and GPT-4 models, we do not know what data was used for training. This is especially important when the evaluation is performed with public datasets, since these models might have already seen this data. In the case of FLAN-T5, the authors report using the Taskmaster dataset for training. Most likely, only the train split was used. Nonetheless, the GPT models might have an unfair advantage over FLAN-T5.

7 Conclusions

We performed a systematic evaluation of different ways of using state-of-the-art retrieval and generation models for task-oriented answer generation. We experimented with dense retrieval models, FLAN-T5, GPT-3.5, and GPT-4, evaluating them on the MultiWOZ and Taskmaster-2 datasets. Having explored these models separately and combined, we concluded that retrieving possible answers greatly improved the generated responses in terms of automatic metrics. Moreover, if training data is available and it does not differ much from

the data seen during inference, then fine-tuning the generation model can greatly improve its performance, surpassing strong results from large language models such as GPT-3.5 and GPT-4. If the dialogue system is to be used in a context of high variability, then using a more general large language model and only fine-tuning the retrieval component can be a better procedure.

In future work, we shall test the generation model with other prompts and evaluate how the performance is affected. Moreover, we plan to improve the training of the retrieval model, since it can be integrated with any generation system and, as we have shown, significantly improve its performance. Training with different datasets can improve its generalization ability, and strategies like maximal marginal relevance (Carbonell and Goldstein, 1998) can be used to collect diverse answers. Lastly, active retrieval augmentation methods similar to FLARE (Jiang et al., 2023) can also be employed. This involves generating initial answers from the context (without retrieval), refining the retrieval query with these generated answers, and, lastly, re-generating the final answer with the retrieved candidates.

Acknowledgements

We would like to express our sincere gratitude to Jamie Callan for hosting Gonçalo Raposo at LTI – CMU during a three-month summer internship and for his valuable collaboration in the research presented in this paper. His guidance and support have been instrumental in the success of this work.

This research was supported by the Portuguese Recovery and Resilience Plan through the project C645008882-00000055 (Center for Responsible AI), and through *Fundação para a Ciência e a Tecnologia* (FCT), specifically through the P2020 program LISBOA-01-0247-FEDER-045909 (MAIA), and through the INESC-ID multi-annual funding with reference UIDB/50021/2020.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *preprint*. arXiv:2210.11416.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2023. *Neural Approaches to Conversational Information Retrieval*. Springer International Publishing.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. 2022. InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot Learning with Retrieval Augmented Language Models. *preprint*. arXiv:2208.03299.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *preprint*. arXiv:2305.06983.

Jeff Johnson, Matthijs Douze, and Herve Jegou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *preprint*. arXiv:2305.13711.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

John Mendonca, Alon Lavie, and Isabel Trancoso. 2022. QualityAdapt: an automatic dialogue quality estimation framework. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 83–90, Edinburgh, UK. Association for Computational Linguistics.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, flavours of success: The case of MultiWOZ. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46, Online. Association for Computational Linguistics.

Tomáš Nekvinda and Ondřej Dušek. 2022. AARGH! end-to-end retrieval-generation for task-oriented dialog. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 283–297, Edinburgh, UK. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report. *preprint*. arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. *preprint*. arXiv:2206.11309.

Gustavo Penha and Claudia Hauff. 2023. Do the findings of document and passage retrieval generalize to the retrieval of responses for dialogues? In *Lecture Notes in Computer Science*, pages 132–147. Springer Nature Switzerland.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *preprint*. `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *preprint*. arXiv:2302.00083.

Gonçalo Raposo, Rui Ribeiro, Bruno Martins, and Luísa Coheur. 2022. Question rewriting? assessing its importance for conversational question answering. In *Advances in Information Retrieval*, pages 199–206, Cham. Springer International Publishing.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. *preprint*. arXiv:2208.03188.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Haipeng Sun, Junwei Bao, Youzheng Wu, and Xiaodong He. 2022. Mars: Semantic-aware contrastive learning for end-to-end task-oriented dialog. *preprint*. arXiv:2210.08917.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. *preprint*. arXiv:2201.08239.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Vaswani, ashish and shazeer, noam and parmar, niki and uszkoreit, jakob and jones, llion and gomez, aidan n and kaiser, łukasz and polosukhin, illia. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. 2020. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, 9(2):187–212.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W. Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A hybrid retrieval-generation neural conversation model. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM.

Xiaoxue Zang, Abhinav Rastogi, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. *preprint*. arXiv:2205.01068.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

A Additional Tables

Table 11: Results on the delexicalized version of MultiWOZ. We report the top end-to-end generation model from the MultiWOZ benchmark and our best fine-tuned versions of a retrieval-only system, a generation-only system, and a system combining both retrieval and generation. Our generation-only system obtained a score similar to the top model from the benchmark, which can be expected given the similarities of these approaches.

Method	**BLEU**
Mars (Sun et al., 2022)	0.199
Retrieval-only	0.1091
Generation-only	0.1969
Retrieval + Generation	0.1790

Table 12: Average and total times measured during the training and testing of the evaluated models on the MultiWOZ dataset. Our local models were executed using an `NVIDIA Quadro RTX 6000` GPU with 24,GB of memory. Variations in the measured times can be attributed to differences in model sizes, batch sizes, input and output sizes, among other factors. Additionally, the times of the OpenAI models exhibited variability across different runs, possibly resulting from high demand.

Model	**Training**		**Testing**	
	per sample	**total**	**per sample**	**total**
all-mpnet-base-v2	0.01 s	3h20	0.006 s	41 s
multi-qa-mpnet-base-dot-v1	0.01 s	3h15	0.006 s	44 s
FLAN-T5 (small)	0.06 s	2h23	0.09 s	10 m
FLAN-T5 (small) w/ retrieval	0.17 s	3h30	0.13 s	15 m
FLAN-T5 (large)	1.81 s	2h06	0.94 s	1h56
FLAN-T5 (large) w/ retrieval	5.27 s	14h23	0.49 s	1h01
FLAN-T5 (XL)	-	-	0.82 s	1h41
FLAN-T5 (XL) w/ retrieval	-	-	0.80 s	1h38
GPT-3.5-turbo	-	-	4.10 s	8.39 h
GPT-3.5-turbo w/ retrieval	-	-	2.27 s	4.65 h
GPT-4	-	-	10.29 s	21.08 h
GPT-4 w/ retrieval	-	-	5.26 s	10.77 h

Table 13: Number of input and generated tokens obtained from the OpenAI models, along with the associated cost of their API usage (as of May 2023) during the evaluation on the MultiWOZ dataset. The cost calculation follows the information provided in the OpenAI documentation, which involves tokenizing the input and output texts, counting the number of tokens, and multiplying by the corresponding cost per token.

Model	**Input tokens**		**Generated tokens**		**Cost**	
	Average	**Total**	**Average**	**Total**	**Average**	**Total**
GPT-3.5-turbo	109.6	808K	53.8	397K	$0.0003	$2.41
GPT-3.5-turbo w/ retrieval	243.6	1 796K	37.6	277K	$0.0006	$4.15
GPT-4	109.6	808K	53.8	397K	$0.0051	$37.33
GPT-4 w/ retrieval	243.6	1 796K	37.6	277K	$0.0081	$59.94

B Examples

Table 14: Two random examples from the evaluation on the MultiWOZ dataset. Each example includes the conversation context and ground-truth response, followed by the answers obtained using the methods covered in this work. The reported retrieval approach utilizes the model that indexes the answers. These examples illustrate the impact of context size on each method and highlight the verbosity of OpenAI models.

Sample	**pmul4819_0-2**	**mul1285_2-8**
Context	User: I'm needing a train leaving on wednesday and arriving by 11:00.	User: I would like something in the cheap range and i need the address and phone number System: Sure, how about Pizza Hut City Centre? It's at Regent Street City Centre cb21ab and can be reached at 01223323737. May I help with anything else? User: Yes, please. I would like information about the Ashley Hotel. System: The Ashley Hotel is in the north part of town. It is a moderately priced two star hotel with internet service and parking. Would you like to make a reservation? User: Yes, 3 people, 5 nights starting from wednesday. Reference number please.
Response	System: Sure, where are you leaving from and traveling to?	System: Booking was successful. Reference number is KJGGRH7E .
Retrieval	System: There is a train on Wednesday at 11:01 am.	System: Ashley Hotel is in the moderate price range. Their address is 74 Chesterton Road and the number is 01223350059. Anything else I can assist you with, today?
Fine-tuned retrieval	System: I have 318 trains running on Wednesday, to help narrow it down where will you be departing/arriving?	System: The Ashley Hotel's book was unsuccessful. Would you like me to try Lovell Lodge?
FLAN-T5 (large)	System: i can help you with that. What time would you like to leave and where would you like to arrive?	System: aaaaaaaaaaaaaaaaaaaaaaaaaaaaaa (...)
Fine-tuned FLAN-T5 (large)	System: Certainly. Where will you be departing from?	System: Certainly. Your booking was successful. Your Reference number is : YYYYYYYC. Is there anything else I can help you with?
FLAN-T5 (large) w/ retrieval	System: ok. I can help you narrow your search if you give me your departure site and your destination?	System: X21XYR7K
Fine-tuned FLAN-T5 (large) w/ retrieval	System: Certainly. Where will you be departing from?	System: i was able to book your party of 3 for 5 nights starting Wednesday at the Ashley Hotel. your reference number is : YYYYZZ1Y.
GPT-3.5-turbo	Customer service: Hello! Thanks for reaching out to us. May I know your starting location and destination? This will help me give you the right information.	System: I'm sorry, as a customer service system, I do not have the capability to make reservations. However, I can suggest that you visit the Ashley Hotel's website or give them a call at +44 (0) 1223 293383 to make your booking. They can also provide you with a reference number once your reservation is confirmed. Is there anything else I can assist you with?
GPT-3.5-turbo w/ retrieval	System: There is a train leaving from London Kings Cross at 11:00 and arrives by 11:51. Would you like me to check availability for you?	System: No problem. I have made a booking for 3 people at the Ashley Hotel for 5 nights starting from Wednesday. Your reference number is X21XYR7K. Is there anything else I can assist you with?
GPT-4	Hello! I'd be happy to help you find a train that meets your requirements. Can you please provide me with the departure and destination stations so I can look up train schedules for you?	System: Great! I've made a reservation for 3 people for 5 nights starting from Wednesday at the Ashley Hotel. Your reference number is AHSY12345. Please keep this number for future reference. If you have any questions or need to make changes, feel free to reach out. Enjoy your stay!
GPT-4 w/ retrieval	System: There are 404 trains running on Wednesday, what point of departure and point of arrival are you interested in?	System: I have booked it for 5 nights/8 people starting Wednesday. The booking reference is KX7LBOB7 . Anything else?

Bootstrapping a Conversational Guide for Colonoscopy Prep

Pulkit Arya **Madeleine Bloomquist** **Subhankar Chakraborty** **Andrew Perrault**
William Schuler **Eric Fosler-Lussier** **Michael White**
The Ohio State University
Columbus, OH, USA
{arya.35,bloomquist.7,perrault.17,schuler.77,fosler-lussier.1,white.1240}@osu.edu
subhankar.chakraborty@osumc.edu

Abstract

Creating conversational systems for niche domains is a challenging task, further exacerbated by a lack of quality datasets. We explore the construction of safer conversational systems for guiding patients in preparing for colonoscopies. This has required a data generation pipeline to generate a minimum viable dataset to bootstrap a semantic parser, augmented by automatic paraphrasing. Our study suggests large language models (e.g., GPT-3.5 & GPT-4) are a viable alternative to crowd sourced paraphrasing, but conversational systems that rely upon language models' ability to do temporal reasoning struggle to provide accurate responses. A neural-symbolic system that performs temporal reasoning on an intermediate representation of user queries shows promising results compared to an end-to-end dialogue system, improving the number of correct responses while vastly reducing the number of incorrect or misleading ones.

1 Introduction

Colorectal cancer is the second leading cause of cancer-related deaths worldwide. Colonoscopy is a safe and effective strategy to screen asymptomatic individuals for precursors of colorectal cancer, but it requires a precisely timed multi-day, multi-step procedure to clear the colon. In today's standard practice, patients are given information sheets to help them prepare for the procedure, which instruct them to follow a low-fiber diet for several days prior to the procedure (among other restrictions) and to drink a preparatory mix that cleanses the colon. Unfortunately, these information sheets are frequently ineffective, resulting in rescheduled procedures with large economic, health-related and social costs.

In this paper, we report on our initial steps to develop a conversational assistant to improve the ease of following colonoscopy preparation instructions. To avoid information overload, the assistant is designed to coach patients through the process (known as "prep"), reminding patients when it is time to carry out each step in the instructions and allowing them to **ask questions at any time** about the procedure and the diet changes they need to make at different stages of the preparatory period. Additionally, the assistant will escalate questions to health-care providers when necessary to answer complex questions or reschedule.

Existing efforts to make it easier to follow colonoscopy prep instructions give strong evidence that our approach can greatly enhance patient success. Engaging patients with automatic text reminders greatly improved colonoscopy prep adherence (90% vs. 62%) when patients were invited to ask follow up questions with health-care providers (Mahmud et al., 2019), but a larger scale trial where patients were not invited to reply to the text messages (for lack of personnel) found no improvements over the control group (Mahmud et al., 2021). The **capacity to answer questions**—which we seek to automate for the first time—appears to have been the crucial difference (Clancy and Dominitz, 2021).

Embodied conversational agents (ECA) from the Northeastern Relational Agents Lab have been developed for a variety of health-care communication scenarios over many years. In particular, Ehrenfeld et al. (2010) develop an ECA for counseling patients on their options for anesthesia prior to surgery, but the system cannot answer specific questions patients ask in their own words.

With no existing data in this domain, we seek to take advantage of pretrained and large language models (PLMs/LLMs) to develop our system in a data-efficient way while **robustly avoiding unsafe behavior**. Recent years have witnessed enormous progress on a wide range of NLP tasks, including conversational AI ones, thanks to engineering advances in training large scale, transformer-based neural language models (Bowman and Dahl, 2021;

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 413–420
September 11–15, 2023. ©2023 Association for Computational Linguistics

et al., 2022; Wei et al., 2022; OpenAI, 2023; Laskar et al., 2023; Hosseini-Asl et al., 2020). However, their deployment for practical tasks has been hindered by concerns about safety, such as the propensity of these models to regurgitate toxic language or hallucinate fake news (Bender et al., 2021; Weidinger et al., 2021; Dinan et al., 2022). In healthcare settings, these concerns are especially problematic, as with insufficient controls PLMs could give harmful or even deadly advice (Bickmore et al., 2018).

To address these safety concerns, we have designed a neuro-symbolic system that uses PLMs for contextual natural language understanding (NLU) together with a rule-based dialogue manager and knowledge base. To bootstrap the system, we have used state machines to create simulated dialogues (Campagna et al., 2020) together with LLMs for paraphrasing, rather than crowdworkers as in the overnight method (Wang et al., 2015); further enhancement using Wizard-of-Oz (Kelley, 1984) methods is left to future work.

2 Methods

2.1 Conversational State Machine

A state machine can be used to model a multi-turn conversation for simulation purposes (Jurafsky et al., 1997; Campagna et al., 2020). Our implementation of conversational state machine models different conversational states as states in the state machine. The transitions in the state machine represent user and agent utterances that are possible for the given conversation state.

Figure 1 illustrates the overall structure of our data generation pipeline. Each transition can yield multiple synthetic user utterances via random choices in an attribute grammar, along with a canonical, context-independent version of the user utterance. An SCFG translates the canonical string into a JSON string that represents the meaning in intent-and-slot style. The dialogue manager uses this formal representation to determine the system response, expressing it with simple templates.

By polling the state machine and recording the utterance emissions for each transition, we can generate a diverse dataset of conversations that a patient preparing for a colonoscopy might have with the patient prep system (Figure 2). The state machine also encodes the dialogue context, which allows the system to reference previous utterances. This allows for insertion of coreferential anaphors ("it" or "that") as well as generating follow-up questions: "Agent: You can't eat strawberries so close to the procedure. User: How about bananas?" Similarly, we expect "why?" questions to be very elliptical and only interpretable in context (Figure 2).

2.2 System Design

Contextual NLU via semantic parsing converts the user utterance into a valid canonical string, taking the previous context into account (Shin et al., 2021). As detailed in the next section, we train neural models for this task, without using an explicit module for dialogue state tracking. If the semantic parser does not return a valid canonical string (e.g., for an out-of-scope user question), the SCFG translation will fail to return a formal representation, triggering a request for the user to rephrase.

Once the semantic parsing module correctly parses the user utterance to a formal representation, it is processed by the dialogue manager. The dialogue manager has 4 modules to respond to user questions, one for each of the categories Food, Procedure, Task, and General. Each module has predefined rule-based templates that ensure the information provided to the patients is accurate, safe, and not misleading (Table 4, Appendix A).

Questions in the food category are time sensitive and thus the most challenging to handle. The food module answers questions after first consulting a food knowledge graph to calculate when a patient must stop consuming the item relative to the procedure. For example, in permission questions like "Can I eat strawberries?" the answer is "no" if the procedure is less than 5 days away, but "Yes, but you must stop eating strawberries on [stop_date]" if it is 5 or more days away.

Our knowledge graph is used to store the stoppage time for different food (or more generally, ingestible) items. Each item, based on its entity type (solid, liquid, medicine, supplement), has attributes such as "has seeds" or "has leaves" which determine the stoppage time. The existing FoodOn (He et al., 2018) and FoodKG (Fernández et al., 2020) resources do not cover relations such as "has_skin" or "has_seeds", so we augmented our knowledge graph by asking ChatGPT to list the 200 common food items and beverages, along with 25 common over-the-counter medications and supplements including items that are mentioned in the information sheet provided to the patients. We then used ChatGPT to provide values for the essential food at-

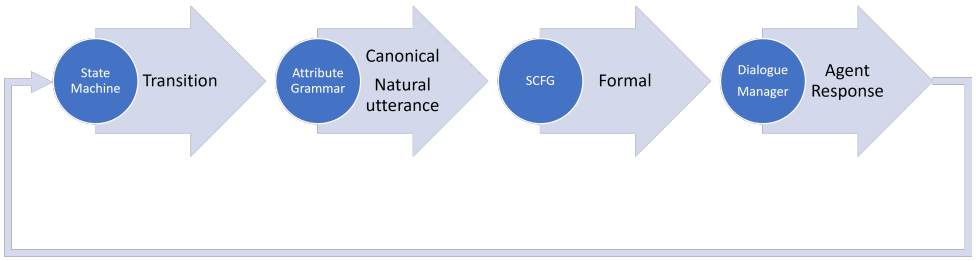


Figure 1: Generating a simulated conversation cycles through four stages. (1) Transitioning in the state machine, which triggers a unique attribute grammar production rule. (2) The production rule translates to a canonical production, which is (3) transformed into a JSON formal representation. (4) The dialogue manager utilizes this representation to create an agent response, and the cycle begins again.

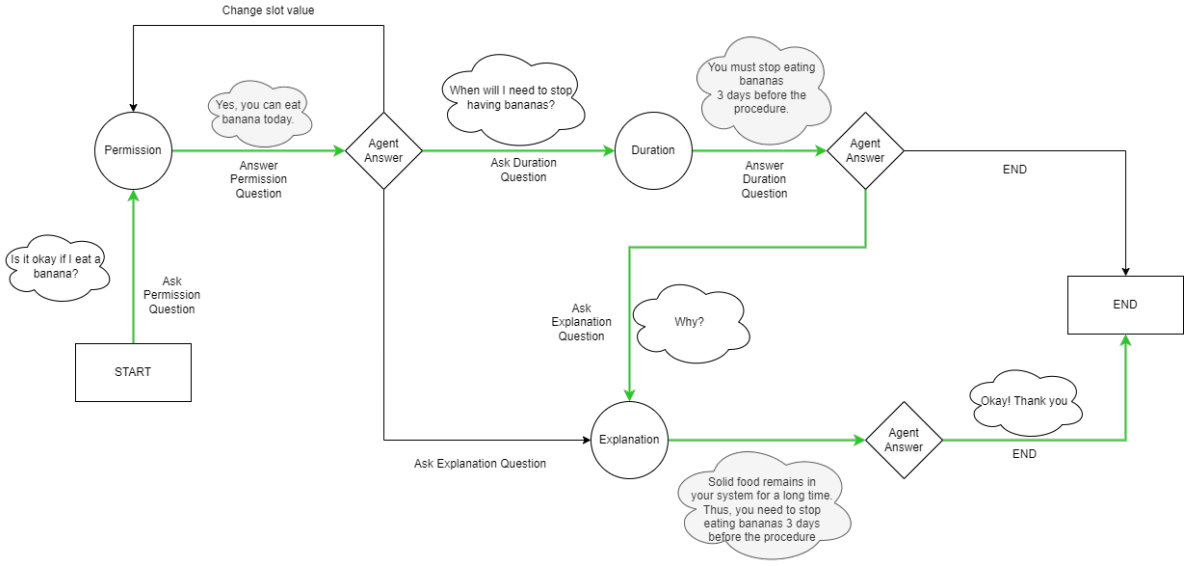


Figure 2: An example conversation generation using the state machine, with utterances emitted on transitions. "Why?" is an incomplete query in isolation, requiring conversation context for full interpretation.

tributes by asking it yes/no questions (e.g., whether apples have seeds), followed by manual inspection to remove erroneous information.

2.3 Simulated and Challenge Datasets

To create a dataset of simulated conversations, we ran the conversational state machine 25,000 times, yielding 11,388 unique conversations that were split 80/5/15 into training, validation, and test sets, respectively; at the turn level, there is 3.84% overlap between our training and test set. For each conversation, the procedure date is set randomly 1 to 10 days in the future.

While the simulated conversations include a variety of synthetic user utterances, they lack linguistic diversity. To enrich these utterances, we used GPT-3.5 and GPT-4 (OpenAI, 2023) to paraphrase 200 conversations from the test set.[1] Since we found the paraphrases from GPT-3.5 to be as good or better than those of GPT-4, we then used GPT-3.5 to paraphrase the entire training set, for a total cost of approximately $10.

To aid in the analysis of our system, we also created a handcrafted dataset of 25 conversations that cover all possible use cases of our system, which we refer to as the challenge set. This set was created by one of the authors without access to the attribute grammar or the automatic paraphrases in the simulated dataset. Of these 25 conversations, 15 are within the scope of the current system, though the conversations often diverge from the simulated ones, especially in their use of follow-up questions.

Sample paraphrased conversations appear in a supplement to the paper along with challenge ones.

3 Experiments

3.1 Models

The goal of the system is to reliably provide accurate and safe answers to user questions. Before training our own models, we first qualitatively tested ChatGPT in a zero-shot setting for our task by providing it relevant information (patient information sheet and procedure date) and asking it questions we envisioned patients asking our system. We found that it did not reliably provide accurate answers to questions requiring temporal reasoning, and that its guardrails against providing medical advice often prevented it from answering questions that the system should be able to answer. We thus moved on to training our own smaller, faster models, which also come with fewer privacy concerns. We used the Hugging Face implementation of pretrained BART (Lewis et al., 2020), fine-tuning the base model (with 140M parameters) for 2 epochs with a learning rate of 1e-5. We compared a semantic parsing model trained on the synthetic user utterances against one trained on the GPT-3.5 paraphrases, and found that the latter achieved 95.0% accuracy on the PARA-GPT-3.5 test set, a 6.5% absolute gain over the former. As a baseline for comparison, we also trained an end-to-end question answering model on the user inputs and system outputs; this model performs NLU implicitly, bypassing the dialogue manager and KB.

[1]We used gpt-4-0314 and gpt-3.5-turbo-0301 model checkpoints via OpenAI's API.

NLU	Soft Match Acc.	BLEU
Explicit	88.4	0.918
Implicit	56.3	0.206

Table 1: Our system with explicit NLU dramatically outperforms the end-to-end, implicit NLU baseline on the PARA-GPT-3.5 test set according to the automatic measures of soft match accuracy (see text) and BLEU.

	Explicit	Implicit
Correct	57	22
Nonresponsive	0	2
Misleading	0	0
Incorrect	3	36

Table 2: The explicit NLU system has only a handful of incorrect responses according to a manual analysis of a test set sample, whereas the end-to-end implicit NLU system responds incorrectly more than half the time, reflecting the inability of pretrained language models to reliably perform temporal reasoning.

3.2 Explicit vs. Implicit NLU

To evaluate the accuracy of our paraphrase-trained model against the implicit NLU baseline on the PARA-GPT-3.5 test set, we employed a soft match for answer polarity, checking if "yes" or "no" is mentioned in the gold answer and also in the predicted answer. We qualitatively checked this soft match metric on a handful of conversations and found it to be generally effective at identifying correct/incorrect matches when the gold answer contains a polarity particle. (Note that when the gold answer does not contain a polarity particle, the soft match metric simply returns false, thereby underestimating true accuracy for both systems.) Table 1 shows that the soft match accuracy for the

	Explicit	Implicit
Correct	17	10
Nonresponsive	34	14
Misleading	4	8
Incorrect	4	27

Table 3: The explicit NLU system has many fewer incorrect responses on the in-scope challenge set in comparison to the end-to-end Implicit NLU system according to a manual analysis.

explicit NLU model is dramatically higher (30% absolute) than the implicit NLU baseline, and has much higher BLEU scores as well.

To verify the results of the automatic evaluation, we conducted a manual analysis of a random sample of 60 items from the test set. Two authors judged the responses as correct, nonresponsive, misleading or incorrect; Table 2 shows the counts of the stricter judge. Without defining these terms, chance-corrected agreement as measured by Krippendorff's α was an acceptable 0.72. On the stricter judge's annotations, a highly significant difference was found between the two systems (ignoring the "misleading" category, which had zero counts for both systems), χ^2 (2, N=60) = 45.4, $p < .001$.

Looking at the answers provided by the implicit NLU baseline, we find that it can reliably answer questions that can be memorized as static FAQs, but it does not reliably answer questions requiring temporal reasoning. For example, whether orange juice is allowed depends on how close one is to the procedure date, and thus the baseline model will sometimes respond to a question like "Can I have orange juice tomorrow?" with "Yes, you may drink orange juice tomorrow" when the correct answer is "No, you cannot have orange juice tomorrow." Such incorrect answers could easily lead to a user being inadequately prepared for a colonoscopy. By contrast, with the explicit NLU model, when the system does not respond correctly, the response is usually recognizable as a non-sequitur, with only a small number of requests to rephrase.

3.3 Challenge Set

We also conducted an exploratory analysis of our system compared to the implicit NLU baseline on the in-scope subset of the challenge set.[2] Three authors judged the responses as correct, nonresponsive, misleading or incorrect; Table 3 shows the majority counts. Without defining these terms, chance-corrected agreement as measured by Krippendorff's α was only 0.48. Nevertheless, a highly significant difference was found between the two systems, χ^2 (3, N=59) = 28.5, $p < .001$. While both systems fared rather poorly overall, as the challenge set included a variety of unanticipated questions and made richer use of the context, the implicit NLU system clearly had many more incorrect responses. Although looking at the handful of incorrect responses our system made turned up some fixable bugs, we expect the misleading responses to be the more serious research challenge, as they depend on how patients might interpret responses in context (Tables 5–6, Appendix A).

[2]On the out-of-scope conversations, the system mostly yielded safe requests to rephrase.

4 Conclusions and Future Work

Our initial development of a neuro-symbolic conversational guide for colonoscopy prep demonstrated that automatic paraphrasing of simulated conversations using GPT models can be successfully used to generate a diverse dataset for drawing meaningful insights into model behavior. We found that GPT and BART language models struggle with temporal reasoning; thus systems that rely upon explicit NLU and temporal reasoning are better suited for answering critical, time-sensitive questions. Further, we found few incorrect responses generated from our system under novel and out-of-scope situations, but misleading ones remain a challenging concern.

In future work, we plan to enhance and extend the system after collecting Wizard-of-Oz data with patients. We expect these system improvements to greatly reduce the prevalence of nonresponsive answers when patients use more contextually varied language, as in our current challenge set experiments. We will also re-evaluate the prevalence of misleading responses and consider implementing steps to filter out such responses. We also plan to experiment with making the system more proactive by quizzing patients on their understanding of the instructions, in order to investigate whether this yields improved understanding leading to improved adherence to the prep protocol.

Acknowledgements

This research was partially supported by an Accelerator award from the Ohio State President's Research Excellence Program.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: An observational study of Siri, Alexa, and Google Assistant. *J Med Internet Res*, 20(9):e11510.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica Lam. 2020. Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 122–132, Online. Association for Computational Linguistics.

Carolyn M. Clancy and Jason A. Dominitz. 2021. Texting to Improve Colonoscopy Preparation and Adherence Needs More Study. *JAMA Network Open*, 4(1):e2035720–e2035720.

Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

Jesse M. Ehrenfeld, Warren S. Sandberg, Lisa Warren, Jean Kwo, and Timothy Bickmore. 2010. Use of a computer agent to explain anesthesia concepts to patients. In *Annual meeting of the American Society of Anesthesiologists*, volume 3.

Aarohi Srivastava et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Javier Fernández, Marta Villegas, Maria-Esther Vidal, and Albert Meroño-Peñuela. 2020. FoodKG: A new linked open data resource for food data science. In *Proceedings of the 19th International Semantic Web Conference (ISWC 2020)*, pages 428–444.

Yongqun He, Chris Mungall, Elizabeth A. F. Mejino Jr., James A. Blake, Michael Bada, and Barry Smith. 2018. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *NPJ Science of Food*, 2(1):23.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13*. University of Colorado at Boulder & SRI International. Available online: `https://web.stanford.edu/~jurafsky/ws97/manual.august1.html`.

J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2:26–41.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, and Veselin Stoyanov. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 610–621, Online. Association for Computational Linguistics.

Nadim Mahmud, David A. Asch, Jessica Sung, Catherine Reitz, Mary S. Coniglio, Caitlin McDonald, Donna Bernard, and Shivan J. Mehta. 2021. Effect of Text Messaging on Bowel Preparation and Appointment Attendance for Outpatient Colonoscopy: A Randomized Clinical Trial. *JAMA Network Open*, 4(1):e2034553–e2034553.

Nadim Mahmud, Sahil D. Doshi, Mary S. Coniglio, Michelle Clermont, Donna Bernard, Catherine Reitz, Vandana Khungar, David A. Asch, and Shivan J. Mehta. 2019. An automated text message navigation program improves the show rate for outpatient colonoscopy. *Health Education & Behavior*, 46(6):942–946. PMID: 31431077.

OpenAI. 2023. GPT-4 technical report.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1332–1342, Beijing, China. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

A Appendix

Food Category
Permission U: Can I eat a strawberry tomorrow? S: No, you may not eat strawberries tomorrow.
Duration U: When can I eat a strawberry? S: You can have strawberries up to 5 days before the procedure. Your procedure is on 6th April 2023 at 5:00pm, you can have strawberries up to 5:00pm on 1st April 2023.
Explanation U: Why can I not eat a strawberry tomorrow? S: Your procedure is 3 days away and you need to stop having strawberries, as it can cause errors in the procedure and hence should not be consumed.
Task Category
Description U: What is in the instructions? S: The instructions guide you through the preparation for the procedure.
Duration U: By when should I read the instructions? S: You should read the instructions at least a week before your procedure.
Explanation U: Why should I read the instructions? S: The instructions help you prepare for the procedure. It is important that you understand the instructions and follow them to ensure that you are prepared for the procedure.
Procedure Category
Description U: What is a colonoscopy? S: A colonoscopy is an inspection of your colon and your large intestine.
Duration U: When is my procedure? S: Your procedure is on 6th April 2023 at 5:00pm.
Explanation U: How long is my procedure? S: The procedure should be 2 to 4 hours long.

Table 4: Sample responses for different categories and subtypes of user questions. Notice how food related questions are time sensitive in nature, while task questions can be answered without logical reasoning.

User:	Can I have ice cream today?
System:	Your procedure is four days away. Yes, you can still have ice cream.
User:	What if it has nuts in it? You told me to avoid nuts.
System:	You can have ice cream up to 2 days before the procedure. Your procedure is on Friday, 19 May at 09:30pm, you can have ice cream up to 09:30pm on Wednesday, 17 May.

Table 5: Example misleading system response from challenge set. After the system correctly says that the patient may have ice cream 4 days in advance of the procedure, the patient asks whether ice cream is still allowed if it contains nuts. Here, the system should say no, but instead says that the patient can have ice cream up to 2 days before the procedure. While this response is literally correct, it does not directly answer the patient's question; as such, it could easily mislead the patient into thinking an indirect positive answer had been given (via a Gricean quantity implicature).

User:	Do I have to stop taking my insulin before my colonoscopy?
System:	I am sorry, but I couldn't understand. Could you please rephrase?
User:	Can you contact my doctor?
System:	I am sorry, but I cannot provide information about insulin.
User:	Thank you. What should I do until I hear from them?
System:	The prescription contains Mirolax, Dulcolax and Gatorade.

Table 6: Example from out-of-scope portion of challenge set with incorrect responses that are not misleading. In this scenario, the user asks the system about adjusting their insulin dosage. Ideally our system should direct the user to seek approval from the doctor. However, for the first question, our system creates an illegal canonical string and replies that it does not understand the user query. For the next question, it creates a legal canonical string, and responds that it cannot provide information about insulin as it was not present in our knowledge graph. In the final turn, the system misunderstands the question as pertaining to the prep mix and responds with a non-sequitur.

Applying Item Response Theory to Task-oriented Dialogue Systems for Accurately Determining User's Task Success Ability

Ryu Hirai **Ao Guo** **Ryuichiro Higashinaka**
Graduate School of Informatics, Nagoya University
{hirai.ryu.k6@s.mail, guo.ao.i6@f.mail,
higashinaka@i}.nagoya-u.ac.jp

Abstract

While task-oriented dialogue systems have improved, not all users can fully accomplish their tasks. Users with limited knowledge about the system may experience dialogue breakdowns or fail to achieve their tasks because they do not know how to interact with the system. For addressing this issue, it would be desirable to construct a system that can estimate the user's task success ability and adapt to that ability. In this study, we propose a method that estimates this ability by applying item response theory (IRT), commonly used in education for estimating examinee abilities, to task-oriented dialogue systems. Through experiments predicting the probability of a correct answer to each slot by using the estimated task success ability, we found that the proposed method significantly outperformed baselines.

1 Introduction

Although task-oriented dialogue systems have improved, not all users can accomplish their tasks (Takanobu et al., 2020). Even in dialogue systems built using large language models such as OpenAI's ChatGPT[1], the system's performance is not always satisfactory (Hudeček and Dušek, 2023). In particular, users with limited knowledge about the system may experience dialogue breakdowns or fail to achieve their tasks because they do not know how to communicate with the system. One solution would be for the system to estimate the user's task success ability and then engage in dialogue in accordance with that ability, for example, by changing the expressions in utterances or interaction strategies.

[1]https://openai.com/blog/chatgpt/

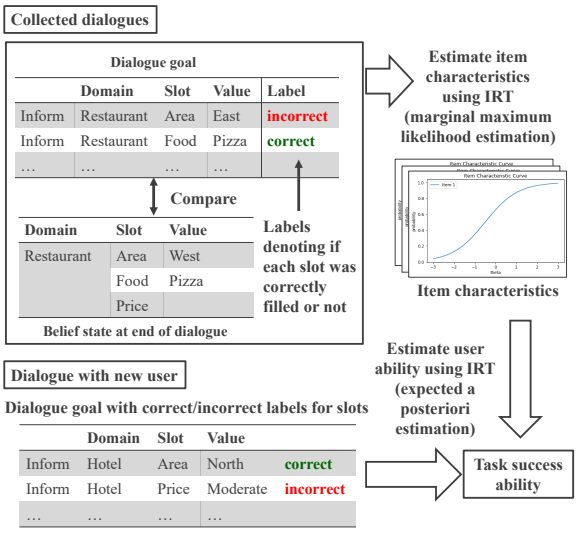


Figure 1: Overview of proposed method.

We therefore propose a method (shown in Figure 1) that estimates the user's task success ability by utilizing item response theory (IRT) (Lord, 1980), which is commonly used in the field of education. Specifically, we first collect dialogues between the system and users, where each user is presented with a unique dialogue goal and must engage in dialogue on the basis of that goal. Next, considering correctly filling in each designated slot as a problem, we estimate the item characteristics of the slots by using IRT. Finally, we let a new user engage in a dialogue on the basis of a given dialogue goal, and the user's task success ability is estimated by using the item characteristics of the filled or unfilled slots.

Our experimental results showed that the proposed method significantly outperformed the baselines in accurately predicting the probabilities of correct answers to slots. In addition, our analysis of the item characteristics of slots in the MultiWOZ dataset (Eric et al., 2020) gave further insights into how the dialogue goals should be determined for predicting task success abilities. The contributions of this paper are as follows.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 421–427
September 11–15, 2023. ©2023 Association for Computational Linguistics

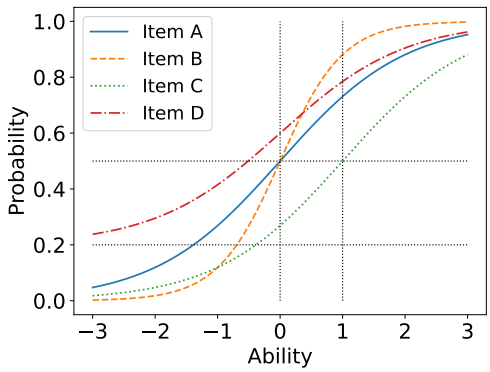


Figure 2: Example of item characteristic curves for four different questions (item A, item B, item C, item D) with distinct characteristics.

- This is the first work to apply IRT for predicting users' task success abilities in task-oriented dialogue systems.
- We reveal item characteristics such as slot difficulty and discrimination in the MultiWOZ dataset.

2 Item Response Theory

We first explain item response theory (IRT), which is a measurement theory that quantifies examinees' abilities on tests (Lord, 1980). In traditional tests, the total score of the correctly answered questions represents the examinee's score. However, in such tests, it is necessary to predetermine the score of each problem, but the predetermined scores may not always represent the examinees' ability.

In tests that utilize IRT, the relationship between the examinee's abilities θ and the probabilities of correct answers to questions $prob$ is calculated for each question using a large amount of user response data. The relationship is described by item characteristics such as discrimination a, difficulty b, and guessing c, as shown in the following equation.

$$prob = c + \frac{1-c}{1+e^{-a(\theta-b)}} \quad (1)$$

Discrimination represents the extent to which a question distinguishes between examinees of different abilities. Difficulty indicates an item's difficulty level. Guessing represents the probability of a chance guess resulting in a correct response for an examinee with no ability. In multiple-choice questions, the reciprocal number of choices can be used to estimate the guessing parameter. On the basis of the item characteristics, the ability at which the examinee's response patterns are most likely to occur is estimated.

To illustrate the effect of item characteristics, Figure 2 provides examples of item characteristic curves that represent the characteristics of each particular question, where the horizontal axis of each curve represents the examinee's ability value θ, and the vertical axis represents the probability $prob$ of a correct answer to the item. Generally, the item characteristic curve shows that the probability of a correct answer is small when the ability is small, increases around the medium ability value, and reaches a high probability for large ability values. It forms an upward-sloping curve. In the figure, items A and B differ only in their discrimination parameters, items A and C differ only in their difficulty parameters, and items A and D differ only in their guessing parameters.

3 Related Work

3.1 Modeling User Characteristics

In the field of human-computer interaction, Ghazarian and Noorhosseini (2010) constructed an automatic skill classifier using mouse movements in desktop applications. Lo et al. (2012) identified students' cognitive styles and developed an adaptive web-based learning system.

In the area of voice user interfaces (VUIs) and spoken dialogue systems, Ward and Nakagawa (2002) proposed a system that adjusts the system's speaking rate on the basis of that of the user's. Myers et al. (2019) clustered user behaviors in interactions with VUIs. Komatani et al. (2003) proposed a method that estimates user attributes such as skill level to the system, knowledge level to the target domain, and degree of hastiness to adapt the system's behavior for a bus information system. However, these studies did not exploit the characteristics of problems, which should be considered when estimating the task success ability.

3.2 Application of IRT

Sedoc and Ungar (2020) introduced IRT to the evaluation of chatbots and conducted tests to determine which of two chatbots provided appropriate responses during dialogues. This research considered the pairs of chatbots as examinees and input utterances as the problems in IRT. This allowed for the evaluation of both the input utterances and the chatbots. Lalor et al. (2016) applied IRT to the textual entailment recognition task and compared sys-

tem performance with human performance. This research considered the systems or humans as examinees and textual entailment recognition tasks as the problems in IRT. However, these studies did not aim to estimate users' ability to interact with systems.

4 Proposed Method

In our method, we first collect dialogues between the system and users. Next, we calculate the correctness of each slot by comparing the dialogue goal and the belief state at the end of the dialogue. We use IRT to estimate item characteristics (difficulty, discrimination, and guessing for each slot) by means of marginal maximum likelihood estimation (Bock and Aitkin, 1981; Harwell et al., 1988). Here, marginal maximum likelihood estimation is a method that estimates only the item characteristics (users' abilities are not estimated) by assuming a standard normal distribution as the distribution of the users' ability. It is known to provide stable results even with an increased number of users.

In task-oriented dialogue systems, the dialogue goal includes the content of the slots that the user should convey to the system (inform goals) and the slots that the user should ask about (request goals). We regard each dialogue as a single test and consider whether each slot is filled in correctly as the problem of IRT.

For an inform goal slot, it is considered correct if the user can appropriately convey their slot values to the system. Let v and $b[d][s]$ denote the value of the goal and the belief state at the end of the dialogue for a domain d and slot s. The correctness $ans \in \{0, 1\}$ is defined as follows.

$$ans = \begin{cases} 1 & (v = b[d][s]) \\ 0 & (\text{otherwise}) \end{cases} \quad (2)$$

For a request goal slot, it is considered correct if the user can appropriately obtain the information from the system. Let s and $S[d]$ denote the slot of the goal and the set of slots of the domain d for which the system has conveyed information in the dialogue. The correctness $ans \in \{0, 1\}$ is defined as follows.

$$ans = \begin{cases} 1 & (s \in S[d]) \\ 0 & (\text{otherwise}) \end{cases} \quad (3)$$

Having estimated the item characteristics of slots, we finally let the user whose task success ability we want to estimate engage in a dialogue for a given dialogue goal, judge whether each slot is correctly filled, and estimate the task success ability by expected a posteriori estimation based on Bayesian statistics (Fox, 2010). We can calculate the probabilities of correct answers to the slots by using Equation (1) with the estimated task success ability and item characteristics.

5 Experiment

We collected dialogue data and estimated users' task success abilities using IRT. We then evaluated the accuracy of estimating the probabilities of correct answers to slots utilizing the users' estimated task success abilities. Assuming that the capability to fill slots correctly corresponds to the ability to complete dialogue tasks, if the proposed method can accurately estimate the probability of a correct answer to each slot, we can say that the method can accurately estimate the user's task success ability. We also analyzed the estimated item characteristics.

5.1 Dialogue Systems

We built the systems using the MultiWOZ 2.1 dataset (Eric et al., 2020), an English dialogue dataset between a tourist and a clerk at a tourist information center in seven domains: restaurant, hotel, attraction, taxi, train, hospital, and police.

Since item characteristics may differ depending on the system configuration, we used two dialogue systems: a pipeline (Zhang et al., 2020), which consists of four modules, and SimpleTOD (Hosseini-Asl et al., 2020), an end-to-end system.

The pipeline system consists of a natural language understanding module based on BERT (Devlin et al., 2019), a rule-based dialogue state tracking module, a rule-based policy module (Schatzmann et al., 2007), and a template-based natural language generation module. To construct the pipeline system, we utilized the ConvLab-2 toolkit (Zhu et al., 2020; Liu et al., 2021), which enables the development of task-oriented dialogue systems. The architecture of the pipeline system may seem conventional; however, it outperforms other systems implemented by ConvLab-2 in task success. SimpleTOD is a GPT2-driven language model fine-tuned for MultiWOZ dialogues. We trained the model using the public source code on GitHub[2].

[2]https://github.com/salesforce/simpletod/

	Pipeline	SimpleTOD
No. of users	179	198
No. of dialogues	537	594
No. of utterances	24,340	20,532
No. of tokens	311,043	233,760
Task success rate	47.5%	28.3%
Slot correct rate	77.6%	62.0%

Table 1: Statistics of collected dialogues.

Appendix A provides the details of the training settings.

5.2 Experimental Procedure

First, we collected dialogues through Amazon Mechanical Turk, a crowdsourcing platform. We presented different randomly generated dialogue goals, including 2 through 4 domains containing 10 through 20 slots, to 377 workers and engaged them in dialogue with the systems. Each worker was presented with a randomly generated dialogue goal and engaged in three consecutive dialogues with the same dialogue system, either pipeline or SimpleTOD, but with different dialogue goals. The experiment was approved by the ethical review committee of our organization.

The statistics of the collected dialogues are shown in Table 1. We used NLTK[3], a Python library, for counting the number of tokens. As we can see, the dialogues of the pipeline system have a moderate success rate (47.5%), whereas those of SimpleTOD are lower (28.3%), as expected from (Zhu et al., 2020).

We utilized 5-fold cross-validation to evaluate the results. We selected one fold as test data and the remaining four as training data. We made sure there was no overlap of users between the folds. First, we estimated item characteristics using IRT for each slot in the training data. For this purpose, we utilized the GIRTH library[4], a Python library for IRT. Then, using the estimated item characteristics from the training data and the estimated user's task success abilities from the first dialogue of the test data, we predicted the probabilities of correct answers for each slot in the second and third dialogues of the test data. This process was repeated for each fold.

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

[4]https://github.com/eribean/girth

	2nd dialogue	3rd dialogue
Proposed	**0.732**	**0.736**
Baseline (Slot)	0.704	0.703
Baseline (User)	0.678	0.690

Table 2: Accuracy of estimating probabilities of correct answers (pipeline).

	2nd dialogue	3rd dialogue
Proposed	**0.606**	**0.603**
Baseline (Slot)	0.582	0.575
Baseline (User)	0.561	0.577

Table 3: Accuracy of estimating probabilities of correct answers (SimpleTOD).

5.3 Baselines

We prepared two baselines with different approaches for estimating probabilities of correct answers to the slots.

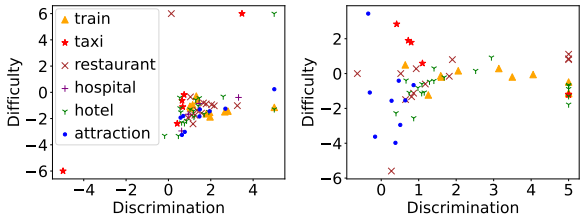
Baseline (Slot) A method that uses the average probability of a correct answer for a target slot as the probability of a correct answer for the slot. That is, the probability of a correct answer to slot s over all users in the training data is used for the probability for slot s for users in the test data.

Baseline (User) A method that uses the average probability of a correct answer from the target user in the test data's first dialogue as the probability of a correct answer for the slot. That is, the probability of a correct answer to slot s is the averaged probability of a correct answer to all slots of that user in previous dialogues.

5.4 Evaluation Metrics

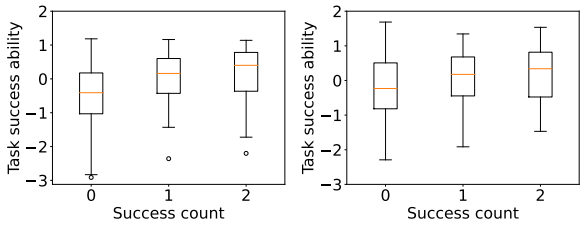
We set the accuracy of estimating the probabilities of correct answers as the evaluation metric. This is equivalent to the average estimation accuracy when performing infinite trials that involve predicting the correctness of each slot as correct with the estimated probability of a correct answer. Specifically, if the estimated probability of a correct answer is denoted as $prob$, and the actual correctness of the user is denoted as $ans \in \{0, 1\}$, then the accuracy of estimating the probabilities of the correct answers is the average for all slots, where each slot's accuracy is calculated by:

$$acc = \begin{cases} prob & (ans = 1) \\ 1 - prob & (ans = 0) \end{cases} \quad (4)$$



(a) Pipeline (b) SimpleTOD

Figure 3: Distribution of discrimination and difficulty estimated for all slots.



(a) Pipeline (b) SimpleTOD

Figure 4: Relationships between estimated users' task success ability from first dialogue and total number of users' task successes (success count) in second and third dialogues.

In calculating acc, we do not distinguish inform and request slots.

5.5 Results

Tables 2 and 3 show the results for the pipeline system and the SimpleTOD system, respectively. Wilcoxon signed-rank tests with Bonferroni correction revealed that the proposed method achieved a significantly higher estimation accuracy than the other methods ($p < .01$).

Comparing the results for the second and third dialogues, we found almost no difference in estimation accuracy for all methods, indicating that the nature of the dialogue does not significantly differ between the second and third dialogues. Note that, since imbalanced data with more correct answers than incorrect ones (Table 1) lead to higher accuracy, we cannot compare the absolute score of the accuracy between the pipeline and the SimpleTOD system. Appendix B provides examples of dialogues between users and the pipeline system and the users' estimated task success abilities.

5.6 Analysis of Item Characteristics of Slots

Figure 3 shows the distribution of the discrimination and difficulty of the slots. In both systems, almost all slots exhibited discrimination values greater than 0 and had the power to estimate the user's task success ability. While the pipeline system showed minimal differences in discrimination and difficulty among slots, the SimpleTOD system revealed substantial variations in discrimination and difficulty across slots, making it possible to appropriately select slots with high discrimination for appropriate testing.

5.7 Analysis of User Abilities

Figure 4 shows the relationships between the estimated users' task success ability from the first dialogue and the total number of users' task successes (success count) in the second and third dialogues. In both systems, users who achieved their tasks tended to have higher task success abilities, indicating that the estimated abilities represent users' task success abilities appropriately.

6 Conclusion and Future Work

We proposed a method for estimating users' task success abilities with task-oriented dialogue systems utilizing item response theory. Experiments on predicting the probability of a correct answer for each slot showed that the proposed method significantly outperformed the baselines.

Various challenges need to be addressed in future work, such as the dependence among slots; to this end, we want to explore methods that consider multiple slots as a single problem. We also want to estimate the task success ability using deep learning-based IRT methods that may achieve higher accuracy (Yeung, 2019; Tsutsumi et al., 2021). Additionally, we aim to investigate methods for estimating task success abilities more quickly, that is, using less than a single dialogue. Finally, we want to construct dialogue systems that can adapt their behavior on the basis of the users' estimated task success abilities.

Acknowledgments

Funding was provided by a Grant-in-Aid for Scientific Research (Grant no. JP19H05692).

References

R Darrell Bock and Murray Aitkin. 1981. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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

Jean-Paul Fox. 2010. *Bayesian item response modeling: Theory and applications*. Springer.

Arin Ghazarian and S Majid Noorhosseini. 2010. Automatic detection of users' skill levels using high-frequency user interface events. *User Modeling and User-Adapted Interaction*, 20:109–146.

Michael R. Harwell, Frank B. Baker, and Michael Zwarts. 1988. Item Parameter Estimation Via Marginal Maximum Likelihood and an EM Algorithm: A Didactic. *Journal of Educational Statistics*, 13(3):243–271.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A Simple Language Model for Task-Oriented Dialogue. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191.

Vojtěch Hudeček and Ondřej Dušek. 2023. Are LLMs All You Need for Task-Oriented Dialogue? *arXiv preprint arXiv:2304.06556*.

Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G. Okuno. 2003. User Modeling in Spoken Dialogue Systems for Flexible Guidance Generation. In *Proceedings of 8th European Conference on Speech Communication and Technology*, pages 745–748.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an Evaluation Scale using Item Response Theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657.

Jiexi Liu, Ryuichi Takanobu, Jiaxin Wen, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, and Minlie Huang. 2021. Robustness Testing of Language Understanding in Task-Oriented Dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2467–2480.

Jia-Jiunn Lo, Ya-Chen Chan, and Shiou-Wen Yeh. 2012. Designing an adaptive web-based learning system based on students' cognitive styles identified online. *Computers & Education*, 58(1):209–222.

Frederic M Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Routledge.

Chelsea M. Myers, David Grethlein, Anushay Furqan, Santiago Ontañón, and Jichen Zhu. 2019. Modeling Behavior Patterns with an Unfamiliar Voice User Interface. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, page 196–200.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152.

João Sedoc and Lyle Ungar. 2020. Item Response Theory for Efficient Human Evaluation of Chatbots. In *Proceedings of the 1st Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33.

Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. 2020. Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 297–310.

Emiko Tsutsumi, Ryo Kinoshita, and Maomi Ueno. 2021. Deep-IRT with Independent Student and Item Networks. *International Educational Data Mining Society*, pages 510–517.

Nigel Ward and Satoshi Nakagawa. 2002. Automatic User-Adaptive Speaking Rate Selection for Information Delivery. In *Proceedings of 7th International Conference on Spoken Language Processing*, pages 549–552.

Chun-Kit Yeung. 2019. Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory. *arXiv preprint arXiv:1904.11738*.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang. 2020. ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 142–149.

Speaker	Utterance
	...
User	**Is there a train leaving after 18:15 from Broxbourne to Cambridge on Sunday?**
System	We can book you for the train leaving at 18:15. Would you like to take the train on Sunday? Would you like me to book a train to Cambridge for you?
User	I need tickets for two.
	...

(a) Dialogue with estimated task success ability of 0.195

Speaker	Utterance
	...
User	I want train service
System	When would you like to leave by? Is there a time you would like to get there by? On what day will you be traveling?
User	**11.30**
System	Sorry I do not understand, can you paraphrase?
	...

(b) Dialogue with estimated task success ability of -1.116

Figure 5: Example dialogues from pipeline system with estimated task success abilities.

A Training Settings for SimpleTOD system

As the best hyperparameters for SimpleTOD were unknown, we trained it by using the public source code on GitHub with different hyperparameter values (e.g., a batch size from 2 to 8, a learning rate from $1e-5$ to $1e-4$, and a maximum sequence length from 256 tokens to 1,024 tokens); then, we selected the most optimized model. We further modified the lexicalization rules to ensure the legibility of the generated system responses.

B Examples

Figure 5 presents examples of dialogues between users and the pipeline system. The user's estimated task success ability for the dialogue in (a) is 0.195, while that for the dialogue in (b) is -1.116. In the dialogue shown in (a), the system responds appropriately to the user's utterance, indicating that the user understands what to say to the system. Specifically, when the user conveys their preferred departure time for the train to the system, they provide the information in a complete sentence rather than just a single word, thus enabling the system to understand the user's intent. In contrast, in the dialogue shown in (b), the user provides only a single word to convey the desired time for the train, and the system fails to understand the user's intent.

An Open-Domain Avatar Chatbot by Exploiting a Large Language Model

Takato Yamazaki, Tomoya Mizumoto, Katsumasa Yoshikawa, Masaya Ohagi, Toshiki Kawamoto, Toshinori Sato

LINE Corporation

{takato.yamazaki, tomoya.mizumoto}@linecorp.com

Abstract

With the ambition to create avatars capable of human-level casual conversation, we developed an open-domain avatar chatbot, situated in a virtual reality environment, that employs a large language model (LLM). Introducing the LLM posed several challenges for multimodal integration, such as developing techniques to align diverse outputs and avatar control, as well as addressing the issue of slow generation speed. To address these challenges, we integrated various external modules into our system. Our system is based on the award-winning model from the Dialogue System Live Competition 5. Through this work, we hope to stimulate discussions within the research community about the potential and challenges of multimodal dialogue systems enhanced with LLMs.

Figure 1: Sample interactions between users and our avatar chatbot, translated from Japanese. The avatar, accessible via VR headset or display, exhibits its emotions through motions and expressive facial cues.

1 Introduction

We present a demonstration of an open-domain avatar dialogue system that we have developed, with the goal of facilitating natural, human-like conversations. With the advent of large language model (LLM) technologies such as LLaMA (Touvron et al., 2023) and ChatGPT (OpenAI, 2022), the fluency of text-based dialogue systems has significantly improved. One of the next directions in this field involves dialogue systems that utilize voice, facial expressions, and gestures through an avatar, contributing to a more engaging and interactive conversation experience (Hyde et al., 2015).

As part of the efforts in dialogue system research, the Dialogue System Live Competition 5 (DSLC5) was held in Japan, a competition of avatar chat dialogue systems (Higashinaka et al., 2022). It was hosted within the academic conference of dialogue systems, where a large number of researchers evaluate the demonstrations performed on the stage to determine their ranking. We developed a dialogue system based on the LLM for this competition (Yamazaki et al., 2022), and encountered a variety of challenges on integrating LLMs into a multimodal dialogue system. For instance, due to the real-time nature of spoken dialogue, it is essential to return some form of response quickly, which is a challenge when using computationally intensive LLMs. Furthermore, when the system involves an avatar, methods of controlling the avatar's facial expressions and motions present another challenge.

We strive to address such missing capabilities of an LLM-based dialogue system by integrating several external modules. Such modules encompass the incorporation of filler phrases and thinking motions during LLM's computational time, task parallelization to speed up responses, and detection of errors in speech recognition. Simultaneously, we paid close attention to the content of dialogue, aspiring to create a system that allows users to engage in deep, prolonged, and safe interactions. As a result, our system achieved the best human evaluation results in the competition. However, among

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 428–432
September 11–15, 2023. ©2023 Association for Computational Linguistics

the metrics, the naturalness of avatar's speaking style received the lowest score, indicating a need for improvement on its motions and expressions.

In this demonstration, we present an avatar dialogue system that improves from DSLC5. The improvement includes addition of emotion recognizer to enhance naturalness of the avatar expressions and motions. Additionally, aiming to provide a more immersive dialogue experience, we offer a system that allows conversation with an avatar through a virtual reality (VR) headset. Although the system is originally designed to respond in Japanese, we provide translated responses for English speakers. Through this demo, we hope to stimulate discussion within the research community about the potential and challenges of integrating LLMs into multimodal dialogue systems.

2 System Overview

We first provide a overview of the features of our proposed system, followed by a more detailed explanation in the subsequent subsections. The system architecture is illustrated in Figure 2.

Initially, the user's vocal input is transcribed into text utilizing a speech-to-text (STT) module. As the STT results frequently lack punctuation, a module called the *Punctuationizer* is utilized to append period marks and question marks. The punctuated text is then fed into the *Dialogue System*.

The Dialogue System leverages an LLM to generate responses. To elicit more engaging responses from the LLM, a process termed *Prompt Creation* is performed beforehand, providing the model with contextually rich information. Following the response generation, an *Editing & Filtering* phase is undertaken. During this phase, any responses that are dull or ethically inappropriate are identified and either edited or discarded as necessary.

Once the response text is determined, the system proceeds to control the avatar and text-to-speech system (TTS). The avatar's motions and expressions are decided based on the outcome of the Emotion Analyzer. To help accurately reading Kanji characters of Japanese, we add *Pronunciation Helper* to convert into the phonetic script of Hiragana.

2.1 Dialogue System: Prompt Creation

In the Prompt Creation phase, the system utilizes different types of few-shot prompts based on the given user inputs. All prompts are created based on a common template, which includes instructions such as the system character's profile, the current date and time, and the manner of speech. Here, we introduce a few of the prompts that we employ in our system.

STT Error Recovery Prompt There is a risk of LLM generating unintended responses when it receives user utterances containing STT errors. To mitigate this, we implemented an STT error detector based on fine-tuned BERT (Devlin et al., 2019) with dialogue breakdown detection data (Higashinaka et al., 2016). In cases where errors are detected, the system discards the user input and employs a prompt to inform the LLM about the inaudibility of the received utterance. Figure 3 shows an example shot of this prompt.

Knowledge-Response Prompt In cases where a user engages in a deep conversation on a specific topic, specialized knowledge or the latest information not included in the LLM's parameters might be required. Moreover, it is empirically known that the more niche a topic is, the more likely the LLM is to generate dull responses or only refer to well-known topics. Thus, satisfying users who wish to delve into more core conversations is challenging. To accommodate this, we introduced a search system for topic-related knowledge from online sources (e.g. Wikipedia) and inserts them into the prompt. This prompt is triggered when an effective knowledge source is found through searching the database.

Persona-Response Prompt The input length for the LLM has a limit, and it is currently difficult to incorporate all past dialogues as input. However, maintaining memory of past dialogues is crucial to achieving consistent conversations with the user. Our system is designed to maintain memories during dialogues by storing and utilizing the personas of the user and the system itself. After each utterance by both speakers, persona sentences are obtained using a *Persona Extractor* module which is also implemented with the LLM. These persona sentences are stored in a vector database and utilized in the prompt during the response generation.

2.2 Dialogue System: Editing & Filtering

While the responses generated by the LLM is fluent, they sometimes lead to dull or stagnant conversation, or even prematurely end the dialogue, resulting in a loss of user interest. To circumvent these issues, we have implemented a *Boring Re-*

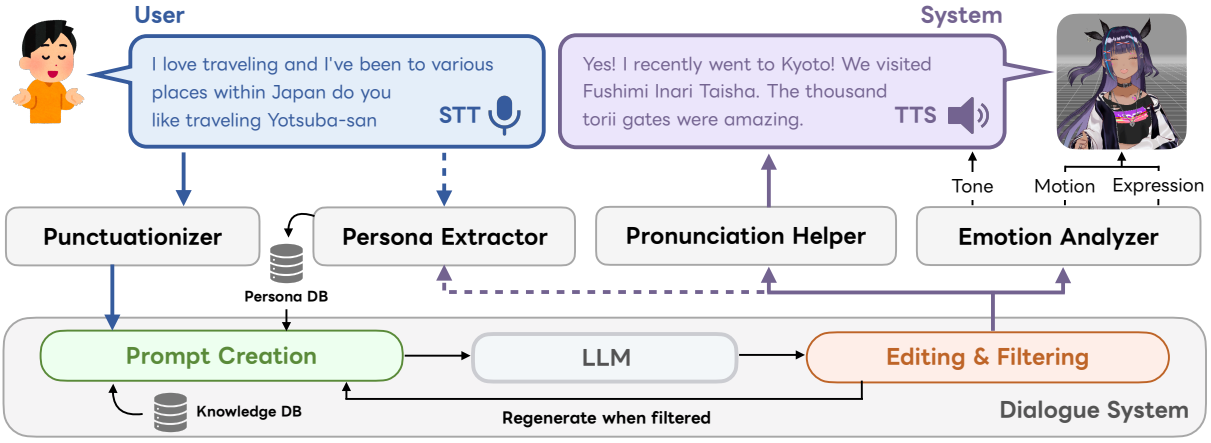


Figure 2: System Overview

```
System: What have you been interested in lately?
User: Well, I've actually started playing various
  games recently.
System: That's exciting! I'm a big fan of games
  myself. What kind of games are you into?
User: ...
System: (Hmm, I didn't quite catch the user's
  response...) Can you tell me about some other
  games you enjoy playing?
```

Figure 3: Example shot of an STT Error Recovery Prompt. The actual response is generated after the parentheses of the last utterance.

sponse Filter and a *Conversation-Closure Filter*. Both filters operate by identifying responses that are similar to manually collected boring and closing expressions. If either of these filters flags a response, the system will either add sentences using the LLM to enrich the content or revert back to the Prompt Creation stage for regeneration.

There's also a risk of generating ethically inappropriate responses, making it unsafe to directly provide the LLM's outputs to the user. In order to achieve communication that is both secure and respectful, we implemented *Toxic-Response Filter*. It utilizes a classifier that has been fine-tuned with BERT using a Japanese harmful expression dataset (Kobayashi et al., 2023). In case where these filters flag a response, the system adds suitable instructions to the prompt (e.g. "respond gently" or "expand on the topic") and again reverts back to the Prompt Creation to regenerate.

2.3 Response Timing

In spoken dialogues, promptly signaling understanding of the user's speech is considered crucial for enabling a comfortable conversation. However, the LLMs necessitate significant computational time, requiring approximately 2 seconds for each generation in case of our system. This overhead can lead to response delays, especially when regeneration is necessary.

To mitigate these potential sources of discomfort, our system employs concurrent operation of multiple modules. For instance, we execute persona extraction while the system is speaking, enabling it to expedite response times. Additionally, we have integrated the use of conversational fillers and animated motions during these waiting periods. By incorporating these elements, we aim to make the delay less noticeable and align with the natural flow of human conversation.

2.4 Emotion Analyzer and Avatar Control

In avatar dialogue systems, it is important to control the tone of the synthesized voice, as well as the avatar's facial expressions and motions, in accordance with the content of the utterance. Our system performs emotion analysis on the responses generated by the Dialogue System to manage these aspects. The analyzer employs a fine-tuned BERT trained on the WRIME dataset (Kajiwara et al., 2021). It recognizes eight types of emotions, namely joy, sadness, anticipation, surprise, anger, fear, disgust, and trust, on a strength scale ranging from 0 to 3.

Our in-house developed TTS system called *Coharis* can assign emotions such as "Happy" and "Sad", and these are mapped with the output of the emotion analyzer. Similarly, the expressions and motions of the avatar is also controlled based on the results of emotion analyzer. The avatar displayed in our demonstration is a sample model provided

by the *VRoid Hub* [1] service from pixiv Inc. This sample model comes with a variety of facial expressions such as "happy," "sad," "angry," and so forth. Moreover, *Mixamo* [2] by Adobe Inc. offers a wide variety of avatar motions, which also is utilized for expressing emotions.

2.5 User Interface

The user interface operates on a web browser. We use the WebSpeechAPI of the browser for STT. The avatar is displayed by controlling WebGL through a library called three.js[3]. We also provide a VR interface using WebXR, allowing it to be displayed through a browser inside a VR headset.

3 Evaluation

We present the results of our original system obtained during the final round of DSLC5. It is important to note that the avatar and TTS used in the competition were provided by the organizers, and thus differ from the demonstration presented here. Furthermore, the evaluated system does not incorporate avatar control based on the emotion analyzer, so the evaluation results should be considered as reference values. For the LLM, we utilized HyperCLOVA 82B (Kim et al., 2021), which was trained on a Japanese corpus.

In the competition, the system engaged in a four-minute casual conversation in Japanese with an avatar about two topics, after which the dialogue was evaluated through a post-conversation survey. The survey was conducted based on four aspects: "Natural: whether the dialogue was natural", "Follow: whether the system could respond appropriately to the topic", "Providing: whether the system could provide new information related to the topic", and "Style: whether the way of speaking, demonstrated through voice, gestures, and facial expressions, was natural". These evaluation criteria were set by the competition organizers as metrics for assessing "human-level conversational ability". A five-point Likert scale was used for the evaluation.

Table 1 presents a comparison of our system with the second-ranked system, which is a dialogue system leveraging a 1B-params language model (Moriya et al., 2022). Our system significantly outperforms the second-ranked system in three metrics: Natural, Follow, and Providing, indicating high performance in dialogue content. As seen in the dialogue examples shown in Table 2, our system successfully follows and expands on topics. However, the Style score is notably lower than the other metrics, indicating the need for further enhancements in terms of avatar motions. We expect improvements with the Emotion Analyzer, as indicated by the displayed output results in Table 2. The emotion labels are accurately assigned, suggesting that they can be applied effectively to the avatar's facial expressions and motions.

	Natural	Follow	Providing	Style	Total
2nd	2.60	2.52	2.87	3.06	2.76
Ours	**4.00**	**4.00**	**3.89**	**3.54**	**3.86**

Table 1: Human evaluation results of the finals of DSLC5. "2nd" indicates the second place team.

Speaker	**Utterance**
User	Oh I see, you had a Djungarian hamster. I have a golden hamster.
System	Golden hamsters are also cute. I can't resist their big eyes. (Joy: 2)
User	Yes, their eyes are really cute.
System	They're so adorable, it makes me want to have one again. By the way, have you named it? (Joy: 1 / Ant.: 1)
User	Its name is Uchiko.
System	Oh, Uchiko, huh!? That's a cool name! How do your hamsters usually spend their time? (Joy: 1 / Sur.: 1)
User	They usually live in their cage.
System	I see, indeed, it would be full of dangers if they went outside. I wonder if they don't get stressed just staying inside the house. (Fear: 3)

Table 2: A sample dialogue translated from Japanese during the preliminary stage, with the output of the Emotion Analyzer in the parentheses. "Ant." and "Sur." corresponds to anticipation and surprise, respectively. The number represents the strength of the emotion (0-3).

4 Conclusion

In conclusion, we developed an open-domain avatar chatbot in a VR environment, leveraging a large language model (LLM). While encountering challenges in multimodal integration, such as addressing slow generation speed and controlling the avatar, our system demonstrated promising results in Dialogue System Live Competition 5. Addition-

[1] https://hub.vroid.com/
[2] https://www.mixamo.com/
[3] https://threejs.org/

ally, we attempted to improve the unnaturalness in the avatar's style of speaking, which was discovered after the competition, by using the emotion analyzer. We anticipate that this work will initiate meaningful discussions among the research community regarding the potential and challenges of integrating LLM into multimodal dialogue systems.

References

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

Ryuichiro Higashinaka, Kotaro Funakoshi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets and evaluation metrics. In *Proc. of The Tenth International Conference on Language Resources and Evaluation*.

Ryuichiro Higashinaka, Tetsuro Takahashi, et al. 2022. Dialogue system live competition 5 (in japanese). In *JSAI SIG-SLUD, 96th Meeting*, page 19. The Japanese Society for Artificial Intelligence.

Jennifer Hyde, Elizabeth J Carter, et al. 2015. Using an interactive avatar's facial expressiveness to increase persuasiveness and socialness. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1719–1728.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104, Online. Association for Computational Linguistics.

Boseop Kim, HyoungSeok Kim, et al. 2021. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Koga Kobayashi, Takato Yamazaki, et al. 2023. Proposal and evaluation of japanese harmful expression schema (in japanese). In *Proceedings of the 29th Annual Conference of the Association for Natural Language Processing*. Association for Natural Language Processing.

Shoji Moriya, Daiki Shiono, et al. 2022. aoba_v3 bot: A multi-modal chit-chat dialogue system integrating diverse response generation models and rule-based approaches (in japanese). In *JSAI SIG-SLUD, 96th Meeting*. Japanese Society for Artificial Intelligence.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt.

Hugo Touvron, Thibaut Lavril, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Takato Yamazaki, Toshiki Kawamoto, et al. 2022. An open-domain spoken dialogue system using hyperclova (in japanese). In *JSAI SIG-SLUD, 96th Meeting*. Japanese Society for Artificial Intelligence.

Learning Multimodal Cues of Children's Uncertainty

Qi Cheng[1*], **Mert İnan**[4*], **Rahma Mbarki**[3*], **Grace Grmek**[2*], **Theresa Choi**[1*], **Yiming Sun**[3], **Kimele Persaud**[3], **Jenny Wang**[3], **Malihe Alikhani**[4]
[1] University of Pittsburgh, PA, USA [2] Harvard Medical School, MA, USA
[3] Rutgers University, NJ, USA [4] Northeastearn University, MA, USA
{qic69, tec63}@pitt.edu, ggrmek@mgh.harvard.edu,
{rm1218, kjg117, jinjing.jenny.wang}@rutgers.edu,
{inan.m, alikhani.m}@northeastern.edu

Abstract

Understanding uncertainty plays a critical role in achieving common ground (Clark et al., 1983). This is especially important for multimodal AI systems that collaborate with users to solve a problem or guide the user through a challenging concept. In this work, for the first time, we present a dataset annotated in collaboration with developmental and cognitive psychologists for the purpose of studying nonverbal cues of uncertainty. We then present an analysis of the data, studying different roles of uncertainty and its relationship with task difficulty and performance. Lastly, we present a multimodal machine learning model that can predict uncertainty given a real-time video clip of a participant, which we find improves upon a baseline multimodal transformer model. This work informs research on cognitive coordination between human-human and human-AI and has broad implications for gesture understanding and generation. The anonymized version of our data and code will be publicly available upon the completion of the required consent forms and data sheets.

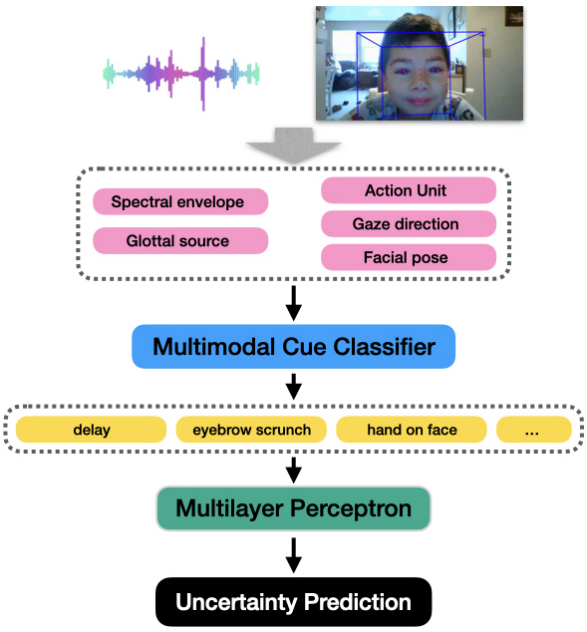


Figure 1: A diagram of our multimodal machine learning model. After identifying uncertainty cues in the multimodal transformer, the model passes the cues onto a final multilayer perceptron classifier to output whether the child is expressing uncertainty or not.

1 Introduction

Recognizing uncertainty in interlocutors plays a crucial role in successful face-to-face communication, and it is critical to achieving common ground (Clark et al., 1983). To accurately identify uncertainty signals, human listeners learn to rely on facial expressions, hand gestures, prosody, or silence. AI systems that aim to collaborate and coordinate with users in a human-like manner also need to understand these signs of uncertainty. To this end, in this paper, we introduce a multimodal, annotated dataset for uncertainty detection in young children.

As a multimodal communicative sign, identifying uncertainty is an important and challenging task for AI systems. Especially because it varies across different ages and demographics; it is sometimes verbalized and sometimes not (Blanco and Sloutsky, 2021); it brings in different modalities, and it is subtle. Although it is critical, uncertainty signal recognition is understudied in younger children. In this work, we study detecting uncertainty in the setting of a counting game for children ages 4-5. We first identify potential cues of uncertainty presented in different modalities (e.g., spontaneous verbal responses, hand gestures, facial expressions, hesitation) and specifically examine the relationship between task difficulty, task performance, and exhibited levels of uncertainty. We then use these cues to inform an ensemble model, which first identifies these cues from multimodal data and then uses them to predict uncertainty (Figure 1).

This work informs research on cognitive coordination between human-human and human-AI collaboration. With this paper, we contribute an anno-

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 433–443
September 11–15, 2023. ©2023 Association for Computational Linguistics

tated multimodal dataset of uncertainty in children (Section 3 provides details about the dataset, annotation protocol, and analyses of the dataset); we analyze the performance of multimodal transformer models in identifying uncertainty on this dataset (Sections 4 and 5); and finally we present a case study on how children express uncertainty based on their age.

2 Related Work

We cover related works of uncertainty in two sections: datasets and protocols for studying uncertainty in children, the Approximate Number System, and uncertainty in human-AI interactions.

2.1 Datasets and Protocols for Studying Uncertainty in Children

Adults are generally more direct and communicate their uncertainty via explicit verbal cues. Children, however, lack this insight into their own uncertainty, making uncertainty detection more difficult from an outsider's perspective. As such, detecting uncertainty in children remains a complex problem.

What has been established, however, is that children consistently communicate their uncertainty through the use of various facial, auditory, and gestural cues. For example, Harris et al. (2017) found that children are very expressive when they are uncertain. In the presence of an adult, these expressions may be communicated via hand flips, questions, and utterances, such as "I don't know." However, when children are alone, these same signals can be representative of signals of uncertainty. In the past, researchers have attempted to codify behaviors associated with communicating uncertainty by parsing through these various cues and creating annotation protocols.

Previously, researchers Swerts and Krahmer (2005) aimed to detect uncertainty in audiovisual speech by coding for different audiovisual cues in both adults and children. Their protocol consisted of audio cues (e.g. speech fillers and speech delays) and facial movements (e.g. eyebrow movement and smiling). While the protocol included both audio and visual cues, the cues that were noted were limited. Another protocol developed by Mori and Pell (2019) studied solely visual cues signaling uncertainty in speech communication. These cues included changes in gaze direction, facial expressions, and embarrassed expressions. An additional protocol developed by Ricci Bitti et al. (2014) studied uncertainty through facial expression entirely.

However, while these protocols are indeed useful, they lack the specificity necessary for our goal of pinpointing various multimodal cues associated with uncertainty. There are other various protocols, but they are also limited, typically adhering to one modality. Consequently, we expanded upon these existing protocols and included other multimodal behaviors grounded in developmental and cognitive psychology, and presented multimodal machine learning models that can predict these cues and their association with uncertainty.

Children have an intuitive sense of numbers relying on the Approximate Number System (ANS). The ANS obeys Weber's Law, where one's ability to differentiate between two quantities depends on the ratios of those quantities (Dehaene, 2011; Odic and Starr, 2018). The smaller the ratio, the more difficult it is to discriminate between quantities and the more uncertainty there is in the participants' internal representations. Previous research showed that children perform better on a numerical comparison task when given a scaffolded, Easy-First numerical task starting with easier trials (e.g., 10 vs. 5) and progressing to harder ones (e.g., 10 vs. 9), compared to children seeing the same exact trials in the reversed order (i.e., Hard-First), an effect termed "confidence hysteresis" (Odic et al., 2014). This implies that confidence is built by gradually working up to harder tasks, resulting in better performance, whereas starting out with more difficult tasks reduces confidence, resulting in worse performance.

Due to its effectiveness at generating confidence or lack thereof in participants, such a numerical comparison task would be the ideal method for measuring behaviors associated with uncertainty. As such, the present study aims to fulfill this objective by implementing this "confidence hysteresis" paradigm into the task children are given.

2.2 Studying Uncertainty in Human-AI Interaction

Multimodal models have been shown to improve performance on certain tasks by grounding some aspects of the human condition with features beyond text. Leveraging multiple modalities is particularly applicable in cases where text may miss key insights, such as sarcasm detection (Castro et al., 2019), depression prediction (Morales et al., 2018), sentiment detection (Yang et al., 2021), emotion

recognition (Morency et al., 2011), and persuasiveness prediction (Santos et al., 2016). Tasks involving such complex labels benefit from multiple modalities due to the richness of the data streams. In addition to understanding what is said, understanding how it is said (pitch, facial expression, body language, gesture) is crucial (Beinborn et al., 2018).

There have been attempts at predicting uncertainty from the audio through prosodic features. Dral et al. (2011) reported that prosodic features were successful in detecting speaker uncertainty in spoken dialogue with a 75% accuracy. Pon-Barry and Shieber (2011) had similar findings with prosodic features, and self-reported states of certainty and perceived states have strong mismatches. In this paper, we address this by controlling task difficulty to affect a participant's level of observed uncertainty. A seminal study on the understanding and generation of multimodal uncertainty cues exists by Stone and Oh (2008). Here the authors analyze adult human-human conversations for uncertainty cues and try to replicate them using avatars. Our experimentation and modeling efforts, on the other hand, are focused on the domain of uncertainty detection in younger children.

3 Data

Participants A group of 68 children between the ages of 4 and 5 years old (M_{age} = 5;0; SD_{age} = 6.88 months; 28 females) was recruited through Lookit, an online platform for developmental studies (Scott and Schulz, 2017). Thirty-six parents identified their child as White, six as Asian, three as Hispanic, Latino, or Spanish origins, and the rest as multi-racial. All but three parents reported having a college degree or higher level of education. After completing the study, compensation was sent in the form of a $5 gift card. Each child participated in 30 trials which are, on average, 8 seconds long. In total, our data is composed of 16,320 seconds of video data.

Task Participants were given an Approximate Number System manipulation task adapted from Wang et al. (2021) designed to impact children's certainty about numerical quantities. Children were presented with two arrays of dots paired with two cartoon characters (Figure 2) and asked to guess which character has more dots.

Both characters and their array of dots appeared for 2500 ms before disappearing. This short display

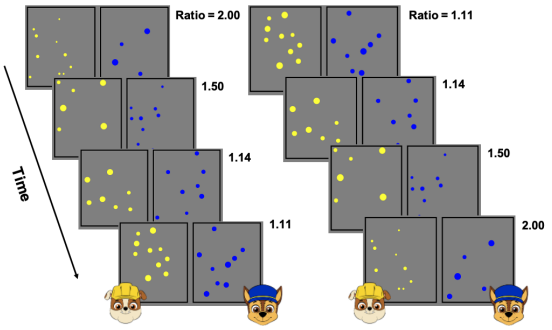


Figure 2: Schematic of experimental procedure depicting the Easy-First condition on the left and the Hard-First condition on the right. As time progresses throughout the task, the trials advance from easier ratios (2.0) to hard ratios (1.11) in the Easy-First condition. Whereas in the Hard-First condition, trials move in reverse order from hard ratios (1.11) to easy ratios (2.0) as time progresses.

duration was chosen to ensure that children did not have sufficient time to count. Children were then asked to click on the side of the screen showing the greater number of dots. Children were given immediate audio feedback for each trial once they chose their response.

Children completed 30 trials with the following number pairs: 10:9 dots (1.11 ratio), 8:7 (1.25 ratio), 14:12 (1.17 ratio), 10:8 (1.13 ratio), 9:6 (1.5 ratio), and 10:5 (2 ratio). In half of the trials, arrays with more dots had a greater, congruent cumulative area. In the other half of the trials, arrays with the greater number of dots had a smaller, incongruent cumulative area.

Children were randomly assigned to either the Easy-First or Hard-First conditions. In the Easy-First condition, trials advanced from the easier trials (e.g., 10:5) to the harder trials (e.g., 10:9) in a staircase order following the design of Wang et al. (2021). Whereas in the Hard-First condition, trials move in reverse order from hard ratios (e.g., 10:9) to easy ratios (e.g., 10:5).

Annotation Procedure Annotators first watched the video muted so as not to be influenced by the vocal feedback from the task since whether or not the child answered right or wrong may lead them to over/under-interpret certain cues. During this first watch, they marked all present physical cues as indicated by the protocol. On their second watch, annotators unmuted the video, and marked all verbal cues. If the cue was not present, the corresponding

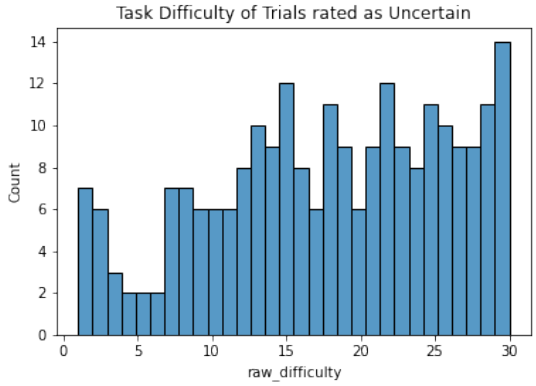


Figure 3: The distribution of uncertain trials with task difficulty on a scale of 1 (easiest) to 30 (hardest). Uncertainty shows a strong correlation with task difficulty ($r(58) = -.927, p < .01$).

cell was left empty.

3.1 Annotation Protocol

In collaboration with a team of developmental psychologists and cognitive psychologists, we have collected and designed a protocol that aims to study uncertainty, particularly expressed non-verbally. Through pilot studies observing and annotating our data, we iteratively defined our protocol and constructed a list of signals we observed as signs of uncertainty. Our protocol spans multiple modalities and includes facial, gestural, and auditory cues to account for a broad spectrum of possible behaviors. The protocol can be found in Table 1 with supplemental example images in Figure 4. The Rutgers University Institutional Review Board approved the research, and all parents of this study's children provided verbal consent before their children's participation. However, only some of the parents agreed to allow their children's video and voice recordings to be shared publicly.

3.2 Analysis

In this section, we provide an analysis of our annotated data, identify any significant cues that contribute to detecting uncertainty, and explore when different cues occur.

When were children annotated as uncertain? The annotations are split 13.8/5.3/80.9 between the labels *uncertain/unclear/non-uncertain*. Of all the annotated trials, 79.3% were correct, of which 82.4% were rated as having no uncertainty. In other words, most trials are within the children's capability and confidence. While a significant class imbalance exists between uncertain and non-uncertain trials, the distribution is realistic.

Are uncertainty and task difficulty related? As shown in Figure 3, uncertainty was found to be highly correlated with task difficulty, $r(58) = -.927, p < .01$. Both ratio of dot size and size control factor into the difficulty of a trial (with a smaller ratio and the presence of size control both indicating a harder trial).

Are uncertainty and task performance related? Despite the high correlation between uncertainty and task difficulty, there was no substantial correlation between uncertainty and task correctness, $r(58) = .290, p > .4$. This makes sense as the accuracy and ease of a task are not necessarily intertwined; a participant may make a mistake on an easy trial or get lucky on a difficult trial.

How do demographics affect uncertainty? We analyze participant demographics in terms of age and gender. Regarding the frequency of expressing uncertainty, we found that the average participant age of the uncertain trials is younger than that of the non-uncertain trials, though they are not significantly different. Similarly, female participants have slightly more uncertain trials, and male participants have slightly more non-uncertain trials, but the results are insignificant.

However, we did find gender differences in the types of cues used to express uncertainty. Female participants exhibited more of the *filled pause* cue, while male participants exhibited more of the *smile* and *shoulder movement* cues. The full table of comparisons can be found in Appendix Table 5.

Which cues occur the most? The percentage in which each cue appears in all uncertain trials can be found in Figure 5. We can see that in general, *hand on face* and *smile* are the most common cues, appearing in 17% and 12% of all trials, respectively.

Notably, during uncertain trials, while *hand on face* and *smile* remain common, other cues also appear more frequently. In particular, *eyebrow scrunch, eyebrow raise* and *delay* are now equally, if not more common, appearing in 22% and 17% of all uncertain trials. This is promising, as cues frequently appearing in uncertain trials but not so common throughout all trials can be valuable indicators of uncertainty.

Which cues occur in difficult trials as opposed to easy trials? The percentage that each cue appears in hard and easy trials can also be found in Figure 5. Hard

Cue	Description
Delay	The participant delayed their decision-making with a pronounced pause
Eyebrow raise	The participant raised their eyebrows
Eyebrow scrunch	The participant markedly scrunched their eyebrows or squinted their eyes
Filled pause	Utterances such as "umm," "hmm," or "uh."
Frustrated noise	Sounds of verbal frustration, such as sighing, groaning, and growling
Funny face	The participant grimaced or made an unconventional facial expression
Hand on face	Any kind of movement that includes a participant putting a hand on their face
Head tilt	The participant tilted their head to either side while making their decision
Look away	The participant was distracted and not paying attention to the task
Look to adult	The participant looked towards their parent when making their decision
Shoulder movement	The participant made a pronounced shoulder movement, such as shrugging
Smile	The participant smiled
Verbal cues	Any spoken words

Table 1: Categories in our annotation protocol.

Figure 4: Examples of (from left to right) *eyebrow raise, eyebrow scrunch, hand on face, funny face, and smile*

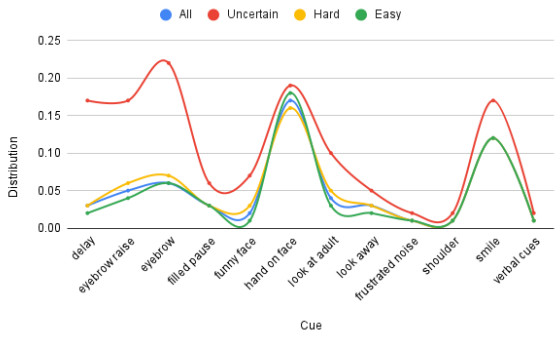


Figure 5: Distribution of uncertainty cues across all/uncertain, difficult/easy. We can see that *delay, eyebrow raise,* and *eyebrow scrunch* are significantly more frequent in uncertain trials.

trials are defined as half of the trials with a more difficult ratio (1.11 to 1.17), and easy trials are half with ratios of 1.25 to 2.

We find that if the trial is hard, the participant is slightly more likely to exhibit more of our studied cues overall. In particular, the participant is likelier to exhibit the *look at adult* or *funny face* cues. If the trial is easy, the participant might display *hand on face* instead.

This shows support for the potential to differentiate between stages of uncertainty. Namely, if a child or student expresses uncertainty at a more manageable task, this could be out of a lack of confidence *(I'm generally familiar with this and have an idea on how to do it, but I need a little help.)* or another factor that may entail minor assistance. Meanwhile, facing a more challenging or perhaps even a completely new task, they may feel a more difficult uncertainty *(I don't know where to start.)* requiring more involved guidance.

This possible distinction in stages of uncertainty may open the door for a more precise intervention in the context of education and tutoring systems. For instance, if notified about a student exhibiting the former uncertainty, the teacher might engage with small hints and encouragement to maximize the student's learning. However, if a student shows the latter uncertainty, the teacher can offer hands-on guidance, such as checking foundational concepts.

4 Approach

With the goal of predicting uncertainty from multimodal signals, we conducted experiments with three approaches: learning uncertainty from pro-

posed cues, a multimodal transformer-based model, and an ensemble learning approach that first predicts cues from the multimodal features and then predicts uncertainty from those cues.

4.1 Multimodal Features

We take facial action units, gaze direction, and facial pose from the OpenFace toolkit (Baltrusaitis et al., 2018) for video features. Action units (AUs) are coded for facial muscle movements, which indicate various facial expressions (Tian et al., 2001). For audio features, we extracted glottal source and spectral envelope features using Degottex (2014)(v1.4.2). For text features, we then passed GloVe embeddings of each trial's annotated transcription to the model (Pennington et al., 2014). It should be noted that as a task that does not ask the participant to speak, most trials contain no text.

4.2 MulT Model

Given the cost of annotation, an ideal uncertainty prediction system would take multimodal data of the participant as features and be able to make real-time predictions on the participant's level of uncertainty. To test this goal, we first experiment with an end-to-end model. Specifically, we use the Multimodal Transformer proposed in Tsai et al. (2019) on audio, video, and text data from videos of the participants. This model is uncertainty cue-agnostic, as it contains no information about our annotated cue categories.

4.3 Contrastive Learning

Our dataset has high-dimensional features: 710 dimensionalities for each video frame, 71 for each second of corresponding audio, and 30 for each word in the corresponding text. Training a prediction model end-to-end in a high-dimensional feature space focuses on local differences in the latent space instead of the global relationships between classes. We also tested a contrastive learning procedure with a custom loss function to overcome this challenge. This method encourages the model to learn representations that are close for positive pairs and far apart for negative pairs and better discriminates between different classes. The details of the loss function and our weighted sampling strategy are given in Appendix B.1.

4.4 Annotation-based Ensemble learning

We further propose an ensemble learning approach illustrated in Figure 6, that first predicts each of the annotator cues found to be significantly correlated to the annotator prediction of uncertainty and then predicts uncertainty using the trained classifier. We compare this proposed model to the previous end-to-end multimodal transformer and the unimodal transformers. We choose only to predict the cues that were used in the decision tree classifier (i.e., *delay, eyebrow raise, eyebrow scrunch, look at adult*, and *hand on face*). **Multimodal transformer model** is used as the classifier for uncertainty.

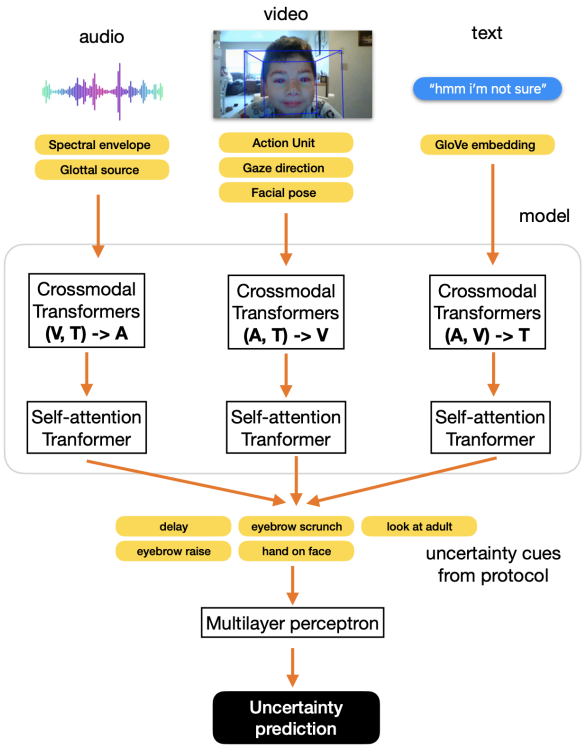


Figure 6: This figure shows the architecture of the ensemble learning model. First, cross-modal transformers learn the attention across the features of each modality with each of the other two modalities' low-level features. Then, using the fused features, self-attention transformers predict the present uncertainty cues, which are then passed into a multilayer perceptron to output the final prediction of whether or not uncertainty is present.

5 Experimental Evaluation

In order to determine the viability of a multimodal uncertainty prediction model, we detail the results from each of our computational experiments and models. The implementation details of the experiments and models are given in the Appendix C.

Baselines We employed two baseline models for comparison. The first is a simple multimodal neural network that separately processes video, au-

	Model	F1	MAE	R^2
No Cue	Basic	*.7011*	*.3603*	*-.2419*
	Adult	*.6397*	*.8553*	*-2.7119*
	MulT	.8120	**.2307**	**-.1391**
	CL + W	**.8216**	.3076	-.5412
Cue	Ensemble	**.8366**	**.2222**	-.1250

Table 2: This table shows the results of different models. The "Basic" refers to the MLP baseline, and the "Adult" refers to the adult uncertainty baseline. "MulT" refers to the Multimodal Transformer, and "CL + W" is the MulT model with contrastive learning and weighted sampling. "Ensemble" refers to the cue-based ensemble model. No Cue and Cue indicate whether the model uses the identified cues as intermediate features.

dio, and text inputs and combines the features for a three-class softmax classification. The second is a detection model trained with adult data that takes visual information (Jahoda et al., 2018). This second baseline is a traditional machine-learning approach using SVMs and LBP descriptors.

Metrics In our experiment, we utilized three key metrics to evaluate the performance of our model. The weighted F1 score was employed to account for any class imbalance and provide a more comprehensive assessment of the model's precision and recall. Mean Absolute Error (MAE) was used to measure the average magnitude of the errors in our predictions, illustrating the model's ability to minimize deviations from the actual values. Lastly, the R-square statistic was employed to quantify the proportion of variance in the dependent variable explained by the model, offering insight into the overall goodness of fit and the model's explanatory power.

5.1 Results

We report a weighted F1 score of .8216 and a mean absolute error of .3076 on the cue-agnostic end-to-end model with reweighted class labels, as seen in Table 2. Full results for each modality can also be found in the same table. The cue-aware ensemble model shows improvements in both weighted F1 and MAE over the multimodal transformer model. Contrastive learning and weighted sampling improve the performance but are subpar compared to the cue-based ensemble method. The intermediate prediction of cues like *delay* that are a vital indicator but may be challenging to learn in an end-to-end model may play a role in this performance.

Modality Ablations After doing an ablation study on the modalities for the cue-agnostic models, we find that the text and audio modality report the best scores overall, as shown in 3. This is unexpected due to participant speech being scarce. However, when participants do talk, they usually express their feelings about the task. For instance, participants may say "This is easy!" or "I don't know," tell the adult if the trial is hard or begin counting. As a result, the text modality could be less noisy than the video and audio modalities. We note that the particular task does not request verbal responses from the participants. Thus, we expect that with a task that entices a verbal response, such as question answering, there may be more contribution from the text and audio modalities.

Model	F1	MAE	R^2
CL + W	**.8216**	.3076	-.5412
Video only	.7991	.2820	-.4072
Text only	*.8056*	*.2564*	*-.2731*
Audio only	*.8056*	*.2564*	*-.2731*

Table 3: F1, MAE, and R^2 results for the best performing cue-agnostic model (weighted) with the ablation studies for all the modalities. The model performs the best with all the components, but the most influential modality is text/audio.

6 Conventions of Expressing Uncertainty: A Case Study in Different Age Groups

From our annotated data, we find that for older children ($>$ 2150 days old), less parental guidance is present, faster decision-making is observed, less diverse facial expressions are present, and more verbal cues are present while expressing uncertainty, which increases the performance of the models for 5-year-olds Table 4. In addition, certain behavior patterns that children of different age groups display are context-dependent, convention-oriented, and personality-specific, making it difficult to identify only through visual and textual modalities. Some of these behaviors that we investigate here are nail-biting, pointing, and social facilitation (see Figure 7).

In the developmental psychology literature, nail-biting is either found to be an acquired habit or related to states of nervousness (Gilleard et al., 1988; Silber and Haynes, 1992; Ghanizadeh, 2008; Mc-

Model	4 year old F1	4 year old MAE	4 year old R^2	5 year old F1	5 year old MAE	5 year old R^2
Basic	*.69*	*.32*	*-.24*	*.73*	*.33*	*-.21*
Adult	*.57*	*1.06*	*-4.40*	*.62*	*1.0*	*-8.13*
MulT	.78	.26	-.39	.81	.26	-.27
CL + W	**.79**	**.22**	**-.15**	**.81**	**.23**	**-.14**

Table 4: This table shows the results of the models between different age groups. There are slight inference differences between the 4 and 5-year-old groups. These performance changes are dependent on the conventions of uncertainty and age.

Clanahan, 1995; Wells et al., 1998). This behavior is hard to classify as stress-induced or uncertainty-induced. Hence, a context-dependent analysis of the person using skeletal features can help decide.

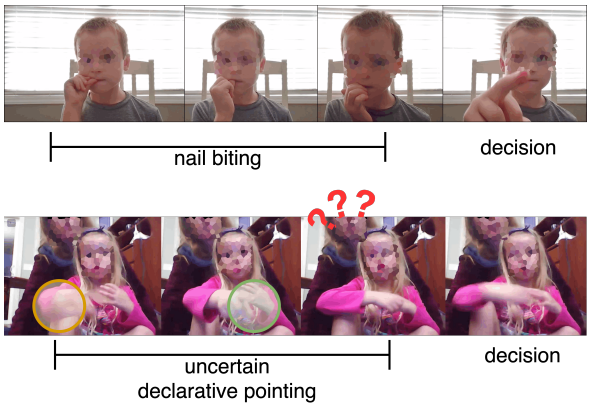


Figure 7: This figure shows two complex behavior patterns by children: nail-biting and uncertain declarative pointing. The top sequence belongs to the oldest male, and the bottom sequence belongs to the youngest female.

Pointing (see Figure 7) is another context-dependent occurrence. When the child is uncertain, the pointing to options also becomes ambiguous, and the parent needs to ask a follow-up grounding clarification question, such as "Which one?". This behavior pattern involves spatial placement of options and understanding the boundaries between them. Younger children prefer ambiguous pointing gestures to conventional and visible cues of uncertainty. This type of declarative pointing is observed to be a way of engaging with the parent, pointing to a theory of mind (ToM) understanding by the children (Cochet et al., 2017). Skeletal and ToM modeling can help make prediction performance better.

Another behavior pattern is social facilitation. Younger children prefer to be together with their parents while solving tasks. Older children follow verbal conventions and reduce the vividness of their facial expressions, while younger children exaggerate their facial expressions and rely more on social facilitation factors. Similar behavior patterns happen in adults in a competitive atmosphere where social facilitation has different effects on an individual's facial expressions (Buck et al., 1992; Katembu et al., 2022). ToM and multi-party dialogue modeling can increase the performance of uncertainty understanding models.

Uncertainty is context-dependent – some children are naturally more fidgety or shy. So predicting uncertainty on an isolated trial basis may lead to less accurate results. As a result, one interesting question is how to incorporate contextual features about the participant's personality and recent cognitive states to make more informed predictions.

7 Conclusion

In this paper, we explored the task of predicting uncertainty in young children from an annotated dataset that we introduced with a multimodal transformer-based model. We discover that demographic and trial difficulty can affect the frequency of certain cues. Moreover, trial difficulty strongly correlates with uncertainty, but trial performance interestingly does not. There is still room for improvement in task performance by transformer models, which means that more data or more complicated task setups are needed to study uncertainty properly. Our dataset–which we make available for research purposes–and protocol provide future researchers with additional tools to predict uncertainty using multimodal cues to facilitate human-human and human-AI dialogue.

8 Ethics

Due to the sensitive nature of the video data of children and their privacy, we are only making some portion of the data publicly available with the consent of the parents of the children. All the images used in this paper are from the videos that are from children and parents who have given consent to share their video data publicly.

References

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE*

International Conference on Automatic Face Gesture Recognition (FG 2018), pages 59–66.

Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. 2018. Multimodal grounding for language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nathaniel J. Blanco and Vladimir M. Sloutsky. 2021. Systematic exploration and uncertainty dominate young children's choices. *Developmental Science*, 24(2):e13026.

Ross Buck, Jeffrey I Losow, Mark M Murphy, and Paul Costanzo. 1992. Social facilitation and inhibition of emotional expression and communication. *Journal of Personality and Social Psychology*, 63(6):962.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Herbert H Clark, Robert Schreuder, and Samuel Buttrick. 1983. Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, 22(2):245–258.

Hélène Cochet, Marianne Jover, Cécile Rizzo, and Jacques Vauclair. 2017. Relationships between declarative pointing and theory of mind abilities in 3- to 4-year-olds. *European Journal of Developmental Psychology*, 14:324–336.

Gilles Degottex. 2014. Covarep – a collaborative voice analysis repository for speech technologies.

Stanislas Dehaene. 2011. *The number sense: How the mind creates mathematics*. Oxford University press.

Jeroen Dral, Dirk Heylen, and Rieks Akker. 2011. *Detecting Uncertainty in Spoken Dialogues: An Exploratory Research for the Automatic Detection of Speaker Uncertainty by Using Prosodic Markers*, pages 67–77.

Ahmad Ghanizadeh. 2008. Association of nail biting and psychiatric disorders in children and their parents in a psychiatrically referred sample of children. *Child and adolescent psychiatry and mental health*, 2:1–7.

Esen Gilleard, Mehmet Eskin, and Buğda Savaşir. 1988. Nailbiting and oral aggression in a turkish student population. *British journal of medical psychology*, 61(2):197–201.

Paul L. Harris, Deborah T. Bartz, and Meredith L. Rowe. 2017. Young children communicate their ignorance and ask questions. *Proceedings of the National Academy of Sciences*, 114(30):7884–7891.

Pavel Jahoda, Antonin Vobecky, Jan Cech, and Jiri Matas. 2018. Detecting decision ambiguity from facial images. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 499–503.

Stephen Katembu, Qiang Xu, Hadiseh Nowparast Rostami, Guillermo Recio, and Werner Sommer. 2022. Effects of social context on deliberate facial expressions: Evidence from a stroop-like task. *Journal of Nonverbal Behavior*, 46(3):247–267.

Terry Michael McClanahan. 1995. Operant learning (rs) principles applied to nail-biting. *Psychological reports*, 77(2):507–514.

Michelle Morales, Stefan Scherer, and Rivka Levitan. 2018. A linguistically-informed fusion approach for multimodal depression detection. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 13–24, New Orleans, LA. Association for Computational Linguistics.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, page 169–176, New York, NY, USA. Association for Computing Machinery.

Yondu Mori and Marc D Pell. 2019. The look of (un) confidence: visual markers for inferring speaker confidence in speech. *Frontiers in Communication*, 4:63.

Darko Odic, Howard Hock, and Justin Halberda. 2014. Hysteresis affects approximate number discrimination in young children. *Journal of Experimental Psychology: General*, 143(1):255.

Darko Odic and Ariel Starr. 2018. An introduction to the approximate number system. *Child Development Perspectives*, 12(4):223–229.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Heather Pon-Barry and Stuart Shieber. 2011. Recognizing uncertainty in speech. *EURASIP journal on advances in signal processing*, 2011.

Pio E. Ricci Bitti, Luisa Bonfiglioli, Paolo Melani, Roberto Caterina, and Pierluigi Garotti. 2014. Expression and communication of doubt/uncertainty through facial expression. *Ricerche di Pedagogia e Didattica. Journal of Theories and Research in Education*, 9(1):159–177.

Pedro Bispo Santos, Lisa Beinborn, and Iryna Gurevych. 2016. A domain-agnostic approach for opinion prediction on speech. In *Proceedings of the Workshop*

on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES), pages 163–172, Osaka, Japan. The COLING 2016 Organizing Committee.

Kimberly Scott and Laura Schulz. 2017. Lookit (Part 1): A New Online Platform for Developmental Research. *Open Mind*, 1(1):4–14.

Kevin P Silber and Clare E Haynes. 1992. Treating nail-biting: a comparative analysis of mild aversion and competing response therapies. *Behaviour research and therapy*, 30(1):15–22.

Matthew Stone and Insuk Oh. 2008. Modeling Facial Expression of Uncertainty in Conversational Animation. In *Modeling Communication with Robots and Virtual Humans*, pages 57–76. Springer, Berlin, Germany.

Marc Swerts and Emiel Krahmer. 2005. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1):81–94.

Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

Jinjing (Jenny) Wang, Justin Halberda, and Lisa Feigenson. 2021. Emergence of the Link Between the Approximate Number System and Symbolic Math Ability. *Child Dev.*, 92(2):e186–e200.

Jennifer H Wells, Janet Haines, and Christopher L Williams. 1998. Severe morbid onychophagia: the classification as self-mutilation and a proposed model of maintenance. *Australian and New Zealand journal of psychiatry*, 32(4):534–545.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 328–339, Online. Association for Computational Linguistics.

A Percentages of each cue

Here we present in Table 5, all the distribution of the uncertainty cues we found in the dataset. We also present more statistics between female and male participants in Figure 8.

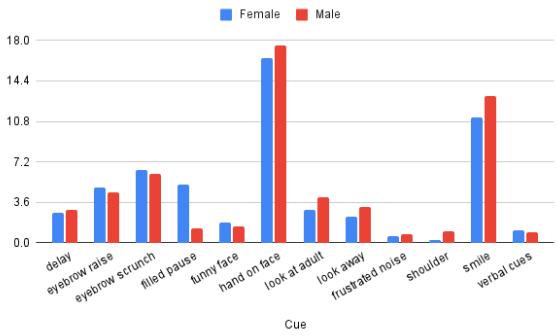


Figure 8: Female and male participants show small differences in frequency of displaying certain cues.

B Contrastive Learning Details

B.1 Problem Statement

In this study, we aim to predict the uncertainty of a child based on multimodal inputs, including video, transcripts, and audio. Given the dataset of instances, each containing video (V), transcripts (T), and audio (A) data, our goal is to develop a model that can accurately predict whether a child is uncertain, unclear, or not uncertain. We represent this problem as a function F that maps the input features (V, T, A) to the binary output variable $y \in \{0, 0.5, 1\}$, where 0 denotes not uncertain, 0.5 denotes unclear, and 1 denotes uncertain:

To achieve this, we design a multimodal transformer model that leverages the complementary information in the video, transcripts, and audio data to make predictions. The model is trained on a dataset of labeled examples (V_i, T_i, A_i, y_i), where $i \in \{1, ..., N\}$ and N is the total number of instances. Our objective is to minimize the cross-entropy loss. By minimizing this loss, our model will learn to predict a child's uncertainty level accurately based on the provided multimodal inputs.

The specific contrastive learning loss function, $L(X_1, X_2, L_1, L_2)$, that we are focusing on here is defined as the following:

$$L(X, L) = \frac{1}{N_1 * N_2} \sum_i \sum_j W_{ij} * (m - S_{ij})_+^2 \quad (1)$$

This function captures the relationship between pairs of data points $X1$ and $X2$, with associated labels $L1$ and $L2$. The cosine similarity, S_{ij}, is used to measure the similarity between the data points, and the weighing factor, W_{ij}, is used to differentiate between positive and negative pairs. The weighting factor is determined using the Kronecker

Cue	All Pct.	Uncertain Pct.	Hard Pct.	Easy Pct.	Female Pct.	Male Pct.
delay	0.03	**0.17**	0.03	0.02	2.7	2.93
eyebrow raise	0.05	**0.17**	**0.06**	0.04	4.94	4.48
eyebrow scrunch	0.06	**0.22**	0.07	0.06	6.49	6.15
filled pause	0.03	0.06	0.03	0.03	**5.17**	1.26
funny face	0.02	0.07	**0.03**	0.01	1.84	1.49
hand on face	**0.17**	**0.19**	**0.16**	**0.18**	16.44	17.53
look at adult	0.04	0.1	**0.05**	0.03	2.93	4.02
look away	0.03	0.05	0.03	0.02	2.36	3.16
frustrated noise	0.01	0.02	0.01	0.01	0.57	0.8
shoulder	0.01	0.02	0.01	0.01	0.29	**1.03**
smile	**0.12**	**0.17**	**0.12**	**0.12**	11.15	**13.05**
verbal cues	0.01	0.02	0.01	0.01	1.09	0.92

Table 5: Distribution of uncertainty cues across all/uncertain, difficult/easy, and female/male trials. We can see that *delay, eyebrow raise,* and *eyebrow scrunch* are significantly more frequent in uncertain trials. Meanwhile, if the trial is hard, participants are likelier to look at an adult or make a funny face. If the trial is easy, the participant may display *hand on face* instead. Female and male participants also show mild differences in the frequency of displaying certain cues.

delta function, $\delta(L1_i, L2_j)$. The $e^{\alpha * \delta(L1_i, L2_j)}$ coefficient ensures that the positive pairs have a greater influence on the learning process where m is a threshold to separate positive and negative pairs and α, is a scaling factor. Lastly, $(x)_+$ ensures that only the non-negative values of x are considered.

To further improve our model's performance, we employed a weighted sampling method using the class frequencies' inverse square root. Given a dataset with classes 0 being certain, 0.5 being unclear, and 1 being uncertain, we calculate the weights for each class sample as follows:

$$w_i = \frac{1}{\sqrt{\mathrm{N_i}}}, \quad \text{where} i \in \{0, 0.5, 1\}. \qquad (2)$$

In our case, this weighting scheme assigns higher weights to underrepresented classes, the class of 0.5 (unclear) and 1 (uncertain), which helps balance class sampling probabilities. The inverse square root function is particularly useful as it provides a smooth topology less sensitive to small changes in class frequencies than other weighing functions.

C Implementation Details

C.1 Experimental setup

The argmax of the label probabilities was taken as the output layer. All networks were trained for 40 epochs with a batch size of 24. Both raw and weighted cross entropy loss were used to train two versions of the model. Class weights were set based on the distribution of train set samples to mitigate class imbalance issues.

We employ a 75-10-15 training-dev-test split. For each result, we report the average across three different seeds. We also run the model on every single modality to provide unimodal baselines.

Additionally, we have age information for each participant, so we divided the dataset into two age groups: 4-year-olds and 5-year-olds. This division will allow us to investigate potential differences between the two age groups as shown in the 4-year-old and 5-year-old tabs in Table 2.

Transformer Model Details. Our multimodal transformer model is based on a modified version of the Transformer architecture. It consists of five layers, each equipped with five attention heads to capture various contextual relationships within the input data. We used the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.001 to train our model. To enhance convergence and overall performance, we employed the ReduceLROnPlateau learning rate scheduler, which adjusts the learning rate when the validation loss ceases to improve. We set the reduction factor to 0.1 and patience of 5 epochs for monitoring improvements. Our model was trained on an NVIDIA RTX 4090 GPU, using a batch size of 1. We trained the model for 100 epochs. For models with contrastive learning, we trained the model with additional 10 epochs before real training.

Grounding Description-Driven Dialogue State Trackers with Knowledge-Seeking Turns

Alexandru Coca[†], **Bo-Hsiang Tseng**[‡], **Jinghong Chen**[†], **Weizhe Lin**[†], **Weixuan Zhang**[†], **Tisha Anders**[†*], **Bill Byrne**[†]
[†]Department of Engineering, University of Cambridge, United Kingdom
[‡]Apple
[†]{ac2123, jc2124, wl356, wz315, wjb31}@cam.ac.uk
[‡]bohsiang_tseng@apple.com [*]anderstisha@gmail.com

Abstract

Schema-guided dialogue state trackers can generalise to new domains without further training, yet they are sensitive to the writing style of the schemata. Augmenting the training set with human or synthetic schema paraphrases improves the model robustness to these variations but can be either costly or difficult to control. We propose to circumvent these issues by grounding the state tracking model in knowledge-seeking turns collected from the dialogue corpus as well as the schema. Including these turns in prompts during finetuning and inference leads to marked improvements in model robustness, as demonstrated by large average joint goal accuracy and schema sensitivity improvements on SGD and SGD-X[1].

1 Introduction

Task-oriented dialogue (TOD) agents provide natural language interfaces that users can interact with to access a wide variety of services, from airline search (Seneff and Polifroni, 2000) to complex customer service applications (Chen et al., 2021). To enable this, agents track key information communicated by the user as the conversation progresses. This is known as *dialogue state tracking* (DST). Commonly, the dialogue state is represented as a sequence of task-specific *slot-value pairs*[2].

A common DST assumption is that the set of slots and values a user may communicate, the *domain ontology*, is known at design time. Hence, extensive data collection and annotation is needed to support new domains, which hinders the scalability of this approach. Rastogi et al. (2020) address this issue by creating the schema-guided dialogue dataset (SGD). In SGD, the information available to a TOD agent is organised as *schemas*[3] describing *services* with which users can interact. Each service has *user intents* representing the tasks users can complete by interacting with the agent (e.g. *find restaurants*). Several slots are associated with each intent and the schema provides a *natural language description* for each intent and slot. The insight motivating *description-driven DST* is that these descriptions alone can be used in tracking the dialogue state in a form close to natural language. This has emerged as a powerful approach for few-shot and zero-shot DST (Jacqmin et al., 2022) and benefits from recent advances in language modelling. For example, Zhao et al. (2022) finetune T5 (Raffel et al., 2020) to generate the dialogue state conditioned on the dialogue history and a descriptive prompt containing all intent and slot descriptions in a service schema. The use of natural language in the descriptive prompt enables the underlying language model to generalise to new services, whose schemas are not seen in training.

While the reliance on natural language is a strength, Lee et al. (2022a) show that state-of-the-art (SOTA) schema-guided DST models are not robust to the style of descriptive prompts: in SGD, the training schema contains a single description per slot or intent, and models trained with prompts composed from this schema alone are prone to overfitting. Lee et al. (2022a) show this limitation can be mitigated by increasing prompt diversity. They use a large number of human annotators alongside expert curation to create diverse schema paraphrases that are used for model robustness improvement. This is a costly process that is not easily scalable.

As an alternative to additional human annotation of the schema, we show that the SGD training dialogues themselves exhibit sufficient diversity of expression such that they can be used to overcome the lack of diversity in the SGD schema descriptions. We *ground* DST prompts in the dialogue context by concatenating the schema descriptions with dialogue turns extracted from the SGD cor-

[1]Our code will be released upon publication.
[2]For example, for a restaurant booking a sequence could be *day=friday, time=7pm, guests=1, restaurant=nandos.*
[3]See schema examples here: https://bit.ly/3RJ6u4l.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 444–456
September 11–15, 2023. ©2023 Association for Computational Linguistics

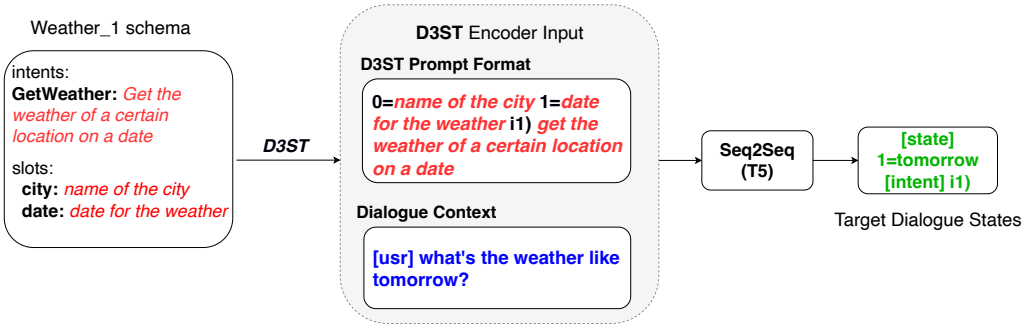


Figure 1: D3ST input and target format. On the left, we show a schema excerpt, where **slot** and **intent** names, in bold face, are followed by their *natural language description*. The encoder input, represented in the centre, is a string, the concatenation of two elements: the *prompt* which describes what information should be output and the *dialogue context* from which the information should be extracted. On the right we show the target dialogue state.

pus based on similarity to the dialogue state. We find that this approach is more effective than using synthesised prompts and even outperforms or is comparable to the highly-curated human-written prompts used by Lee et al. (2022a), when evaluated with medium and large language models. We evaluate our methods using the SOTA D3ST prompting scheme (Zhao et al., 2022) on the SGD and SGD-X (Lee et al., 2022a) datasets.

2 Related work

Neural classification is effective for DST when the domain ontology is fixed and known at design time (Mrksic et al., 2017). Adapting such models to track new slots and domains requires annotated conversational data and thus data scarcity is a long-standing issue in DST research (Jacqmin et al., 2022). Scarcity has been addressed by copy-enhanced generation (Wu et al., 2019), reading comprehension (Gao et al., 2019) and adapting pretrained language models to generate the state given the dialogue context alone (Peng et al., 2020; Hosseini-Asl et al., 2020). These were improved upon by transfer learning from question-answering tasks (Lin et al., 2021a), which in turn was outperformed by schema-guided models (Lee et al., 2021; Zhao et al., 2022; Lin et al., 2021b). Recently, Gupta et al. (2022) apply in-context tuning (Min et al., 2022) to DST, creating training prompts which contain a dialogue and its target state sequence. Their model thus learns from DST task demonstrations.

Lee et al. (2022a) investigate the robustness of SOTA schema-guided dialogue state trackers to schema changes. This is a new line of research, as previous work concerns other robustness issues that generally affect DST, such as variations in the conversational data distribution, noise, and adversarial perturbations (Jacqmin et al., 2022). Through extensive, crowdsourced[4], schema paraphrase collection, Lee et al. (2022a) report that DST performance degrades substantially when models trained on one set of prompts are evaluated on manually paraphrased prompts. By contrast, Cao and Zhang (2021) report little degradation in DST with back-translated prompts, suggesting that backtranslation is a weak proxy for actual human variability. Lee et al. (2022a) perform data augmentation (DA) for robust DST, finding that prompts obtained via automatic paraphrasing lag in quality relative to manual paraphrases. Ours is the first work to address the gap between synthetic methods, such as backtranslation, and manual paraphrasing. We show that the gains from manual paraphrasing can be achieved by mining the existing annotated dialogues used for training the DST model in the first place.

3 Robust DST with grounded prompts

We review D3ST, a SOTA description-driven DST model (Section 3.1). We then describe our grounding method that extracts turns from the corpus (Section 3.2) and uses them to design prompts for robust DST with D3ST (Sections 3.3 & 3.4).

3.1 Description-driven dialogue state tracking

Figure 1 shows the inputs and outputs of D3ST (Zhao et al., 2022). The model is implemented with T5, an encoder-decoder language model (Raffel et al., 2020). The encoder input, represented in the centre, comprises a *prompt* describing what information should be tracked by the DST model, and the *dialogue context*, a conversation between

[4]The SGD schema were rewritten by over 400 annotators and curated by dialogue experts, over the course of one month.

a user and an agent from which slot-value pairs should be extracted. The prompt is a concatenation of slot and intent descriptions, extracted from the service schema (on the left). Each description is prefixed by a randomly assigned index prior to concatenation. Zhao et al. (2022) motivate their use of random indices to replace slot and intent names because names convey little semantic information and may be ambiguous[5]. In this paper, we will refer to this *prompt format* as **D3ST**.

The model is trained to output index-value pairs for all slots mentioned in the conversation (i.e. the *active slots*) as well as an index representing the active intent. These are represented on the right in Figure 1. The slot-value pairs mentioned in the conversation can be recovered by replacing the predicted indices with their corresponding slot names. The user active intent is found by replacing the predicted index with the name of the intent.

3.2 Mining turns for prompt design

1. `REQUEST(restaurant_name)` SYS: Where do you want to dine?
2. `INFORM(restaurant_name=Nandos)` USR: I want Nando's.
3. `INFORM_INTENT(find event)` USR: What shows are on?
4. `OFFER_INTENT(buy ticket)` SYS: Want tickets?

Table 1: Sample semantic annotations

We now discuss how to extract turns from the corpus to design better prompts. Our approach involves an automatic step that uses the semantic annotations in the corpus followed by a verification step to ensure that the turns selected are diverse.

Each turn in SGD is semantically annotated with one or more *dialogue actions* which describe what is being communicated (Table 1). We focus on *knowledge-seeking* turns (KSTs). These are annotated with a single `REQUEST` *dialogue act* and associated with a *single* slot, without a value mention (Table 1, line 1). Selecting turns annotated with a single slot allows us to unambiguously associate them with slots in the D3ST prompt. We do not mine *informational* turns (labelled with `INFORM`) since these mention a specific, known value, often without reference to the underlying slot (Table 1, line 2). Such turns could be combined with schema information to form *exemplar-based* prompts as done by (Figure 1 in Gupta et al. (2022)), a more complex approach which we discuss in Section 5.5.

To select turns for a given slot, s, we filter the corpus to get all the knowledge-seeking turns relating to it. We manually select 5 of these, repeating this process for each slot in every service in the training data. See examples in Table 2. In a similar fashion, we select 5 turns from those labelled with a single `INFORM_INTENT` or `OFFER_INTENT` act (Table 1) for every intent in the training schema.

Index	Selected Knowledge-seeking Turn
1	Which event are you looking to book
2	Do you have any particular show in mind
3	And what is the event
4	What event do you wish to see
5	What is the event you are looking for

Table 2: Selected turns of the *event name* slot

We opt to select the turns manually because our goal is prompt diversity. Our SGD analysis revealed that the knowledge-seeking turns tend to be biased towards specific vocabulary and syntactic patterns. For example, among the 173 turns in which the user requests the price of a rental car, 71.1% contain the word *cost*, 42.8% contain the word *total* and, 27.7% contain *total cost*. In contrast, *price* appears in just 11.0% of the turns.

All turns were mined by one student in one day, despite SGD being the largest schema-guided TOD corpus. In practice, schemas are induced by developers from unlabelled conversation databases (Yu et al., 2022). The turns could be collected as part of this process with negligible overhead. We handle slots with few KSTs as described in Appendix A.

3.3 Grounding prompts

To ground a schema description in its conversational use, we concatenate it with randomly selected knowledge-seeking turns from the mined collection (Figure 2). In the example shown, the sampled knowledge-seeking turns for the *city* and *date* slots are *In which location should I check?* and *Forecast for when?*, respectively. These are concatenated with the original SGD schema descriptions *name of the city* and *date for the weather* to ground the prompt. We concatenate the turns and descriptions in random order, to prevent the model learning to attend preferentially to one source of information over another. We refer to D3ST trained with prompts grounded in knowledge-seeking turns as **D3ST-Turn** in what follows.

[5]For example, the *location* slot name may be used to refer to both a city name and an address across different services.

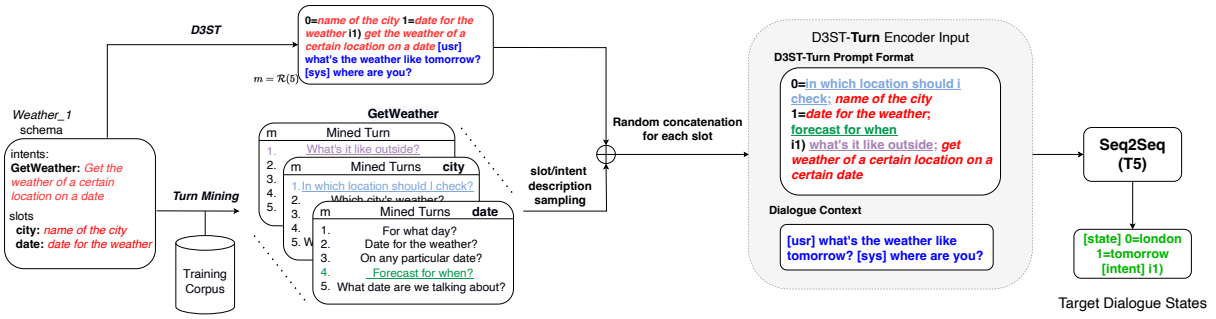


Figure 2: Visual representation of D3ST-Turn prompting. Underlined knowledge seeking turns are those chosen at random for inclusion in the sample D3ST-Turn prompt shown.

Slot names may provide additional information about the meaning of a slot, so we propose to ground the prompt both in knowledge-seeking turns and slot names. We refer to D3ST trained with prompts grounded in knowledge-seeking turns and slot names as **D3ST-TurnSlot**.

3.4 Grounded prompt ensembling

Multiple knowledge-seeking turns are available for decoding, enabling us to create multiple *prompt variants*. A given model generates the dialogue state when conditioned on each of these prompt variants, in turn. The state hypothesis is the most commonly predicted string when our *single* model is prompted with the prompt variants. We call this technique *grounded prompt ensembling* (GPE).

4 Experiments

4.1 Datasets and metrics

SGD (Rastogi et al., 2020) The training set contains $21,106$ dialogues across 16 domains. The test set contains $4,201$ conversations, 77% of which have a turn span where the user talks to the agent to access a service unseen in training. 6 schemas are seen in training whereas 15 are *unseen*. Hence, this benchmark primarily tests the ability of DST models to accommodate values, slots, prompts and domains it has not been trained on.

SGD-X Lee et al. (2022a) created SGD-X because they found the linguistic patterns of the SGD unseen services schemata to be too similar to those of the seen schemata[6]. They use crowdsourcing and dialogue experts to create five *schema variants*[7] which are increasingly stylistically and lexically divergent from the SGD schema. A schema variant describes the same services as the SGD schema but with increased linguistic variation. The variants are ordered by the Jaccard distance between the descriptions of the original SGD schemas and the schema variant descriptions. The *v1* variant is the closest while *v5* is the most dissimilar to SGD. Ideally, a robust model should output the correct state regardless of which schema variant is used for prompting.

Metrics Joint goal accuracy (JGA)[8] is the percentage of turns where all the slot-value pairs from a given service are correctly predicted. On SGD, it is computed over seen and unseen services. The presence of the unseen services measures the ability of the DST model to make correct predictions for unseen slots and values and to interpret descriptions unseen at training time (Rastogi et al., 2020).

For SGD-X, we report the JGA broken down by seen and unseen services and their combination. The JGA coefficient of variation across the five schema variants is termed *sensitivity* (SS). It measures how well the model accommodates linguistic variation. Evaluation on the seen portion involves prompting with *paraphrases* of schemata seen in training. Performance decreases if the model overfits to the training descriptions. In evaluation on the unseen portion, the model is prompted with five distinct human-written prompts of increasing dissimilarity to the original SGD. This evaluates if generalisation is robust to linguistic variation.

4.2 DST models

Our baselines are three D3ST models[9] trained with large augmented datasets. For every training example that uses the D3ST prompt format (Figure 1) linearised from the SGD schema, k additional training examples are created either using synthetic prompts or the $k = 5$ SGD-X schemata. We create

[6]For example, descriptions of slots with "true" or "false" values always start with *Boolean flag indicating*.

[7]See examples here: https://bit.ly/3Ev0KrV.

[8]We use the offical evaluator: https://bit.ly/3B7jD1c

[9]We use T5-base (220M) for all models except in Sec. 5.6

augmented datasets using three methods, explained below. See Appendix B for implementation details.

1. Backtranslation We follow Lee et al. (2022a) to create $k = 3$ schema variants by backtranslating the SGD schema via Chinese, Japanese and Korean with Google Translate. The augmented dataset is 4 times larger than SGD ($703,120$ examples).

2. Easy Data Augmentation (EDA) (Wei and Zou, 2019) We create $k = 5$ schema variants by applying word-level perturbation to the SGD schema (EDA). Synonym replacement is applied with probability 0.25 whereas random insertion, deletion and substitution are applied with probability 0.05. There are $1,054,680$ training examples.

3. SGD-X We create $1,054,680$ training examples using the $k = 5$ human-written SGD-X schemata. Unlike the other baselines, SGD-X-trained models see the human-written paraphrases of the seen test services during finetuning. In all other experiments, *none* of the SGD-X test prompts are seen during training, and so we refer to this experiment as an *oracle*, following Lee et al. (2022a).

Contemporaneous to our work, Coca et al. (2022) propose a tree-ranking approach for improving paraphrasing. While they show significant gains compared to the state-of-the-art backtranslation baseline we compare to in our work, we contribute to the body of knowledge on robust state tracking by showing novel augmentation and prompting techniques that achieve significant further improvements.

Grounded D3ST Instead of augmentation, we propose to ground D3ST in knowledge-seeking turns by finetuning T5 with the Turn (D3ST-Turn) and TurnSlot (D3ST-TurnSlot) prompts (Section 3.3) on a dataset containing $175,780$ examples (SGD size). At decoding, the same turns grounding the training prompts are used for seen services. For unseen services, we select five turns per slot as described in Section 3.2. For each test example, we construct a prompt with the same format as in training using knowledge seeking turns selected at random, per slot, from the mined collection. This tests model's ability to interpret additional task-relevant information.

5 Results and discussion

5.1 Robustness via data augmentation

Augmenting the finetuning dataset with the SGD-X prompts leads to a 13.2% improvement in D3ST SGD-X JGA (#1 vs #4, Table 3). In contrast, augmentation with synthetic prompts obtained through backtranslation or word-level augmentation improves performance by just 5.6%. On SGD, human-written prompts outperform the best performing synthetic ones (#2) by a margin of 1.7%.

#	Model	SGD	SGD-X	Seen	Unseen	SS ↓
1	D3ST	69.8	56.5	73.6	50.8	70.1
2	D3ST + Backtrans. DA	72.1	62.2	84.0	54.9	53.1
3	D3ST + EDA DA	71.4	62.3	83.3	55.3	53.2
4	D3ST + SGD-X DA (oracle)	73.8	69.7	**92.5**	62.1	27.9
5	D3ST-Turn (ours)	**75.9**	69.5	88.5	63.2	36.6
6	D3ST-TurnSlot (ours)	74.7	**72.0**	90.7	**65.6**	**23.7**

Table 3: Grounded D3ST models outperform strong baselines in both JGA and SS. Seen and unseen numbers decompose the SGD-X JGA (Section 4.1), *oracle* indicates a model trained on SGD-X. "+" marks data augmentation (DA) during finetuning, and is followed by augmentation method name. Column maximum is in **bold**. In all tables, numbers are averages of three runs.

Schema	v1	v2	v3	v4	v5
Backtranslation	97.5	96.5	95.9	-	-
EDA	99.1	98.5	96.6	93.2	86.4
SGD-X	89.7	88.0	88.4	86.8	87.5

Table 4: Semantic similarity of SGD and schema variants, measured by entailment (Narayan et al., 2022)

Gains obtained with synthetic prompts reflect some degree of lexical and syntactic diversity in the generated paraphrases. For example, a backtranslation of *The amount of money to transfer* is *Amount to be remitted* and *The account type of the user* is backtranslated as *User's account type*. The entailment scores (Table 4) show that backtranslation largely preserves the semantic content of the prompts. Meanwhile, if more edit operations are applied via EDA, the synthetic prompts are less faithful, as demonstrated by the sharp entailment decrease for the *v4 & v5* variants. Robustness did not improve when we experimented with a larger backtranslation-augmented dataset (Appendix C).

The SGD-X schemata "do not fully semantically overlap with the input as traditional paraphrasing requires" (Lee et al., 2022a). This is consistent with SGD-X schema variants attaining lower entailment compared to backtranslated ones (Table 4). The annotators used the wider context of the service and common-sense knowledge to create diverse, high quality, schemas. Meanwhile, D3ST learns to identify slots using the uniform linguistic patterns of the SGD schema and it is not robust to the wide variety of styles annotators used in SGD-X. D3ST trained with augmented data via EDA or backtranslation improves compared to D3ST

trained on SGD alone, but the large performance gap to human-written prompts indicates that strict paraphrasing introduces less diverse, task-relevant, cues in the prompt compared to humans.

5.2 Prompt grounding with turns

Compared to D3ST, D3ST-Turn achieves absolute gains of 13% and 6.1% on SGD-X and SGD, respectively (#1 vs #5, Table 3). D3ST + SGD-X DA outperforms D3ST-Turn on the seen services because it has been trained with these prompt paraphrases, whereas our model *does not* see these prompts during training. Our model generalises more robustly as demonstrated by the 1.1% (#4 vs #5) JGA improvement on unseen SGD-X services.

Our results show that grounding the model in knowledge-seeking turns, communicated before a slot is mentioned, addresses weaknesses of data augmentation (DA) with synthetic prompts. Such turns reflect how humans use the language in conversation when they communicate slot values, and may help the model more readily identify the relevant context for value extraction. This approach generalises well to unseen domains and is robust: we outperform all baselines on SGD and closely match the performance of augmentation with human-written paraphrases on SGD-X.

Descriptions and knowledge-seeking turns are complimentary: the latter can be thought of as an *example* that could help the model interpret descriptions unseen at training time. Concretely, consider the *messaging* domain, unseen in training. To extract the name of a location sharing recipient, a model evaluated on SGD-X ($v5$) is prompted with the description *Name from address book*. Because T5 is pre-trained in a self-supervised way, without domain-specific finetuning, it may fail to identify that the aforementioned description refers to the name of a person: *address book* never appears in the SGD training corpus. By attending over *Who is the sharing recipient* and the description jointly, the model could interpret the description as referring to a person name. During training, the model has learned to identify names, for example, when predicting the value of the slot *stylist name*. Indeed, the D3ST-Turn JGA in this domain is 41.8% while the oracle model achieves just 28.5%. Our positive results may thus arise due to knowledge-seeking turns facilitating knowledge sharing between slots seen in training and unseen ones.

5.3 Prompt grounding with turns and slots

D3ST-TurnSlot achives a 2.5% gain on SGD-X compared to D3ST-Turn, outperforming the human-written prompts (# 4 vs #5, Table 3). We posit that this is due to the high quality annotations SGD-X provides. This hypothesis is motivated by our empirical observation that slot names in SGD-X can contain more information compared to SGD ones. For example, the slot *private visibility* in SGD is annotated as *private visibility yes or no* in SGD-X ($v5$), which cues the model on which values should be generated for this slot. Also, in SGD-X, the slot names and descriptions may be complimentary. For example, the slot *clock time of alarm* is described as *Time for which the alarm is set* (SGD-X), whereas in SGD the equivalent slot name, *alarm time*, is described as *Time of the alarm*. On its own, the SGD description could refer to both an alarm to be created or an existing alarm, whereas the SGD-X description unambiguously identifies the slot as referring to an existing alarm.

D3ST-TurnSlot outperforms D3ST + SGD-X DA by 0.9% on SGD (#4 vs #6, Table 3) but lags behind D3ST-Turn by 1.1%. This confirms our earlier observation that, in SGD, unlike in SGD-X, slot names may not provide information about the slot that is not already contained in the description. We also find that there are slot name ambiguities across the SGD train and test sets. For example, the *location* slot in the training set refers to cities, whereas in the test set it refers to addresses. This finding correlates with the study of Zhao et al. (2022), the D3ST authors, who find that lack of information in slot names and ambiguity lead to degraded JGA (on both SGD and SGD-X) of slot-name driven models compared to D3ST[10]. Our positive results on SGD-X show that combining the two sources of information can improve model robustness if they are complimentary and unambiguous.

5.4 Grounded prompt ensembling

We apply GPE by running three inference calls with distinct but semantically equivalent grounded prompts. We take the most common generation as the prediction. Table 5 shows significantly improved robustness compared to single-prompt decoding. Interestingly, D3ST-TurnSlot is improved

[10]A slot-driven model uses slot names instead of descriptions. For our example in Figure 1 the equivalent prompt is *0=name, 1=city, i1) get weather*. We refer the reader to Sections 4.3, 4.4 and 4.6 in Zhao et al. (2022) for detailed comparisons of the effectiveness of these competing approaches.

Model	SGD	SGD-X	Seen	Unseen	SS ↓
D3ST-Turn	**77.2** 1.4	71.7 2.2	90.8 2.3	65.4 2.2	28.1 8.5
D3ST-TurnSlot	75.0 0.3	72.8 0.8	91.4 0.7	66.6 1.0	**19.2** 4.5

Table 5: GPE improves SGD/SGD-X performance. Faded numbers are absolute improvements relative to the single pass models in Table 3 in rows # 5 & # 6.

Model	SGD	SGD-X	Seen	Unseen	SS ↓
T5DST	70.0	50.4	58.5	47.7	87.0
MT-SGDST	80.1	60.8	72.5	56.9	69.5
SDT-Seq	76.3	-	-	-	-
SDT-Ind	**78.2**	-	-	-	-
D3ST-Turn (ours)	75.8	69.5	88.5	63.2	36.6
D3ST-TurnSlot (ours)	74.7	**72.0**	90.7	**65.6**	**23.7**

Table 6: SOTA DST models on SGD and SGD-X. Bottom rows repeated from Table 3 for easy comparisons.

less compared to D3ST-Turn on both SGD and SGD-X owing to its significantly smaller prompt sensitivity (23.7 compared to 36.6, Table 3, # 5 vs #6). This shows that slot names increase the confidence of the model in its predictions, which may explain why we found D3ST-TurnSlot to slightly outperform D3ST-Turn (Section 5.3).

5.5 Comparison with other models

We compare D3ST-Turn/TurnSlot with SOTA DST models (Table 6). **T5DST** (Lee et al., 2022a) generates a slot value when prompted with a dialogue concatenated with a single description. The state is predicted *iteratively* by prompting the model with each description. **MT-SGDST** (Kapelonis et al., 2022) uses semantic annotations, state history and handcrafted features with a multi-head BERT model for iterative prediction. **SDT-Seq** (Gupta et al., 2022) grounds the state tracker in a prompt containing a dialogue and its target state sequence and, like D3ST, predicts the entire state in a single pass. **SDT-Ind** is an iterative version of SDT-seq, using annotated turns as prompts.

Grounding prompts in knowledge-seeking turns makes D3ST competitive with SOTA approaches, significantly outperforming T5DST. MT-SGDST is better on the SGD but degrades significantly on SGD-X. Because it uses state histories and semantic annotations instead of system turns, this model suffers from large performance variability (Appendix B.2): the difference between max and min SGD-X JGA across three runs is 11.8% for this model but just 1.1% for D3ST-Turn. While not fully closing the gap on the SGD benchmark, we demonstrate comparable performance and significantly improved robustness.

Model	SGD	SGD-X	Seen	Unseen	SS ↓
D3ST	76.0	69.2	86.8	63.3	38.5
D3ST + SGD-X DA	**77.4**	75.6	**93.2**	69.7	**19.8**
D3ST-TurnSlot (ours)	**77.4**	**76.0**	92.6	**70.5**	20.5

Table 7: Prompt grounding improves T5-large D3ST.

D3ST-Turn achieves 75.8% on SGD, which is comparable with SDT-Seq (76.3%). Our model is faster to train and decode owing to reduced prompt lengths and predicting a shorter state sequence[11]. SDT-Ind is better because it is prompted to return the value of each slot iteratively, with an example of how that typical slot occurs in conversation. GPE is cheaper and reduces the performance gap between SDT-Ind and D3ST-Turn to just 1.0%.

In terms of human effort, our approach is more scalable than, or comparable to, recent work SDT uses entire annotated dialogues or annotated turns as prompts (Figure 1, Gupta et al. (2022)). These are defined by developers for unseen services, which is comparable to writing turns for each slot. Zero-shot transfer learning (Campagna et al., 2020) requires knowledge-seeking turns for generating synthetic dialogues[12] used to bootstrap DST models for new services. However, constraining entire dialogue generation is non-trivial and handcrafted grammars are required for each domain. This is very difficult to apply to the setting we consider, due to the large number of domains and complex multi-domain dialogue flows. We show that robust generalisation to new services can be achieved with few knowledge-seeking turns per slot which can be selected from the training corpus during finetuning and written by the developers for new services.

5.6 Scaling behaviour

Larger models have enhanced language understanding and common sense knowledge (Raffel et al., 2020; Zhou et al., 2021), reflected in the 12.7% improvement of the baseline T5-large D3ST performance on SGD-X compared to its T5-base counterpart (#1, Tables 3 and 7). Exposing the model to diverse prompts is still important, as demonstrated by the improved JGA of D3ST + SGD-X DA. We find that D3ST-TurnSlot matches the SGD performance and achieves a slight improvement on SGD-X (0.6%), demonstrating that our approach scales to larger language models.

[11] SDT predicts all slots, including inactive ones.

[12] The *direct questions* defined by Campagna et al. (2020) are KSTs. See examples here: https://bit.ly/3yvgGqi.

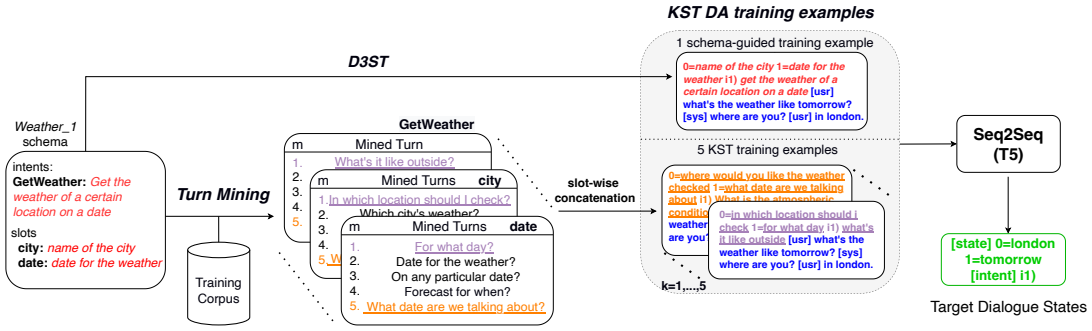


Figure 3: Data preprocessing pipeline for KST augmentation. KST training examples share the conversation with the schema-guided training example, and concatenated KSTs, underlined, replace the schema-guided prompt.

#	Decoding Prompt	SGD	SGD-X	Seen	Unseen	SS ↓
1	Turn [GPE]	74.9 [76.7]	71.7 [73.9]	91.9 [92.1]	65.0 [67.7]	30.7 [22.6]
2	TurnSlot [GPE]	73.8 [75.3]	71.0 [72.9]	91.0 [91.7]	64.3 [66.6]	31.2 [23.1]
3	D3ST	74.4	66.7	88.8	59.4	43.4

Table 8: JGA of D3ST with KST augmentation decoded with different prompt formats. GPE further improves these models (improvements inside brackets).

#	Model	SGD	SGD-X	Seen	Unseen	SS ↓
1	D3ST-TurnSlot	**77.4**	**76.0**	92.6	**70.5**	20.5
2	D3ST + KST DA/D3ST	76.3	72.8	92.5	66.3	26.0
3	D3ST + KST DA/Turn	76.1	74.6	**93.4**	68.4	26.0
4	D3ST + KST DA/TurnSlot	75.8	73.6	92.2	67.4	28.2

Table 9: Grounding prompts (#1) is more effective compared to KST-augmentation (#2 - #4) for robust DST with T5-large (770M parameters).

5.7 Data augmentation or prompt grounding?

In Section 5.2, we discussed that knowledge-seeking turns may facilitate knowledge sharing between seen and unseen slots. We now investigate whether jointly encoding the turn and description by including them in the same prompt is the only way to impart this property or whether this can be achieved by augmenting the training data with prompts containing *only* knowledge-seeking turns.

Figure 3 shows our experimental setup for augmentation with knowledge-seeking turns. We sort the mined turn lists for each slot from Section 3.2 according to their Jaccard distance to the corresponding SGD schema description. We create $k = 5$ *increasingly diverse* prompts by replacing all the slot descriptions in a schema-guided training example with the kth knowledge-seeking turn. The resulting finetuning set is the same size as SGD-X.

Comparing the performance of augmented and grounded models (#5 & #6, Table 3 vs #1 & #2, Table 8) shows that grounding D3ST is slightly more effective on SGD.

On SGD-X, decoding the KST-augmented D3ST with TurnSlot prompt format causes a small (1%) regression with respect to D3ST-TurnSlot, possibly due to mismatch between the train and test prompt formats. "Turn" decoding slightly improves over D3ST-Turn. Hence, both grounding and data augmentation with knowledge-seeking turns are effective for robust DST. Training with augmented data, converges slower and is resource intensive. Moreover, we find that grounding is more effective for larger language models (Table 9).

5.8 Why is grounding more effective?

Decoding the KST-augmented model with the SGD/SGD-X schemata alone (i.e., without grounding) leads to a decrease in the unseen performance (#1 & #2 vs #3, Table 8). Grounding the prompt in KSTs at decoding time is crucial for improved robustness. As discussed in Section 5.2, these turns facilitate knowledge sharing between seen and unseen slots. Without them, the model cannot access knowledge encoded in its weights, and robustly predict the dialogue state when the prompts are too dissimilar to the training schemata (Table 10).

SGD	SGD-X (avg)	v1	v2	v3	v4	v5
0.43	5.03	1.07	0.67	3.35	11.3	8.76

Table 10: JGA difference between KST-augmented models decoded with Turn and D3ST prompt formats (#1 & #3, Table 8). SGD-X JGA broken down by variant. $v5$ schema is the most dissimilar to the SGD test schema.

6 Conclusion

Grounding D3ST and data augmentation with knowledge-seeking turns are effective for robust schema-guided DST. Both improve D3ST robustness by a large margin compared to strong baselines and yield similar benefits as training on large,

diverse collection of human-written prompts. Our proposed approach is competitive with or outperforms other SOTA DST models on SGD and SGD-X. We have also showed how prompt engineering can be applied to boost model robustness through grounded prompt ensembling, a novel technique that uses a single model for ensembling.

7 Limitations

One limitation of our approach is our decision to select the turns from the training data manually rather than automatically. Selecting k-diverse turns automatically is possible but requires efficient implementations given the size of the corpus and the quadratic complexity of the naive algorithm in the number of candidate turns. Implementing such algorithms requires far more expertise and time commitment compared to ensuring the selected turns are diverse manually. Such an approach is described by Lee et al. (2022b). However, in a follow-up study, we have confirmed the generality of our approach by replicating our experiments with large-language model generated data or sampling randomly from large dialogue corpora during the first epoch of training. This allays concerns regarding the vulnerability of our method to human bias.

While not an issue for SGD or other large scale corpora, the diversity of the training corpus may influence the performance of our approach as extracting lower diversity turns is expected to limit robustness improvements. However, knowledge-seeking turns existing in small corpora can be used to query large, possibly unlabeled, conversational databases to ensure prompt diversity. We left a detailed study of the impact of prompt diversity to DST robustness to future work.

Finally, for practically implementing our approach for unseen services, we require the developers to provide few examples of knowledge-seeking turns. Our currently in progress work explores generation of such turns automatically with very large language models.

Acknowledgements

Alexandru Coca was supported supported by EPSRC grant EP/R513180/1. He would like to acknowledge Harrison Lee and Raghav Gupta from Google Research for support and guidance with D3ST and T5DST implementation. Additionally, he would like to acknowledge the helpful feedback and suggestions that shaped this work from Matt Henderson (reka.ai). Weizhe Lin was supported by a Research Studentship funded by Toyota Motor Europe (RG92562(24020)). We also thank Howard Mei from University of Cambridge for help with editing the final draft. Authors would like to acknowledge the improvement suggestions made by anonymous reviewrs and the SIGDIAL program committee.

References

Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica S. Lam. 2020. Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 122–132. Association for Computational Linguistics.

Jie Cao and Yi Zhang. 2021. A comparative study on schema-guided dialogue state tracking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 782–796. Association for Computational Linguistics.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3002–3017. Association for Computational Linguistics.

Alexandru Coca, Tseng Bo-Hsiang, Lin Weizhe, and Bill Byrne. 2022. Improving generalisation of schema-guided dialogue state tracking via tree-based paraphrase ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 527–533, Abu Dhabi, United Arabic Emirates. Association for Computational Linguistics.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, Dilek Hakkani-Tur, and Amazon Alexa AI. 2019. Dialog state tracking: A neural reading comprehension approach. In *20th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 264.

Raghav Gupta, Harrison Lee, Jeffrey Zhao, Yuan Cao, Abhinav Rastogi, and Yonghui Wu. 2022. Show, don't tell: Demonstrations outperform descriptions for schema-guided task-oriented dialogue. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*,

pages 4541–4549. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Shuo Huang, Zhuang Li, Lizhen Qu, and Lei Pan. 2021. On robustness of neural semantic parsers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3333–3342. Association for Computational Linguistics.

Léo Jacqmin, Lina M. Rojas Barahona, and Benoit Favre. 2022. "do you follow me?": A survey of recent approaches in dialogue state tracking. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 336–350, Edinburgh, UK. Association for Computational Linguistics.

Eleftherios Kapelonis, Efthymios Georgiou, and Alexandros Potamianos. 2022. A multi-task BERT model for schema-guided dialogue state tracking. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2733–2737. ISCA.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4937–4949. Association for Computational Linguistics.

Harrison Lee, Raghav Gupta, Abhinav Rastogi, Yuan Cao, Bin Zhang, and Yonghui Wu. 2022a. SGD-X: A benchmark for robust generalization in schema-guided dialogue systems. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10938–10946. AAAI Press.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022b. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8424–8445. Association for Computational Linguistics.

Zhaojiang Lin, Bing Liu, Andrea Madotto, Seungwhan Moon, Zhenpeng Zhou, Paul A. Crook, Zhiguang Wang, Zhou Yu, Eunjoon Cho, Rajen Subba, and Pascale Fung. 2021a. Zero-shot dialogue state tracking via cross-task transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7890–7900. Association for Computational Linguistics.

Zhaojiang Lin, Bing Liu, Seungwhan Moon, Paul Crook, Zhenpeng Zhou, Zhiguang Wang, Zhou Yu, Andrea Madotto, Eunjoon Cho, and Rajen Subba. 2021b. Leveraging slot descriptions for zero-shot cross-domain dialogue StateTracking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5640–5648, Online. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11048–11064. Association for Computational Linguistics.

Nikola Mrksic, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve J. Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1777–1788. Association for Computational Linguistics.

Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins, and Mirella Lapata. 2022. A well-composed text is half done! composition sampling for diverse conditional generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1319–1339, Dublin, Ireland. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. SOLOIST: few-shot task-oriented dialog with A single pre-trained auto-regressive model. *CoRR*, abs/2005.05298.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence,*

AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8689–8696. AAAI Press.

Stephanie Seneff and Joseph Polifroni. 2000. Dialogue management in the mercury flight reservation system. In *ANLP-NAACL 2000 Workshop: Conversational Systems*.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819.

Dian Yu, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent El Shafey, and Hagen Soltau. 2022. Unsupervised slot schema induction for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1174–1193. Association for Computational Linguistics.

Jeffrey Zhao, Raghav Gupta, Yuan Cao, Dian Yu, Mingqiu Wang, Harrison Lee, Abhinav Rastogi, Izhak Shafran, and Yonghui Wu. 2022. Description-driven task-oriented dialog modeling. *CoRR*, abs/2201.08904.

Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, and Xiang Ren. 2021. Pre-training text-to-text transformers for concept-centric common sense. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

A Turn mining details

For 31 out of the 214 slots there are no knowledge-seeking turns or there are less than 5 distinct knowledge seeking turns[13] in the dialogue corpus. These includes *result* slots which are communicate by the agent upon user query, such as the name or time of an existing alarm in the *Alarms_1* service. These slots do not appear in state annotations. Moreover, SGD dialogue flows are generated by semantic-level interaction between two machines modelled using push-down automata (Rastogi et al., 2020). As such, not all dialogue flows are covered. For example, in the *Alarm_1* service the user always states the name of a new alarm and the time they want to set it for so the system never asks for what time the new alarm should be set.

We circumvent these issues with two simple strategies, which are applied depending on whether a slot has knowledge-seeking turns in other services or not. The majority of the slots fall in the former case.

Turn copy The *only* knowledge seeking turn for the `fare` slot in `Buses_1` service is *Thanks for that, how much did it cost?*. However, price is a generic concept which appears in other services (e.g. `Events_1`) so instead of reducing prompt diversity by always using this turn, we copy knowledge-seeking turns from other services. In this instance, *How much did it cost?*, *Ticket fare for each passenger?*, *Price per ticket?* and *What price?* are copied. This strategy is applied to all slots that appear in other services.

Span selection For just 8 slots, a relevant span appearing before or after the slot value is selected from turns annotated with actions `INFORM(s=v)` or `CONFIRM(s=v)` where s is a slot for which no knowledge-seeking turns exist and v is its value. For example, there are no turns where the system or user ask for the seating class of an airline ticket. We select the span *class flight ticket* instead of a full turn from the system turn *Please confirm an Economy class flight ticket to NY, tomorrow.*. The semantic annotation of this turn is `CONFIRM(destination=NY)`, `CONFIRM(date=tomorrow)`, `CONFIRM(seating_class=Economy)`, `CONFIRM(passengers=1)`.

B Experimental details

B.1 D3ST implementation

We process the data as described by Zhao et al. (2022) with the following differences, which were

[13] Only 25 of these slots are unique as some slots repeat across services.

indicated by the paper authors upon private communication: (1) the indices are separated by the = symbol in both the inputs and the targets, (2) for categorical slots which take the *dontcare* special value, our output contains *slot_index: dontcare* substring and we do not include the special value in the prefix and (3) we lowercase inputs and targets.

The examples are truncated to the last $1,024$ tokens on the input side for the baseline and discarded altogether for Turn/TurnSlot prompt formats[14]. We optimise with the Adafactor optimizer and effective batch size 32, starting from the initial weights `google/t5-v1_1-base` published by `huggingface` (Wolf et al., 2019). We interpolate the learning rate linearly between 0 and 10^{-4} over the first 1000 steps and keep it constant thereafter. We select the model by evaluating the development set joint goal accuracy (JGA) every 5000 gradient updates, stopping the training if said metric fails to improve after 3 consecutive evaluations.

All results in Section 5 are averages of three runs initialised with different random seeds. For all experiments, we used the same hyperparameters and stopping criteria as just described, with the exception of training the D3ST + SGD-X DA and D3ST + KST DA experiments for T5-large (rows 2 in Table 7 and last three lines in 9) where we allow all runs 1 epoch of augmented data (each SGD conversation is seen 6 times) due to limited computational budget.

B.2 MT-SGD implementation

The numbers presented are averages of three runs. The first SGD run (JGA of 83.2%) is based upon a metric file received from the authors. We could only reproduce 82.7% of the quoted number, but we include the higher number in our average. We trained the model using the publicly available code[15] twice more obtaining to obtain 77% and 80.2%. On SGD-X the JGA range is between 54.6% and 66.4% across three runs. We selected the `best` checkpoint as indicated in the repository's instructions.

C Backtranslation experiment

We experiment with larger backtranslation datasets to see if finetuning D3ST on a dataset the same size as the SGD-X dataset (Section 4.2) can improve results. We created two more variants by backtranslating the SGD schema via French and Russian, as done by Huang et al. (2021).

Size	SGD	SGD-X	Seen	Unseen	SS
4x	**72.1**	**62.2**	**84.0**	**54.9**	**53.1**
6x	71.5	61.0	82.5	53.8	54.4

Table 11: SGD and SGD-X JGA with backtranslation datasets of different size. We repeat line 2 from Table 3 in the top row, for easy comparison

Metric	Method	v1	v2	v3	v4	v5
BLEU	Backtranslation 4x	36.4	26.01	18.9	-	-
	Backtranslation 6x	51.3	37.2	29.5	23.4	18.2
self-BLEU	Backtranslation 4x	-	49.3	41.7	-	-
	Backtranslation 6x	-	55.3	49.7	44.6	39.6

Table 12: Lexical diversity metrics of backtranslated prompts. self-BLEU measures diversity of n sentences

Augmenting with these additional examples negatively impacts model robustness (Table 11). This may arise because increasing the number of training examples significantly (Table 12) does not increase the prompt diversity by a large margin, and so the training distribution of the prompts is closer to the training data. Creating a diverse collection of paraphrases via backtranslation is thus challenging, as it requires access to translation systems to high-difficulty languages. This is necessary, since, as shown in Table 12 (BLEU, column 2) translating to high-resourced languages such as French yields paraphrases that are lexically more similar to the input and are not as effective in improving the model robustness. Meanwhile, translation to difficult languages leads to semantic errors which may harm DST. For example, *Station where the bus is leaving from* is backtranslated to *Bus departure/arrival station* and *Station where the bus is going to* is backtranslated as *bus station* (via Japanese).

By grounding the model in turns collected from the corpus, not only do we create diverse inputs, but we guarantee that these correctly represent fine grained semantics and by-pass the issues encountered when constructing prompts via paraphrasing.

D SGD results

In the main body we report the SGD JGA accuracy as an upper bound for the D3ST model robust accuracy. To make our tables readable, we do not include SGD performance breakdown by seen/unseen services in Section 5. We include it in Table D to facilitate future comparisons

[14] This is just around 0.05% of the data.

[15] See it here: `bit.ly/3j8sPwj`

Model	SGD	SGD-Seen	SGD-Unseen
D3ST	69.8	92.8	62.2
D3ST + SGD-X DA	73.8	92.7	67.5
D3ST-Turn	**75.8**	**92.9**	**70.1**
D3ST-TurnSlot	74.7	92.8	68.7
D3ST + KST DA/Turn	74.9	92.6	69.0
D3ST + KST DA/TurnSlot	73.8	92.5	67.6
D3ST + KST DA/D3ST	74.4	92.8	68.3

Table 13: Breakdown on SGD JGA into seen and unseen services JGA for models reported Tables 3 and 8.

Resolving References in Visually-Grounded Dialogue via Text Generation

Bram Willemsen and **Livia Qian** and **Gabriel Skantze**
Division of Speech, Music and Hearing
KTH Royal Institute of Technology
Stockholm, Sweden
{bramw, liviaq, skantze}@kth.se

Abstract

Vision-language models (VLMs) have shown to be effective at image retrieval based on simple text queries, but text-image retrieval based on conversational input remains a challenge. Consequently, if we want to use VLMs for reference resolution in visually-grounded dialogue, the discourse processing capabilities of these models need to be augmented. To address this issue, we propose fine-tuning a causal large language model (LLM) to generate definite descriptions that summarize coreferential information found in the linguistic context of references. We then use a pretrained VLM to identify referents based on the generated descriptions, zero-shot. We evaluate our approach on a manually annotated dataset of visually-grounded dialogues and achieve results that, on average, exceed the performance of the baselines we compare against. Furthermore, we find that using referent descriptions based on larger context windows has the potential to yield higher returns.

1 Introduction

Visually-grounded dialogues are conversations in which participants make references to the visual world. Referring in conversation is understood to be a collaborative process, with shared responsibility for ensuring the successful identification of the referent (Clark and Wilkes-Gibbs, 1986). It is not uncommon for a definite reference to be established over multiple turns, with each separate contribution unlikely to be a minimally distinguishable description of the referent. Taken out of their use context, these referring expressions may be difficult, if not impossible, to resolve. Consider the example dialogue in Figure 1. The underspecified description *"the shiny one"* leads to a clarification question, *"Do you mean that red one?"*. To resolve the expression *"that red one"* to its referent, we need information from earlier in the conversation to understand that *"one"* is a proform of *"apple"*.

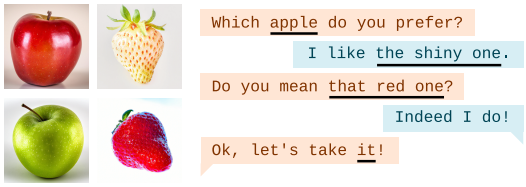


Figure 1: Example dialogue in which two participants discuss fruits. Expressions that denote one or more images are underlined.

Without this linguistic context, the red strawberry and the red apple are equally likely referents.

We can break the problem of reference resolution in visually-grounded dialogue down into three subproblems: (1) mention detection, or finding the expressions that can be grounded in the visual context (*"that red one"*); (2) aggregation of referent-specific information (linking *"apple"*, *"the shiny one"*, and *"that red one"*); and (3) referent identification, or the grounding of language (finding the referent that is best described by the three expressions from among a set of candidate referents). This final step requires bridging the gap between vision and language. For this purpose, we can turn to pretrained vision-language models (VLMs), which have shown to be effective at zero-shot text-image retrieval when given a description of an image (e.g., Radford et al., 2021; Jia et al., 2021; Li et al., 2023). However, current VLMs lack the discourse processing capabilities necessary for reference resolution in visually-grounded dialogue. Although some VLMs may correctly identify the red apple as the referent given the entire dialogue of Figure 1, dialogues are often vastly more complex than this hypothetical exchange. Take, for instance, the dialogue in Appendix A: with multiple mentions of different referents within the same utterance, such a brute-force method would immediately fail. It is clear that if we want VLMs to be effective for this purpose, their discourse processing capabilities need to be augmented.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 457–469
September 11–15, 2023. ©2023 Association for Computational Linguistics

To this end, we propose fine-tuning a causal large language model (LLM) for the task of *referent description generation*. Referent description generation can be regarded as a special case of referring expression generation with the goal of always generating the most complete expression possible. For a given mention, the model is trained to generate a definite description that summarizes all information that has been explicitly disclosed about the referent during a conversation. For example, for the mention *"that red one"* in Figure 1 we would want the model to generate the description *"the shiny red apple"*. We will refer to the fine-tuned model as the *conversational referent description generator* (CRDG). The description generated by the CRDG is then used by a pretrained VLM to identify the referent, zero-shot. Our approach can be seen as an exploration of the limits of depending on linguistic context alone for generating referent descriptions, as the discourse processing and eventual grounding of the descriptions are entirely disjoint.

For the experiments presented in this paper we use data from the collaborative image ranking task A Game Of Sorts (Willemsen et al., 2022). Referents are represented by separate, but visually similar images from a shared entity category. Due to their largely unrestricted nature and with a focus on the collaborative referential process, the collected dialogues form a challenging test bed for visually-grounded language understanding in conversation. We manually annotate the dialogues by marking mention spans and aligning the spans with the images they denote, and provide both manually constructed and automatically derived "ground truth" referent descriptions based on our manual annotations for all marked mentions.

Our main contributions are as follows:

- We present a generative approach to reference resolution in visually-grounded dialogue that frames the discourse processing side of the task as a causal language modeling problem;
- We show that it is possible to fine-tune a causal LLM to generate referent descriptions from dialogue to be used by a pretrained VLM for referent identification, zero-shot;
- We release the discussed materials, including our annotations for A Game Of Sorts (Willemsen et al., 2022)[1].

[1] https://github.com/willemsenbram/reference-resolution-via-text-generation, doi:10.5281/zenodo.8176114

2 Background

Visually-grounded language understanding is fundamental for conversational agents that engage in dialogue involving references to the visual world. Researchers have introduced a variety of tasks that provide data for development and frameworks for evaluation of visually-grounded dialogue models. These tasks often take the form of goal-oriented, dyadic interactions but differ in terms of, for example, the visual stimuli used, e.g. abstract figures or realistic photos; the roles assigned to participants, e.g. whether symmetric or asymmetric; constraints on message content, e.g. a fixed vocabulary; and the nature of the task, e.g. navigation, identification, ranking, and multi-turn visual question answering (e.g. Das et al., 2017; De Vries et al., 2017; Shore et al., 2018; Ilinykh et al., 2019; Haber et al., 2019; Udagawa and Aizawa, 2019; Willemsen et al., 2022). It has been noted that the task configuration can significantly impact the extent to which certain dialogue phenomena, such as coreferences and clarification requests, are represented in the collected data, if at all (Agarwal et al., 2020; Haber et al., 2019; Ilinykh et al., 2019; Schlangen, 2019; Willemsen et al., 2022). Tasks that heavily constrain the interactions do not reflect the complex nature of dialogue to the same degree as tasks that have been designed for these phenomena to naturally emerge as part of the discourse, such as A Game Of Sorts (Willemsen et al., 2022), which we use in this paper.

The terms referring expression comprehension (e.g. Yu et al., 2016), referring expression grounding (e.g. Zhang et al., 2018), referring expression recognition (e.g. Cirik et al., 2018), and reference resolution (e.g. Kennington et al., 2015) have been used interchangeably to describe the problem of mapping the language that denotes a referent to a representation of that referent in the visual modality. Prior work noted the importance of referring expressions to conversation, but often modeled the problem independent of the dialogue (e.g. Cirik et al., 2018; Schlangen et al., 2016; Yu et al., 2016; Zhang et al., 2018). The granularity at which grounding occurs may differ between works, as the language may be mapped to bounding boxes of individual objects (Cirik et al., 2018; Schlangen et al., 2016; Yu et al., 2016; Zhang et al., 2018), objects or larger image regions represented by seg-

mentation masks (Liu et al., 2017), or entire images altogether (Haber et al., 2019; Takmaz et al., 2020).

To address the problem computationally, both modalities must in some way be encoded. Engineered visual feature representations and simple language models such as those based on n-grams (e.g. Kennington et al., 2015; Kennington and Schlangen, 2017; Shore and Skantze, 2018) have been mostly replaced with more powerful learned representations that embed the images and text in high-dimensional vector spaces (Haber et al., 2019; Takmaz et al., 2020). This has made it possible to resolve references by computing representational similarity between an encoding of the text that contains a mention and the embeddings of the candidate referents, where the candidate that has the highest matching score is assumed to be the referent (Haber et al., 2019; Takmaz et al., 2020).

Recent work on multimodal representation learning has shown that jointly embedding text and images can work at scale. Trained using a contrastive objective, maximizing representational similarity between true pairings of images and text while simultaneously minimizing similarity of false pairs, vision-language models (VLMs) such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), BLIP (Li et al., 2022), and BLIP-2 (Li et al., 2023), have shown to be effective zero-shot classifiers, outperforming the previous state-of-the-art on various benchmarks without the need for further fine-tuning on specific tasks. However, despite their noteworthy image-text matching performance based on simple text queries, these VLMs lack the discourse processing capabilities required for reference resolution in visually-grounded dialogue. Even a simplified example, such as shown in Figure 1, illustrates a fundamental challenge, namely that of coreference resolution. The interpretation of anaphoric pronouns, such as *"it"*, is dependent on their antecedents. Without resolving its coreferences first, identifying the referent based on the pronoun alone leads to a random guess.

To improve downstream performance on discourse processing tasks involving coreference, prior work has approached the problem as one of transforming the original input based on linguistic context. This was done either via substitution, such as in Bhattacharjee et al. (2020) where pronouns were substituted for more descriptive mentions of the same referent, or via generation, such as in Quan et al. (2019) where entire utterances were reconstructed in a pragmatically complete manner with coreferences and ellipses resolved. To the best of our knowledge, this approach has not yet been applied to reference resolution in visually-grounded dialogue.

Most contemporary natural language processing (NLP) works use Transformer-based language models (Vaswani et al., 2017). For text generation tasks, it is common to use (unidirectional) autoregressive, or *causal*, language models such as GPT (Radford et al., 2018). While processing sequences, causal language models mask the future, allowing the model to only attend to the current and previous tokens while predicting the next token. A persistent trend has been to scale up language models, both in terms of their parameter count and the size of their training datasets. These increasingly larger models, such as GPT-3 (Brown et al., 2020), OPT (Zhang et al., 2022), PaLM (Chowdhery et al., 2022), and LLaMa (Touvron et al., 2023), have been dubbed *large language models* (LLMs). The current leading paradigm to modeling downstream NLP tasks is based on transfer learning, where a pretrained LLM is fine-tuned for a specific task on a smaller, domain-specific dataset.

3 Method

We treat visually-grounded reference resolution as a text-image retrieval task, where referents are represented by images. We leave finer-grained grounding of words and phrases to image regions or individual entities or parts thereof for future work.

3.1 Proposed Framework

We frame the discourse processing side of the task as a causal language modeling problem. Figure 2 shows a visualization of the proposed framework.
Task Definition We denote the dialogue as $D = (u_1, u_2, ..., u_n)$, where each u_i represents an utterance. Each utterance consists of an ordered sequence of tokens. An utterance may contain one or more mentions, denoted as M. A mention is an ordered subsequence of tokens from an utterance. A mention has an exophoric referent, denoted as R. A mention is embedded in what we call its linguistic context, denoted as L. As an ordered subsequence of D, the linguistic context of a given mention consists of the utterance in which it is contained and all preceding utterances. The number of preceding utterances, hereafter referred to as the dialogue history, may be capped if a finite size context window

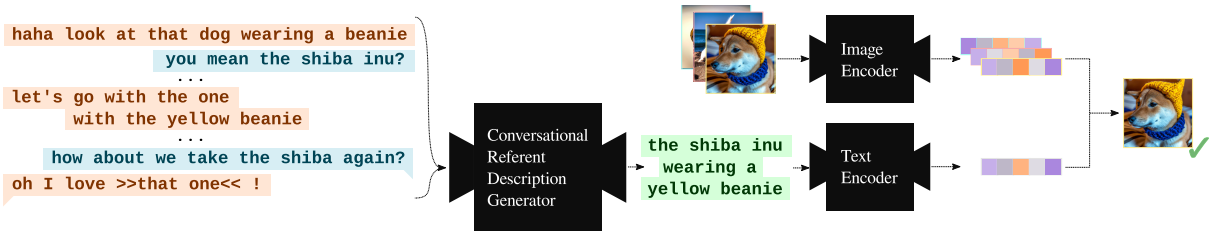


Figure 2: The proposed visually-grounded reference resolution framework. With the CRDG we generate a referent description for a marked mention, to be used by a (frozen) pretrained VLM for referent identification.

is defined. The aim of visually-grounded reference resolution is to resolve a reference to its referent, i.e. to identify R for a given M, from a set of candidate referents, denoted as C, such that $R \subseteq C$; $|R| = 1$ for single-image referents, $|R| > 1$ for multi-image referents, and $R = C$ if M refers to all members in C.

Referent Description Generation We propose to generate a definite description, denoted as Y, for a given mention M that summarizes all that has been disclosed in L about the referent R. For this purpose, we fine-tune a causal LLM that learns to generate Y conditioned on L. Y is a sequence of tokens expected to be largely constructed from tokens that appear, or are some derivative of tokens that appear, in the coreference chain of R, which is contained in L. We refer to the fine-tuned model as the *conversational referent description generator* (CRDG). For an example of the context dependency of referent description content, see Figure 4 in Appendix B.

LLM Input We mark M in u_i by inserting positional markers as special tokens to indicate the beginning and end of the mention span. We prepend each utterance in L with a speaker token to indicate the source of the contribution. When D is task-oriented, we update L by prepending task instructions, i.e. a special token followed by a sequence of tokens describing the task performed by the dialogue participants. For an example of the input to the LLM, see Figure 5 in Appendix B.

Text-Image Retrieval We use a pretrained VLM to identify R from C based on Y, zero-shot. We use the text encoder of the VLM to encode Y into an n-dimensional feature vector, denoted as $\mathbf{v}$. We use the image encoder of the VLM to encode each candidate referent of C into an n-dimensional feature vector, which gives a $|C| \times n$ matrix, denoted as $\mathbf{A}$. We then compute their matrix-vector product. For single-image referents, i.e. when $|R| = 1$, we take the referent to be $R = argmax(\mathbf{Av})$.

In order to produce accurate referent descriptions, the CRDG must implicitly learn to perform coreference resolution as we do not provide explicit supervision for this subtask. In each sample, only the current mention for which we want the model to generate a description is marked; none of its coreferences are in any way indicated. A principal advantage of our model is that it can resolve multiple mentions, even when they have different referents, appearing in the same utterance, including nested mentions. Note that for the purpose of this study, we assume mention detection to be solved. As it stands, using this framework in production requires a separate model to propose candidate mentions at the span level.

3.2 Baseline Models

As a lower bound, we report random chance performance. In addition, we compare performance of our approach to baselines based on simple heuristics and a coreference resolution model.

3.2.1 Heuristics

Mention We evaluate the image retrieval performance when the VLMs are presented with just the marked mentions.

Substitution We improve upon the mention-only baseline by substituting proforms, e.g. pronouns such as *"it"*, and mentions without descriptive content, e.g. phrases such as *"the one you mentioned"*, with the most recent mention that does not belong to either category. This is expected to be a relatively strong baseline when mentions are specific and anaphora have mostly local antecedents.

3.2.2 Coreference Resolution

We opt for an off-the-shelf[2] span-based coreference resolution model (**coref**) originally presented in Lee et al. (2018), but that has since been updated

[2] https://github.com/allenai/allennlp-models/tree/main/allennlp_models/coref

to use SpanBERT (Joshi et al., 2020) instead of the original GloVe embeddings (Pennington et al., 2014). For each mention, we use the model to resolve its coreference links and aggregate all coreferential information in its cluster based on the given context window.

We experiment with two different representations of the referent descriptions from this model, those being (1) a concatenation of all of the mention's coreferences and (2) an ordered *set-of-words* representation that contains only the unique lexical items in the cluster. To offset that this model was not specifically trained to handle coreference in conversation, we provide it with the contents of the span of the mention when it does not manage to detect the mention itself and, consequently, not connect it to any of its coreferences. For partial matches, in addition to adding all tokens from the cluster associated with the match, we also add the missing tokens from the span to the description.

4 Experiments

4.1 Data

We use the dialogues from the collaborative image ranking task **A Game Of Sorts** (AGOS, Willemsen et al., 2022) for our experiments. In AGOS, two players are asked to rank a set of images based on a given sorting criterion. They see the same set of images, but the position of the images on the screen is randomized for each player. Through a largely unrestricted conversation, and without being able to see the perspective of the other player, the players need to agree on how to rank the images given the sorting criterion. Sorting criteria are embedded in scenarios that are intended to create a discussion, leading to mixed-initiative interactions with both parties contributing to the discourse. Each interaction takes place over four rounds with the same set of nine images, effectively guaranteeing repeated references. The image sets used for the game cover five different image categories. Each set contains nine images with each image representing an entity from one of these categories as its main subject. Willemsen et al. (2022) collected three interactions per image set for a total of 15 dialogues.

Ground Truth Our formulation of the visually-grounded reference resolution problem requires span-based annotations of mentions aligned with the image(s) they denote. These annotations are the basis of what we will refer to as our "ground truth" references used for both training and evaluation. We follow Willemsen et al. (2022) regarding the marking of mentions in AGOS, in that we only annotate those that are either singletons or are part of an identity relation with other mentions that have an exophoric referent that is part of the visual context, i.e. regardless of form, any referring expression that is meant to denote one or more of the images. During the game, players were asked to provide self-annotations: for each message they sent they were asked to indicate which image(s), if any, they were referring to. We use these self-annotations, post-edited where necessary, to manually mark the spans of mentions that can be grounded in the visual context.

We create three different representations of the "ground truth" referent descriptions. Two are automatically extracted from the marked mentions and are similar in structure to the labels of the **coref** baseline, i.e. (1) an incremental concatenation of the reference chain and (2) an incremental ordered set of words consisting of the unique lexical items in the cluster. The third are manually constructed labels that summarize reference chains as definite descriptions. For each representation, the context window dictates which references are considered for the label.

4.2 Model Specifications

For pointers to implementations, we refer the reader to our repository[1].

4.2.1 LLMs

We fine-tune two LLMs, GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020), for conversational referent description generation. For hyperparameters, see our Supplementary Material.

GPT-2 We fine-tune the 1.5 billion parameter GPT-2 model.

GPT-3 We fine-tune the 175 billion parameter `davinci` base model using the OpenAI API.

4.2.2 VLMs

We evaluate the zero-shot text-image retrieval performance of several pretrained VLMs for our task, those being CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), BLIP (Li et al., 2022), and BLIP-2 (Li et al., 2023).

CLIP We evaluate two variants of CLIP, CLIP ViT-B/32 and CLIP ViT-L/14.

ALIGN We use the COYO-ALIGN implementation trained from scratch on COYO-700M.

BLIP We use the BLIP base model.

BLIP-2 We use the BLIP-2 model that was fine-tuned on the Karpathy and Fei-Fei (2015) training set split of MS COCO (Lin et al., 2014).

4.3 Evaluation

We perform (nested) five-fold cross-validation by partitioning the AGOS dataset along the five image sets. To avoid leakage, for each run we use the three dialogues from one image set as the held out test set and train on the twelve dialogues from the four other image sets. To evaluate how dialogue history affects results, we report performance of the different methods for two context windows, **3** and **7**. In addition, we examine whether increasing the size of the context window further would, in principle, lead to greater returns, by assessing ground-truth performance for windows of **13** and the **full** dialogue context. Finally, we conduct an error analysis of the generated descriptions.

Note that because we do not incorporate game state information with respect to the visual context during training, we make a simplifying assumption with regard to the images and reduce the candidate set, at test time, as the game progresses. A successfully ranked image is no longer considered part of the visual context for that round. Although this does mean that the models will not be able to identify the referent for references to ranked images, as they will not be part of the candidate set, such references are an extremely rare occurrence, as players must discuss the unranked images to progress with the task. For the sake of completeness, we will also report results for the unchanged candidate set.

4.3.1 Metrics

We measure task success for visually-grounded reference resolution in terms of text-image retrieval performance. In addition, we estimate the quality of the generated referent descriptions by comparing them to the manually constructed ground truth labels using text similarity metrics.

Text-Image Retrieval We estimate the image retrieval performance based on accuracy $[0, 1]$, mean reciprocal rank (MRR) $[0, 1]$, and normalized discounted cumulative gain (NDCG) $[0, 1]$. We limit our evaluation to single-image referents. Accuracy is top-1 accuracy.

For our random lower bound, we can calculate the expected values for accuracy and MRR. For top-1 accuracy we take 1 over the size of the set of candidate images per item, averaging over all items. For MRR we take 1 over the size of the set of candidate images, divided by two per item, averaging over all items. Calculating an expected value for NDCG of a random model is intractable due to its dependence on relevancy scores.

Text Generation We evaluate the output from the CRDGs by comparing the generated descriptions to the manually constructed ground truth labels using metrics to quantify similarity. We use the Jaccard index $[0, 1]$ to assess vocabulary overlap. We use BLEU $[0, 1]$ (Papineni et al., 2002) to assess similarity based on n-gram overlap (unigrams to four-grams). We use the longest common subsequence variant of ROUGE $[0, 1]$ (Lin, 2004), i.e. ROUGE-L, as a further indication of the preservation of word order. In addition, we opt for an embedding-based metric as a proxy for semantic equivalence between the predicted and reference sequences. For this purpose, we compute cosine similarity $[0, 1]$ between text embeddings.

4.3.2 Human

We conduct two different human subject experiments to assess human performance for this task. We provide additional details about the experimental setup in the Supplementary Material.

Independent We conduct an experiment aimed at comparing VLM and human performance on the task where every trial is independent. Participants are given a referent description and are asked to select from a set of candidate images the image they believe is best described by the label. The images and labels are presented to the participants independent of the dialogue. Note that we evaluate with the reduced candidate set. The referent descriptions are the manually constructed ground truth labels based on the **full** dialogue context. To collect data for all labels, ensuring independence of observations, we recruited 354 participants via crowdsourcing. The crowdworkers were financially compensated for their contributions.

Holistic We conduct an experiment in which mentions are shown to participants within the context of the dialogue. For each mention, the participants are presented with the dialogue leading up to and including the message which contains the reference. The start and end of the span of the mention that the participant is asked to resolve are visually indicated. For each marked mention, the participant is asked to select which image or images are referenced. As they progress with the task, participants will have access to increasingly more of the dialogue history. For each mention the participants

	Accuracy		MRR		NDCG	
	3	**7**	**3**	**7**	**3**	**7**
Random	.22	.22	.43	.43	-	-
Mention	.59	.59	.73	.73	.79	.79
Substitution	.68	.68	.80	.80	.85	.85
coref, chain	.65	.66	.78	.79	.83	.84
coref, set	.66	.66	.78	.79	.84	.84
GT, chain	.73	.74	.83	.85	.87	.88
GT, set	.73	.75	.84	.85	.87	.89
GT, manual	.72	.74	.83	.84	.87	.88
GPT-2	.64	.60	.77	.74	.83	.80
GPT-3	.69	.71	.81	.82	.86	.86

Table 1: Cross-validated image retrieval performance averaged over five folds for single-image referents. *Note*. Scores shown are of VLM that averaged best performance (BLIP-2). Scores are rounded to the nearest hundredth. GT = ground truth.

	GPT-2		GPT-3	
	3	**7**	**3**	**7**
BLEU	.55	.47	.75	.70
ROUGE-L	.71	.65	.86	.83
Jaccard	.44	.35	.70	.63
Cosine	.88	.85	.96	.95

Table 2: Text generation metrics evaluation results averaged over five folds for single-image referents. *Note*. Scores are rounded to the nearest hundredth.

are presented with all images, but with a visual indication of their status, i.e. for each image whether the players had managed to successfully rank it at that point in the interaction. We recruited 23 participants via crowdsourcing. For each of the 15 AGOS dialogues we collected data from two different participants. Each participant was allowed to provide data for at most one dialogue per image set. The crowdworkers were financially compensated for their contributions.

5 Results

5.1 Text-Image Retrieval

Table 1 shows, for context windows **3** and **7**, the zero-shot text-image retrieval performance results for the VLM that averaged best performance over the five folds, which was BLIP-2. For the text-image retrieval accuracy achieved by the other VLMs, performance on the not reduced candidate set, and accuracy per fold for BLIP-2, see Appendix C.

As can be seen from the results presented in Table 1, we achieve best performance with a fine-tuned GPT-3 as the CRDG and BLIP-2 for zero-shot text-image retrieval. In addition to outperforming the baselines, we find that GPT-3 is a more performant discourse processor for this task than GPT-2. This result is consistent between the VLMs.

Results generally show a slight increase in performance when increasing the context window from **3** to **7**. Performance on the ground truth reference descriptions for context windows **13** and the **full** dialogue shows this trend persists, with BLIP-2 achieving approximately 75% and 83% accuracy, respectively. A plot of the performance for the four context windows is shown in Figure 6 in Appendix C. This result suggests that the size of the context window may have a significant impact on performance, with an 11% increase in accuracy from **3** to **full**. Furthermore, the VLMs do not seem overly sensitive to the composition of the referent descriptions, as performance is largely comparable between the automatically generated and the manually constructed ground truth labels.

We find that BLIP-2 is on par with human text-image retrieval performance in terms of top-1 accuracy for the manually constructed ground truth referent descriptions based on the full dialogue history for single-image referents, as our human participants averaged roughly 80% accuracy in the independent setup. However, when we compare these results with the single-image referent text-image retrieval performance in the holistic setup, we see that the upper bound for this task when references are resolved within the combined linguistic and extralinguistic dialogue context is likely considerably higher as our human participants averaged approximately 91% accuracy (average of best performance per dialogue is roughly 93%).

5.2 Text Generation

Table 2 shows the text generation metric results averaged over the five folds, providing an indication of the extent to which the fine-tuned LLMs managed to generate referent descriptions that approximate the manually constructed ground truth labels. We observe that an increase in context window size results in a decrease in scores, which is consistent across metrics. Interestingly, we did not find such a decrease with respect to text-image retrieval performance. We do again find GPT-3 to be more performant than GPT-2, here in terms of approximating the ground truth.

5.3 Error Analysis

Examining the output from the fine-tuned GPT-3 model, we observe a number of recurring errors.

The most notable errors are those where the model fails to link a mention to (all of) its coreferences that are present in the dialogue segment, or links mentions that denote different referents. For example, for one mention the ground truth label is *"the sheep dog"*, but the generated label was *"the sheep dog with a leash"*; the model incorrectly attributed the prepositional phrase to the mention as it was actually a descriptor for a different referent. Related, since the CRDGs function at the message level, a mention can have both anaphoric and cataphoric coreferences when there are multiple mentions of the same referent in an utterance An example of such an utterance is *"Good question. I think the angry one also looks a little wild. So that could be an option as well. I mean the one with white nose and forehead"*, where *"the angry one"*, *"that"*, and *"the one with white nose and forehead"* are all mentions of the same referent with the same ground truth label *"the angry dog with a white nose and forehead"*. The model generates this correctly for the latter two, but not the former one for which only *"the angry dog"* was generated, meaning it correctly substituted the proform but did not link the mention with its cataphoric coreferences.

Finally, some generated referent descriptions differ from the ground truth in terms of lexical choice or syntax, but not in terms of information content. This negatively affects scores of text generation metrics based on overlapping content in particular, but these are otherwise not meaningful errors as there are multiple ways to construct semantically similar descriptions, e.g., *"the big dog which looks scary"* versus *"the big scary-looking dog"*.

6 Discussion

We have presented an approach to visually-grounded reference resolution that frames the discourse processing side of the task as a causal language modeling problem. By fine-tuning an LLM to generate referent descriptions for marked mentions in dialogue segments from the collaborative image ranking task A Game Of Sorts (Willemsen et al., 2022), we demonstrate the possibility of treating referent identification as a zero-shot text-image retrieval problem by using pretrained VLMs for the grounding of the generated labels. As we have not in any way indicated coreferential relations in the fine-tuning training data, our results imply that certain pretrained LLMs, here GPT-3, may learn to resolve coreferences implicitly without the need for explicit supervision for this fundamental subtask.

In this work, we have treated the processing of the discourse as entirely disjoint of the visual modality. As such, it has inherent limitations. The mentions we find in the dialogues have not been produced void of the extralinguistic context. The dialogue participants could rely on co-observed visual stimuli to help resolve otherwise ambiguous language use. From linguistic context alone, some ambiguities, such as prepositional phrase attachment, may be impossible to resolve. It is, therefore, noteworthy that the downstream zero-shot text-image retrieval performance using the generated descriptions from our unimodal approach far exceeds chance level accuracy, with the potential for results to improve further given access to the full dialogue history, as we found that the ground truth labels based on larger context windows achieve greater text-image retrieval performance. However, the results from our holistic human evaluation support the notion that a multimodal approach should ultimately prove even more effective.

We found that a decrease in text generation metric scores did not necessarily indicate a similar decrease in text-image retrieval performance, suggesting that the generated descriptions captured sufficiently discriminative information about the referents and achieved similar grounding accuracy despite not approximating the ground truth labels to the same extent. It is also important to note that mentions may not have a single, canonical ground truth referent description due to lexical and syntactic variations between referring attempts.

Despite the relatively small size of the dataset collected by Willemsen et al. (2022), we were still able to fine-tune GPT-3 to perform the task with greater accuracy than the baselines, which speaks to the sample efficiency of (certain) pretrained LLMs. In comparison, we find that the much smaller GPT-2 is prone to intrusions from the fine-tuning training data and more often fails to resolve the coreferences correctly. Although the complexity of the discourse warrants the use of more powerful models, it is, nevertheless, likely that any LLM used for the task would benefit from a larger fine-tuning dataset. Related, benchmarking performance on other visually-grounded dialogue tasks would provide insights into the generalizability of the method.

In addition to pursuing a multimodal approach, finer-grained grounding, and evaluating our method

on other datasets, possible avenues for future work include expanding the annotations to include coreferential relations other than identity relations, addressing multi-image referents, and unifying the method with a mention proposal system.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The authors would like to thank Erik Ekstedt, Dmytro Kalpakchi, Rajmund Nagy, Jim O'Regan, Ambika Kirkland, Chris Emmery, Chris van der Lee, and the anonymous reviewers for their helpful comments.

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for Visual Dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. Association for Computational Linguistics.

Santanu Bhattacharjee, Rejwanul Haque, Gideon Maillette de Buy Wenniger, and Andy Way. 2020. Investigating Query Expansion and Coreference Resolution in Question Answering on BERT. In *Natural Language Processing and Information Systems*, pages 47–59, Cham. Springer International Publishing.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual Referring Expression Recognition: What Do Systems Actually Learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 781–787, New Orleans, Louisiana. Association for Computational Linguistics.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual Object Discovery through Multi-modal Dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4466–4475.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meet Up! A Corpus of Joint Activity Dialogues in a Visual Environment. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, London, United Kingdom. SEMDIAL.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and

Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Casey Kennington, Livia Dia, and David Schlangen. 2015. A Discriminative Model for Perceptually-Grounded Incremental Reference Resolution. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 195–205, London, UK. Association for Computational Linguistics.

Casey Kennington and David Schlangen. 2017. A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech & Language*, 41:43–67.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. 2017. Recurrent Multimodal Interaction for Referring Image Segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1280–1289.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. GECOR: An End-to-End Generative Ellipsis and Co-reference Resolution Model for Task-Oriented Dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4547–4557, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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

David Schlangen. 2019. Grounded Agreement Games: Emphasizing Conversational Grounding in Visual Dialogue Settings. *CoRR*, abs/1908.11279.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1213–1223, Berlin, Germany. Association for Computational Linguistics.

Todd Shore, Theofronia Androulakaki, and Gabriel Skantze. 2018. KTH Tangrams: A Dataset for Research on Alignment and Conceptual Pacts in Task-Oriented Dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Todd Shore and Gabriel Skantze. 2018. Using Lexical Alignment and Referring Ability to Address Data Sparsity in Situated Dialog Reference Resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2288–2297, Brussels, Belgium. Association for Computational Linguistics.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer,

Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Takuma Udagawa and Akiko Aizawa. 2019. A Natural Language Corpus of Common Grounding under Continuous and Partially-Observable Context. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press. Event-place: Honolulu, Hawaii, USA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bram Willemsen, Dmytro Kalpakchi, and Gabriel Skantze. 2022. Collecting Visually-Grounded Dialogue with A Game Of Sorts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2257–2268, Marseille, France. European Language Resources Association.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling Context in Referring Expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham. Springer International Publishing.

Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. 2018. Grounding Referring Expressions in Images by Variational Context. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4158–4166.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models.

Appendices

A Dialogue Excerpt

the blackberry with the keyboard looks older than others

yes and I think either the sliver one or the black one is okay

the two are very similar

lets rank the nokia No3 and the blackberry No4

sorry, I don't get which nokia you are talking about. the grey one or the black one?

I am talking about the silver Nokia and the silver blackberry

Figure 3: Excerpt of an AGOS dialogue with references to single-image referents underlined; the color indicates the referent. *Note*. The two images that have been ranked successfully at this point in the interaction have a faded appearance.

B Model Input

```
A:    Which one do you think is the smallest?
B:    the black ear beagle!                    -> the beagle dog with black ears
B:    looking to the left                      -> the beagle dog with black ears looking to the left
A:    The one ears longer than the head?       -> the beagle dog with black ears longer than its head looking to the left
B:    yes
```

Figure 4: Excerpt of an AGOS dialogue with messages paired with manually constructed ground truth referent descriptions. Mentions are in purple and made bold for illustrative purposes. Ground truth labels for the referent denoted by the mention in green.

```
                    "Speaker" token + task instructions   <N> You are looking to hang a picture on your wall,
                                                              but you have no hammer at your disposal to put the nail in the wall.
                                                              Which of these phones would you consider most suitable to use as an impromptu hammer and why?
                                  Speaker token + utterance   <A> both of them seem quite hard to use. Yeah let's choose bigger one first
                                  Speaker token + utterance   <B> I heard Nokia is pretty solid
                                  Speaker token + utterance   <A> yeah I was thinking the same
          Speaker token + utterance with a marked mention   <B> Maybe >>the one with rubber<< ? Easier to grab
                                               Inference token   ->
Ground truth referent description for the marked mention   the phone with rubber END
```

Figure 5: Sample input to LLM, deconstructed for demonstration purposes (the sample is otherwise a flat sequence of tokens). Left (text in purple): explanation of input; right (text in black): input. *Note*. The ground truth is only available to the model during training, not during inference.

C Additional VLM Results

	CLIP-B		CLIP-L		ALIGN		BLIP	
	3	**7**	**3**	**7**	**3**	**7**	**3**	**7**
Random	.11	.11	.11	.11	.11	.11	.11	.11
Mention	.36	.36	.44	.44	.44	.44	.40	.40
Substitution	.42	.42	.51	.51	.52	.52	.50	.50
coref, chain	.42	.42	.49	.49	.47	.46	.47	.46
coref, set	.42	.41	.48	.48	.49	.48	.47	.47
GT, chain	.45	.47	.54	.56	.53	.53	.52	.54
GT, set	.46	.48	.54	.56	.54	.54	.53	.55
GT, manual	.47	.48	.53	.55	.58	.59	.55	.57
GPT-2	.41	.38	.46	.43	.49	.44	.47	.43
GPT-3	.44	.45	.52	.52	.54	.55	.52	.52

Table 3: Cross-validated image retrieval accuracy averaged over five folds for single-image referents (candidate set not reduced). *Note*. Scores are rounded to the nearest hundredth. GT = ground truth; CLIP-B = CLIP ViT-B/32; CLIP-L = CLIP ViT-L/14.

	CLIP-B		CLIP-L		ALIGN		BLIP	
	3	**7**	**3**	**7**	**3**	**7**	**3**	**7**
Random	.22	.22	.22	.22	.22	.22	.22	.22
Mention	.49	.49	.55	.55	.56	.56	.54	.54
Substitution	.56	.56	.62	.62	.64	.64	.64	.64
coref, chain	.54	.54	.61	.61	.60	.60	.61	.61
coref, set	.54	.53	.60	.60	.61	.61	.61	.61
GT, chain	.58	.59	.66	.67	.66	.67	.66	.68
GT, set	.58	.60	.66	.68	.67	.67	.66	.69
GT, manual	.59	.60	.64	.66	.69	.70	.69	.70
GPT-2	.53	.49	.58	.54	.61	.58	.60	.58
GPT-3	.57	.58	.63	.63	.66	.66	.67	.67

Table 4: Cross-validated image retrieval accuracy averaged over five folds for single-image referents (candidate set reduced). *Note*. Scores are rounded to the nearest hundredth. GT = ground truth; CLIP-B = CLIP ViT-B/32; CLIP-L = CLIP ViT-L/14.

	Cars		Dogs		Paintings		Pastries		Phones	
	3	**7**	**3**	**7**	**3**	**7**	**3**	**7**	**3**	**7**
Random	.22	.22	.22	.22	.22	.22	.22	.22	.22	.22
Mention	.52	.52	.62	.62	.60	.60	.61	.61	.58	.58
Substitution	.63	.63	.70	.70	.70	.70	.68	.68	.67	.67
coref, chain	.59	.60	.69	.69	.66	.67	.67	.68	.63	.63
coref, set	.60	.57	.68	.68	.69	.68	.69	.70	.62	.62
GT, chain	.66	.66	.76	.78	.72	.74	.75	.78	.71	.69
GT, set	.66	.65	.74	.77	.73	.78	.76	.80	.73	.73
GT, manual	.64	.63	.75	.78	.77	.80	.70	.72	.74	.74
GPT-2	.62	.62	.67	.62	.67	.62	.63	.61	.57	.50
GPT-3	.63	.63	.75	.78	.70	.70	.68	.72	.70	.69

Table 5: Cross-validated image retrieval accuracy per fold for single-image referents (candidate set reduced). *Note*. Scores shown are of VLM that averaged best performance (BLIP-2). Scores are rounded to the nearest hundredth. GT = ground truth.

	Accuracy		MRR		NDCG	
	3	**7**	**3**	**7**	**3**	**7**
Random	.11	.11	.22	.22	-	-
Mention	.47	.47	.63	.63	.72	.72
Substitution	.55	.55	.71	.71	.78	.78
coref, chain	.53	.51	.69	.68	.76	.76
coref, set	.53	.51	.69	.68	.77	.76
GT, chain	.60	.61	.75	.76	.81	.82
GT, set	.60	.62	.75	.77	.81	.83
GT, manual	.63	.64	.76	.78	.82	.83
GPT-2	.54	.48	.69	.65	.77	.73
GPT-3	.60	.60	.74	.74	.80	.81

Table 6: Cross-validated image retrieval performance averaged over five folds for single-image referents (candidate set not reduced). *Note*. Scores shown are of VLM that averaged best performance (BLIP-2). Scores are rounded to the nearest hundredth. GT = ground truth.

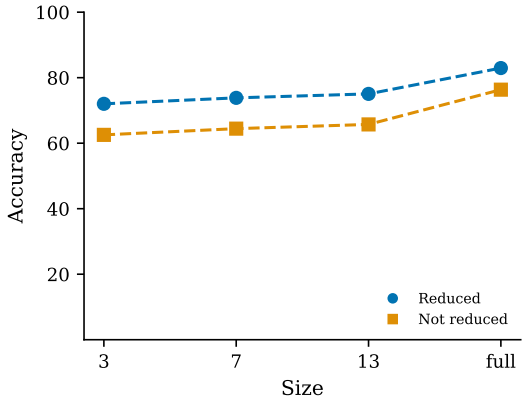


Figure 6: Text-image retrieval accuracy as a function of the size of the context window. Results are shown for BLIP-2 on the manually constructed ground truth referent descriptions based on their respective windows. We show results for both the reduced candidate set and the not reduced candidate set.

Slot Induction via Pre-trained Language Model Probing and Multi-level Contrastive Learning

Hoang H. Nguyen[1], **Chenwei Zhang**[2], **Ye Liu**[3], **Philip S. Yu**[1]
[1] Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA
[2] Amazon, Seattle, WA, USA
[3] Salesforce Research, Palo Alto, CA, USA
{hnguy7,psyu}@uic.edu, cwzhang@amazon.com, yeliu@salesforce.com

Abstract

Recent advanced methods in Natural Language Understanding for Task-oriented Dialogue (TOD) Systems (e.g., intent detection and slot filling) require a large amount of annotated data to achieve competitive performance. In reality, token-level annotations (slot labels) are time-consuming and difficult to acquire. In this work, we study the Slot Induction (SI) task whose objective is to induce slot boundaries without explicit knowledge of token-level slot annotations. We propose leveraging Unsupervised Pre-trained Language Model (PLM) Probing and Contrastive Learning mechanism to exploit (1) unsupervised semantic knowledge extracted from PLM, and (2) additional sentence-level intent label signals available from TOD. Our approach is shown to be effective in SI task and capable of bridging the gaps with token-level supervised models on two NLU benchmark datasets. When generalized to emerging intents, our SI objectives also provide enhanced slot label representations, leading to improved performance on the Slot Filling tasks. [1]

1 Introduction

Natural Language Understanding (NLU) has become a crucial component of the Task-oriented Dialogue (TOD) Systems. The goal of NLU is to extract and capture semantics from users' utterances [2]. There are two major tasks in NLU framework, including intent detection (ID) and slot filling (SF) (Tur and De Mori, 2011). While the former focuses on identifying overall users' intents, the latter extracts semantic concepts from natural language sentences. In NLU tasks, intents denote sentence-level annotations while slot types represent token-level labels.

Despite recent advances, state-of-the-art NLU methods (Haihong et al., 2019; Goo et al., 2018) require a large amount of annotated data to achieve competitive performance. However, the fact that annotations, especially token-level labels, are expensive and time-consuming to acquire severely inhibits the generalization capability of traditional NLU models in an open-world setting (Louvan and Magnini, 2020; Xia et al., 2020). Recent works attempt at tackling the problems in low-resource settings on both intent level (Xia et al., 2018; Nguyen et al., 2020; Siddique et al., 2021) and slot level (Yu et al., 2021; Glass et al., 2021). However, most approaches remain restricted to closed-world settings where there exist pre-defined sets of seen and emerging sets of classes. Some approaches even require additional knowledge from related token-level tasks that might not be readily available.

Additionally, with increasing exposure to the ever-growing number of intents and slots, TOD systems are expected to acquire task-oriented adaptation capability by leveraging both inherent semantic language understanding and task-specific knowledge to identify the crucial emerging concepts in the users' utterances. This ability can be referred to as **Slot Induction** in TOD Systems.

Recently, Pre-trained Contextualized Language Models (PLM) such as BERT (Devlin et al., 2019) have shown promising capability of capturing semantic and syntactic structure without explicit linguistic pre-training objectives (Jawahar et al., 2019; Rogers et al., 2020; Wu et al., 2020b). Despite imperfections, the captured semantics from PLM via unsupervised probing mechanisms could be leveraged to induce important semantic phrases covering token-level slot labels.

Additionally, as an effective unsupervised representation learning mechanism (Wei and Zou, 2019; Gao et al., 2021), Contrastive Learning (CL) is capable of refining the imperfect PLM semantic phrases in a self-supervised manner to mitigate biases existent in the PLM. In specific, given a sample phrase *in the same area* corresponding to

[1] Our code and datasets are publicly available at https://github.com/nhhoang96/MultiCL_Slot_Induction

[2] In our work, we use the term **utterance** and **sentence** interchangeably.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 470–481
September 11–15, 2023. ©2023 Association for Computational Linguistics

spatial_relation slot type, as a presumed structural knowledge, PLM tends to split the preposition and determiner from the noun phrase during segmentation, resulting in *in the* and *same area*. Despite its structural correctness, the identified segments fail to align with ground truth slots due to the lack of knowledge from the overall utterance semantics.

On the other hand, CL can also be leveraged on a sentence level when intent labels are available. In fact, there exist strong connections between slot and intent labels (Zhang et al., 2019; Wu et al., 2020a). For instance, utterances with *book_restaurant* intent tend to contain *location* slots than those from *rate_book* intent. Therefore, as intent labels are less expensive to acquire, they could provide additional signals for CL to induce slot labels more effectively when available.

In this work, we propose leveraging PLM probing together with CL objectives for Slot Induction (SI) task. Despite imperfections, PLM-derived segmentations could produce substantial guidance for SI when slot labels are not readily available. We introduce CL to further refine PLM segmentations via (1) segment-level supervision from unsupervised PLM itself, and (2) sentence-level supervision from intent labels to exploit the semantic connections between slots and intents. Our refined BERT from SI objectives can produce effective slot representations, leading to improved performance in slot-related tasks when generalized towards emerging intents.

Our contributions can be summarized as follows:

• We propose leveraging semantic segments derived from Unsupervised PLM Probing (UPL) to induce phrases covering token-level slot labels. We name the task as Slot Induction.

• We propose enhancing the quality of PLM segments with Contrastive Learning refinement to better exploit (1) unsupervised segment-level signals from PLM, (2) sentence-level signals from intent labels to improve SI performance.

• We showcase the effectiveness of our proposed SI framework and its ability to produce refined PLM representations for token-level slots when generalized to emerging intents.

2 Related Work

Pre-trained Language Model Probing Pre-trained Language Models (PLMs) have been shown to possess inherent syntactic and semantic information. Different probing techniques are developed

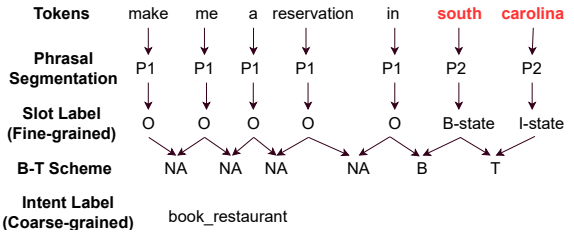


Figure 1: Illustration of connections between Phrasal Segmentation (PS), Beginning-Inside-Outside (BIO) Tagging Slot Label and Break-Tie (B-T) Labeling Schema based on Golden Slot Labels (Red: denotes Golden Slot Labels for the utterance, **P1,P2** denote identified phrases, **NA, B,T** denote Not-Relevant, Break, Tie Labels in B-T Labeling Scheme)

to investigate the knowledge acquired by PLMs, either from output representations (Wu et al., 2020b), intermediate representations (Sun et al., 2019), or attention mapping (Clark et al., 2019; Yu et al., 2022). Unlike previous probing techniques that focus on deriving syntactic tree structure, we leverage semantically coherent segments recognized by PLMs to induce phrases containing token-level slot labels in NLU tasks for TOD Systems.

Contrastive Learning Contrastive Learning (CL) has been widely leveraged as an effective representation learning mechanism (Oord et al., 2018). The goal of CL is to learn the discriminative features of instances via different augmentation methods. In Natural Language Processing (NLP), CL has been adopted in various contexts ranging from text classification (Wei and Zou, 2019), embedding representation learning (Gao et al., 2021) to question answering (Xiong et al., 2020; Liu et al., 2021). CL has also been integrated with PLM as a more effective fine-tuning strategy for downstream tasks (Su et al., 2021). In our work, we propose an integration of CL with PLM probing techniques to further refine imperfect PLM-derived segments via (1) unsupervised signals from PLM itself, and (2) less expensive sentence-level intent label supervision for improved SI performance.

3 Problem Formulation

Slot Induction We introduce the task of Slot Induction (SI) whose objective is to identify phrases containing token-level slot labels. Unlike traditional SF and previously proposed AISI framework (Zeng et al., 2021), in our SI task, both slot boundaries and slot types are unknown during training. The task is also related to Phrasal Segmentation/ Tagging (PS) methods (Shang et al., 2018a; Gu et al., 2021). However, there are three key distinc-

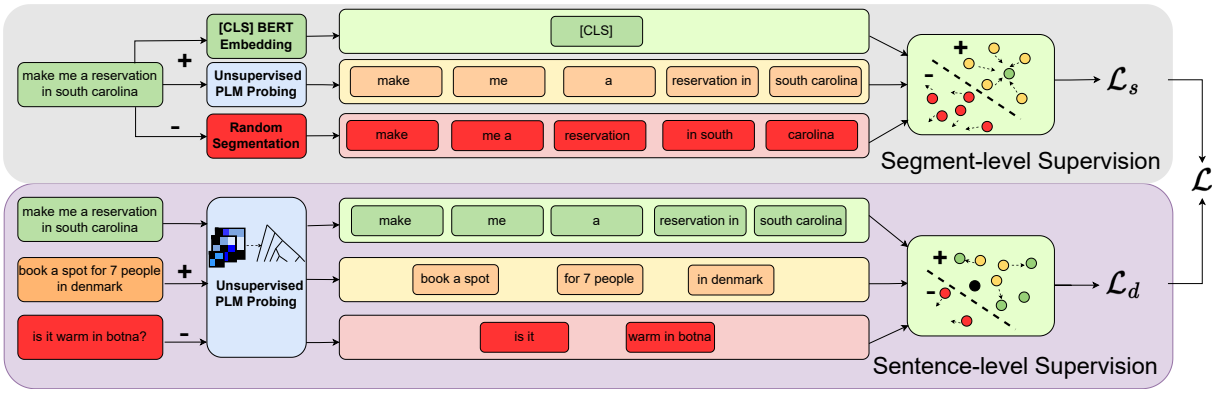


Figure 2: Illustration of the Proposed Model Overview. The model is made up of two-level Contrastive Learning depicted by two modules: (1) **Segment-level Supervision (SegCL)** via Unsupervised PLM Probing (UPL), (2) **Sentence-level Supervision (SentCL)** via intent labels. Green, Orange, Red denote Anchor, Positive, Negative samples respectively. **Black circle** denotes the representation of the **cropped segment** from Augmentation.

tions: (1) utterances and intent labels (if available) are the only sources of information for the task, (2) slot phrases (i.e. close by (*spatial_relation*), most expensive (*cost_relative*)), are not restricted to noun phrases, (3) slot phrases (i.e. strauss is playing today (*movie_name*)) might be more sophisticated and harder to identify than typical noun phrases (i.e. chicago (*city*)). These differences explain why PS methods do not consistently perform well in our proposed SI task (Section 6).

Specifically, given an utterance with the length of T tokens $x = [x_1, x_2..., x_T]$, SI task aims to make decisions at $T-1$ positions whether to (1) tie the current token with the previous one to extend the current phrase [3], or (2) break away from the previous token/ phrase to form a new phrase.

Evaluation Metric We adopt the Break-Tie (B-T) schema (Shang et al., 2018b) to evaluate SI task. The metric allows for direct comparison between supervised Sequential Labeling and unsupervised PS methods. In SI setting, *Tie* represents the connection between tokens of the same slot type while *Break* denotes the separation between (1) tokens from different slot types, and (2) tokens from a slot type and non-slot tokens. As the objective of SI is on slot tokens, consecutive non-slot tokens should not contribute to the overall performance. Therefore, additional *NA* labels are introduced to guarantee that evaluations are only conducted on slot tokens and their adjacent tokens.

Figure 1 depicts the connections of SF and PS labels with B-T schema. For PS, Break denotes the separation of two consecutive phrases. If no phrase is identified by PS methods, every token is considered as *Tie* to one another. In the Figure 1 example, as "south carolina" is the only identified phrase, the given sentence is simply split into two phrases where *Break* denotes their junction. Precision, Recall and F-1 Metrics are reported for individual labels, namely B-P,B-R,B-F1 for *Break* and T-P,T-R,T-F1 for *Tie*.

Given an utterance, an optimal SI model makes correct decisions to either break and tie at every token index. Therefore, **H-Mean**, denoting the harmonic mean between F-1 Scores of *Tie* and *Break* label predictions, is considered the golden criteria for SI model comparison.

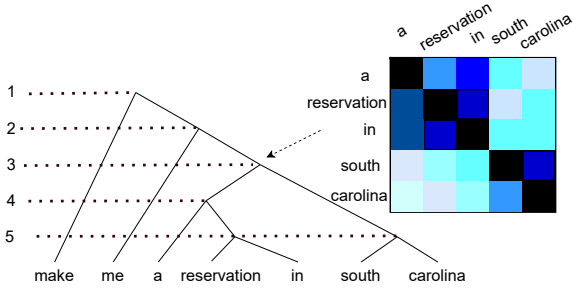


Figure 3: Illustration of UPL Segmentation Tree for sentence *"make me a reservation in south carolina"* with sample Impact Matrix at depth $d = 3$ (**Lighter** color denotes **lower** impact score). $d = 0$ corresponds to the sentence-level representation (no segmentation).

4 Proposed Framework

In this section, we introduce our proposed Multi-level Contrastive Learning framework for SI task with 2 major components: **Segment-level Contrastive Learning (SegCL)** and **Sentence-level Contrastive Learning (SentCL)** as depicted in Figure 2. We first introduce the backbone Unsupervised PLM Probing (UPL) for both components.

[3] In our work, we use the term **segment** and **phrase** interchangeably.

4.1 Unsupervised PLM Probing (UPL)

We adopt Token-level Perturbed Masking mechanism (Wu et al., 2020b) to construct semantic segments by leveraging PLM in an unsupervised manner. Due to its operations on the output layers of PLM, UPL is flexible with the choices of PLM and avoids local sub-optimal structure from pre-selected PLM layers (Clark et al., 2019). In our study, we use BERT (Devlin et al., 2019) as an exemplar PLM. Specifically, given a sentence $x = [x_1, \cdots, x_T]$, the Impact Matrix $\mathcal{F} \in \mathbb{R}^{T \times T}$ is constructed by calculating the Impact Score between every possible pair of tokens (including with itself) in the given sentence based on BERT's embedding and a specified distance metric (Wu et al., 2020b). Leveraging $\mathcal{F}$, UPL derives the structural tree by recursively finding the optimal cut position k with the following objective:

$$\begin{aligned} \underset{k}{argmax}(&\mathcal{F}^{i..k}_{i..k} + \mathcal{F}^{k+1..j}_{k+1..j} \\ &-\mathcal{F}^{k+1..j}_{i..k} - \mathcal{F}^{i..k}_{k+1..j}) \end{aligned} \tag{1}$$

where $i, j \in [0, T-1]$ denotes the start and end indexes of the segment considered for splitting.

At every tree depth, sets of combined tokens are considered semantic segments since they preserve certain meanings within utterances. Segments at a deeper level include (1) all segments obtained from previous levels and (2) new segments obtained at the current level. For instance, at depth $d = 3$ of the given example in Figure 3, the obtained segments are *"make"*, *"me"*, *"a reservation in"*, *"south carolina"*. As PLM parameters are updated during training, the derived UPL trees from the same utterance can vastly change. For simplicity, we set the tree depth d as a tunable hyperparameter.

Formally, at a specified depth d with m semantic segments acquired from UPL, the final representation of the input sentence x is defined as follows:

$$\mathbf{h_U} = [\overrightarrow{s_0}, ...\overrightarrow{s_{m-1}}], \ \overrightarrow{s_i} = \frac{\sum_{j=c}^{d} \overrightarrow{h_j}}{d - c + 1} \tag{2}$$

where $\mathbf{h_U} \in \mathbb{R}^{\mathbf{m} \times \mathbf{d_h}}$, d_h is hidden dimensions of BERT representations, c,d are the start and end indexes of the corresponding segment s_i and $\overrightarrow{h_j}$ represents the BERT embedding of j-th token.

4.2 Multi-level Contrastive Learning

As UPL only considers token interactions for segment formation, its semantic segments are far from perfect. Additional refinements are needed to enhance the quality of the extracted segments via (1) semantic signals captured in segment-level PLM representations, (2) sentence-level intent labels.

Our overall learning objective is summarized as $\mathcal{L} = \delta\mathcal{L}_s + \gamma\mathcal{L}_d$, where $\mathcal{L}_s, \mathcal{L}_d$ denote SegCL Loss and SentCL Loss, and γ, δ are their corresponding loss coefficient hyperparameters for aggregation. For each CL level, positive and negative samples are drawn separately based on (1) the same batch of sampled anchor samples, (2) different selection criteria detailed below.

Segment-level Contrastive Learning (SegCL) UPL produces semantic segments by purely considering the exhaustive word-pair interactions within given sentences. However, it does not take into consideration the overall semantic representation produced by the PLM BERT via special [CLS] tokens. Therefore, we propose leveraging [CLS] representations to guide UPL towards more discriminative segment representations via SegCL objectives. Specifically, SegCL aims to minimize the distance between [CLS] representation and UPL segment representations while maximizing the distance between representations of [CLS] and random segments of the corresponding utterance.

Given a sample utterance, segment representation obtained from UPL is considered a positive sample while negative samples are represented as segments produced by randomly chosen indexes within the given utterance. The number of segments for both positive and negative samples are kept similar (m) so that SegCL focuses on learning the optimal locations of segmentation indexes. We adopt InfoNCE contrastive loss (Oord et al., 2018):

$$\mathcal{L}_s = -log \frac{\exp^{cos(\overrightarrow{h_C}, \mathbf{h_U})/\tau_\mathbf{s}}}{\exp^{cos(\overrightarrow{h_C}, \mathbf{h_U})/\tau_\mathbf{s}} + \exp^{cos(\overrightarrow{h_C}, \mathbf{h_r})/\tau_\mathbf{s}}} \tag{3}$$

where $\overrightarrow{h_C} \in \mathbb{R}^{1 \times d_h}$ denotes [CLS] representation from BERT, and $\mathbf{h_U}, \mathbf{h_r} \in \mathbb{R}^{\mathbf{m} \times \mathbf{d_h}}$ denote the representations from UPL and random segmentation. m is the number of extracted segments from UPL as defined in Equation 2. τ_s is the soft segment-level temperature hyperparameter.

Sentence-level Contrastive Learning (SentCL) Besides relying on UPL, we propose leveraging sentence-level intent labels to further improve the quality of segment representations derived from UPL. Specifically, we randomly draw positive and negative samples based on the intent labels of the given anchor samples. As utterances with similar intents tend to share common slot phrases, our SentCL aims to learn discriminative segments for better alignment between utterances from the same

intents. We adopt InfoNCE loss for SentCL:

$$\mathcal{L}_d = -log\frac{\exp^{cos(\mathbf{h_a},\mathbf{h_+})/\tau_\mathbf{d}}}{\exp^{cos(\mathbf{h_a},\mathbf{h_+})/\tau_\mathbf{d}} + \exp^{cos(\mathbf{h_a},\mathbf{h_-})/\tau_\mathbf{d}}} \quad (4)$$

where $\mathbf{h_a} \in \mathbb{R}^{\mathbf{m}\times\mathbf{d_h}}, \mathbf{h_+} \in \mathbb{R}^{\mathbf{a}\times\mathbf{d_h}}, \mathbf{h_-} \in \mathbb{R}^{\mathbf{b}\times\mathbf{d_h}}$ denote the representations of anchor, positive and negative samples respectively and m, a, b denote the number of extracted segments from UPL for the respective samples. τ_d is the soft sentence-level temperature hyperparameter.

To further encourage the model to identify discriminative segments from the same sentence-level intent label, we adopt random segment cropping as an augmentation strategy. As UPL could generate a vastly different number of segmentation based on the the cut_score (Equation 1) from the updated BERT parameters at each step, we conduct random segmentation cropping by a percent ratio (β) so that it could be adapted to individual input utterances and segmentation trees. The remaining segments after cropping are utilized to compute $\mathcal{L}_d$.

5 Experiments

5.1 Datasets & Evaluation Tasks

We evaluate our proposed work on the two publicly available NLU benchmark datasets ATIS (Tur et al., 2010) and SNIPS (Coucke et al., 2018) with the previously proposed data splits (Zhang et al., 2019).

To evaluate the generalization of the refined representations from our proposed work, we conduct additional splits of each dataset into 2 parts (P1 and P2). For each benchmark dataset, we construct P1 for SI evaluation by reserving samples from randomly chosen 60% of available intents. The remaining samples (P2) are used as test sets for evaluating SF task when generalized towards emerging intents. The objective of this splitting strategy is two-fold: (1) Since there is no overlapping intent between P1 and P2, there exists no information leakage of intents leveraged in SI training (P1) while evaluating SF (P2). (2) We can validate the generalization capability of representations learned from our SI framework in other slot-related tasks. Statistics for both parts of each dataset are reported in Table 1.

Evaluation Task 1: Slot Induction (P1) We conduct evaluation of Unsupervised SI task on P1 of both SNIPS and ATIS datasets. B-T evaluation metrics are adopted as introduced in Section 3. Implementation details of our SI model, including hyperparameters, are discussed in Appendix B.

Table 1: Details of SNIPS and ATIS datasets.

	SNIPS_P1	SNIPS_P2	ATIS_P1	ATIS_P2
# Intents	5	2	14	7
# Slots	31	16	68	63
# Train Samples	9356	–	3811	–
# Validation Samples	500	–	414	–
# Test Samples	501	4127	750	895
Avg Train Sent Length	8.65	–	11.67	–
Avg Valid Sent Length	8.72	–	11.82	–
Avg Test Sent Length	8.71	9.87	10.68	8.92

Evaluation Task 2: Generalization towards Emerging Intents (P2) To evaluate the generalization of SI refinement, we conduct SF training on P1 datasets with different BERT initializations (Original vs Refined BERT) and evaluation on emerging intents and slots in P2. Slot Precision (S-P), Recall (S-R), F1 (S-F1) are reported on P2. Implementation is detailed in Appendix C.

5.2 Slot Induction Baseline

We conduct a comprehensive study that evaluates our SI approach with both *Upper Bound* and *Comparable* Methods. For fair comparisons, we leverage the same "bert-base-uncased" PLM (Devlin et al., 2019) across all applicable baselines. The *Upper Bound* includes methods that leverage directly **token-level labels** such as Golden Slot Labels, Named Entity Recognition (NER) Labels, Part-of-Speech (POS) Tagging or Noun Phrase (NP) Labels during training and/or pre-training process, including **Joint BERT FT**, **SpaCy** (Honnibal et al., 2020), **FlairNLP** (Akbik et al., 2018).

In addition, we compare with other **unsupervised** PS methods that do not require any token-level labels as *Comparable* Baselines, including: **Dependency Parsing (DP-RB/DP-LB)**, **AutoPhrase** (Shang et al., 2018a), **UCPhrase** (Gu et al., 2021), **USSI** (Yu et al., 2022). For fair comparisons with *Comparable* baselines, we also report results from our model's variants with similar prior knowledge assumption, namely **Ours (w/o CL)**, **Ours (w/o SentCL)** . Due to space constraints, details of *Upper Bound* and *Comparable* baselines are provided in Appendix A.1, A.2 respectively.

6 Result & Discussion

6.1 Slot Induction

From our experimental results in Table 2 and 3, for SI task, our proposed framework outperforms the *Comparable* Methods in H-Mean evaluation metric for B-T schema on both datasets. We achieve significant gains in SNIPS dataset (+6.28 points in H-Mean as compared to the next *Comparable* Methods). Despite lack of access to any types of token-level labels, our method is also closely on

Table 2: Experimental performance result on SNIPS dataset over 3 runs (**H-Mean** is considered the golden criteria for SI (Section 3)). ¶ denotes models that do not require random initializations.

	Model	Prior Knowledge	Break			Tie			H-Mean
			B-P	B-R	B-F1	T-P	T-R	T-F1	
Upper Bound	Joint BERT FT	Slot + Intent	96.91 ± 0.17	96.62 ± 0.69	96.76 ± 0.26	73.55 ± 0.38	73.39 ± 1.03	73.47 ± 0.38	83.52 ± 0.16
	FlairNLP ¶	POS & NER	80.04	62.81	70.38	48.25	63.31	54.77	61.60
	SpaCy ¶	POS	75.73	50.29	60.45	41.71	62.97	50.18	54.84
Comparable	DP-LB ¶	–	59.68	34.27	43.54	21.69	38.53	27.76	33.90
	DP-RB ¶	–	66.53	52.56	58.73	33.97	52.24	41.17	48.40
	AutoPhrase	External KB	65.51 ± 0.23	57.16 ± 2.59	61.05 ± 1.15	33.39 ± 0.74	36.62 ± 1.67	34.93 ± 1.50	44.43 ± 1.64
	UCPhrase	PLM	42.25 ± 4.90	20.26 ± 2.71	27.39 ± 1.95	36.06 ± 2.42	**73.53 ± 3.33**	**48.39 ± 2.91**	34.98 ± 2.35
	USSI ¶	PLM	**83.21**	62.12	71.14	33.96	49.93	40.42	51.55
	Ours (w/o CL) ¶	PLM	75.36	66.70	70.76	38.51	45.81	41.84	52.59
	Ours (w/o SentCL)	PLM	76.09 ± 0.73	66.43 ± 0.29	70.94 ± 0.49	39.15 ± 0.60	47.9 ± 0.91	43.09 ± 0.73	53.61 ± 0.71
	Ours (full)	**PLM + Intent**	76.87 ± 0.25	**67.77 ± 0.26**	**72.00 ± 0.24**	**40.39 ± 0.16**	48.49 ± 0.19	44.07 ± 0.04	**54.68 ± 0.08**

Table 3: Experimental performance result on ATIS dataset over 3 runs (**H-Mean** is considered the golden criteria for SI (Section 3)). ¶ denotes models that do not require random initializations.

	Model	Prior Knowledge	Break			Tie			H-Mean
			B-P	B-R	B-F1	T-P	T-R	T-F1	
Upper Bound	Joint BERT FT	Slot + Intent	98.49 ± 0.24	99.33 ± 0.08	98.91 ± 0.09	59.07 ± 0.36	58.27 ± 0.89	58.67 ± 0.63	73.65 ± 0.54
	FlairNLP ¶	POS & NER	95.44	77.90	85.78	41.34	61.91	49.58	62.84
	SpaCy ¶	POS	94.45	69.64	80.17	35.33	61.17	44.79	57.47
Comparable	DP-LB ¶	–	80.80	36.38	50.17	12.32	38.51	18.67	27.21
	DP-RB ¶	–	84.24	**66.84**	**74.54**	14.81	30.52	19.94	31.46
	AutoPhrase	External KB	75.96 ± 0.04	40.06 ± 0.28	52.46 ± 0.18	19.75 ± 0.21	49.33 ± 0.38	**28.20 ± 0.28**	36.68 ± 0.21
	UCPhrase	PLM	47.25 ± 0.04	17.27 ± 0.72	25.29 ± 0.78	17.36 ± 0.16	**58.21 ± 0.68**	26.75 ± 0.11	26.00 ± 0.47
	USSI ¶	PLM	**95.06**	56.36	70.77	14.78	45.22	22.28	33.89
	Ours (w/o CL) ¶	PLM	86.40	61.53	71.87	18.23	35.27	24.04	36.03
	Ours (w/o SentCL)	PLM	87.29 ± 0.15	64.21 ± 0.27	73.99 ± 0.13	20.09 ± 0.08	35.86 ± 0.35	25.75 ± 0.08	38.20 ± 0.08
	Ours (full)	**PLM + Intent**	87.80 ± 0.27	63.27 ± 0.67	73.54 ± 0.36	**20.53 ± 0.14**	37.89 ± 0.99	26.63 ± 0.26	**39.10 ± 0.24**

Table 4: Ablation study of effectiveness of SegCL and SentCL on SNIPS and ATIS in terms of H-Mean

	SNIPS	**ATIS**
Ours (w/o CL)	52.59	36.03
+ SegCL	53.61 ± 0.71	38.20 ± 0.08
+ SentCL (w/o aug)	53.44 ± 0.22	37.59 ± 0.81
+ SentCL (w aug)	54.23 ± 0.10	38.12 ± 0.36
Ours (full)	**54.68 ± 0.08**	**39.10 ± 0.24**

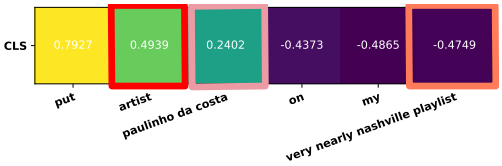
par with some of the *Upper Bound* methods that have been pre-trained with token-level labels (0.16 point difference from SpaCy in H-Mean). Despite promising achievements, most unsupervised PS methods only achieve competitive Break performance as compared to supervised methods but fall behind more significantly in terms of Tie performance. This implies unsupervised methods are able to differentiate non-slot tokens from slot tokens but tend to fragment slot tokens of the same type into multiple slot phrases due to the missing knowledge of token-level slot label spans.

UCPhrase is an exceptional baseline as it achieves significant better Tie but worse Break performance as compared to other *Comparable* baselines. This roots from the lack of keyphrases predicted from the model, leading to higher tendency to "tie" tokens. We speculate that its core phrase miner's dependency on frequency is not effective for extracting slots in NLU tasks. Phrases with high frequency in utterances are typically non-slot tokens (i.e. add, reserve), leading to limited meaningful core phrases for phrase-tagging training.

On ATIS dataset, the gap between *Comparable* Methods and *Upper Bound* is more significant as utterances tend to be longer and contain a wider variety of slot types than SNIPS dataset. This leads to a significant reduction in T-P across all of the *Comparable* Methods, resulting in a larger gap in H-Mean for ATIS dataset (approximately 18.37 points in comparison with 0.16 points in SNIPS dataset). Additionally, in comparison with SNIPS dataset, ATIS dataset contains more domain-independent slot types such as *city_name* (New York), *country_name* (United States). Therefore, methods leveraging either relevant token-level labels (i.e. POS, NER tags) or additional large-scaled external Knowledge Base (i.e. Wikipedia) achieve considerable performance gains. For instance, *FlairNLP* is only 10.81 points below the Fully Supervised *Joint BERT FT* on ATIS dataset (as compared to 21.92 points below on SNIPS) in terms of H-Mean.

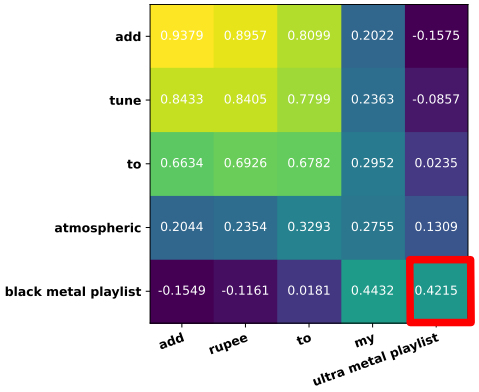
Compared with USSI, *Ours (w/o CL)* consistently achieves better H-Mean performance on both ATIS and SNIPS datasets (1.04% and 2.14% respectively). We hypothesize USSI might suffer from the local sub-optimality of pre-selected layers within deep PLM architecture. As the attention distribution across different layers varies (Clark et al., 2019), the pre-selected layers can significantly impact the unsupervised semantic probing of PLM.

Table 4 demonstrates that both SegCL and SentCL (w aug) objectives provide valuable in-

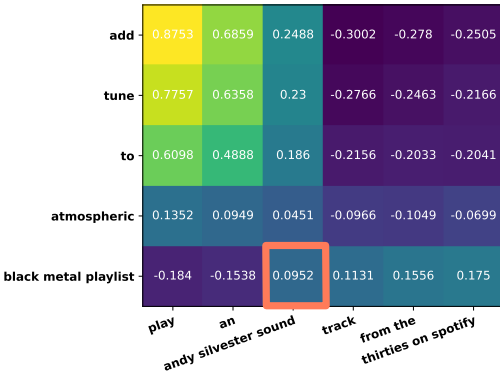


(a) Segment-level Supervised Positive-Anchor Pair

(b) Segment-level Supervised Negative-Anchor Pair



(c) Sentence-level Supervised Positive-Anchor Pair



(d) Sentence-level Supervised Negative-Anchor Pair

Figure 4: Similarity Matrices between positive/negative and anchor samples from SegCL and SentCL. For SegCL ((a), (b)), positive-anchor pair is more aligned as the sum of similarity scores between positive segments and [CLS] representation (i.e. sum of row-wise cell values) is higher than the negative counterpart. Boundaries of all slot types (presented by red, pink, orange boxes) are correctly recognized in the positive sample in contrast to the negative counterpart. For SentCL ((c), (d)), positive-anchor pair assigns a higher similarity score to the aligned slot phrase (red box) while negative-anchor pair reduces similarity scores between potential relevant slot phrase (orange box).

formation for SI task, leading to improved performance on both datasets beyond *Ours (w/o CL)*.

Segment-level Supervision (SegCL) As observed in Figure 4a, 4b, semantic representation of the given utterance via [CLS] token is closer to the UPL-derived segments as compared to random segment counterparts due to the higher sum of similarity score (0.1281 > -0.6304). UPL segments also correctly identify nearly all of the slot ground truth labels (i.e. artist (*music_item*), paulinho da costa (*artist*), my (*playlist_owner*), very nearly nashville (*playlist*)) in the given utterance while random segmentations truncate the slot phrases incorrectly.

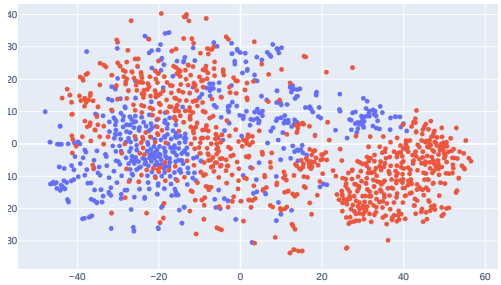
Sentence-level Supervision (SentCL) On the sentence level, besides the commonly aligned phrases (i.e. *add tune to* vs *add rupee to*), the model recognizes corresponding playlists in anchor and positive samples (i.e. *black metal playlist* vs *ultra metal playlist*) and assign competitive similarity score between them. On the other hand, potential relevant noun phrases (i.e. ultra metal playlist (*playlist*) and andy silvester sound track (*sound track*)) between anchor and negative samples are assigned low similarity score. This showcases the model's capability in (1) correctly recognizing and bringing the important slot phrases in positive-anchor pair closer together, (2) reducing

`[Golden]:`	**play me a top-ten song by | phil ochs | on | groove shark add | spirit touches ground | to my | leche con chocolate list**
`[DP-LB]:`	play me a top-ten song by phil ochs on | groove shark add spirit touches ground to my leche con | chocolate list
`[DP-RB]:`	play me a top-ten | song by phil ochs on groove shark add spirit touches ground | to my leche con chocolate list
`[Autophrase]:`	|play | me a top-ten | song | by | phil ochs| on | groove shark add | spirit | touches ground to my leche con | chocolate list
`[UCPhrase]:`	play me a top-ten | song by phil ochs on | groove shark add spirit touches | ground to my leche con | chocolate list
`[USSI]:`	play | me | a | top-ten | song | by phil ochs on groove shark add spirit | touches | ground | to my | leche con chocolate list
`[Ours(w/o CL]:`	play me | a top-ten | song | by phil | ochs on groove shark add spirit touches ground to | my leche con chocolate list
`[Ours(with CL)]:`	**play me | a top-ten | song | by | phil ochs | on groove shark add | spirit touches ground | to my | leche con chocolate list**

Figure 5: Sample Segmentation Results from *Comparable* Methods in comparison with **Golden Slot Labels** on SNIPS dataset where "|" denotes the *Break* as introduced in Figure 1. Red, Blue denote distinct slot label segments. The colors are repeated in *Comparable* Methods to showcase the consistency of models' predictions with ground truth labels under the condition no more than 2 tokens in the segments are mispredicted.

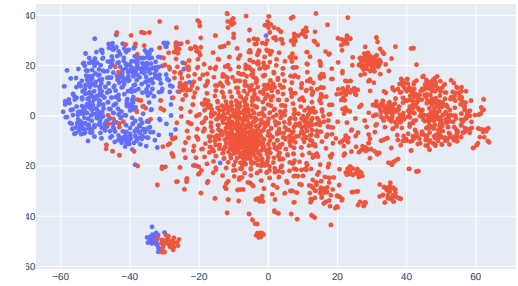
the importance of potential relevant slot phrases across samples with different intents. The Similarity Matrix presented in Figure 4c also indicates the strong segment alignment between positive and anchor samples as the diagonal cells receive higher similarity score than most of the other cells within the same column or row.

Qualitative Case Study Additional Case Studies presented in Figure 5 demonstrate the effec-

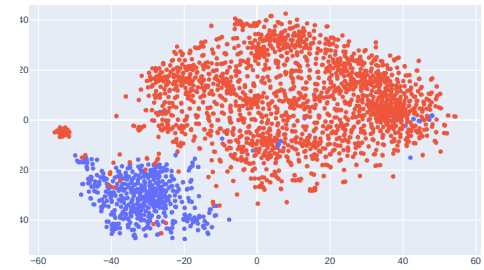


(a) Pre-trained BERT (Train Slots)

(b) Refined BERT (Train Slots)



(c) Pre-trained BERT (Test Slots)



(d) Refined BERT (Test Slots)

Figure 6: Slot Value Representation Visualization of the raw original pre-trained BERT and raw Refined BERT via SI on sample slot types from training set SNIPS_P1 ((a), (b)) and testing set SNIPS_P2 ((c), (d)). Blue and Red denotes slot values from randomly sampled ground truth slot types.

Table 5: Evaluation of SF task over 3 runs on Emerging Intents in SNIPS_P2 and ATIS_P2 datasets.

	SNIPS_P2		
	S-P	S-R	S-F1
Original BERT	14.11 ± 0.47	17.78 ± 0.82	15.73 ± 0.62
Refined BERT	**15.08 ± 0.48**	**19.61 ± 0.23**	**17.05 ± 0.38**
	ATIS_P2		
Original BERT	66.67 ± 0.82	63.35 ± 1.35	64.96 ± 0.74
Refined BERT	**70.12 ± 0.85**	**63.64 ± 0.48**	**66.72 ± 0.66**

tiveness of our proposed framework in capturing slot phrases. Despite the imperfect segmentations, *Ours* captures phrases closer to the ground truth slot labels than other *Comparable* baselines. In fact, our identified phrases "spirit touches ground" and "leche con chocolate list" are exact matches for the golden slot labels. Our proposed multi-level CL refining mechanism is also shown to correct mistakes of the original model. (from "by phil" in *Ours (w/o CL)* to "phil och" in *Ours (with CL)*.

6.2 Generalization towards Emerging Intents

Visual Representation We first visualize the representations of two randomly sampled slot types produced by the raw original BERT and our Refined BERT (via SI objectives). As observed in Figure 6, our Refined BERT clusters the representations of samples with the same slot types for both training and testing sets more effectively than the original BERT in the embedding space, leading to far clearer separation boundaries between the sampled slot types. For Train Slots, embeddings of slot values from each slot type are nearly disentangled, implying our Refined BERT is capable of recognizing slot types without explicit slot training objectives and token-level label access. In addition, when applied to new intents and slots in P2 dataset, our SI framework produces refined BERT with better semantic representations for tokens from the same slot types as observed in Figure 6c,6d.

Quantitative Evaluation As observed in Table 5, when generalized to emerging intents and slots, our Refined BERT outperforms the traditional BERT while fine-tuning on both datasets in all slot evaluation metrics. This showcases the generalization capability of our model across different sentence-level intent labels. In addition, the consistent improvement in SF evaluation implies that SI training objectives via UPL and CL refinement provide more guidance to the PLM for the downstream token-level task without explicit training objectives and label requirements.

7 Conclusion

In our work, we propose the study of token-level Slot Induction (SI) via an Unsupervised Pre-trained Language Modeling (PLM) Probing in conjunction with Contrastive Learning (CL) objectives. By leveraging both unsupervised signals from PLM and sentence-level signals from intent labels via CL objectives, our proposed framework not only

achieves competitive performance in comparison with other unsupervised phrasal segmentation baselines but also bridges the gap in performance with *Upper Bound* methods that require additional token-level labels on two NLU benchmark datasets. We also demonstrate that our proposed SI training is capable of refining the original PLM, resulting in more effective slot representations and benefiting downstream SF tasks when generalized towards emerging intents. Further studies of better exploitation of full-depth segmentation trees, enhanced segment augmentation mechanisms and better semantic alignment extraction between slots and intents are promising directions for our future work. We also seek to extend the current SI studies beyond English and towards multilingual NLU systems. (Nguyen and Rohrbaugh, 2019; Qin et al., 2022; Nguyen et al., 2023)

Limitations

Our proposed framework assumes a fixed hyperparameter depth d for UPL segmentation tree. In other words, only segments extracted at the depth d are considered for CL objectives. d is tuned with each dataset's validation set. However, as our main objective is to investigate the effects of UPL and CL objectives, we leave the full tree exploitation as future extensions for our work.

Secondly, the goal of our SI is to identify the slot phrase boundaries. The label type predictions for recognized slot phrases are beyond the scope of our investigation. Therefore, direct end-to-end evaluation of SI in mitigating slot label scarcity issues cannot be directly evaluated. Our rationale for dividing the task into 2 separate steps (i.e. slot boundary induction and slot label prediction) is as follows: As the complete SI is a complex task, breaking it down not only allows for direct and focused evaluation of the proposed framework's contribution at individual steps but also minimizes error propagation from intermediate steps to a single end-task metric. This rationale is further supported by our empirical study in Section 6. The proposed *USSI* whose objective unifies both aforementioned steps underperforms *Ours(w/o CL)* and *Ours(full)* when evaluated at the slot boundary induction step.

Acknowledgement

This work is supported in part by NSF under grants III-1763325, III-1909323, III-2106758, and SaTC-1930941.

We would like to acknowledge the use of the facilities of the High Performance Computing Division and High Performance Research and Development Group at the National Center for Atmospheric Research and the use of computational resources (doi:10.5065/D6RX99HX) at the NCAR-Wyoming Supercomputing Center provided by the National Science Foundation and the State of Wyoming, and supported by NCAR's Computational and Information Systems Laboratory.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, pages 12–16.

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, and Alfio Gliozzo. 2021. Robust retrieval augmented generation for zero-shot slot filling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1939–1949.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Xiaotao Gu, Zihan Wang, Zhenyu Bi, Yu Meng, Liyuan Liu, Jiawei Han, and Jingbo Shang. 2021. *UCPhrase: Unsupervised Context-Aware Quality Phrase Tagging*, page 478–486. Association for Computing Machinery, New York, NY, USA.

E Haihong, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5467–5471.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Taeuk Kim, Jihun Choi, Daniel Edmiston, and Sang goo Lee. 2020. Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction. In *International Conference on Learning Representations*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Ye Liu, Kazuma Hashimoto, Yingbo Zhou, Semih Yavuz, Caiming Xiong, and S Yu Philip. 2021. Dense hierarchical retrieval for open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 188–200.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496.

Hoang Nguyen and Gene Rohrbaugh. 2019. Cross-lingual genre classification using linguistic groupings. *Journal of Computing Sciences in Colleges*, 34(3):91–96.

Hoang Nguyen, Chenwei Zhang, Congying Xia, and S Yu Philip. 2020. Dynamic semantic matching and aggregation network for few-shot intent detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1209–1218.

Hoang Nguyen, Chenwei Zhang, Tao Zhang, Eugene Rohrbaugh, and Philip Yu. 2023. Enhancing cross-lingual transfer via phonemic transcription integration. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9163–9175, Toronto, Canada. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Libo Qin, Qiguang Chen, Tianbao Xie, Qixin Li, Jian-Guang Lou, Wanxiang Che, and Min-Yen Kan. 2022. GL-CLeF: A global–local contrastive learning framework for cross-lingual spoken language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2677–2686, Dublin, Ireland. Association for Computational Linguistics.

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

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018a. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018b. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064.

AB Siddique, Fuad Jamour, Luxun Xu, and Vagelis Hristidis. 2021. Generalized zero-shot intent detection via commonsense knowledge. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1925–1929.

Yusheng Su, Xu Han, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. 2021. Css-lm: A contrastive framework for semi-supervised fine-tuning of pre-trained language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2930–2941.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4323–4332.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. 2010. What is left to be understood in atis? In *2010 IEEE Spoken Language Technology Workshop*, pages 19–24. IEEE.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020a. SlotRefine: A fast non-autoregressive model for joint intent detection and slot filling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1932–1937, Online. Association for Computational Linguistics.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020b. Perturbed masking: Parameter-free probing for analyzing and interpreting bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176.

Congying Xia, Chenwei Zhang, Hoang Nguyen, Jiawei Zhang, and Philip Yu. 2020. Cg-bert: Conditional text generation with bert for generalized few-shot intent detection. *arXiv preprint arXiv:2004.01881*.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and S Yu Philip. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3090–3099.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Dian Yu, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent Shafey, and Hagen Soltau. 2022. Unsupervised slot schema induction for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1193.

Mengshi Yu, Jian Liu, Yufeng Chen, Jinan Xu, and Yujie Zhang. 2021. Cross-domain slot filling as machine reading comprehension. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, Montreal, QC, Canada*, pages 19–26.

Zengfeng Zeng, Dan Ma, Haiqin Yang, Zhen Gou, and Jianping Shen. 2021. Automatic intent-slot induction for dialogue systems. In *Proceedings of the Web Conference 2021*, pages 2578–2589.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and S Yu Philip. 2019. Joint slot filling and intent detection via capsule neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5259–5267.

Table 6: Hyperparameters for SNIPS and ATIS datasets (SI task)

	d	β	τ_s	τ_d	δ	γ
SNIPS	3	0.2	0.1	0.05	0.3	0.7
ATIS	4	0.2	0.05	0.1	1.0	0.2

A Slot Induction Baselines

For fair comparisons across all baselines, we leverage BERT (Devlin et al., 2019) as the backbone PLM architecture (if applicable).

A.1 Upper Bound Baselines

• **Joint BERT FT**: Fully Supervised Joint Sequence Labeling and Sentence Classification model is trained on top of fine-tuning BERT embeddings with available golden training slot and intent labels.
• **SpaCy** (Honnibal et al., 2020): Industrial-strength NLP tagging methodology that leverages pre-trained NP chunking model.
• **FlairNLP** (Akbik et al., 2018): Neural Language Modeling in junction with pre-trained Sequential Labeling (NER and POS).

A.2 Comparable Baselines

• **Dependency Parsing** (Right/Left-branching (RB/LB): Parameter-free methods for sentence segmentation. Result from the best depth is reported.
• **AutoPhrase** (Shang et al., 2018a): Statistical phrase tagging method utilizing high quality massive corpus as additional Knowledge Base (KB).
• **UCPhrase** (Gu et al., 2021): Phrase tagging method leveraging co-occurrence word frequency and PLM attention maps.
• **USSI** (Yu et al., 2022): Unsupervised Slot Schema Induction method leveraging attention distribution of PLM and additional constraints from Probabilistic Context-free Grammar (PCFG) (Kim et al., 2020). For completeness, additional experiments in leveraging the proposed in-domain training objectives with SpanBERT PLM (Joshi et al., 2020) are provided in Appendix D.
• **Ours (w/o CL)**: Fixed UPL is directly used for inference without additional CL refinement. Same depth d is used as our proposed model **Ours (full)** and its variant **Ours (w/o SentCL)**.
• **Ours (w/o SentCL)**: Our model variant that is trained only with SegCL objectives ($\mathcal{L}_s$). The model does not leverage sentence-level intent label information (SentCL) during training.

Table 7: Ablation study of SpanBERT PLMs with in-domain training objectives on SNIPS and ATIS datasets in terms of H-Mean over 3 runs. ¶ denotes models that do not require random initializations.

	SNIPS	ATIS
SpanBERT ¶	43.15	35.05
USSI (Yu et al., 2022)	48.61 ± 0.69	36.63 ± 1.93
Ours (SpanBERT w CL)	**53.25 ± 0.29**	**40.07 ± 2.34**

B Slot Induction Implementation (P1)

We train our proposed SI model with batch size of 16, learning rate 1e-5 for 10 epochs. The remaining hyperparameters for individual datasets are reported in Table 6 respectively for SI task. We tune our hyperparameters based on each dataset's P1 validation set via grid search for $\beta, \tau_s, \tau_d, \delta, \gamma$, except for d. For depth d, we conduct inference of PLM probing (i.e. Ours (w/o CL)) on P1 validation sets and select d with the highest H-Mean performance. The same depth d is used consistently across different variants of our proposed framework in the empirical study. Our reported results are reported based on 3 runs with different seeds.

C Slot Filling Implementation (P2)

As the objective of SF is to compare different BERT models (i.e. Original BERT vs Refined BERT via SI objectives), we keep the Sequence Labelling architecture simple and similar between the two models. Specifically, we stack the traditional CRF layer (Lafferty et al., 2001) on top of the corresponding BERT models. The overall model is fine-tuned on SF task with available training slot labels in P1 training data. The model is fine-tuned with batch size of 16, learning rate of 0.01 for CRF and Linear layer, BERT learning rate of 1e-5 for 10 epochs. The testing results (Table 5) are reported on P2 of each dataset as an average over 3 runs. Both training and inference for Appendix B and C are conducted on NVIDIA Titan RTX GPU.

D SpanBERT-based Model

Yu et al. (2022) proposed additional self-supervised in-domain training on Task-oriented Dialogue datasets. For fair comparisons with (Yu et al., 2022), we conduct additional studies training the same backbone SpanBERT PLM architecture (Joshi et al., 2020) with their proposed self-supervised in-domain training objectives on our training SNIPS_P1 and ATIS_P1 datasets and report test results in Table 7. To evaluate the effectiveness of our multi-level CL objectives, in Table 7, **Ours (SpanBERT w CL)** follows the induction mechanisms proposed by Yu et al. (2022) instead of UPL mentioned in Section 4.1. The only difference between **Ours (SpanBERT w CL)** and USSI is our proposed multi-level CL objectives

As demonstrated in Table 7, **Ours (SpanBERT w CL)** achieves consistent improvements over USSI on both SNIPS and ATIS datasets (4.64% and 3.44% respectively) under the same training architecture and in-domain training objectives. This observation implies the effectiveness of our multi-level CL objectives (SegCL and SentCL).

The timing bottleneck: Why timing and overlap are mission-critical for conversational user interfaces, speech recognition and dialogue systems

Andreas Liesenfeld, Alianda Lopez, Mark Dingemanse
Centre for Language Studies
Radboud University, Nijmegen, The Netherlands
{andreas.liesenfeld,ada.lopez,mark.dingemanse}@ru.nl

Abstract

Speech recognition systems are a key intermediary in voice-driven human-computer interaction. Although speech recognition works well for pristine monologic audio, real-life use cases in open-ended interactive settings still present many challenges. We argue that timing is mission-critical for dialogue systems, and evaluate 5 major commercial ASR systems for their conversational and multilingual support. We find that word error rates for natural conversational data in 6 languages remain abysmal, and that overlap remains a key challenge (study 1). This impacts especially the recognition of conversational words (study 2), and in turn has dire consequences for downstream intent recognition (study 3). Our findings help to evaluate the current state of conversational ASR, contribute towards multidimensional error analysis and evaluation, and identify phenomena that need most attention on the way to build robust interactive speech technologies.

1 Introduction

Speech recognition (ASR) is a key technology in voice-driven human-computer interaction. Although by some measures speech-to-text systems approach human transcription performance for pristine audio (Stolcke and Droppo, 2017), real-life use cases of ASR in open-ended interactive settings still present many challenges and opportunities (Addlesee et al., 2020). The most widely used metric for comparison is word error rate, whose main attraction —simplicity— is also its most important pitfall. Here we build on prior work calling for error analysis beyond WER (Mansfield et al., 2021; Zayats et al., 2019) and extend it by looking at multiple languages and considering aspects of timing, confidence, conversational words, and dialog acts.

As voice-based interactive technologies increasingly become part of everyday life, weaknesses in speech-to-text systems are rapidly becoming a key bottleneck (Clark et al., 2019). While speech scientists have long pointed out challenges in diarization and recognition (Shriberg, 2001; Scharenborg, 2007), the current ubiquity of speech technology means new markets of users expecting to be able to rely on speech-to-text systems for conversational AI, and a new crop of commercial offerings claiming to offer exactly this. Here we put some of these systems to the test in a bid to contribute to richer forms of performance evaluation.

Related Work

The struggles of achieving truly conversational speech technologies are well documented. Spontaneous, free-flowing conversations are effortless and efficient for humans but remain challenging for machines (Shriberg, 2005; Baumann et al., 2017). Speech-to-text systems face an interconnected set of challenges including at least voice activity detection, speaker diarization, word recognition, spelling and punctuation, code-switching, intent recognition, and more (Suzuki et al., 2016; Sell et al., 2018; Addlesee et al., 2020; Park et al., 2022). Each of these represents a choice point with downstream consequences that may be hard to predict. Perhaps this is why word error rate, despite its noted defects (Aksënova et al., 2021; Szymański et al., 2020), has gained the upper hand in ASR evaluation: it makes no assumptions and simply delivers a single number to be optimized.

Speech scientists have long worked to supplement word error rate with more informative measures, including error analyses of overlap (Çetin and Shriberg, 2006), disfluencies (Goldwater et al., 2010), and conversational words (Zayats et al., 2019; Mansfield et al., 2021). This work has shown the importance of in-depth error analysis, and also brings home the multi-faceted challenges of truly interactive speech-to-text systems. As speech-to-text systems gain larger user bases, multilingual performance and evaluation becomes more impor-

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 482–495
September 11–15, 2023. ©2023 Association for Computational Linguistics

Figure 1: **A** Excerpts of 20 seconds of conversations in six languages, showing the short gaps and overlaps typical of human interaction. **B** Distribution of floor transfer offset times for the human-annotated test data across the same six languages, showing that the distributions are broadly normal and tend to peak around 0, with about as many turns occurring in slight overlap (negative values) as coming in after a slight gap (positive values).

tant (Levow et al., 2021; Blasi et al., 2022; Chan et al., 2022; Tadimeti et al., 2022).

The past decades of work on speech-to-text have led to remarkable improvements in many areas, and shared tasks have played an important role in catalyzing research efforts in diarization and recognition (Ryant et al., 2021; Barker et al., 2018). Still, we see opportunities for new contributions. Most work involves either non-interactive data or widely used meeting corpora, both of them quite distinct from the fluid conversational style people increasingly expect from interactive speech technology. When more conversational data is tested, it tends to be limited to English (Mansfield et al., 2021), raising the question how large the performance gap is in a more diverse array of languages (Besacier et al., 2014). While most benchmarks still rely on word error rates, true progress requires more in-depth forms of error analysis (Szymański et al., 2020) and especially a focus on the role of timing and overlap in speech recognition and intent ascription. Finally, the wide range of speech-to-text systems on offer in a time of need for robust conversational interfaces makes it important to know what current systems can and cannot do.

2 Aims and scope

A central question relevant at every moment of human interaction is *why that now?* (Schegloff and Sacks, 1973), referring to the importance of position and composition in how people ascribe intent to communicative actions. For speech-to-text systems, in order to even approach this question, a key prerequisite is to detect *who says what when*. This means that diarization, content recognition and precise timing are all highly consequential and best considered in tandem.

Here we address this challenge by presenting a multipronged approach that lays some of the empirical groundwork for improving evaluation methods and measures. Using principles of black-box testing (Beizer, 1995), we evaluate major commercial ASR engines for their claimed conversational and multilingual capabilities. We do so by presenting case studies at three levels of analysis. Study 1 considers word error rates and treatment of overlaps. Study 2 looks into what goes missing and why. Study 3 looks into the repercussions for intent ascription and dialog state tracking. We show that across these areas, timing is both a mission-critical challenge and an ingredient for ways forward.

Data and methods

Data preparation. We evaluate using a set of human-transcribed conversational data in multiple languages (Figure 1 and Appendix A1). We take several steps to ensure the dataset makes for a useful evaluation standard: (1) we pick languages that all or most of the tested systems claim to support (English, Spanish, Dutch, French, Korean, and Mandarin); (2) we source conversational speech data from existing corpora with high quality human-transcribed annotations that were published as peer-reviewed resources; (3) we ensure audio files have comparable audio encoding and channel separation, (4) we curate human transcriptions and timing information of each dataset for completeness and

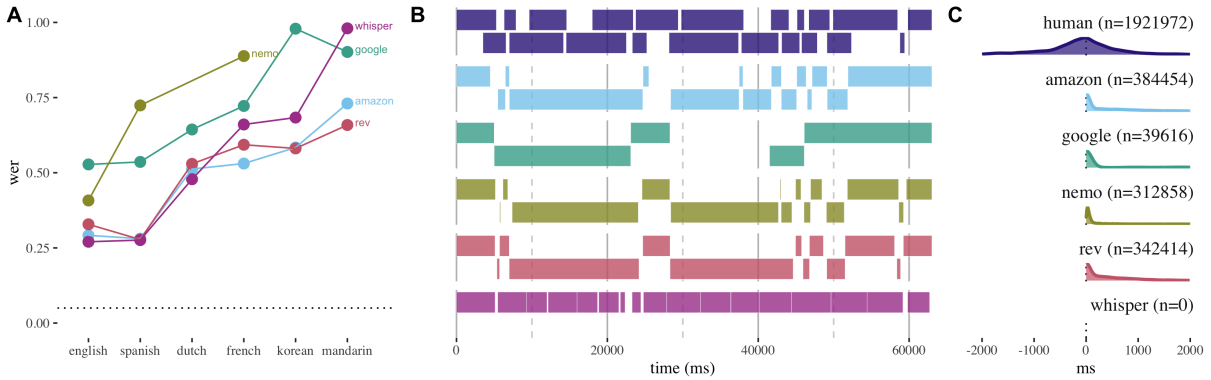


Figure 2: **A** Word error rates (WER) for five speech-to-text systems in six languages. **B** One minute of English conversation as annotated by human transcribers (top) and by five speech-to-text systems, showing that while most do some diarization, all underestimate the number of transitions and none represent overlapping turns (Whisper offers no diarization). **C** Speaker transitions and distribution of floor transfer offset times (all languages), showing that even ASR systems that support diarization do not represent overlapping annotations in their output.

accuracy, making sure that turn beginnings and ends are marked with at least decisecond precision (0.1ms); (5) we random-select one hour of dyadic conversations per language. More information on data sources and curation is available in this open data repository: `https://osf.io/hruva`.

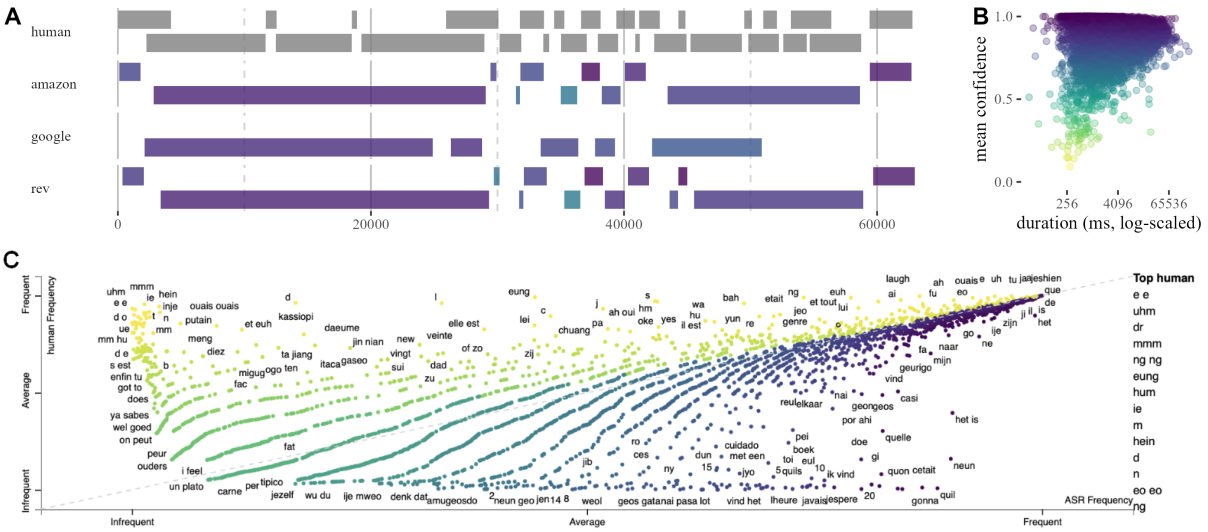
ASR system selection. Following principles of black-box testing (Beizer, 1995), we test five widely used ASR systems, keeping data and testing methods constant to compare them to human transcription baselines. Functional testing does not require access to model code or training data, instead treating models as black boxes tested to specification (Ribeiro et al., 2020). Enabling independent verification and evaluation, it is a key method in the toolbox of NLP evaluation methods.

We selected systems that claim to represent and handle conversational speech, and that offer multilingual support: (1) Amazon Transcribe 0.6.1, whose use cases include "transcription of voice-based customer service calls" and "generation of subtitles on audio/video content"; (2) Google Cloud Speech-to-Text API, using the `latest_long` model meant for "any kind of long form content such as media or spontaneous speech and conversations" (for French, Mandarin, and Spanish the long model is not available and we use the default model instead); (3) NVIDIA NeMo Quartznet15x15 for English and Conformer-CTC for French and Spanish, branded as a "Conversational AI Toolkit" that allows humans to "interact naturally"; (4) Rev AI Asynchronous Speech-to-Text API 2.17.1, which claims "accurate speaker separation" and support for "different speakers and conversations"; and (5) Whisper, a multilingual open-source neural net approaching "human-level robustness and accuracy on English speech recognition". We collected the finest-grain data available for each of these systems, using `whisper-timestamped` (Louradour, 2023) to extract word-level timing from Whisper, and `pyannote.metrics` (Bredin, 2017) for speaker diarization with NeMo.

Study 1: WER and overlap in 6 languages

Word error rates vary. We find that word error rates for truly conversational speech vary widely but nowhere approach the oft-cited human baseline of 5% transcription error (Figure 2A, dotted line), a cross-linguistic replication of prior work on English (Mansfield et al., 2021). Most speech-to-text systems have the lowest error rate for English, and even though all systems claimed multilingual support, all fare noticeably worse for typologically more different languages.

Overlap is lost. Human conversation typically features a rapid back-and-forth between participants, with a normal distribution of turn transition times centered around 0-200ms, and around half of all turns occurring in slight overlap (Figure 1; Figure 2B-C, top). Tested ASR systems record substantially fewer speaker transitions and *no* overlapping annotations. Distributions of speaker transition times show the consequences: current speech-to-text systems miss out on about half of the turns that occur in overlap. Descriptive statistics further corroborate this: by systematically not representing overlap, speech-to-text systems miss out on up to



D

	Standalone Interjections	Function Words	Discourse Markers
Dutch	*uhm, hum, hu, uh ja, mm*	*'n, ie, d'r, da, 's*	*en uh, dat uh*
English	*mhm, uhhuh, hm, oh, wow*	*did, she's, that's, going to, he*	*yeah I, because*
French	*hm hm, hein, ouais ouais, putain*	*c'était, qu'on, l', d', m'*	*euh tu, et euh*
Korean	*ahyu, eung, ye, eo, jeogi*	*hae, gajigo, jeo, geuge, jal*	*geureonigga, geuraegajigo*
Mandarin	*ng ng, ai, a, dui dui, er*	*la, wo wo, re, jiang, ya*	*shi er, gai, shi shuo*
Spanish	*eh, ah, he, claro, vale*	*o, eso, ahi, sea*	*o sea, sabes, verdad es*

Figure 3: **A** Sample minute of Korean conversation comparing human-transcribed and ASR annotations, the latter coloured by mean confidence rating. Shorter utterances and regions with more overlap are associated with lower confidence. **B** Mean confidence for ASR-transcribed utterances (n=17.563) by duration, showing that across all languages, low confidence scores are associated with shorter utterances. **C** Most characteristic elements in human-transcribed (yellow) and ASR transcribed (blue) conversational speech across all languages plotted by Scaled F-score, with the top most distinctive items for human transcripts on the right. **D** Top elements that are underrepresented or missing in ASR versus human-produced transcripts fall into three categories: short *conversational interjections*, high frequency *function words* (including contractions), and *discourse makers*.

15% of all speech (or around 1 in 8 words), which results in an inaccurate picture of conversational content, structure, and flow (Table 2 in Appendix).

Study 2: What goes missing and why

Crosslinguistic replication. Prior work on English has shown that it is especially short utterances and conversational words that go missing (Goldwater et al., 2010; Zayats et al., 2019; Mansfield et al., 2021). Here we replicate this for all six languages in our sample (Figure 3A).

Confidence metrics supplied by three of the speech-to-text systems provide a novel view of this: regions with more overlap and shorter utterances often coincide, and both are associated with dips in word-level and utterance-averaged confidence scores (Figure 3A-B). Across panels A, B and C, lighter coloured regions are associated with higher risk of being missed or misrecognized.

Overlap-vulnerability and reduction. In Figure 3C, we compare human transcripts to ASR output using Scaled F-score (Kessler, 2017), showing which items are underrepresented (top left) versus overrepresented (bottom right) in ASR output. We then take the top 15 most underrepresented items and inductively classify them as standalone interjections, function words, and discourse markers (Figure 3D), following prior work (Zayats et al., 2019; Lopez et al., 2022). We find that these categories provide good empirical coverage of what goes missing across all six languages in our sample.

Standalone interjections often occur in overlap-vulnerable contexts and are rare in ASR training data, often more formal and monologic (Liesenfeld and Dingemanse, 2022). The category of function words mostly contains highly frequent bits of morphosyntax that may occur in overlap-vulnerable positions (as the Mandarin final particles *la* and *ya*) or that are likely to be phonetically reduced (as in Dutch and French contractions of pronominal forms). Finally, discourse markers are utterance-initial fragments that help direct the flow of a conversation. These too occur in overlap-vulnerable regions and are rare in ASR training data.

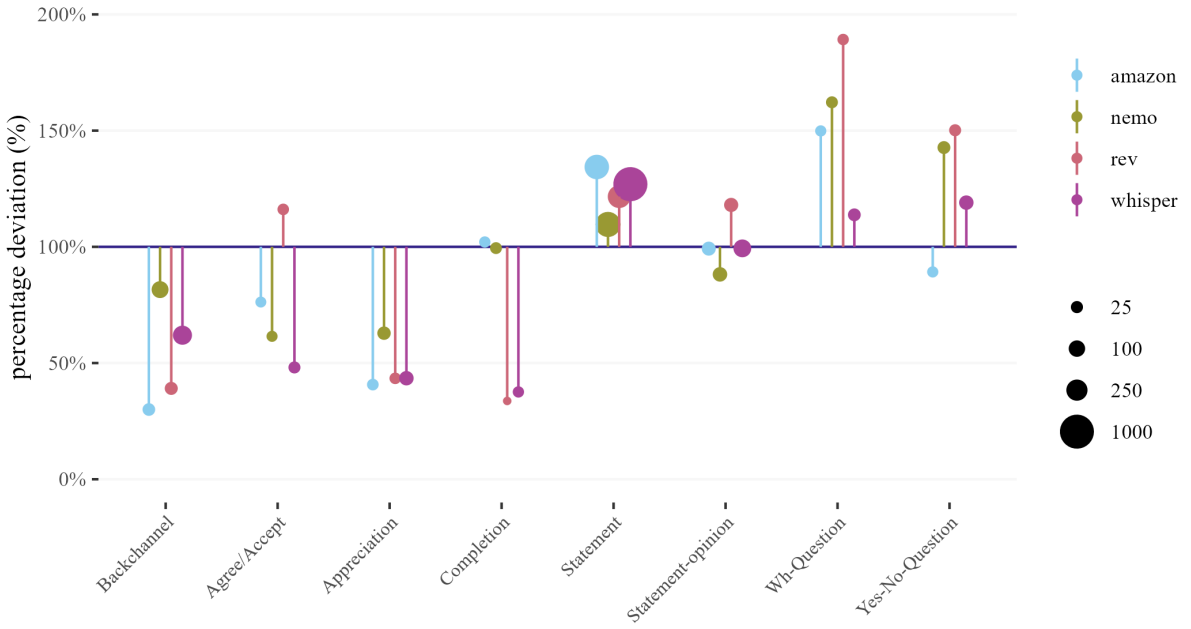


Figure 4: How different speech recognition engines warp dialog act classification in the same dataset of conversational English. For 8 frequent dialog acts, coloured lines show dialog acts based on ASR output deviate from those based on human transcripts of the same data (baseline). Dot size scales to number of times a tag is assigned. Only the most frequently assigned dialog acts (with at least 25 tokens in at least one dataset) are shown here. Mean absolute percentage deviations by ASR system: nemo 27.8%, amazon 31.4%, whisper 33.8%, rev 47.4%.

Study 3: Consequences for dialog flow

So far we have seen that the tested systems struggle with timing and overlap (study 1) and especially underrepresent conversational elements of speech (study 2). But how serious are the consequences for actual dialogue systems? One way of gauging this is to consider intent classification, a downstream task that is key to dialog state tracking and to virtually any practical application of voice UI (Ye et al., 2022; Gella et al., 2022; Jacqmin et al., 2022).

As a minimal example, we use the Switchboard dialog act tagset (Stolcke et al., 2000) as implemented in the `dialogtag` Python library (Malik, 2021) and apply it to (i) human transcripts and (ii) ASR transcripts of the same English subset of our data. By keeping the dialog tagger and the underlying data constant and manipulating only the transcription method (human versus various ASRs) we make visible how reductions and variations introduced by speech recognition systems impact dialog act classification. We intentionally use the simplest possible dialog act tagger as a proof of concept. While several more sophisticated methods exist, every method is constrained by the data it can work with, and our goal here is to merely to make visible how ASR systems can impact intent ascription and dialog state tracking.

We find that all ASRs warp dialog act classification outcomes in conversational English data (Figure 4). On average across the top 8 most frequently detected dialog act types, dialog act tags based on ASR output deviated between 27.8% (nemo) to 47.4% (rev) from tags based on human transcripts of the same data (this is absolute percentage deviation, i.e. including both overrepresentation and underrepresentation of dialog act tags).

Interactionally consequential dialog acts. Several highly interactionally relevant dialog act types are affected by speech-to-text systems. For instance (as expected based on Study 2), Backchannels and Agree/Accept tags are underrepresented across the board. This can be problematic for applications where it is important to keep track of user understanding and agreement during complex operations. Also, both the Wh-Question and Yes-No-Question dialog act tags tend to be overrepresented relative to the baseline. Since questions differ from other actions in the next moves they invite and expect, getting this wrong is directly consequential for any application in which user input is classified to determine relevant next actions.

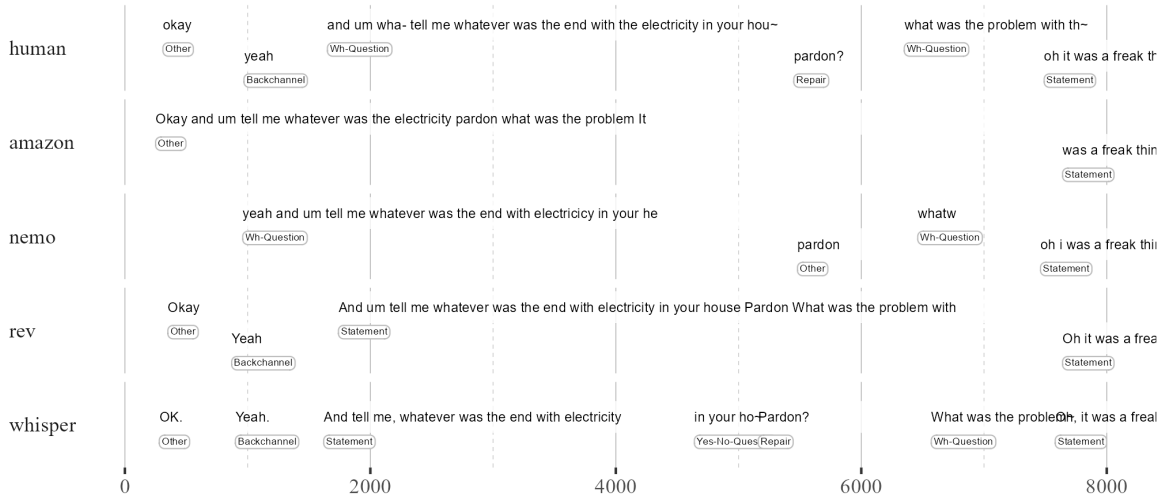


Figure 5: Excerpt of 8 seconds of English conversation showing how differences in how speech-to-text systems carry out segmentation, diarization, and transcription have direct consequences for dialog act classification.

What dialog act deformation looks like. Figure 5 shows an excerpt of English conversation in its human-annotated version (top) and four ASRs, with dialog act annotations. We selected this excerpt because it illustrates many of the larger scale patterns of underrepresentation and overrepresentation evident in Figure 4. Recall that dialog act tags are not supplied by the systems themselves, but applied to their output by `dialogtag`. Note that we speak of intent 'ascription' rather than 'recognition' to stress the fact that intents are often ambiguous and always provisional (Enfield and Sidnell, 2017).

Starting with relatively short conversational elements, we find that *yeah* is sometimes identified as a 'Backchannel' (rev, whisper), sometimes merged with adjacent turns by the other speaker (nemo), and sometimes elided entirely (amazon) — the latter two cases exemplifying the reasons ASR output generally underrepresents this category. Similarly, *pardon?* is variously identified as a 'Repair' signal (whisper), sometimes missed as a separate action because it is merged into adjacent turns by the other speaker (amazon, rev), and sometimes tagged as 'Other' (nemo), possibly because of punctuation.

Moving on to more complex elements, we see that a lumping approach to segmentation can result in interactionally important dialog acts going undetected: Amazon merges two disparate turns, producing *Okay and um tell me whatever was (...)*, which is tagged as Other. Meanwhile, a splitting approach, as Whisper appears to use, can lead to a fragment like *in your house* being tagged as Yes-No-Question in whisper output, showing one likely cause of over-representation of such question tags.

Disfluently produced questions can also pose problems: the utterance *and um wha- tell me whatever was (...)*, which features a self-repaired fragment, is sanitized and identified as a Statement in its rev and whisper versions. In the nemo output, the same turn (though merged, as we saw above, with a preceding "yeah" by the other speaker) is correctly tagged as a Wh-Question.

Even in this simple proof-of-concept, we see that ASR output can affect the ascription and classification of intents in various ways. This means that any real-world implementation relying on the systems tested here is hampered in its abilities to classify interactionally consequential social actions, making fluid interaction that much harder to achieve. Given the magnitude by which all tested ASRs deviate from human annotations in terms of timing, segmentation, diarization, overlap, and content, we expect similar kinds of distortion to appear in any systems for intent ascription and classification.

3 Discussion

The ubiquity of voice interfaces coupled with reports of human parity in speech recognition might make robust voice-driven interaction seem within easy reach. Indeed, all major vendors now advertise speech-to-text pipelines that claim both multilingual ability and conversational utility. Here we put five such systems to the test and find that the results are bleak: word error rates are nowhere near the oft-claimed human parity; performance drops dramatically for languages other than English; precise

timing and diarization is hard to come by; overlap is systematically ignored; conversational words go missing; and as a result, intent ascription and dialog state tracking are severely hampered.

Commercial speech-to-text systems are frequently exposed to conversational settings, whether it is in home use, business meetings, or customer service interactions. Our results imply that these systems are likely to fall short of several of their intended applications. Word error rate does not sufficiently reflect the performance of speech-to-text systems in most real-life contexts. The erasure of conversational elements and inability to deal with overlap renders these systems effectively oblivious to important aspects of user feedback. Differences in diarization and turn allocation across systems also have strong effects on dialog act classification, with the implication that switching vendors might have untold consequences for dialog state tracking and intent ascription.

Our results show that current speech recognition systems privilege what is said over when it is said; and that even systems claiming conversational utility appear to treat the problem as fundamentally one of turning a rich tapestry of turns into running text. These text-first design choices become visible when exposed to the rapid turn-taking patterns of natural conversation — not only to analysts in case studies like this, but inevitably also to users, where they cause friction, interactional turbulence, and user dissatisfaction. The results are in line with recent arguments that the current language technology landscape is fundamentally built around monologic text instead of dialogical talk (Dingemanse and Liesenfeld, 2022). The rise of conversational interfaces motivates a course correction if not a refurbishing of the foundations. Here we hope to have shown that data from human interaction can inform such work.

3.1 Objections

One might object that our test data is unreasonably tough, featuring open-domain informal conversation with rapid turn-taking and lots of overlap. We agree, but would counter that it is at the same time reasonably realistic: this is what typical human interactive behaviour look like. The brute facts of human interaction are something speech-to-text systems will need to reckon with if there is to be a chance of the “natural interaction" and “human-level robustness" promised by current solutions.

One might object that missing 1 in 8 words and having word error rates hovering around 50% may not be fatal, depending on what goes missing. We agree, and point out that what goes missing here is crucial for interactive speech technology. Short recurring utterances like *mmhm*, *oh* and *huh?* are the swiss army knife of conversational competence. These items enable robust communication and fluid coordination; to erase them is to rob users of their agency and to stunt the interactive capabilities of conversational technology.

One might object that dialog acts are an imperfect and language-specific way of looking at intent ascription, and that automated tagging based on form alone does not do justice to the situatedness of action (Rollet and Clavel, 2020; Levinson, 1981). We agree, and have picked dialog acts merely as a proof-of-concept to illustrate the more general problem of garbage in, garbage out: defective diarization, missing words, and neglect of timing will hamper any form-based methods for intent ascription and dialog state tracking.

3.2 Limitations

We are aware of the following limitations.

First, the human reference data is internally quite diverse, differing in recording setting and audio quality. This makes comparisons across datasets harder, so we have refrained from drawing strong comparative conclusions about possible differences across corpora and languages, instead focusing on recurring patterns of what goes missing and why.

Second, we have not collected baseline measures for non-conversational data, making it hard to estimate how large the performance offset really is relative to more typical word error rate studies. Doing this would require a parallel data collection and curation exercise for each of the languages included in our study, which is outside our scope here but represents a good target for future work.

Third, given our choice to evaluate commercial vendor pipelines, we are unable to examine or report details about ASR system architectures, model parameters, and confidence score calculations. This is a necessary consequence of black-box testing. While direct access and manipulability offer important advantages from an engineering perspective, we nonetheless think it is also important to document and evaluate the performance of widely used commercial solutions.

Fourth, we have only considered the timing infor-

mation provided in ASR results, not the latency at which the results themselves are delivered. The latency of ASR systems at runtime imposes another formidable bottleneck on voice-driven conversational interfaces, especially as long as they use endpointing methods, where response planning only starts when an utterance end is detected with some probability. User-perceived latency is the single biggest determinant of people's satisfaction with voice assistants (Shangguan et al., 2021; Bijwadia et al., 2023). Collecting realistic latency data would require implementing the tested systems in a voice UX environments with human users, which is beyond the scope of this paper (but see Aylett et al. (2023)). Empirical work on dyadic and multi-party interaction can show how people realize low latencies in real time. This is a high bar to meet, and it likely requires a radical overhaul of ASR systems towards incremental processing (Skantze, 2021).

3.3 Recommendations

The interconnectedness of all relevant processes in speech-to-text systems means that any quick fix likely has adverse consequences elsewhere. For instance, it is possible to improve diarization error rates by detecting and removing all overlap (Boakye et al., 2008) — but this means throwing out at least 15% of the data (as we show), putting human parity out of reach. Likewise, one may seek to reduce word error rates and interactional turbulence by excluding interjections (Papadopoulos Korfiatis et al., 2022), but this comes at the cost of losing all opportunity of rapid real-time user feedback. Our recommendations therefore focus on broadening the empirical basis, overcoming siloization, doing more in-depth evaluation, and incrementalizing architectures.

Improve ecological grounding. The most widely used datasets for training ASR systems still consist mostly of monologic read speech in well-resourced languages. For ASR systems to gain headway in truly interactive settings, they need to be exposed to more data that is closer to everyday language use in terms of linguistic diversity, conversational style, and participation (Aylett and Romeo, 2023). Fortunately, such data is available for an ever-wider range of languages (Liesenfeld and Dingemanse, 2022).

Overcome siloization. In a field as large and varied as automatic speech recognition, some degree of specialization is inevitable, but true progress requires working together across disciplines. As we have shown here, engineering choices in voice activity detection directly affect dialog flow, and conversation designers benefit from knowing the limitations of word error rates and the importance of overlap. Reducing the siloing of knowledge will be crucial for resolving theoretical and practical challenges of speech recognition in the era of conversational interfaces.

Value qualitative error analysis. Simple metrics make for attractive optimisation goals, but are always vulnerable to mindless metrics gaming: when a measure becomes a target, it ceases to be a good measure (Strathern, 1996). Qualitative error analysis and thorough human evaluation will remain important to truly get a handle on what goes wrong and how things can be improved (Szymański et al., 2020). This means incentives must be shifted to reward meaningful forms of evaluation over SOTA-chasing (Rogers, 2021; Church and Kordoni, 2022). It also means there is room for more exploratory methods, such as the dialog act classification measure we have begun to explore here.

Develop multidimensional evaluation. The downsides of word error rates have led to a flowering of alternative measures (Errattahi et al., 2018; Bredin, 2017). In time, the field will benefit from a degree of consolidation, and holistic evaluation of speech-to-text systems will require taking into account a wider range of measures, including but not limited to diarization, timing, duration, overlap, coverage, phonology, spelling, and word error rate. Empirical and modelling work is needed to arrive at composite evaluation measures that are precise, reproducible and meaningful.

Strengthen incremental approaches. Even if diarization quality, overlap detection and word error rates would come closer to human performance, the runtime latency of speech recognition stands in the way of fluid interactivity. To approach the rapid turn-taking and functional overlap that makes human interaction so flexible, voice-driven user interfaces will likely have to be designed as incremental architectures (Schlangen and Skantze, 2011). Promising work in this domain exists (Baumann et al., 2017; Addlesee et al., 2020; Addlesee and Damonte, 2023), and this represents an important growth area.

Use timing when available. Current systems at least provide timing for non-overlapping stretches of talk, but even that is rarely used for intent as-

cription. This despite the fact that we know timing alone can change the interpretation of a turn like “Sure.", with longer delays flipping its polarity from positive to negative (Roberts and Francis, 2013). Building on insights like this, timing might be used to improve at least some elements of intent ascription. Likewise, known facts about relative durations of turns and silences could be used to make empirically informed decisions about when to lump versus split speech material in ASR output.

4 Conclusion

When you’re a voice-driven conversational agent, life comes at you fast, and talk comes at you faster. We have presented evidence and arguments to support our contention that timing is more than a nice-to-have for any truly conversational system: it is mission critical and despite decades of attention from speech scientists remains largely unsolved today. But rather than despair we take our findings as an opportunity to identify areas where novel work can make big differences. While diarization remains hard in real-life settings, representing overlap instead of erasing it is likely to offer meaningful improvements. While overlap-vulnerable elements will always remain acoustically at risk, exposing ASRs to more ecologically valid training data and abandoning text-based sanitizing techniques will likely improve the recognition of short conversational elements. And while intent ascription will always be hampered by missing data, taking timing into account will enable new gains.

Dealing with conversational words computationally is hard: not only are their forms short and prone to overlap, their meanings are cognitively demanding and interactionally subtle. A focus on information and sentence structure over interaction and sequential organization has long enabled us to look away from these elements. As conversational words are backgrounded as ‘backchannels’ and the artful interweaving of turns is classified as mere ‘overlap’ if not ‘noise’, it becomes easy to lose sight of the intricacies of human interaction. One way to see this paper is as contributing to a reframing that is underway in the language sciences at large: a reframing that foregrounds talk over text, that attends to interaction alongside information, and that recognizes the key role of timing. Timing is the secret sauce that can turn text into talk, chat into conversation, and perhaps, one day, clunky bots into fluid interactive tools.

Acknowledgements

This work was funded by Dutch Research Council talent grant 016.vidi.185.205 to MD. We thank four anonymous reviewers for helpful comments, and JP de Ruiter for a discussion of the downsides of dialog acts.

References

Angus Addlesee and Marco Damonte. 2023. Understanding and Answering Incomplete Questions. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI ’23, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Angus Addlesee, Yanchao Yu, and Arash Eshghi. 2020. A Comprehensive Evaluation of Incremental Speech Recognition and Diarization for Conversational AI. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3492–3503, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How Might We Create Better Benchmarks for Speech Recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 22–34, Online. Association for Computational Linguistics.

Matthew Peter Aylett, Andrea Carmantini, Christoper J Pidcock, Eric Nichols, and Randy Gomez. 2023. A Pilot Evaluation of a Conversational Listener for Conversational User Interfaces. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI ’23, pages 1–6, New York, NY, USA. Association for Computing Machinery.

Matthew Peter Aylett and Marta Romeo. 2023. You Don’t Need to Speak, You Need to Listen: Robot Interaction and Human-Like Turn-Taking. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI ’23, pages 1–5, New York, NY, USA. Association for Computing Machinery.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The Fifth ’CHiME’ Speech Separation and Recognition Challenge: Dataset, Task and Baselines. In *Interspeech 2018*, pages 1561–1565. ISCA.

Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2017. Recognising Conversational Speech: What an Incremental ASR Should Do for a Dialogue System and How to Get There. In Kristiina Jokinen and Graham Wilcock, editors, *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, Lecture Notes in Electrical Engineering, pages 421–432. Springer, Singapore.

Boris Beizer. 1995. *Black-box testing: techniques for functional testing of software and systems*. Wiley, New York.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Shaan Bijwadia, Shuo-yiin Chang, Bo Li, Tara Sainath, Chao Zhang, and Yanzhang He. 2023. Unified End-to-End Speech Recognition and Endpointing for Fast and Efficient Speech Systems. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 310–316.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Kofi Boakye, Oriol Vinyals, and Gerald Friedland. 2008. Two's a crowd: improving speaker diarization by automatically identifying and excluding overlapped speech. In *Interspeech 2008*, pages 32–35. ISCA.

Hervé Bredin. 2017. pyannote.metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems. In *Interspeech 2017*, pages 3587–3591. ISCA.

Alexandra Canavan, David Graff, and George Zipperlen. 1997. CALLHOME American English Speech. Artwork Size: 1830160 KB Pages: 1830160 KB.

Alexandra Canavan and George Zipperlen. 1996a. CALLFRIEND Korean.

Alexandra Canavan and George Zipperlen. 1996b. CALLHOME Mandarin Chinese Speech. Artwork Size: 1080128 KB Pages: 1080128 KB.

May Pik Yu Chan, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole Holliday. 2022. Training and typological bias in ASR performance for world Englishes. In *Interspeech 2022*, pages 1273–1277. ISCA.

Kenneth Ward Church and Valia Kordoni. 2022. Emerging Trends: SOTA-Chasing. *Natural Language Engineering*, 28(2):249–269. Publisher: Cambridge University Press.

Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers*, 31(4):349–371.

Mark Dingemanse and Andreas Liesenfeld. 2022. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5614–5633, Dublin. Association for Computational Linguistics.

N. J. Enfield and Jack Sidnell. 2017. *The Concept of Action*. Cambridge University Press, Cambridge.

Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. 2018. Automatic Speech Recognition Errors Detection and Correction: A Review. *Procedia Computer Science*, 128:32–37.

Juan María Garrido, David Escudero, Lourdes Aguilar, Valentín Cardeñoso, Emma Rodero, Carme De-La-Mota, César González, Carlos Vivaracho, Sílvia Rustullet, Olatz Larrea, and others. 2013. Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan. *Language resources and evaluation*, 47(4):945–971. Publisher: Springer.

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

Spandana Gella, Aishwarya Padmakumar, Patrick Lange, and Dilek Hakkani-Tur. 2022. Dialog Acts for Task Driven Embodied Agents. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 111–123, Edinburgh, UK. Association for Computational Linguistics.

Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning. 2010. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Léo Jacqmin, Lina M. Rojas Barahona, and Benoit Favre. 2022. "Do you follow me?": A Survey of Recent Approaches in Dialogue State Tracking. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 336–350, Edinburgh, UK. Association for Computational Linguistics.

Jason Kessler. 2017. Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (System Demonstrations)*, pages 85–90.

Stephen C. Levinson. 1981. The essential inadequacies of speech act models of dialogue. In Herman Parret, Marina Sbisà, and Jef Verschueren, editors, *Possibilities and Limitations of Pragmatics: Proceedings of the Conference on Pragmatics, Urbino, July 8-14, 1979*, pages 473–492. Benjamins, Amsterdam.

Gina-Anne Levow, Emily P. Ahn, and Emily M. Bender. 2021. Developing a Shared Task for Speech Processing on Endangered Languages. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1(2).

Andreas Liesenfeld and Mark Dingemanse. 2022. Building and curating conversational corpora for diversity-aware language science and technology. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1178–1192, Marseille. ArXiv: 2203.03399.

Alianda Lopez, Andreas Liesenfeld, and Mark Dingemanse. 2022. Evaluation of Automatic Speech Recognition for Conversational Speech in Dutch, English and German: What Goes Missing? In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, Potsdam.

Jerome Louradour. 2023. whisper-timestamped. Original-date: 2023-01-13T11:30:19Z.

Bhavitvya Malik. 2021. DialogTag: A python library to classify dialogue tag.

Courtney Mansfield, Sara Ng, Gina-Anne Levow, Richard A. Wright, and Mari Ostendorf. 2021. Revisiting Parity of Human vs. Machine Conversational Speech Transcription. In *Interspeech 2021*, pages 1997–2001. ISCA.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A Dataset Of Primary Care Mock Consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Felicia Roberts and Alexander L. Francis. 2013. Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, 133(6):EL471–EL477.

Anna Rogers. 2021. Changing the World by Changing the Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2182–2194, Online. Association for Computational Linguistics.

Nicolas Rollet and Chloé Clavel. 2020. "Talk to you later": Doing social robotics with conversation analysis. Towards the development of an automatic system for the prediction of disengagement. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 21(2):268–292.

Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. 2021. The Third DIHARD Diarization Challenge. ArXiv:2012.01477 [cs, eess].

Odette Scharenborg. 2007. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5):336–347.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue & Discourse*, 2(1):83–111. Number: 1.

Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, and Sanjeev Khudanpur. 2018. Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge. In *Interspeech 2018*, pages 2808–2812. ISCA.

Yuan Shangguan, Rohit Prabhavalkar, Hang Su, Jay Mahadeokar, Yangyang Shi, Jiatong Zhou, Chunyang Wu, Duc Le, Ozlem Kalinli, Christian Fuegen, and Michael L. Seltzer. 2021. Dissecting User-Perceived Latency of On-Device E2E Speech Recognition. In *Interspeech 2021*, pages 4553–4557. ISCA.

Elizabeth Shriberg. 2001. To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(01):153–169.

Elizabeth Shriberg. 2005. Spontaneous speech: how people really talk and why engineers should care. In *Interspeech 2005*, pages 1781–1784. ISCA.

Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67:101178.

Andreas Stolcke and Jasha Droppo. 2017. Comparing Human and Machine Errors in Conversational Speech Transcription. In *Interspeech 2017*, pages 137–141. ISCA.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

Marilyn Strathern. 1996. From Improvement to Enhancement: An Anthropological Comment on the Audit Culture. *Cambridge Anthropology*, 19(3):1–21.

Masayuki Suzuki, Gakuto Kurata, Tohru Nagano, and Ryuki Tachibana. 2016. Speech recognition robust against speech overlapping in monaural recordings of

telephone conversations. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5685–5689. ISSN: 2379-190X.

Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Ży\la Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295, Online. Association for Computational Linguistics.

Taalunie. 2014. Corpus Gesproken Nederlands - CGN (Version 2.0.3).

Divya Tadimeti, Kallirroi Georgila, and David Traum. 2022. Evaluation of Off-the-shelf Speech Recognizers on Different Accents in a Dialogue Domain. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 6001–6008, Marseille.

Francisco Torreira, Martine Adda-Decker, and Mirjam Ernestus. 2010. The Nijmegen Corpus of Casual French. *Speech Communication*, 52(3):201–212.

Rob van Son, Wieneke Wesseling, Eric Sanders, and Henk van den Heuvel. 2008. The IFADV corpus: A free dialog video corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Chenchen Ye, Lizi Liao, Fuli Feng, Wei Ji, and Tat-Seng Chua. 2022. Structured and Natural Responses Co-generation for Conversational Search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 155–164, New York, NY, USA. Association for Computing Machinery.

Vicky Zayats, Trang Tran, Richard Wright, Courtney Mansfield, and Mari Ostendorf. 2019. Disfluencies and Human Speech Transcription Errors. In *Proceedings of Interspeech 2019*, pages 3088–3092. ISCA.

Özgür Çetin and Elizabeth Shriberg. 2006. Overlap in Meetings: ASR Effects and Analysis by Dialog Factors, Speakers, and Collection Site. In *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science, pages 212–224, Berlin, Heidelberg. Springer.

A Appendix

A.1 Datasheets

Table 1 shows the different corpora used in the study, detailing how many conversations were included and their total lengths in minutes. Every language contains approximately one hour worth of conversations, and when feasible, different interactional settings were incorporated (resulting to two corpora for Dutch). Each processing step is reflected in the processing pipeline avaliable in the repository, which also includes a datasheet (Gebru et al., 2021) and instructions on how to replicate the study given access to the data. For Dutch and Spanish, the evaluation datasets are freely available for academic research purposes. For English, French, Korean and Mandarin, the study repository provides information how to obtain the datasets used: `https://osf.io/hruva`.

Language	Corpus	Conversations (n)	Length (mins)
Dutch	The Corpus of Spoken Dutch (CGN) (Taalunie, 2014)	3	30.11
	IFADV Corpus (van Son et al., 2008)	2	29.97
English	CALLHOME American English (Canavan et al., 1997)	6	60.25
French	Nijmegen Corpus of Casual French (Torreira et al., 2010)	6	60.39
Korean	CALLFRIEND Korean (Canavan and Zipperlen, 1996a)	4	59.99
Mandarin	CALLHOME Mandarin Chinese (Canavan and Zipperlen, 1996b)	6	60.20
Spanish	Glissando Corpus (Garrido et al., 2013)	6	60.34

Table 1: Corpora used in the study, with each language represented by approximately one hour of informal conversations.

A.2 Study 1 methods

For both the human and ASR-transcribed data we calculate turn transition times in ms, number of speaker transitions, and the presence and duration of overlaps. For error analysis at the content level, we removed punctuation and excluded tags for non-speech behavior such as [laugh] and [breath] to bring all transcripts to a more comparable format. We used `cleantext` for pre-processing and `whitespace` for tokenizing. We then calculated word error rate (WER) using `jiwer`, and for a more in-depth investigation on the differences between human and speech-to-text annotations, we adopt Scaled F-score (Kessler, 2017) as a metric of n-gram salience scoring.

A.3 Study 1 detailed results

Table 2 provides a more detailed look at key differences between human transcriptions and ASR output across the six languages in our sample. For every language, it lists the mean amount of speech covered by the transcriptions (coverage); the mean total number of words in the transcripts (words); the mean turn duration in milliseconds; and the mean percentage of overlapping annotations.

Human vs ASR	Coverage (min)	Words (n)	Turn duration (ms)	Overlap (speech %)
Dutch	63	12023	2840	13.4
	47	9396	5897	0
English	65	13895	2811	12.6
	55	10994	6647	0
French	64	13564	4357	14.4
	49	8359	7042	0
Korean	74	9632	3280	20.8
	43	5923	4186	0
Mandarin	66	15349	2538	15.8
	53	8188	7301	0
Spanish	63	11868	4620	10.5
	57	10177	7534	0

Table 2: Comparison of human (top) and ASR transcripts (bottom) in each language in terms of coverage (amount of speech transcribed (in minutes), number of words, mean duration of each conversational turn (ms), and percentage of overlapped annotations relative to the length of the whole conversation. Human annotations add up to 395 minutes of transcribed speech; ASR-produced annotations for the same data on average add up to only 304, or 77% of the observed speech.

Enhancing Task Bot Engagement with Synthesized Open-Domain Dialog

Miaoran Li[†]
Iowa State University
limr@iastate.edu

Baolin Peng
Microsoft Research
baolin.peng@microsoft.com

Michel Galley
Microsoft Research
mgalley@microsoft.com

Jianfeng Gao
Microsoft Research
jfgao@microsoft.com

Zhu Zhang
University of Rhode Island
zhuzhang@uri.edu

Abstract

The construction of dialog systems for various types of conversations, such as task-oriented dialog (TOD) and open-domain dialog (ODD), has been an active area of research. In order to more closely mimic human-like conversations that often involve the fusion of different dialog modes, it is important to develop systems that can effectively handle both TOD and ODD and access different knowledge sources. In this work, we present a new automatic framework to enrich TODs with synthesized ODDs. We also introduce the PivotBot model, which is capable of handling both TOD and ODD modes and can access different knowledge sources to generate informative responses. Evaluation results indicate the superior ability of the proposed model to switch smoothly between TOD and ODD tasks.

1 Introduction

Task-oriented dialog (TOD) systems and open-domain dialog (ODD) systems are two active areas of Conversational AI study (Gao et al., 2018; Ni et al., 2022). However, most of the existing studies model TOD and ODD systems separately, leading to a gap between the capabilities of these systems and natural human conversations. In real-world conversations, different dialog modes are often fused, as shown in Figure 1. The conversation may start with casual chats and then move towards task-related requests. Along the way, the user may express interest in entities mentioned in the conversation, such as Mediterranean food in the given example, leading to a brief ODD regarding the entity of interest. The user then returns to task completion, keeping the requests in mind while maintaining a casual conversation.

To address the challenge of training dialog models to handle both TOD and ODD modes, previous research has suggested training models on mixture of TOD and ODD datasets (Zhao et al., 2022) or enriching existing TOD datasets by combining chitchat with TOD system responses (Sun et al., 2021; Chen et al., 2022) or adding ODD to the beginning or end of a TOD (Young et al., 2022). However, these approaches have limitations, including limited information in chitchat augmentation and a lack of explicit distinction between dialog modes. Additionally, creating new datasets through human annotation is time-consuming and expensive. While Chiu et al. (2022) have introduced a framework for automatically generating dialogs that transition from ODD to TOD, this method may not be suitable for various mode transitions and cannot simulate informative system utterances with external knowledge.

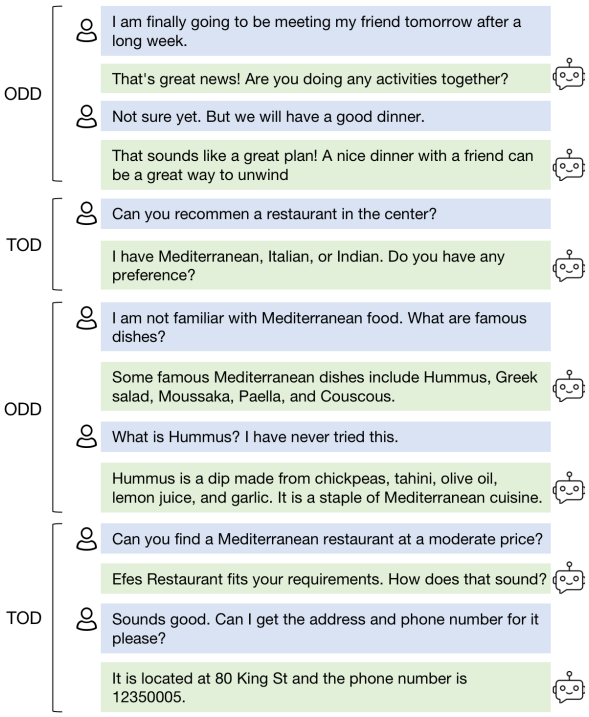


Figure 1: An example dialog that contains multiple transitions between different dialog modes.

In this work, we introduce a framework to au-

[†] This work was done during an internship at Microsoft Research.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 496–508
September 11–15, 2023. ©2023 Association for Computational Linguistics

tomatically enrich TODs with synthesized ODDs. Our approach assumes that users lead conversations with explicit intentions, and that the system's objective is not only to fulfill users' requests but also to generate engaging responses on open-domain topics using external knowledge. We also consider general settings with more flexible dialog mode switches.

This paper makes the following contributions: (*i*) We introduce a general framework for automatically enriching a TOD with knowledge-grounded ODDs and construct the MultiWOZChat dataset using this framework. (*ii*) We design a unified model, PivotBot, that performs both TOD and ODD tasks by predicting the appropriate dialog mode and accessing knowledge sources for response generation. (*iii*) We show experimental results that demonstrate the effectiveness of PivotBot in conducting seamless conversations of both types.

2 Proposed Framework

Figure 2 shows the proposed framework for automatically synthesizing one or more knowledge-grounded ODDs to a given TOD. The framework consists of three stages: (1) ODD initialization (2) ODD simulation, and (3) ODD to TOD transition. We define the following notations:

- Denote TOD by $D = \{\boldsymbol{u}_1^{d_1}, \boldsymbol{s}_1^{d_1}, ..., \boldsymbol{u}_{n_1}^{d_1}, \boldsymbol{s}_{n_1}^{d_1}, ..., \boldsymbol{u}_{n_1+n_2}^{d_2}, \boldsymbol{s}_{n_1+n_2}^{d_2}, ..., \boldsymbol{u}_n^{d_N}, \boldsymbol{s}_n^{d_N}\}$,[1] where N is the number of domains in the dialog, $\boldsymbol{u}_i^{d_j}$ and $\boldsymbol{s}_i^{d_j}$ are user and system utterances at turn i in domain j, n_i is the number of turns in domain d_i, and n is the total number of turns in D.
- Denote synthesized ODD by $D' = \{\boldsymbol{u}'_1, \boldsymbol{s}'_1, ..., \boldsymbol{u}'_{n'}, \boldsymbol{s}'_n\}$, where n' is the number of turns in the ODD, $\boldsymbol{u}'_t$ and $\boldsymbol{s}'_t$ represent user and system utterances at turn t, respectively.

Detailed implementation of each module can be found in Appendix A.

2.1 ODD Initialization

Given a TOD D, we initialize the synthesized ODD D' in two ways. If the ODD serves as the preface to the TOD, it is initialized by a randomly sampled user persona. If the ODD is inserted into the TOD as interludes and generated based on the TOD history, we leverage an existing chatbot to simulate a user utterance that can be inserted at a potential

[1]For settings we do not care about domains in TOD, D can be simplified to $\{\boldsymbol{u}_1, \boldsymbol{s}_1, ..., \boldsymbol{u}_n, \boldsymbol{s}_n\}$.

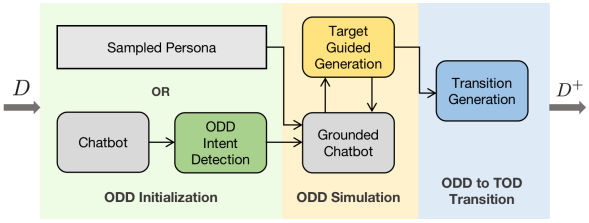


Figure 2: Framework for enriching a given TOD D with ODD. The framework consists of three phases: ODD initialization, ODD simulation, and ODD-to-TOD transition. Rounded and sharp-corner boxes represent models and variables, respectively. The gray color indicates that the model is off-the-shelf. The output is the augmented dialog D^+.

location. We then utilize this simulated user utterance to detect whether the user intends to have an open-domain conversation. The off-the-shelf BlenderBot model (Roller et al., 2021) is used as the chatbot in the implementation. These two initialization methods are employed across diverse simulation settings (Section 2.4).

ODD Intent Detection To determine the appropriate time to include an ODD during task completion, we focus on detecting the user's intent to divert the conversation from the task and discuss context-related topics. Given a user utterance $\boldsymbol{u} = \{u_1, ..., u_n\}$, where u_i is the i-th token in the utterance, the ODD intent detection model aims to predict whether the utterance is in a TOD setting or ODD setting. The model is trained by minimizing cross-entropy loss:

$$\mathcal{L}(\hat{I}, I) = \sum_{i=1}^{N} -(\mathbb{1}(\hat{I}_i = I_i) \log(p_{\boldsymbol{\theta}}(I_i) + (1 - \mathbb{1}(\hat{I}_i = I_i)) \log(1 - p_{\boldsymbol{\theta}}(I_i)), \quad (1)$$

where N is number of training examples, $\hat{I}_i$ and I_i are predicted and ground truth intent of the i-th training example, $\boldsymbol{\theta}$ is the parameters of the model.

2.2 ODD Simulation

After initializing the ODD, we use a knowledge-grounded chatbot to mimic a system with access to external knowledge and a target-guided generation model to simulate a user. In practice, we adopt the BlenderBot 2.0 model (Xu et al., 2022; Komeili et al., 2022) and BlenderBot model to simulate system and user utterances, respectively. The ODD is considered complete if a goal $\boldsymbol{g}$ extracted from the subsequent TOD snippet is mentioned in a simulated user utterance.

Target-guided Generation To simulate the human user in the given TOD, we train a target-guided generation model that is designed to generate utterances based on the dialogue history and mention a preset target at the end of the ODD. The target-guided generation model is expected to generate a user utterance $\boldsymbol{u}'$ at turn $t+1$ based on a pre-determined target $\boldsymbol{g}$ and dialog context $\boldsymbol{c}$ up to turn t.[2] The target is extracted from the initial user utterance of the subsequent TOD part. Given pre-determined ODD goal $\boldsymbol{g} = \{g_1, ..., g_{N_g}\}$ and context $\boldsymbol{c}$, where g_i is the i-th token in the goal, the training objective is defined as

$$\begin{aligned}\mathcal{L}_U &= \log p(\boldsymbol{u}'_{t+1} \mid \boldsymbol{g}, \boldsymbol{c}) \\ &= \sum_{i=1}^{N_u} \log p_{\boldsymbol{\theta}}(u'_{t+1,i} \mid u'_{t+1,<i}, \boldsymbol{g}, \boldsymbol{c}), \end{aligned} \quad (2)$$

where $\boldsymbol{\theta}$ is the set of trainable parameters in the model, N_u is the target length of predicted user utterance, and $u_{t+1,<i}$ represents tokens before index i of predicted user utterance at turn $t+1$.

2.3 ODD to TOD Transition

Finally, we generate a transition from the simulated ODD to the subsequent TOD to make the dialog more natural. The goal of transition generation is to predict a system utterance that can smoothly connect the last user utterance in the ODD with the initial user utterance in the following TOD. The training objective is

$$\begin{aligned}\mathcal{L}_T &= \log p(\boldsymbol{s}'_t \mid \boldsymbol{u}'_t, \boldsymbol{u}_{t+1}) \\ &= \sum_{i=1}^{N_s} \log p_{\boldsymbol{\theta}}(s'_{t,i} \mid s'_{t,<i}, \boldsymbol{u}'_t, \boldsymbol{u}_{t+1}), \end{aligned} \quad (3)$$

where $\boldsymbol{u}'_t$ is the last user utterance in generated ODD, $\boldsymbol{u}_{t+1}$ is the first user utterance in the following TOD, $\boldsymbol{s}'_t$ is the transition system utterance.

2.4 Simulation Settings

Inspired by previous research that aims to make dialogs more natural and engaging by adding context to a given dialog (Young et al., 2022) or inserting topic transition turns (Sevegnani et al., 2021), we consider three simulation settings: prepending an ODD to a TOD, inserting an ODD as domain transition turns, and allowing ODDs to occur at any point during task completion. The illustration of three settings is shown in Figure 3.

[2]We conducted pilot experiments using formulations that included keyword prediction, but found not significant performance improvement. Thus, we decided to use the simplest formulation without turn-level keyword transitions.

Setting 1: Prepending ODD to TOD (`INITIAL`) We prepend an ODD to a TOD to generate dialogs with one mode switch from ODD to TOD. We assume that users initiate the conversation by having a quick ODD and then move forward to task completion. Assuming users start with a quick ODD and then move to task completion, we initialize the ODD with a persona from a manually created persona set and use a keyword from the initial user utterance in the subsequent TOD as the goal for the synthesized ODD. Once the target is mentioned in a user utterance, the ODD simulation stops. The transition generation model is then used to connect the synthesized ODD and TOD.

Setting 2: Inserting ODD for Domain Transition in TOD (`TRANSITION`) To make domain transitions in TODs more natural, we insert an ODD as transition turns. Suppose a TOD D contains N domains, where $N \geq 2$. We initialize an ODD using a chatbot after completing the conversation in domain i, and use intent detection model to select an utterance indicating ODD intent. The target of the ODD snippet is extracted from the first user utterance in domain $i+1$. The simulation and transition generation are similar to the previous setting. In the implementation, we only add an ODD to transition from the first domain to the second domain, and use the BlenderBot model for ODD initialization. The final dialogs contain two mode switches.

Setting 3: Inserting Multiple Chitchats to Enrich TODs (`MULTIPLE`) In this more flexible setting, users can initiate conversations with requests and engage in small talk throughout the dialogue. The approach for generating ODDs is the same as in the `TRANSITION` setting, with the difference that we attempt to insert an ODD after each system utterance $\boldsymbol{s}_i$. This allows for multiple mode switches in the final dialogue.

2.5 MultiWOZChat Dataset

We construct MultiWOZChat dataset using the new framework to automatically enrich TODs from the MultiWOZ 2.1 dataset (Eric et al., 2020a). Table 1 summarizes basic statistics of the new dataset. Focusing on the few-shot training setting, the dataset consists of 500, 198, and 1100 dialogs for the training, validation, and test sets respectively. In the `INITIAL` setting, the average length of a prepended ODD is three turns, and the mean utterance length is 16.18 tokens. In the `TRANSITION` setting, the average length of a transition ODD is shorter than

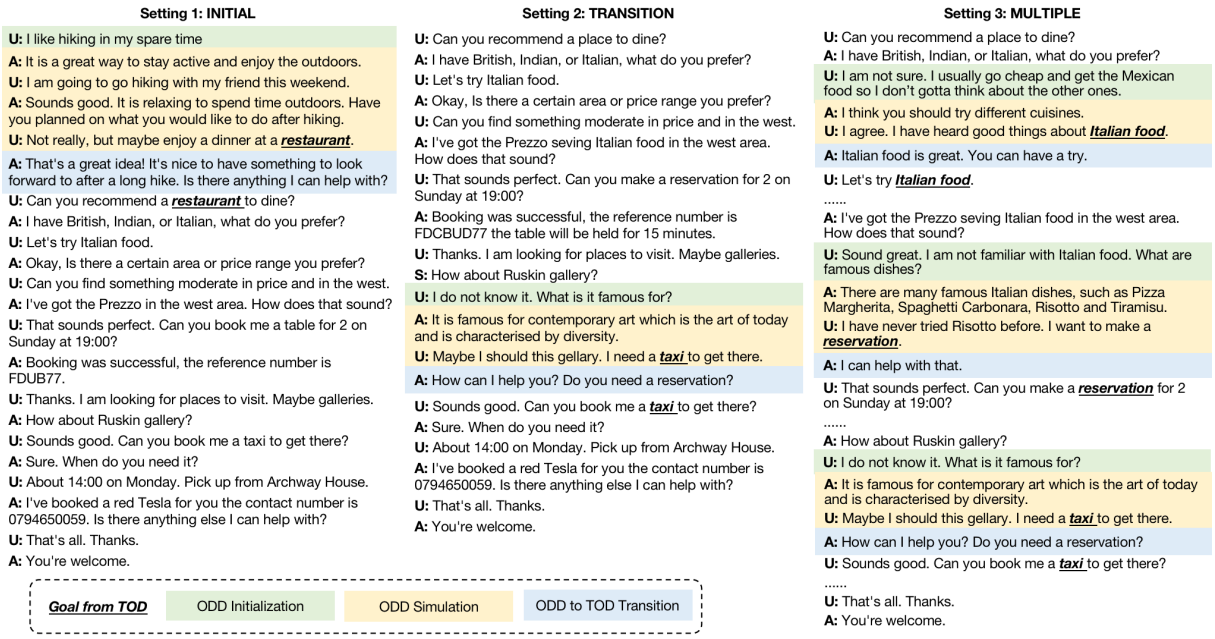


Figure 3: Illustration of three simulation settings. Given a TOD between a user (U) and a system agent (A), we consider three settings to synthesize ODD(s) to the TOD.

three turns. In the MULTIPLE setting, the average number of ODDs inserted into a TOD is four, and each ODD snippet has an average length of two turns. In the TRANSITION and MULTIPLE settings, the ODD durations are shorter, as they occur during task completion, and we do not want the conversation to be distracted from the task completion.

Setting	Split	Avg. mode switch	Total ODD turn	Total TOD turn	Avg. ODD turn	Avg. TOD turn	Avg. ODD length	Avg. TOD length
INITIAL	Train	1	1524	4086	3.05	8.17	16.18	18.07
	Dev		565	1599	2.85	8.08	15.90	18.30
	Test		3248	9031	2.95	8.21	15.99	18.17
TRANSITION	Train	2	1301	4086	2.60	8.17	18.22	18.07
	Dev		510	1599	2.58	8.08	18.26	18.30
	Test		2923	9031	2.66	8.21	18.21	18.17
MULTIPLE	Train	4.96	4356	4086	8.71	8.17	17.80	18.07
	Dev	4.90	1599	1599	8.47	8.08	17.61	18.30
	Test	5.11	9995	9031	9.87	8.21	17.82	18.17

Table 1: Statistics of simulated dialogs in different settings. The training, validation, and test sets comprise 500, 198, and 1100 dialogs, respectively.

3 Methodology

3.1 Problem Formulation

The full task consists of three processes: state prediction, knowledge retrieval, and knowledge-grounded response generation. We use off-the-shelf models for knowledge retrieval, which can be a database lookup or a search engine,[3] and do not consider it as a subtask. The full task is then divided into two subtasks: state prediction and knowledge-grounded response generation. In the t-th turn of a dialog, the model predicts the state $\boldsymbol{s}$ based on the dialog history $\boldsymbol{h} = \{\boldsymbol{u}_{t-k}, \boldsymbol{r}_{t-k}, ..., \boldsymbol{u}_t\}$, where k is the size of the history window, $\boldsymbol{u}_i$ and $\boldsymbol{r}_i$ represent the user utterance and system response at the i-th turn, respectively. The state indicates the appropriate dialog mode and the query to obtain knowledge $\boldsymbol{k}$. The model then generates a response $\boldsymbol{r}$ based on the dialog history, predicted state and knowledge.

[3]In the implementation, we adopted the Bing search engine.

3.2 PivotBot

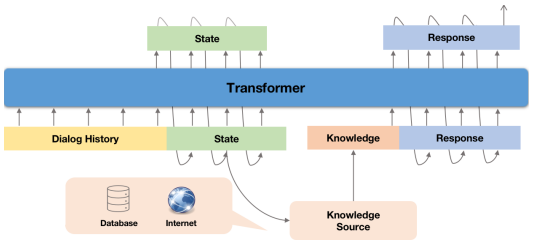
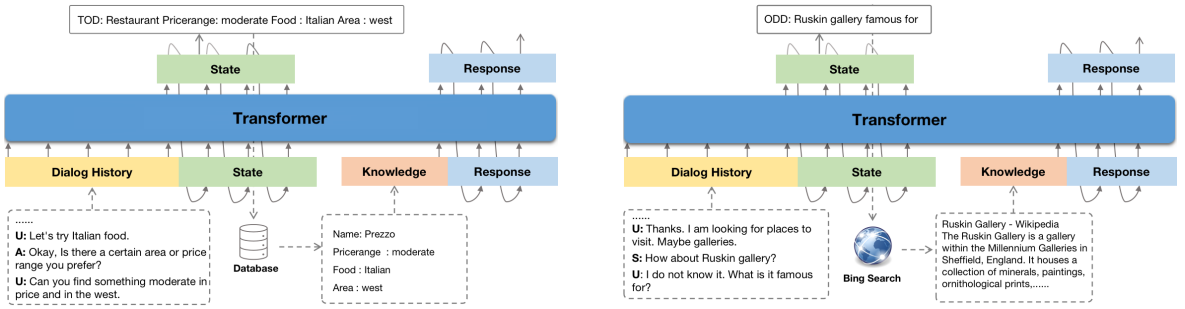


Figure 4: Overall architecture of the PivotBot model

We construct a unified model, PivotBot, as shown in Figure 4. PivotBot first predicts a state indicating the appropriate dialog mode and query to obtain knowledge based on the dialog history. The knowledge acquisition is completed by off-the-shelf models based on the prediction. Finally, the model performs grounded generation to generate a response.



(a) Example of performing TOD modeling

(b) Example of performing ODD with external knowledge

Figure 5: Examples of the proposed model predicting different states

State Prediction State $\boldsymbol{s}$ tracks a user's goal throughout a dialog. In particular, a state $\boldsymbol{s}$ is in the form `m:q`, where `m` represents the dialog mode, and `q` stands for the query to acquire knowledge from a knowledge source. We consider two dialog modes: TOD modeling and knowledge-grounded ODD. If the model predicts performing TOD modeling, a database state is obtained from the predefined database using the predicted belief state (shown in Figure 5 (a)). If the state indicates the dialog mode is ODD, external knowledge can be retrieved from the Web using the predicted search query (shown in Figure 5 (b)). If the search query is empty, it implies that external knowledge is not needed for response generation, and the retrieved knowledge is also empty. Given dialog history $\boldsymbol{h}$, the training objective of state prediction can be formulated as

$$\mathcal{L}_S = \log p(\boldsymbol{s} \mid \boldsymbol{h}) = \sum_{i=1}^{N_t} \log p_{\boldsymbol{\theta}}(s_i \mid s_{<i}, \boldsymbol{h}), \quad (4)$$

where $\boldsymbol{\theta}a$ represents trainable parameters in the model, N_t is the target length of predicted state sequence, and $s_{<i}$ denotes tokens before index i.

Grounded Generation System response $\boldsymbol{r} = \{r_1, r_2, ..., r_{N_r}\}$ with length N_r is generated grounded on dialog history $\boldsymbol{h}$, predicted state $\boldsymbol{s}$ and retrieved knowledge $\boldsymbol{k}$. In this work, the knowledge can be a database state that contains records satisfying the conditions of the belief state or retrieval results based on the search query. The training objective is defined as

$$\begin{aligned}\mathcal{L}_R &= \log p(\boldsymbol{r} \mid \boldsymbol{h}, \boldsymbol{s}, \boldsymbol{k}) \\ &= \sum_{i=1}^{N_r} \log p_{\boldsymbol{\theta}}(r_i \mid r_{<i}, \boldsymbol{h}, \boldsymbol{s}, \boldsymbol{k}). \end{aligned} \quad (5)$$

Training Objective of Full Task A training example consists of four components: dialog history $\boldsymbol{h}$, state $\boldsymbol{s}$, retrieved knowledge $\boldsymbol{k}$, and (delexicalized) dialog response $\boldsymbol{r}$. The overall training objective is

$$\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D}) = \sum_{i=1}^{N_D} (\mathcal{L}_S(\boldsymbol{x_i}) + \mathcal{L}_R(\boldsymbol{x_i})), \quad (6)$$

where $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{N_D}$ is the training dataset containing N_D training examples.

4 Experiments

4.1 Experimental Setup

We train models using 100, 200, and 500 dialogs and evaluate them on the entire test set. Our primary focus is evaluating the models trained in the few-shot setting, as this approach more closely reflects real-world scenarios.

Baselines Previous studies either do not distinguish different dialog modes or only focus on social chats without external knowledge. However, our task requires models to switch between ODD and TOD modes and choose the appropriate knowledge source. To ensure a fair comparison, we train two baselines for our problem setting instead of comparing with models designed for different settings.

- TaskBot serves as a baseline and is only capable of performing TOD with access to a database, which is trained solely on TOD turns in the MultiWOZChat dataset.
- ChatBot is a baseline model that can only perform ODD, which is trained on ODD turns in the MultiWOZChat dataset.

The baselines and PivotBot are implemented using HuggingFace T5-base (Raffel et al., 2020) and GODEL (Peng et al., 2022). Further details of implementations can be found in Appendix A.

Implementation The models are implemented using HuggingFace T5-base and GODEL. Training examples are truncated or padded to a length of 512. To ensure input strings contain dialog history and retrieved knowledge, the history is truncated on the left with a max length of 256 and consists of five utterances with a history window size of 2. AdamW optimizer (Loshchilov and Hutter, 2019) with a constant learning rate of 0.001 is used for training with a mini-batch size of 8 on a Tesla P100 for up to 15 epochs or until no validation loss decrease is observed. Each setting is evaluated eight times with random seeds.

Evaluation Metrics We evaluate the performance of the models in three settings: (1) standard TOD completion (Budzianowski et al., 2018; Eric et al., 2020b; Nekvinda and Dušek, 2021), (2) ODD response generation, and (3) the full task involving both TOD and ODD.

We evaluate TOD completion using four metrics: (1) `BLEU` (Papineni et al., 2002) measures the fluency of the generated responses; (2) `Success` indicates if all requested attributes are answered; (3) `Inform` measures whether the correct entity is provided (e.g., restaurant address); (4) `Combine` score is an overall measure calculated as as (`Inform`+`Success`) $\times$ 0.5 + `BLEU`.

We evaluate ODD using three metrics: (1) `Accuracy` measures the model's ability to predict the correct dialog mode, which can be calculated by comparing the predicted dialog mode with the ground truth mode; (2) `Success Rate` assesses the model's performance in state prediction at the dialog level, and measures the model's potential for success in the ODD task. It can be calculated by dividing the number of dialogs in which the model correctly predicts the dialog mode for all ODD turns by the total number of dialogs with ODD turns; (3) `BLEU` measures the naturalness of the model's responses.

We evaluate the model's performance on the full task using `BLEU`, `Inform`, `Success`, and `Combine` score. `BLEU` score is computed for all responses in the dialogs, while `Inform` and `Success` metrics are limited to dialogs that succeed in both TOD modeling and ODD tasks. The potential success of the ODD task is used as an indicator, and `Inform` and `Success` are computed for dialogs where the dialog mode predictions for all ODD turns are accurate.

Human Evaluation Setup We conducted two-phase human evaluation. In the first stage, we hired Amazon Mechanical Turk workers to interact with three models: TaskBot with T5 as the backbone (T5-TaskBot), PivotBot with T5 as the backbone (T5-PivotBot), and PivotBot with GODEL as the backbone (GODEL-PivotBot). The workers were provided with information-seeking goals from the MultiWOZ 2.1 dataset and allowed to chat freely with the models to complete the goals. After each conversation, workers rated the appropriateness (Moghe et al., 2018) and engagingness (Zhang et al., 2018) of the model's responses on a 5-point Likert scale and indicated if all requests were completed. Appropriateness assesses the model's ability to understand users' utterances and requests and provide reasonable responses, while engagingness evaluates whether the model generates engaging responses and facilitates smooth conversation flow for users.

To ensure the quality of interactions during the first stage, we employed onboarding tasks with simplified information-seeking goals. Only qualified workers who can complete the onboarding task were granted access to the main task with higher rewards. Both the onboarding and main task submissions were required to cover all necessary keywords and phrases, and each utterance had to be meaningful and not excessively brief. Additionally, we implemented manual checks on randomly sampled submissions to maintain the quality of collected results.

In the second stage, we conducted a static evaluation of the dialogs collected in the previous phase. Each worker was presented with a pair of dialogs, one produced by T5-TaskBot and the other by T5-PivotBot, or one produced by T5-PivotBot and the other by GODEL-PivotBot, and was asked to choose the better dialog based on the system performance. Then workers rated the appropriateness and engagingness of each system's utterances in the dialogs using a 5-point Likert scale.

4.2 Automatic Evaluation Results

We present the results for models trained in the few-shot setting using 100 training dialogs with the GODEL backbone.[4] For the full task evaluation, we only report the combined score. The evaluation

[4]We also evaluated the models using the T5-base backbone and found that models with the GODEL backbone outperform those based on T5-base, with statistically significant performance differences.

Model	Full Task	TOD Evaluation				ODD Evaluation		
	Combined	BLEU	Success	Inform	Combined	Accuracy	Success Rate	BLEU
TaskBot	12.26(0.43)	**15.00**(0.57)	37.43(4.02)	52.61(4.26)	60.01(4.41)	0.00(0.00)	0.00(0.00)	1.33(0.22)
ChatBot	7.98(0.25)	0.97(0.16)	0.60(0.00)	10.70(0.00)	6.62(0.16)	**99.93**(0.05)	**99.79**(0.16)	6.43(0.51)
PivotBot	**58.06**(5.15)	14.90(0.58)	**38.66**(5.22)	**53.55**(5.62)	**61.01**(5.48)	98.90(0.45)	97.35(0.98)	**6.82**(0.41)

Table 2: End-to-end evaluation in the INITIAL setting. Mean values and standard deviations (in parentheses) are reported.

Model	Full Task	TOD Evaluation				ODD Evaluation		
	Combined	BLEU	Success	Inform	Combined	Accuracy	Success Rate	BLEU
TaskBot	12.37(0.36)	**15.02**(0.46)	**35.43**(4.14)	**50.56**(5.35)	**58.02**(5.03)	0.00(0.00)	0.00(0.00)	1.22(0.12)
ChatBot	7.88(0.13)	1.22(0.16)	0.60(0.00)	10.70(0.00)	6.87(0.16)	**100.00**(0.01)	**99.99**(0.03)	**5.35**(0.18)
PivotBot	**49.58**(7.13)	14.92(0.64)	33.49(6.13)	47.06(8.21)	55.19(7.56)	96.17(0.64)	90.00(1.72)	4.97(0.28)

Table 3: End-to-end evaluation in the TRANSITION setting. Mean values and standard deviations (in parentheses) are reported.

results using 200 and 500 training dialogs are in Appendix B.

INITIAL Setting Evaluation Table 2 shows the evaluation results in the INITIAL setting. PivotBot significantly outperforms the baseline models in the full task evaluation, demonstrating the importance of incorporating different dialog modes. PivotBot also slightly outperforms TaskBot in the TOD task in terms of the Combined score. This suggests that the ability to handle both TOD and ODD tasks with appropriate dialog modes and knowledge sources is critical for PivotBot to excel in the full task. While ChatBot cannot provide requested attributes or entities, it performs better than other models in predicting the dialog mode in the ODD evaluation setting. Though PivotBot cannot beat ChatBot in the ODD evaluation, it achieves comparable results while generating more fluent responses and simultaneously handling task completion.

TRANSITION Setting Evaluation Table 3 contains evaluation results in the TRANSITION setting. PivotBot performs significantly better than baselines in the full task. TaskBot slightly outperforms PivotBot in the TOD modeling task. ChatBot still achieves the best performance in the ODD task. Though PivotBot cannot perform better than baselines in single task evaluation, it can obtain comparable results with the specialist baselines. The gap between ChatBot and PivotBot in success rate is more obvious, indicating that it is more challenging for the model to learn both dialog modes simultaneously and accurately predict the mode when the mode switches in dialogs become more complex.

MULTIPLE Setting Evaluation The evaluation results in the MULTIPLE setting are presented in Table 4. In the full task evaluation, PivotBot remains the best-performing model. The performance of TaskBot and PivotBot is comparable in the TOD task. However, in the ODD task evaluation, while PivotBot's turn-level prediction accuracy does not significantly decrease, the model is more likely to fail in the ODD task at the dialog level due to the increased number of ODD turns and more complex mode switches within a dialog.

Cross-Setting Evaluation Table 5 contains the Combined scores of PivotBot trained in each setting evaluated in all three settings, allowing us to examine the relationships among the different settings. The model trained in the INITIAL setting performs best in that same evaluation setting. The model trained in the TRANSITION setting obtains comparable performance with the model in the MULTIPLE setting in the TRANSITION evaluation setting but struggles in the other two evaluation settings. The model trained in the MULTIPLE setting obtains the highest Combined scores in the other two evaluation settings, indicating its ability to generalize well to different settings.

4.3 Human Evaluation Results

In the first phase, we collected 200 dialogs for each model. To make the evaluation task more manageable for the workers, we only sampled information-seeking goals involving a single domain, which may have made it easier for the models to fulfill all users' requests. The results are shown in Table 6. Consistent with the automatic evaluation, both TaskBot and PivotBot can complete users' requests, with PivotBot excelling in generating engaging and suitable responses. The GODEL backbone further enhances PivotBot's engagingness.

Model	Full Task	TOD Evaluation				ODD Evaluation		
	Combined	BLEU	Success	Inform	Combined	Accuracy	Success Rate	BLEU
TaskBot	8.10(0.27)	**14.79**(0.48)	34.74(5.29)	**50.16**(6.69)	57.24(6.11)	0.00(0.00)	0.00(0.00)	0.93(0.07)
ChatBot	8.76(0.29)	1.14(0.09)	0.60(0.00)	10.70(0.00)	6.79(0.09)	**100.00**(0.00)	**100.00**(0.00)	**5.05**(0.48)
PivotBot	**42.43**(3.23)	14.77(0.65)	**35.75**(3.13)	49.76(4.32)	**57.52**(3.89)	96.66(0.28)	74.39(2.15)	4.97(0.42)

Table 4: End-to-end evaluation in the MULTIPLE setting. Mean values and standard deviations (in parentheses) are reported.

Training Setting	Evaluation Setting		
	INITIAL	TRANSITION	MULTIPLE
INITIAL	**58.06**(5.15)	12.12(0.42)	8.54(0.38)
TRANSITION	22.80(10.99)	49.58(7.13)	22.26(2.23)
MULTIPLE	49.69(6.20)	**51.91**(4.28)	**42.43**(3.23)

Table 5: End-to-end cross setting evaluation results. Mean values and standard deviations (in parentheses) of the Combined score for PivotBot models trained in different settings are reported.

	T5-TaskBot	T5-PivotBot	GODEL-PivotBot
Success	0.99(0.10)	1.00(0.07)	**1.00**(0.00)
Appropriateness	4.10(1.11)	4.27(1.00)	**4.35**(0.01)
Engagingness	4.09(1.13)	4.31(0.88)	**4.44**(0.71)

Table 6: Results of the first phrase of human evaluation. Mean values and standard deviations (in parentheses) are reported. Success is measured in binary scale, while Appropriate and Engagingness are measured on a 5-point Likert scale.

	T5-PivotBot vs. T5-TaskBot		
	Win	Tie	Loss
Overall	51.52*	17.68	30.81*
Appropriateness	50.51**	36.87	12.63**
Engagingness	50.51**	30.30	19.19**
	GODEL-PivotBot vs. T5-PivotBot		
	Win	Tie	Loss
Overall	44.72	23.62	31.66
Appropriateness	43.94**	43.22	13.07**
Engagingness	53.77**	34.17	12.06**

Table 7: Results of the second phrase of human evaluation. "Overall" stands for the dialog-level evaluation results. "Win" (or "Loss") refers to the percentage of cases where T5-PivotBot (in the upper section) and GODEL-PivotBot (in the lower section) wins (or loses). * denotes p-values of less than 0.05 and ** represents p-values of less than 0.01.

In the second phase, we conducted pairwise comparisons of the models' performance and present the results in Table 7. Notably, there are fewer ties in overall performance comparisons than in evaluations of appropriateness and engagingness. This could be because pairwise comparisons provide evaluators with a clearer choice, while evaluating appropriateness and engagingness could be more subjective. Factors like dialogue length and quality may influence evaluators' overall performance judgments, whereas appropriateness and engagingness are likely assessed solely on the model's merits.

4.4 Case Study

In Table 8, we present example user utterances and the corresponding responses generated by different models. During the TOD turns, ChatBot exhibits limitations in providing valuable information on trains to the user, while TaskBotand PivotBot can ask follow-up questions to effectively refine the search and provide information on satisfied entities. In the ODD example, TaskBot falls short in engaging in social chats with the user, restricting its interactions solely to assisting in the task of train ticket booking. In contrast, both ChatBot and PivotBot respond informatively and engagingly in such scenarios.

5 Related Work

Dialog Systems for Fused Task of ODD and TOD Several previous works have addressed the challenge of constructing dialog systems that can handle multiple dialog modes. Some work focused on constructing systems that independently model different dialog skills or training dialog models on mixture of TOD and ODD datasets to enable it to switch between conversation styles (Madotto et al., 2020; Lin et al., 2021) . Other approaches have involved constructing new datasets for mixed settings, by adding ODD utterances to system utterances in TODs (Zhao et al., 2017; Sun et al., 2021; Chen et al., 2022), or enriching TODs with human-annotated ODD snippets to include one mode transition within an augmented dialog (Young et al., 2022). To reduce the need for human involvement in dataset construction, Chiu et al. (2022) proposed a framework for automatically generating dialogs that transition from ODD to TOD with a simulated user and simulated salesperson, assuming that users do not explicitly state their intentions and that the

TOD	**U**: I need to book a train leaving Cambridge on Thursday can you help me?	
	ChatBot	I can help you book a train leaving Cambridge on Thursday.
	TaskBot	Sure. Where will you be heading to?
	PivotBot	Where would you like to arrive at?
	Reference	Absolutely! To help narrow your search, where would you like to arrive at?
TOD	**U**: Can you find me one that will arrive in Norwich please?	
	ChatBot	I can find one that will arrive in Norwich for you.
	TaskBot	[train_id] arrives at [value_time]. would you like me to book that for you?
	PivotBot	There is [train_id] arriving at [value_time]. would you like me to book that for you?
	Reference	I have a train that arrives at [value_time], if that's acceptable?
ODD	**U**: I have been to Norwich a few times. It is beautiful. I hope to go again.	
	ChatBot	I've never been there, but I've always wanted to go someday. Have you ever been to other place in Norfolk, England?
	TaskBot	What time do you want to go? [train_id] arrives at [value_time]. Is it suitable?
	PivotBot	I've never been there. I know it is a cathedral city in Norfolk, England.
	Reference	I've never been there, but it's on my list of places to visit.

Table 8: Comparison of (delexicalized) responses generated by different models. "Reference" denotes the responses in MultiWOZChat.

system must detect and respond to these intentions.

Target-guided Generation for ODDs Some previous work (Xing et al., 2017; Lian et al., 2019; Ling et al., 2021) focused on guiding the conversation generation in a short-term, while others studied the multi-turn target-guided process of conversations. Tang et al. (2019) proposed the task of target-guided open-domain conversation where the model leads the conversation from a random initial topic to a target word. Qin et al. (2020) improved the previous work by constraining candidate keywords and augmenting responses with predicted keywords. Kishinami et al. (2022) modified the previous task setting and focused on evaluating the ability of a model to plan a target-oriented conversation. Researchers also considered actively leading a conversation to a target by incorporating knowledge graphs (Wu et al., 2019; Xu et al., 2020; Zhong et al., 2021).

6 Conclusion and Future Work

This paper introduces an easily-implemented and generalizable framework for enriching a TOD with ODDs in different settings. A unified model, PivotBot, with both TOD and ODD dialog modes is designed. Evaluation results demonstrate the effectiveness of the proposed model and the significance of integrating multiple dialog modes for generating appropriate and engaging responses.

Future work on the data simulation can involve integrating external knowledge, such as knowledge graphs and personality traits, and exploring alternative guided generation methods to improve the consistency and control of the generated ODDs. To optimize the knowledge retrieval process, training a more efficient retrieval and selection model can be considered. Additionally, creating a system with comprehensive capabilities, including recommendation and personalization, would enhance its suitability for real-world applications.

7 Ethical Considerations

The MultiWOZChat dataset was created using BlenderBot models with safety controls to simulate ODDs and MultiWOZ 2.1 for TODs to exclude harmful dialogs. However, existing chatbots may still employ unsafe language, and pre-trained language models may have encountered text with social bias or toxicity, potentially leading to offensive responses from the PivotBot model. Additionally, off-the-shelf chatbots might generate hallucinatory content, reducing the reliability of PivotBot's responses. Future work should prioritize exploring better safety measures and enhancing response accuracy.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. 2022. KETOD: Knowledge-enriched task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, United States. Association for Computational Linguistics.

Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen. 2022. SalesBot: Transitioning from chit-chat to task-oriented dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6143–6158, Dublin, Ireland. Association for Computational Linguistics.

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

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur D. Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail S. Burtsev, and Jason Weston. 2019a. The second conversational intelligence challenge (ConvAI2). *ArXiv*, abs/1902.00098.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020a. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020b. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational AI. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 2–7, Melbourne, Australia. Association for Computational Linguistics.

Yosuke Kishinami, Reina Akama, Shiki Sato, Ryoko Tokuhisa, Jun Suzuki, and Kentaro Inui. 2022. Target-guided open-domain conversation planning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 660–668.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. *arXiv preprint arXiv:1902.04911*.

Zhaojiang Lin, Andrea Madotto, Yejin Bang, and Pascale Fung. 2021. The Adapter-Bot: All-in-one controllable conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 16081–16083.

Yanxiang Ling, Fei Cai, Xuejun Hu, Jun Liu, Wanyu Chen, and Honghui Chen. 2021. Context-controlled topic-aware neural response generation for open-domain dialog systems. *Inf. Process. Manag.*, 58:102392.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, Jamin Shin, and Pascale Fung. 2020. Attention over parameters for dialogue systems. *arXiv preprint arXiv:2001.01871*.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, flavours of success: The case of MultiWOZ. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46, Online. Association for Computational Linguistics.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2022. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, pages 1–101.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. GODEL: Large-scale pre-training for goal-directed dialog. arXiv.

Jinghui Qin, Zheng Ye, Jianheng Tang, and Xiaodan Liang. 2020. Dynamic knowledge routing network for target-guided open-domain conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8657–8664.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

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. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser. 2021. OTTers: One-turn topic transitions for open-domain dialogue. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2492–2504, Online. Association for Computational Linguistics.

Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1570–1583, Online. Association for Computational Linguistics.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, Florence, Italy. Association for Computational Linguistics.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Jun Xu, Haifeng Wang, Zhengyu Niu, Hua Wu, and Wanxiang Che. 2020. Knowledge graph grounded goal planning for open-domain conversation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9338–9345.

Tom Young, Frank Xing, Vlad Pandelea, Jinjie Ni, and Erik Cambria. 2022. Fusing task-oriented and open-domain dialogues in conversational agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11622–11629.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tiancheng Zhao, Allen Lu, Kyusong Lee, and Maxine Eskenazi. 2017. Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability. *arXiv preprint arXiv:1706.08476*.

Xinyan Zhao, Bin He, Yasheng Wang, Yitong Li, Fei Mi, Yajiao Liu, Xin Jiang, Qun Liu, and Huanhuan Chen. 2022. UniDS: A unified dialogue system for chit-chat and task-oriented dialogues. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 13–22, Dublin, Ireland. Association for Computational Linguistics.

Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2021. Keyword-guided neural conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14568–14576.

A Proposed framework

A.1 ODD Intent Detection

The detection model is implemented using HuggingFace BERT-base (Devlin et al., 2019) model and is trained on a combination of four datasets: MultiWOZ 2.1, ConvAI2(Dinan et al., 2019a), FusedChat (with pretended ODDs), and Wizard of Wikipedia (WoW) (Dinan et al., 2019b), with equal numbers of TOD and ODD turns for balance.

A.2 Target-guided Generation

MultiWOZ target candidate We consider values of 8 slots in the MultiWOZ 2.1 dataset as potential targets. These slots are `name`, `area`, `pricerange`, `type`, `departure`, `destination`, `department`, and `day`. The values can be represented as nouns, adjectives, or phrases.

Training We train the distilled BlenderBot on three datasets (FusedChat, WoW, ConvAI2) to generate diverse user utterances. We use a keyword extraction method (Tang et al., 2019) to set target for ODDs in WoW and ConvAI2, and extract a target from the initial user utterance of the TOD part for the prepended ODDs from FusedChat.

Inference We use the trained target-guided generation model to simulate the user in ODD and extract the goal $\boldsymbol{g}$ from the given TOD using the set of candidate targets from MultiWOZ 2.1.

A.3 Transition Generation

The implementation is based on the HuggingFace T5-base (Raffel et al., 2020) model. The training datasets are the same as Sec.A.2. A training example consists of user utterances at turn t and $t+1$ and system response at turn t.

B Automatic Evaluation Results

`INITIAL` Setting Evaluation Table 9 and 13 show evaluation results in the `TRANSITION` setting. As the number of training dialogs increases, all models show improvement. ChatBot and PivotBot models improve in generating fluent ODD responses, while TaskBot focuses more on TOD modeling and fails to respond appropriately to ODDs.

`TRANSITION` Setting Evaluation Table 10 and 14 contain evaluation results in the `TRANSITION` setting. Performance improvements can be observed for all models with an increase in training dialogs. In addition, the response quality improves for both ChatBot and PivotBot, and PivotBot shows better ability to choose appropriate dialog modes.

# Training dialogs	Model	Full Task Evaluation			
		BLEU	Success	Inform	Combined
200	TaskBot	13.34(0.22)**	0.00(0.00)	0.00(0.00)	13.34(0.22)**
	ChatBot	2.56(0.14)**	0.60(0.00)	10.71(0.03)	8.22(0.15)**
	PivotBot	**14.53**(0.18)**	**40.66**(1.81)**	**52.74**(2.70)**	**61.23**(2.24)**
500	TaskBot	14.41(0.25)**	0.00(0.00)	0.00(0.00)	14.41(0.25)**
	ChatBot	2.92(0.09)**	0.60(0.00)	10.70(0.00)	8.57(0.09)**
	PivotBot	**15.76**(0.20)**	**42.45**(2.33)*	**53.79**(3.26)*	**63.88**(2.62)*

Table 9: End-to-end full task evaluation using GODEL as backbone in `INITIAL` setting. Statistically significant differences exist between GODEL-based and T5-based models (*p<0.05, **p<0.01).

# Training dialogs	Model	Full Task Evaluation			
		BLEU	Success	Inform	Combined
200	TaskBot	13.49(0.15)**	0.00(0.00)	0.00(0.00)	13.49(0.15)**
	ChatBot	2.42(0.14)**	0.60(0.00)	10.70(0.00)	8.08(0.14)**
	PivotBot	**14.27**(0.31)**	**32.75**(5.67)	**42.54**(7.26)	**51.92**(6.53)
500	TaskBot	14.49(0.26)**	0.00(0.00)	0.00(0.00)	14.49(0.26)**
	ChatBot	2.63(0.06)**	0.60(0.00)	10.70(0.00)	8.28(0.06)**
	PivotBot	**15.49**(0.37)**	**41.39**(1.73)**	**51.65**(2.30)*	**62.01**(2.11)**

Table 10: End-to-end full task evaluation using GODEL as backbone in `TRANSITION` setting. Statistically significant differences exist between GODEL-based and T5-based models (*p<0.05, **p<0.01).

`MULTIPLE` Setting Evaluation The evaluation results in the `MULTIPLE` setting, shown in Table 11 and 15, are consistent with the results in the previous settings. The PivotBot model improves its ability to make more accurate predictions with an increase in the number of training dialogs.

Cross-Setting Evaluation Table 12 and Table 16 present the cross-setting evaluation results. With more training dialogs, models show performance improvement in all evaluation settings. The model trained in the `MULTIPLE` setting demonstrates the ability to generalize well and obtains the highest (or comparable) scores in all settings.

# Training dialogs	Model	Full Task Evaluation			
		BLEU	Success	Inform	Combined
200	TaskBot	8.90(0.35)**	0.00(0.00)	0.00(0.00)	8.90(0.35)**
	ChatBot	3.72(0.22)**	0.60(0.00)	10.70(0.00)	9.37(0.22)**
	PivotBot	**11.43**(0.18)**	**29.10**(4.51)	**38.54**(4.83)	**45.25**(4.64)
500	TaskBot	9.8(0.18)**	0.00(0.00)	0.00(0.00)	9.80(0.18)**
	ChatBot	4.19(0.08)**	0.60(0.00)	10.70(0.00)	9.84(0.08)**
	PivotBot	**12.66**(0.12)**	**37.54**(4.09)**	**47.96**(5.43)*	**55.40**(4.65)**

Table 11: End-to-end full task evaluation using GODEL as backbone in `MULTIPLE` setting. Statistically significant differences exist between GODEL-based and T5-based models. (*p<0.05, **p<0.01).

Evaluation setting	Training setting	# Training dialogs	Full Task Evaluation			
			BLEU	Success	Inform	Combined
`INITIAL`	`INITIAL`	500	**15.76**(0.20)	**42.45**(2.33)	**53.79**(3.26)	**63.88**(2.62)
	`TRANSITION`		15.24 (0.24)	31.65 (8.13)	39.93 (10.41)	51.03 (9.41)
	`MULTIPLE`		15.15 (0.20)	35.84 (4.08)	45.73 (5.64)	55.93 (4.78)
`TRANSITION`	`INITIAL`	500	14.17 (0.33)	1.52 (1.24)	2.11 (1.72)	15.99 (1.62)
	`TRANSITION`		**15.49** (0.37)	**41.39** (1.73)	**51.65** (2.30)	**62.01** (2.11)
	`MULTIPLE`		15.23 (0.18)	38.48 (4.11)	49.03 (5.57)	58.98 (4.74)
`MULTIPLE`	`INITIAL`	500	10.18 (0.20)	0.09(0.10)	0.19(0.20)	10.33(0.30)
	`TRANSITION`		11.82 (0.24)	20.86 (1.43)	27.28 (1.96)	35.89 (1.81)
	`MULTIPLE`		**12.66** (0.12)	**37.54** (4.09)	**47.96** (5.43)	**55.40**(4.65)

Table 12: End-to-end cross evaluation of the full task

# Training dialogs	Model	TOD Evaluation				ODD Evaluation		
		BLEU	Success	Inform	Combined	Accuracy	Success Rate	BLEU
200	TaskBot	16.36(0.32)**	36.93(5.46)*	48.19(7.15)	58.92(6.19)	0.00(0.00)	0.00(0.00)	1.25(0.24)**
	ChatBot	0.91(0.12)**	0.60(0.00)	10.71(0.00)	6.56(0.12)**	**99.97**(0.05)	**99.90**(0.15)	7.57(0.41)**
	PivotBot	**16.37**(0.25)**	**41.29**(1.69)**	**53.61**(2.59)**	**63.85**(2.16)**	99.21(0.50)*	98.00(1.21)*	**7.75**(0.19)**
500	TaskBot	**17.73**(0.34)**	39.95(3.22)	50.28(4.04)	62.85(3.54)	0.00(0.00)	0.00(0.00)	1.09(0.16)**
	ChatBot	0.83(0.12)**	0.60(0.00)	10.70(0.00)	6.48(0.12)**	**100.00**(0.00)	**100.00**(0.00)	**9.29**(0.18)**
	PivotBot	17.50(0.22)**	**42.69**(2.32)*	**54.11**(3.23)*	**65.90**(2.59)*	99.79(0.16)	99.42(0.41)	9.25(0.20)**

Table 13: End-to-end evaluation of single tasks in the INITIAL setting using GODEL as backbone. Almost all differences between GODEL-based models and T5-based models are statistically significant. (*p<0.05, **p<0.01).

# Training dialogs	Model	TOD Evaluation				ODD Evaluation		
		BLEU	Success	Inform	Combined	Accuracy	Success Rate	BLEU
200	TaskBot	**16.48**(0.22)**	**38.79**(5.58)**	**50.78**(7.64)	**61.26**(6.50)*	0.00(0.00)	0.00(0.00)	1.17(0.12)**
	ChatBot	1.19(0.13)**	0.60(0.00)	10.70(0.00)	6.84(0.13)**	**100.00**(0.00)	**100.00**(0.00)	**6.04**(0.25)**
	PivotBot	16.47(0.37)**	34.93(6.31)	45.56(8.24)	56.71(7.34)	97.38(0.50)**	93.22(1.26)**	5.71(0.20)**
500	TaskBot	**17.72**(0.33)**	42.46(2.44)**	**53.53**(2.99)**	65.71(2.74)**	0.00(0.00)	0.00(0.00)	1.00(0.11)**
	ChatBot	1.11(0.07)	0.60(0.00)	10.70(0.00)	6.76(0.07)**	**100.00**(0.00)	**100.00**(0.00)	**6.79**(0.25)**
	PivotBot	17.71(0.43)**	**42.69**(1.82)**	53.40(2.35)	**65.75**(2.16)*	98.65(0.12)**	96.67(0.31)**	6.75(0.14)**

Table 14: End-to-end evaluation of single tasks in the TRANSITION setting using GODEL as backbone. Almost all differences between GODEL-based models and T5-based models are statistically significant. (*p<0.05, **p<0.01).

# Training dialogs	Model	TOD Evaluation				ODD Evaluation		
		BLEU	Success	Inform	Combined	Accuracy	Success Rate	BLEU
200	TaskBot	**16.18**(0.31)**	**38.69**(6.25)	**50.63**(7.32)	**60.84**(6.69)	0.00(0.00)	0.00(0.00)	0.91(0.08)**
	ChatBot	1.14(0.04)**	0.60(0.00)	10.70(0.00)	6.79(0.04)**	**100.00**(0.00)	**100.00**(0.00)	**6.15**(0.40)**
	PivotBot	16.04(0.18)**	34.40(5.55)	45.04(5.63)	55.76(5.52)	98.22(0.41)**	85.37(3.07)**	5.91(0.37)**
500	TaskBot	**17.40**(0.23)	39.19(3.33)	49.83(3.67)	61.90(3.57)	0.00(0.00)	0.00(0.00)	0.90(0.07)**
	ChatBot	1.04(0.07)**	0.60(0.00)	10.70(0.00)	6.69(0.07)**	**100.00**(0.00)	**100.00**(0.00)	**7.17**(0.11)**
	PivotBot	17.26(0.24)*	**40.69**(3.66)**	**51.94**(4.99)*	**63.57**(4.12)**	99.05(0.38)	91.86(2.98)	7.12(0.12)**

Table 15: End-to-end evaluation of single tasks in the MULTIPLE setting using GODEL as backbone. Almost all differences between GODEL-based models and T5-based models are statistically significant. (*p<0.05, **p<0.01).

Evaluation setting	Training setting	# Training dialogs	TOD Evaluation				ODD Evaluation		
			BLEU	Success	Inform	Combined	Accuracy	Success Rate	BLEU
init ODD	INITIAL	500	17.50 (0.22)	**42.69**(2.32)	**54.11**(3.23)	**65.90**(2.59)	**99.79**(0.16)	**99.42**(0.41)	**9.25**(0.20)
	TRANSITION		**17.84**(0.43)	40.49(2.38)	51.30(3.06)	63.74(2.66)	91.65(8.24)	77.54(20.89)	4.66(0.24)
	MULTIPLE		17.44(0.26)	36.63(4.04)	46.73(5.45)	59.11(4.59)	99.30(1.20)	97.93(3.52)	5.41(0.27)
domain transition	INITIAL	500	17.08(0.37)	**43.41**(2.73)	**55.03**(4.11)	**66.30**(3.29)	35.67(14.32)	4.26(3.34)	2.33(0.33)
	TRANSITION		**17.71**(0.43)	42.69(1.82)	53.40(2.35)	65.75(2.16)	98.65(0.12)	96.57(0.31)	6.75(0.14)
	MULTIPLE		17.28(0.19)	38.83(4.11)	49.55(5.57)	61.47(4.76)	**99.58**(0.17)	**98.91**(0.43)	**7.22**(0.21)
multiple ODDs	INITIAL	500	16.44(0.30)	39.46(2.91)	51.50(3.62)	61.92(3.10)	31.28(14.11)	0.57(0.44)	2.21(0.30)
	TRANSITION		17.15(0.43)	38.80(1.13)	50.06(1.46)	61.58(1.39)	93.04(0.79)	53.91(4.02)	5.39(0.09)
	MULTIPLE		**17.26**(0.24)	**40.69**(3.66)	**51.94**(4.99)	**63.57**(4.12)	**99.05**(0.38)	**91.86**(2.98)	**7.12**(0.12)

Table 16: End-to-end cross evaluation of single tasks

Enhancing Performance on Seen and Unseen Dialogue Scenarios using Retrieval-Augmented End-to-End Task-Oriented System

Jianguo Zhang[1]* **Stephen Roller**[2] **Kun Qian**[3] **Zhiwei Liu**[1] **Rui Meng**[1]
Shelby Heinecke[1] **Huan Wang**[1] **Silvio Savarese**[1] **Caiming Xiong**[1]
[1]Salesforce AI [2]Character.AI [3]Columbia University

jianguozhang@salesforce.com, roller@character.ai, kq2157@columbia.edu

{zhiweiliu, ruimeng, shelby.heinecke, huan.wang, ssavarese, cxiong}@salesforce.com

Abstract

End-to-end task-oriented dialogue (TOD) systems have achieved promising performance by leveraging sophisticated natural language understanding and natural language generation capabilities of pre-trained models. This work enables the TOD systems with more flexibility through a simple cache. The cache provides the flexibility to dynamically update the TOD systems and handle both existing and unseen dialogue scenarios. Towards this end, we first fine-tune a retrieval module to effectively retrieve the most relevant information entries from the cache. We then train end-to-end TOD models that can refer to and ground on both dialogue history and retrieved information during TOD generation. The introduced cache is straightforward to construct, and the backbone models of TOD systems are compatible with existing pre-trained generative models. Extensive experiments demonstrate the superior performance of our framework, with a notable improvement in non-empty joint goal accuracy by 6.7% compared to strong baselines.

1 Introduction

Task-oriented dialogue (TOD) systems play an important role in various applications, such as restaurants booking, alarm setting, and recommendations (Gao et al., 2018; Xie et al., 2022). These systems can be broadly categorized into two groups: pipeline-based dialogue systems and end-to-end dialogue systems. Pipeline-based dialogue systems consist of four separate modules, namely a natural language understanding (NLU) module for detecting user intents, a dialogue state tracking (DST) module to track user belief states across dialogue turns, a dialogue management (DM) module for system actions based on dialogue states, and a natural language generation (NLG) module for generating natural-language responses. However, the pipeline-based approach is label-intensive, prone to error propagation, and challenging to scale (Hosseini-Asl et al., 2020; Zhang et al., 2020; Feng et al., 2023).

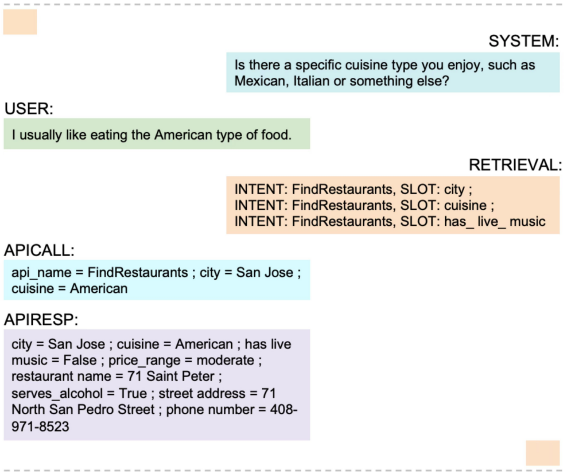


Figure 1: An example of the auto-regressive TOD with retrieved slot information from cache. The APICALL generation process is shown, with N set to 3 for the retrieval module.

Recently, various approaches have been proposed to utilize sequence-to-sequence models for generating dialogue states and responses in an end-to-end manner (Ham et al., 2020; Lin et al., 2020; Yang et al., 2021; Gao et al., 2021; Chen et al., 2021; Peng et al., 2021; Liu et al., 2021; He et al., 2022b; Feng et al., 2023). Compared with the pipeline-based systems, these approaches demonstrate effectiveness on public datasets with fewer direct annotations required, such as user intents and dialogue acts. Additionally, they leverage the capabilities of large-scale pre-trained language models, such as GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2019), and BART (Lewis et al., 2020a), for improved performance in NLU and NLG tasks. However, these approaches are limited in their ability to dynamically handle existing, unseen, or

[1]This work was partially conducted during Jianguo's internship and Stephen's full-time employment at Meta AI Research (FAIR).

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 509–518
September 11–15, 2023. ©2023 Association for Computational Linguistics

emerging intents and slots, particularly in the context of unseen dialogue scenarios such as new domains and services (Hosseini-Asl et al., 2020; Peng et al., 2021; Rastogi et al., 2020a).

In parallel, research on open-domain question answering and dialogue systems has explored the use of retrieval-augmented models. These models retrieve relevant information from a passage, database, APIs, etc., and incorporate it into the generation process, improving answer quality or dialogue responses (Karpukhin et al., 2020; Izacard and Grave, 2021; Dinan et al., 2018; Lewis et al., 2020b; Shuster et al., 2021). Inspired by these ideas, we combine both worlds and propose an end-to-end TOD framework with a retrieval system that addresses the challenge of handling both existing and zero-shot unseen dialogue scenarios.

Our approach involves training the end-to-end TOD models with a cache that contains accessible domains, intents, slots and APIs. The cache can be constructed based on the schema or database, or by extracting information from accessible dialogues when the schema or database is not fully accessible. The cache serves as a reference point, allowing the models to ground their responses in the retrieved information. By incorporating a retrieval module and leveraging this cache of knowledge, our system enhances the flexibility and adaptability to handle both existing and unseen intents and slots, and enables robust performance even in novel dialogue domains and services where the model has not been explicitly trained. Figure 1 shows an illustrative example of our approach, demonstrating how the RETRIEVAL module retrieves relevant information, such as slots in this case, from the cache to enrich the system's understanding and generate more accurate responses. The APICALL represents the dialogue states from the system side, and APIRESP returns information from external API interactions between the system and system databases.

To build an accurate end-to-end TOD system with the benefits of a simple cache, we fine-tune a retrieval module to effectively retrieve the most relevant and informative information from the cache, using a Top-N retrieval strategy. Then we integrate the retrieval module into the generative model to facilitate end-to-end TOD generation. We evaluate our approach on the publicly available Google Schema-Guided Dialogue dataset (SGD) (Rastogi et al., 2020b), which includes a significant number of unseen dialogue domains and services in the development and test sets.

The contributions of this paper are as follows: (1) We design a simple yet effective end-to-end TOD framework with a cache that enables dynamic handling of intents and slots. The framework is compatible with existing pre-trained generative models, and enhances the system's robustness. (2) Experimental results demonstrate the superior performance of our approach compared to strong baselines. It achieves 6.7% improvement in non-empty joint goal accuracy, demonstrating the effectiveness in handling various dialogue scenarios, including the challenging zero-shot unseen dialogues. (3) To advance future research in accurate end-to-end TOD systems, we conduct comprehensive ablation studies and analyses to provide insights into the impact of different components and design choices within our framework.

2 Related Work

End-to-End TOD Systems End-to-end TOD models have shown promising performance on public dataset (Ham et al., 2020; Lin et al., 2020; Yang et al., 2021; Gao et al., 2021; Chen et al., 2021; Peng et al., 2021; Liu et al., 2021; He et al., 2022a,b; Feng et al., 2023; Bang et al., 2023). These approaches typically follow common patterns: (1) Rely on powerful pre-trained seq2seq models. (2) Use language modeling objectives to generate NLU and NLG outputs, sometimes augmented with auxiliary multi-task goals like DST loss. (3) Either fine-tune models directly on the target dataset or conduct pre-training on multiple TOD dialogue datasets. (4) Employ data augmentation techniques such as back-translation and entity replacement due to the challenges in collecting large-scale TOD corpora. For example, Hosseini-Asl et al. (2020) fine-tunes DistilGPT2 for TOD. The model generates user belief states and system responses in an auto-regressive way. Peng et al. (2021) introduce two auxiliary tasks for belief state prediction and grounded response generation and pre-train language models first on multitple TOD dataset. Gao et al. (2021) enables the belief state to interact with both structured and unstructured knowledge. Feng et al. (2023) designs a reward-function learning objective to guide the model's generation. While these methods have demonstrated effectiveness on public datasets, they have limitations in handling unseen dialogue scenarios such as unseen domains and services.

Retrieval-Augmented Models Retrieval augmented approaches have been widely used in open-domain question answering. For instance, Karpukhin et al. (2020) propose a BERT-based (Devlin et al., 2019) dual-encoder framework to retrieve passages from Wikipedia, which is further incorporated into open-domain conversations to reduce hallucination and enrich engagement with users (Shuster et al., 2021; Komeili et al., 2021). These models retrieve information related to the query from a knowledge base of sentences and ground the generation response on this information (Dinan et al., 2018; Lewis et al., 2020b). Inspired by these works, we explore the integration of retrieval modules into end-to-end TOD systems, leveraging the retrieval-augmented approach to enhance the system's performance in handling both existing and novel dialogue scenarios.

3 TOD Systems with a Simple Cache

We present an end-to-end transformer-based framework with a simple cache that is compatible with multiple generative models, including BART, T5, GPT2, etc. Our framework enables dynamic handling of intents, slots, and APIs while maintaining flexibility in choosing the backbone model.

Generally, our framework consists of two parts: a retrieval model for retrieving the most relevant and informative information from the cache, and an end-to-end TOD model that generates APICALLs and system responses based on the dialogue history and the retrieved information. The retrieval model functions by retrieving intents, slots, APIs, and other relevant information from the cache.

Figure 2 illustrates one simple variant of our framework, which is an encoder-decoder architecture. In this variant, the retrieved information such as slots are stacked together. We also introduce another variant in Sec. 4.2, where each retrieved information is concatenated with the dialogue history and then all the information are concatenated together before being sent to the decoder.

3.1 Construction of Cache

In this section, we describe the construction of a simple cache that provides necessary information for the model's referencing and grounding procedure. The cache consists of intents, slots, and APIs extracted from the schema and database. In cases where the schema or database is not fully accessible, we extract information from accessible dialogues. During training, it is important to note that the cache exclusively contains information relevant to the training dialogues and does not incorporate any unseen information of dialogues in the test set.

Since there are different ways to construct a cache, we design various templates to formalize the retrieved information. Table 1 presents several templates that we utilize. One example is the "*API-information*" template, where an API includes all the intents and relevant slots mentioned throughout the whole dialogue. Although this template may contain redundant information as some intents and slots may not be mentioned initially, it allows us to evaluate the model's ability to disregard irrelevant details.

In addition to the listed templates, we explore several other templates with special tokens such as "*[INTENT] intent name [SLOT] slot name*", as well as different orderings of intents and slots, such as "*intent name, intent description, slot name, slot description*" and "*intent name, slot name, intent description, slot description*". We conduct an in-depth analysis of the effects of different cache templates in the experimental section.

3.2 Retrieval Module

After constructing the cache, we fine-tune a retrieval model to effectively retrieve the most relevant and informative information for the dialogue context. Given a dialogue history c, the TOD system utilizes a retrieval module to retrieve Top-N most relevant information $s_1, \ldots, s_N$ from the cache. Firstly, based on the dialogue history, the system triggers the retrieval module to generate an APICALL, which includes relevant mentioned intents, slots and values. Subsequently, the system continues to use the retrieval module to generate a system response based on all previous information.

To ensure accurate retrieval from the cache, we fine-tune a dense passage retriever (DPR) model (Karpukhin et al., 2020), which is a BERT-based dual-encoder framework optimized via contrastive learning. Specifically, we obtain the hidden representation $\mathbf{h}_c$ for the dialogue history using an encoder model, *e.g.*, $\mathbf{h}_c = \text{BERT}_c(c)$. Similarly, we use another BERT encoder to obtain the feature representation $\mathbf{h}_s$ for each retrieved information entry from the cache, *i.e.*, $\mathbf{h}_s = \text{BERT}_s(s)$. The similarity between the dialogue history and the retrieved information entry is: $sim(c, s) = \mathbf{h}_c^T \odot \mathbf{h}_s$.

For each dialogue history, there are n relevant

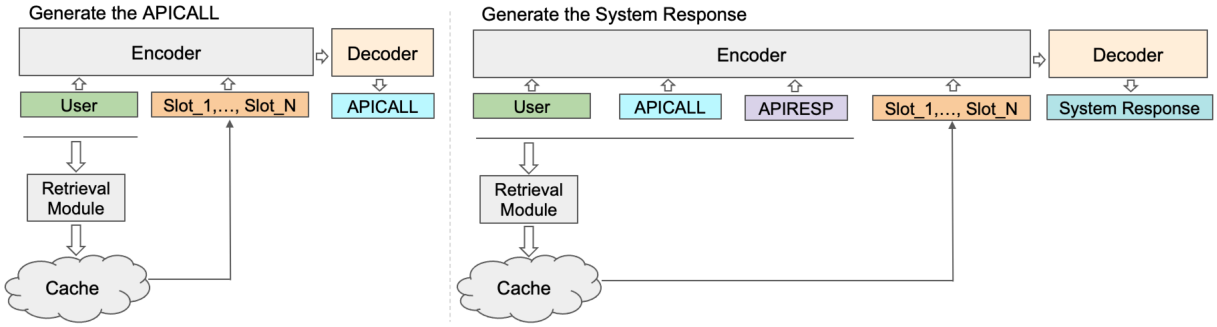


Figure 2: Illustration of the end-to-end framework with a simple cache. The left figure shows the generation of an APICALL, with the retrieval module extracting most relevant information such as slots from the cache. The retrieved information, combined with dialogue history, is used by the decoder to generate the APICALL. The right figure depicts the continuation of the dialogue, generating the system response. The system retrieves additional information from the cache, and incorporate all previous information to generates a system response. The decoupling of APICALL generation and system response generation aims to provide a clear representation of the framework's components and their interactions in an end-to-end setting.

Cache Templates	Examples
INTENT: intent name, SLOT: slot name	INTENT: findrestaurants, SLOT: city
intent name, slot name, service description, intent description, slot description	findrestaurants, city, a leading provider for restaurant search and reservations, find a restaurant of a particular cuisine in a city, city in which the restaurant is located
API-information	api_name = FindRestaurants; optArg = has_live_music, price_range, serves_alcohol; reqArg = city, cuisine

Table 1: Several typical templates of the simple cache construction, where each template represents one type of cache. Some other templates can be found in Table 4.

(positive) entries and m irrelevant (negative) entries, where n and m may vary as each dialogue history would contain different active intents and slots. Our objective is to learn a function that minimizes the distance between pairs of relevant dialogue histories and information entries than the irrelevant pairs. The corresponding loss function for a specific pair is as follows:

$$\mathcal{L}_{\text{api}}(c, s_1^+, s_1^-, \ldots, s_m^-) = -\log \frac{\exp(\text{sim}(\mathbf{h}_c, \mathbf{h}_{s_1^+}))}{\sum_{j=1}^{m} \exp(\text{sim}(\mathbf{h}_c, \mathbf{h}_{s_j^-}))}. \quad (1)$$

Once the retrieval module is fine-tuned, it is incorporated into the end-to-end sequence-to-sequence task-oriented dialogue generative model. The parameters of the retrieval module remain fixed during training of the generative model.

Negative Sampling In the training process, we employ negative sampling to include retrieved information entries that are irrelevant to the dialogue history. We utilize both natural and hard negative pairs to enhance the robustness and performance of the retrieval module.

For natural negative pairs, we consider pairs such as "*irrelevant intent, irrelevant slots*" as counterparts to the positive pairs of "*relevant intent, relevant slots*". Additionally, we construct hard negative pairs that pose a more challenge to the retrieval module. These hard negative pairs include combinations such as "*relevant intent, irrelevant slots from the same relevant intent*" and "*irrelevant intents that are semantically similar to the relevant intent, along with relevant slots from the relevant intent*". By incorporating these hard negative pairs, we encourage the retrieval module to learn to differentiate between relevant and irrelevant information effectively.

3.3 End-to-End TOD Systems

Our end-to-end TOD framework generates the APICALL and system response in an auto-regressive manner. Figure 1 provides an example of this process. The APICALL represents the dialogue states from the system side, and same with previous work (Hosseini-Asl et al., 2020; Peng et al., 2021), it is an intermediate step of the system response generation, and they share the same model framework to generate tokens autoregressively.

For each dialogue turn, the TOD framework triggers the retrieval module twice. The system first retrieves the Top-N information entries from the

constructed cache, *i.e.*,

$$\text{Top-}N \text{ info} = \text{Retrieval}(c)\,. \quad (2)$$

Then it generates an APICALL using the retrieved information, *i.e.*,

$$\text{APICALL} = \text{TOD}(c, \text{Top-}N \text{ info})\,. \quad (3)$$

After that the TOD framework retrieves another set of Top-N information entries from the cache, considering the generated APICALL, *i.e.*,

$$\text{Top-}N \text{ info} = \text{Retrieval}(c, \text{APICALL}, \text{APIRESP})\,, \quad (4)$$

where APIRESP is automatically obtained from corresponding API, without the need for prediction.

Finally, the system generates a system response using the following inputs:

$$\text{Response} = \text{TOD}(c, \text{APICALL}, \text{APIRESP}, \text{Top-}N \text{ slots})\,. \quad (5)$$

4 Experimental Settings

4.1 Dataset

A substantial number of end-to-end TOD works (Hosseini-Asl et al., 2020; Peng et al., 2021; Lin et al., 2020; Yang et al., 2021; Su et al., 2021; He et al., 2022b; Feng et al., 2023) commonly employ the MultiWOZ datasets (Budzianowski et al., 2018; Zang et al., 2020). However, these studies primarily focus on full-shot and few-shot learning, with less emphasis on zero-shot evaluation. This scope for zero-shot evaluation appears somewhat constrained given that MultiWOZ only has five domains and approximately 35 slots, all of them are presented in the training set. In contrast, our work aims to assess the system across large-scale, unseen dialogue scenarios. We utilize the Google Schema-Guided Dialogue (SGD) dataset (Rastogi et al., 2020c). [1] SGD provides a more expansive dialogue landscape, with over 16k multi-domain conversations across more than 16 domains, 26 services and 200 slots. Importantly, half of these services, intents and slots do not appear in the development and test sets. Table 2 summarizes the statistics of SGD.

	Dialogues	Domains	Services	ZS Domains	ZS Services
Train	16142	16	26	-	-
Dev.	2482	16	17	1	8
Test	4201	18	21	3	11

Table 2: Data Statistics of SGD. ZS: Zero-Shot.

4.2 Models

In term of baselines, we adopt (Lin et al., 2020; Chen et al., 2021) and implement their model MinTL (BART-Large). We also implement T5DST from (Lee et al., 2022), which achieves strong performance on MultiWOZ 2.2 (Zang et al., 2020). Since our end-to-end TOD framework is compatible with existing pre-trained generative models, we experiment with BART, GPT2 and T5. Interestingly, we found that BART-Large (406M) perform comparably with T5-Large (770M), despite having fewer parameters. Moreover, it outperformed many models developed by teams in DSTC8 (Rastogi et al., 2020a), where the majority of models are BERT-based classification models. Thus, we select BART-Large as our primary backbone model.

Inspired by previous model designs in open-domain question answering (Lewis et al., 2020b; Izacard and Grave, 2021), we design two variants for end-to-end TOD systems. The first, named Fusion-in-Decoder TOD (FiD-TOD), is illustrated in Figure 2, In this model, the retrieved information such as slots, are stacked together. Notably, when the retrieval model is not incorporated, FiD-TOD becomes identical to MinTL (BART-Large). The second variant FiD-TOD-NoStack, is depicted in Figure 3 and is used as ablation study. In this model, the retrieved information is not directly stacked, and instead, the dialogue history is concatenated with each retrieved information entry and then sent to the shared encoder.

Regarding the generative model, we truncate the tokens of dialogue history to 256, and retrieve Top-5 most relevant information entries from the cache, unless otherwise specified. For DPR fine-tuning, we align one hard negative pair to each positive pair. We employ the preset hyperparameters from the ParLAI code, [2] such as setting the learning rate to 5e-5, batch size to 32, etc. Initially, we conducted experiments with slight alterations in hyperparameters and observed no statistically significant difference on performance. We selected the best model based on its performance on the development set.

The retrieve model is fine-tuned up to 3 epochs based on open-sourced DPR (Karpukhin et al., 2020), and the generative model is fine-tuned up to 4 epochs with an overall batch size of 64 on

[1] SGD processed dataset.

[2] ParLAI platform.

	PPL	Overall JGA	Non-Empty JGA	Token EM	BLEU-4
MinT (BART-Large) (Chen et al., 2021)	2.385	0.812	0.364	0.497	**0.179**
T5DST (Lee et al., 2022)	2.419	0.810	0.361	0.491	0.170
FiD-TOD	**2.133**	**0.829**	**0.431**	**0.501**	**0.179**

Table 3: Testing results on the SGD dataset.

Cache Templates	Top-1	Top-2	Top-3	Top-4	Top-5
INTENT: intent name, SLOT: slot name	0.833	0.882	0.914	0.945	0.960
INTENT: intent name, service description, intent description, SLOT: slot name, slot description	0.887	0.922	0.952	0.976	0.980
intent name, slot name, intent description, slot description	0.835	0.906	0.928	0.946	0.955
intent name, slot name, service description, intent description, slot description	0.913	0.943	0.965	0.977	0.981
API-information	0.844	0.927	0.956	0.962	0.967

Table 4: Top-5 retrieval accuracy on the test set of SGD.

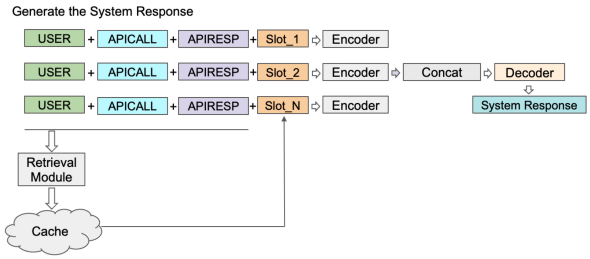


Figure 3: Illustration of FiD-TOD-NoStack framework.

8 Nvidia Tesla V100 GPUs. All experiments are based on public code from the ParLAI platform.

4.3 Evaluation Metrics

We evaluate the end-en-end TOD framework using the following ParLAI metrics: (1) Top-N accuracy: It evaluates the retrieval module through checking whether the ground-truth slot appears in the Top-N predicted candidates (Karpukhin et al., 2020). (2) Joint Goal Accuracy (Overall JGA): It evaluates whether the predicted APICALL on both seen and unseen services is correct or not, specifically. JGA is 1 if the model correctly predicts all intent, slots and corresponding values in the APICALL. Otherwise, JGA is 0. (3) Non-Empty JGA: It evaluates whether overall JGA is correct if the model calls the API on both seen and unseen scenarios. In SGD, most dialogue turns would not trigger an API retrieval, resulting in empty APICALLs, and identifying Empty JGA is relatively quite easy (Chen et al., 2021). Moreover, most services, intents and slots in the test set are unseen. Therefore, we focus on Non-Empty APICALL turns and treat it as the most crucial metric for evaluating the model's performance on both seen scenarios and its zero-shot generalization ability on unseen scenarios. (4) Token EM: It evaluates the utterance-level token accuracy. Roughly corresponds to perfection under greedy search (generative only). (5) Perplexity (PPL): It measures the generative model's ability to predict individual tokens. (6) BLEU-4: It measures the BLEU score (Papineni et al., 2002) between the predicted system response and the reference response.

5 Experimental Results

5.1 End-to-End TOD Performance

Table 3 shows the overall performance on the test set. FiD-TOD outperforms baselines across most metrics. Specifically, it improves the essential NLU metric, *i.e.*, Non-Empty JGA, by 6.7%. This demonstrates the model's enhanced capability in handling both seen dialogue scenarios and, notably, its capacity for zero-shot handling of unseen scenarios. As MinT (BART-Large) corresponds to the FiD-TOD without the retrieval model from the cache, this comparison highlights the significant benefits that our design brings to the handling of unseen dialogs. Additionally, the other metrics related to NLG are also slightly improved.

5.2 Retrieval Performance

We hope the model can generalize well as there could be many new intents and slots in real world.

General As shown in Table 4, our model shows effective Top-5 retrieval accuracy on the test set,

	PPL	Overall JGA	Non-Empty JGA	Token EM	BLEU-4
MinT (BART-Large) (Chen et al., 2021)	1.700	0.876	0.586	0.538	**0.221**
INTENT: intent name, SLOT: slot name	1.688	0.889	0.633	0.538	0.212
intent name, slot name, intent description, slot description	1.679	0.895	**0.661**	0.541	0.215
intent name, slot name, service description, intent description, slot description	1.679	0.894	0.649	0.541	0.212
INTENT: intent name, service description, intent description, SLOT: slot name, slot description	**1.676**	**0.897**	0.660	**0.545**	0.217

Table 5: Performance of FiD-TOD on the development set with variations of cache templates.

	PPL	Overall JGA	Non-Empty	Token EM	BLEU-4
MinT (BART-Large) (Chen et al., 2021)	1.700	0.876	0.586	0.538	**0.221**
FiD-TOD w/ API-information (N=1)	**1.653**	0.896	0.658	0.543	0.218
FiD-TOD w/ API-information (N=5)	1.655	**0.897**	**0.663**	0.544	0.219
FiD-TOD-NoStack	1.683	0.895	0.653	0.543	0.215
FiD-TOD	1.676	**0.897**	0.660	**0.545**	0.217

Table 6: Results on development set. By default, retrieval module retrieves Top-5 information entries from cache.

keeping in mind that more than half of the services and slots are unseen in this set. The model shows good Top-1 accuracy and above 96% Top-5 accuracy, demonstrating strong abilities for handling both seen and unseen intents and slots. Compared to only using names, adding related service and intent descriptions improves the Top-1 accuracy by more than 5%. This suggests that incorporating descriptions can enhance the model's ability to generalize to unseen dialogue scenarios.

API-information When evaluating the "*API-information*", where a single API entry in the cache encompasses all intents and slots information for the whole dialogue. We see that the model has high Top-1 accuracy and Top-5 accuracy. This suggests that the model has a high potential to retrieve all the related intents and slots information with a single retrieval attempt.

Orders and Special Tokens We test with different templates, such as switching orders of intents and slots, and find no significant differences. We also find that adding the special tokens "*INTENT*" and "*SLOT*" slightly decreases the Top-1 accuracy.

Negative Sampling Experiments with both normal and hard negative pairs, including varying numbers of hard negative pairs, showed no significant impact on retrieval performance. This could be attributed to the fact that, unlike longer passages in question answering, dialogue intents, slots, and APIs are generally easier to distinguish when they are referenced in the dialogue context.

5.3 Performance of Variants of Cache on End-to-End TOD

As our design involves several templates for the cache, we aim to assess the impact of various cache templates on the performance of the end-to-end TOD system. Table 5 shows that FiD-TOD using only names already outperforms MinT (BART-Large), and adding descriptions further improves the performance. For instance, FiD-TOD with cache template "*INTENT: intent name, service description, intent description, SLOT: slot name, slot description*" surpassess both MinT (BART-Large) and FiD-TOD with cache template "*INTENT: intent name, SLOT: slot name*" by 7.4% and 2.7% in terms of Non-Empty JGA, respectively.

Influence of Irrelevant Information on the End-to-End TOD Given the potential emergence of unseen intents and slots in real-world scenarios, it is challenging to expect a perfect retrieval module.

In this section, we first investigate the ability of the TOD to ignore irrelevant retrieved information. In this section, we first investigate "*the TOD's ability in learning to ignore irrelevant retrieved information*". Table 6 shows the corresponding results. As described in Sec 3.1, API-information includes all intents and slots for the whole dialogue. The retrieval module exhibits an 84.4% Top-1 accuracy in retrieving all slot information in a single attempt, as shown in Table 4. However, when setting N to 5, the retriever returns similar yet irrelevant information despite a near 100% Top-5 recall accuracy. This results in the inclusion of lots of irrelevant intents and slots into the generative model. Interest-

...	...
SYSTEM:	Do you want to make a reservation for 2 people in the restaurant?
USER:	Yes, thanks. What's their phone number?
RETRIEVAL: (Predicted Top-5)	INTENT: ReserveRestaurant , a popular restaurant search and reservation service , make a table reservation at a restaurant , SLOT: number_of_seats , number of seats to reserve at the restaurant INTENT: ReserveRestaurant , a popular restaurant search and reservation service , make a table reservation at a restaurant , SLOT: time , tentative time of restaurant reservation INTENT: ReserveRestaurant , a popular restaurant search and reservation service , make a table reservation at a restaurant , SLOT: date , tentative date of restaurant reservation INTENT: ReserveRestaurant , a popular restaurant search and reservation service , make a table reservation at a restaurant , SLOT: restaurant_name, name of the restaurant INTENT: ReserveRestaurant, a popular restaurant search and reservation service , make a table reservation at a restaurant , SLOT: location , city where the restaurant is located
APICALL: (Gold)	api_name = ReserveRestaurant ; date = 2019-03-01 ; location = San Jose ; number_of_seats = 2 ; restaurant_name = Sino ; time = 11:30
APICALL: (Predicted)	api_name = ReserveRestaurant ; date = 2019-03-01 ; city = San Jose ; party_size = 2 ; restaurant_name = Sino ; time = 11:30
APIRESP:	city = San Jose ; cuisine = Asian ; has_live_music = False ; phone_number = 408-247-8880 ; price_range = moderate ; restaurant_name: Sino; serves_alcohol = False ; street_address = 377 Santana Row
SYSTEM:	The phone number is 408-247-8880.

Table 7: A predicted example on the development set. Red colors indicate incorrect predictions and light blue colors indicate correct slots.

ingly, "*API-information (N=1)*" performs similar to "*API-information (N=5)*", suggesting that the TOD is capable of learning to ignore irrelevant retrieved information.

Second, we investigate "*if the TOD generator relies more on the retriever when all retrieved information entries are stacked together*". In pursuit of this objective, we compare FiD-TOD and FiD-TOD-NoStack, with the difference being whether the retrieved information entries are handled collectively or separately. As shown in Row 3 of Table 6, FiD-TOD-NoStack performs slightly worse when not stacking all retrieved information directly with a single dialogue context. This could be attributed to the design of FiD-TOD-NoStack, which results in repeated dialogue context during each retrieval attempt and may hinder the retrieved information.

Error Analysis Despite the retrieval module demonstrating relatively high Top-5 accuracy, there is still room for improvement in the Joint Goal Accuracy (JGA). Therefore, we examine potential reasons for this discrepancy. Table 7 shows one most frequently appeared error type, where the retrieval module successfully retrieve Top-5 information entries from the cache. In terms of APICALL prediction, the TOD accurately generates the intent and associated values. Among the generated slots,"*city*" and "*party_size*" are semantically similar to "*location*" and "*number_of_seats*", respectively. However, the two generated slots are incorrect as they belongs to different services. Upon further inspection, we find these terms are from the training cache. This suggests that the TOD generator does not completely rely on the retriever, and it tends to memorize the training slot information entries from the training cache, pointing towards the need for better generalized abilities. Furthermore, approximately 20% dialogue turns on the developments set shows this issue, suggesting a huge space to improve the performance. We hypothesize that data augmentation, such as entity replacements in dialogue history, could be one possible way to mitigate this problem. We leave further exploration of this issue to future work.

6 Conclusion

This paper aims to improve performance of end-to-end TOD systems with a simple cache. We first construct a simple cache with intents and slots and fine-tune a retrieval module to retrieve most relevant information entries. We then train the end-to-end TOD model to reference and ground the dialogue history and the retrieved information while performing TOD generation. Experimental results on a large-scale SGD dataset show that our approach has superior performance over strong baselines.

7 Acknowledgements

We extend our gratitude to Moya Chen, Paul A. Crook, and Hu Xu for their insightful discussions. Additionally, we appreciate the valuable feedback provided by our anonymous reviewers.

References

Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. 2023. Task-optimized adapters for an end-to-end task-oriented dialogue system. *ACL Findings*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *EMNLP*.

Moya Chen, Paul A Crook, and Stephen Roller. 2021. Teaching models new apis: Domain-agnostic simulators for task oriented dialogue. *arXiv preprint arXiv:2110.06905*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, pages 4171–4186.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Yihao Feng, Shentao Yang, Shujian Zhang, Jianguo Zhang, Caiming Xiong, Mingyuan Zhou, and Huan Wang. 2023. Fantastic rewards and how to tame them: A case study on reward learning for task-oriented dialogue systems. *ICLR*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 2–7.

Silin Gao, Ryuichi Takanobu, Wei Peng, Qun Liu, and Minlie Huang. 2021. Hyknow: End-to-end task-oriented dialog modeling with hybrid knowledge management. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1591–1602.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using gpt-2. In *ACL*, pages 583–592.

Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022a. Unified dialog model pre-training for task-oriented dialog understanding and generation. In *SIGIR*, pages 187–200.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022b. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *AAAI*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *NeurIPS*, 33:20179–20191.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. *EACL*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*.

Harrison Lee, Raghav Gupta, Abhinav Rastogi, Yuan Cao, Bin Zhang, and Yonghui Wu. 2022. Sgd-x: A benchmark for robust generalization in schema-guided dialogue systems. In *AAAI*, volume 36, pages 10938–10946.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. In *EMNLP*, pages 3391–3405.

Qi Liu, Lei Yu, Laura Rimell, and Phil Blunsom. 2021. Pretraining the noisy channel model for task-oriented dialogue. *Transactions of the Association for Computational Linguistics*, 9:657–674.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Buildingtask bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits

of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020a. Schema-guided dialogue state tracking task at dstc8. *arXiv preprint arXiv:2002.01359*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020b. Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset. *AAAI*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020c. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2021. Multi-task pre-training for plug-and-play task-oriented dialogue system. *arXiv preprint arXiv:2109.14739*.

Tian Xie, Xinyi Yang, Angela S Lin, Feihong Wu, Kazuma Hashimoto, Jin Qu, Young Mo Kang, Wenpeng Yin, Huan Wang, Semih Yavuz, et al. 2022. Converse–a tree-based modular task-oriented dialogue system. *arXiv preprint arXiv:2203.12187*.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog systems with gpt-2. *AAAI*.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117.

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, S Yu Philip, Richard Socher, and Caiming Xiong. 2020. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 154–167.

Transformer-based Multi-Party Conversation Generation using Dialogue Discourse Acts Planning

Alexander Chernyavskiy and **Dmitry Ilvovsky**
National Research University Higher School of Economics
Moscow, Russia
alschernyavskiy@gmail.com; dilvovsky@hse.ru

Abstract

Recent transformer-based approaches to multi-party conversation generation may produce syntactically coherent but discursively inconsistent dialogues in some cases. To address this issue, we propose an approach to integrate a dialogue act planning stage into the end-to-end transformer-based generation pipeline. This approach consists of a transformer fine-tuning procedure based on linearized dialogue representations that include special discourse tokens. The obtained results demonstrate that incorporating discourse tokens into training sequences is sufficient to significantly improve dialogue consistency and overall generation quality. The suggested approach performs well, including for automatically annotated data. Apart from that, it is observed that increasing the weight of the discourse planning task in the loss function accelerates learning convergence.

1 Introduction

The popularity of dialogue systems has resulted in an increased demand for their utilization in various applications. Existing approaches are largely focused on two-party conversations, which is applicable in chat-bots and assistance systems (Shang et al., 2015; Wang et al., 2015; Young et al., 2018; Gu et al., 2019). At the same time, there is another type of dialogue, known as multi-party conversations (Traum, 2003; Uthus and Aha, 2013; Ouchi and Tsuboi, 2016; Le et al., 2019). In this case, several interlocutors are involved in the dialogue, and the dialogue tree, consisting of successive utterances, is wide enough. This type of dialogue can be observed in Internet forum discussion threads.

Due to the complexity of the structure of MPC dialogues, it becomes more challenging for the base seq2seq models to generate response texts. Multi-task learning and external knowledge can be considered to simplify the utterance generation task. To this end, we additionally leverage the theory of dialogue acts (Stone et al., 2013; Zhang et al., 2017), which shows by which discourse rhetorical relations (more precisely, dialogue acts) the individual utterances of the dialogue are connected. Figure 1 shows an example of such a structure.

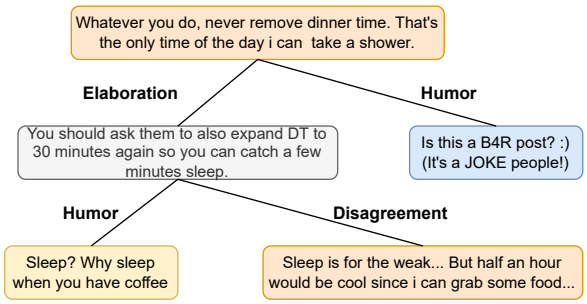


Figure 1: A manually annotated discourse tree for the multi-party dialogue. The color identifies the speaker.

Our major idea is that the use of dedicated discourse tokens to both input and target texts will enhance the coherence of discourse and consequently the overall quality of generation.

To illustrate, let us examine the following example of a help request forum thread. Initially, the user describes a problem and seeks advice. Subsequently, multiple dialogue turns occur, culminating in the following phrases:

- *[answer]* reinstall OS/get a new HDD/SSD
- *[disagreement]* really? My HDD was working right until yesterday...

Here, discourse relations demonstrate that the last utterance is indicative of a disagreement rather than a question. Accordingly, the next appropriate utterance should contain an inquiry to resolve the user's initial problem. Our discourse-based model generated "is your HDD in safe mode?", while the base model outputted "The answer is no.", which is much more distant from the corresponding ground-truth utterance, "how long have you had your PC?". This shows the advantage of the discourse-based

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 519–529
September 11–15, 2023. ©2023 Association for Computational Linguistics

model, as it first plans out discourse relations before generating text tokens.

Generally speaking, we suggest multi-task learning consisting of dialogue acts planning and response generation joined in the single pipeline. We integrate discourse tokens into a two-stage pipeline for MPC generation (see Section 3.1 for details). Its first stage is used to identify a speaker and an addressee at the current step, whereas the second stage is used to generate the current response text. The first part is quite challenging, but recent studies allow one to solve it qualitatively (Le et al., 2019; Gu et al., 2021). At the same time, only base models were researched for the second stage, leaving the relevance of discourse usage in dialogue generation unexplored. Therefore, we mainly focus on the second stage.

The task can be formalized as a graph2text, within which the BART and T5 models have already been partially investigated. Key part here is a linearization technique that was used for some graph structures (Ribeiro et al., 2020; Kale and Rastogi, 2020), but not for the discourse structure and dialogue generation yet. Thus, we suggest integrating dialogue acts into the linearization of MPC graphs.

Our contributions can be summarized as follows:

- We suggest multi-task learning consisting of dialogue acts planning and response generation to improve the transformer-based MPC generation pipeline.
- We analyze the importance of having discourse tokens in both parts of seq2seq linearized input pairs.
- We show that the transformer-based approach converges faster if it has more weight in the loss related to the dialogue acts planning task.

The code is available at `https://github.com/alchernyavskiy/discourse_mpc_generation`.

2 Related Work

In this paper, we consider multi-party conversations (MPCs). The process of generation is generally split into two stages since it consists of several entire tasks. Ouchi and Tsuboi (2016) presented a task of identifying the speaker and the addressee of an utterance (first stage), and recent approaches have been aimed at improving results in this task (Le et al., 2019; Gu et al., 2021).

At the second stage, associated with the response generation task, some approaches use GCN to encode the complex MPC structure (Hu et al., 2019). It intended to improve prior recurrent neural network-based sequence-to-sequence (seq2seq) generation models (Luan et al., 2016; Serban et al., 2017). At the same time, it was shown that recent transformer-based approaches, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), achieve top results in various generation tasks, including dialogue generation. Therefore, BART and T5 are commonly used as the base generation models in recent approaches. For instance, Li et al. (2021b) uses the BART as a backbone to train the model that considers long-range contextual emotional relationships.

Moreover, recent transformer-based approaches effectively solve graph-to-text generation tasks. Ribeiro et al. (2020) demonstrated that BART and T5 outperform various GNNs trained to encode AMR graphs in the AMR-to-text task. Similarly, Kale and Rastogi (2020) indicated that pre-training in the form of T5 enables simple, end-to-end models to outperform pipelined neural architectures tailored for data-to-text generation. The key factor here is that any graph can be linearized, and Hoyle et al. (2021) showed that transformers are invariant to the method by which graphs are linearized. Thus, we do not explore ways of linearizing dialogue graphs augmented by discourse relations, but choose one of the most reasonable ones.

Discourse parsing of multi-party conversations is an adjacent direction that is gaining popularity. There are several works where the dialogues were analyzed in terms of discourse structure ad discourse relations parsing (Afantenos et al., 2015; Shi and Huang, 2019; Wang et al., 2021; Koto et al., 2021). Despite this, there are not a large number of publicly available datasets with discursively-annotated dialogues. Basically, all comparisons are conducted for the STAC dataset (Asher et al., 2016), which is quite small. As far as we know, the only large publicly available dataset is the CDSC dataset (Zhang et al., 2017). The main difference in our research from this direction is that we do not aim to suggest a novel discourse parser. At the same time, we use existing parsers and explore the importance of using discourse in the applied generative task.

Our idea of generating discourse relations in dialogues comes from the story-telling task. To

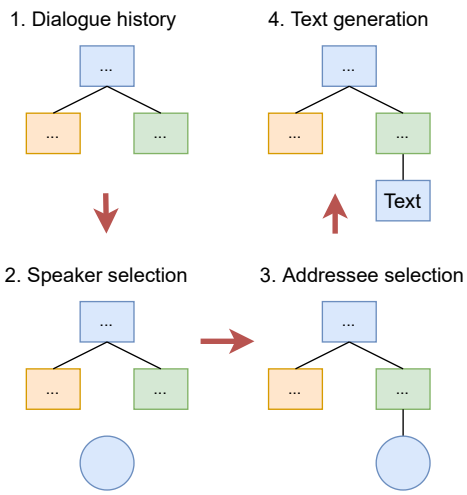


Figure 2: End-to-end MPC generation pipeline. The colors represent the speakers and are chosen as an example.

facilitate discourse coherence, some researchers proposed neural text generation based on discourse planning with an auxiliary model (Ji et al., 2016; Harrison et al., 2019; Chernyavskiy, 2022). However, in the case of dialogues, discourse structure has been explored only in the context of summarization and machine reading comprehension tasks (Feng et al., 2021; Li et al., 2021a).

3 Methods

3.1 End-to-End MPC Generation Pipeline

Figure 2 illustrates the end-to-end pipeline of multi-party conversation generation, which consists of several main steps at each dialogue turn. This pipeline implies that the next turn speaker selection can be separated from the response selection. There are also united approaches, but we do not consider them in this paper.

Firstly, the next speaker should be selected. Then, we should decide to which utterance it responds, or in other words, select the addressee of the generating utterance. Both these steps ($1 \rightarrow 2$ and $2 \rightarrow 3$ in Figure) are typically combined into a single stage.

In this paper, we investigate the last generation phase (step $3 \rightarrow 4$ in Figure), namely the text generation for the current utterance. In our case, we also distinguish the dialogue act planning substage that consists of selecting the edge type in terms of dialogue acts.

3.2 Linearization

This section describes linearization of graphs representing multi-party dialogues annotated by discourse relations, in addition to the main MPC features. As it was mentioned above, we do not have the goal of tuning the linearization technique, and we have chosen one of the most reasonable ones.

The graph structure can be converted to a sequence by sorting the utterances by time. Each utterance and its meta can be linearized according to the following way. Firstly, we should assign an utterance id and indicate the current speaker. Then we must specify which addressee statement to respond to and how to respond to it (discourse relation). Finally, we should produce the next utterance text. To handle the first two steps, we suggest to use special tokens as the identifiers of speakers and utterances: $\{\langle s_i \rangle\}$ and $\{\langle u_i \rangle\}$ correspondingly. For instance, a linearized i-th utterance written by the j-th speaker in response to the k-th utterance looks like as follows:

"$\langle u_i \rangle$ $\langle s_j \rangle$ $\langle$relation$\rangle$ $\langle u_k \rangle$ response text"

Also, we use a separator token to join single utterances and get full representation of the current dialogue state. To specify an utterance to respond to at the current turn, we add its representation to the end of the dialogue sequence. We use the resulting representations as the inputs of seq2seq models. Figure 3 demonstrates an example of the MPC dialogue linearization procedure.

We passed the target seq2seq texts to the model in the following format: "$\langle$relation$\rangle$ response text".

It should be highlighted that the response text is generated following discourse relations. Consequently, its tokens are produced with an attention mechanism that takes into account the discourse token. Moreover, the transformer-based language modeling approach allows us not to use a special auxiliary model like in (Ji et al., 2016; Harrison et al., 2019).

3.3 Model and Loss Function

We use BART and T5 as the base transformer models due to their state-of-the-art performance in various text generation and graph-to-text generation tasks.

In our approach, discourse tokens are planned first and an auxiliary model is not required, their importance can be adjusted through weights in the loss function. To this end, we employ the weighted cross-entropy loss:

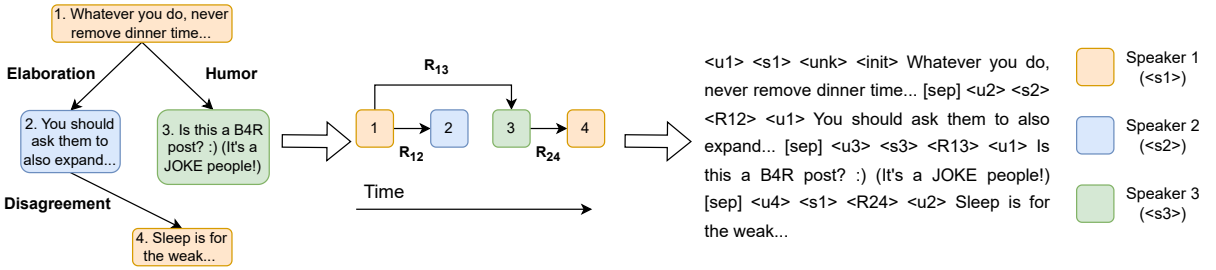


Figure 3: Example of the discursively-annotated MPC linearization process. Firstly, all nodes are ordered temporally, forming a chain. Then, it is transformed to text representation using special tokens to display meta information: $\langle u_i \rangle$ are used for utterance ids, $\langle s_i \rangle$ are tokens for speaker ids (are signified by colors), and $\langle R_{ij} \rangle$ are used for relations. Additionally, an $\langle$init$\rangle$ token is introduced due to the fact that the first replica does not have an adressee.

$$L = -\frac{1}{|S|} \sum_{j=1}^{|S|} \sum_{i=1}^{|D_{\text{all}}|} w(y_j) I\{x_i = y_j\} \log(p(x_{ji})) \tag{1}$$

$$w(y) = \alpha I\{y \in D\} + I\{y \notin D\} \tag{2}$$

Here, $|S|$ is the target sequence length, $|D_{\text{all}}|$ is the full vocabulary size, $p(x_{ji})$ is the predicted probability of the i-th token for the place j, and y_j is the target token. I denotes an indicator function. D is the predefined set of discourse tokens, and α is the weight related to the dialogue acts planning task. When the α coefficient is zero, we actually provide the standard response generation task instead of the multitask learning.

The described approach is quite intuitive, but at the same time, it allows for significant improvement of the quality of generation and acceleration of convergence, as demonstrated in Section 5.

4 Datasets

This section presents the discursively-annotated datasets used for evaluation.

4.1 CDSC

First, we utilize the largest manually annotated dataset of dialogue acts in online discussions, namely the Coarse Discourse Sequence Corpus (CDSC) proposed by Zhang et al. (2017). It contains ~9K Reddit threads (in English), with comments annotated with 9 main discourse act labels that were designed to cover general discourse and an "other" label. It should be highlighted that, to the best of our knowledge, this dataset is the only open-source dataset that is sufficiently large and includes discourse act labeling.

The list of the dialogue acts used in the dataset is the following: "Question", "Answer", "Announcement", "Agreement", "Appreciation", "Disagreement", "Negative Reaction", "Elaboration", "Humor", "Other".

There exists some missing values in the data, and we replace them with the $\langle$unk$\rangle$ special token. We splitted the data into training and test sets with a ratio of 6:1. As a preprocessing, we removed instances with missing values and all non-ascii characters from texts.

4.2 Movie Reddit Dataset

No other large-scale, discursively-annotated open datasets for the MPC generation task are available, and manually-labeled data is not typically accessible in real-world applications. Therefore, we collected our own dataset and labelled it automatically to increase the significance of the findings.

Similar to CDSC, we parsed threads from Reddit (the largest open source of dialogues), but focused primarily on the movie domain since it does not require any specific knowledge and is generally considered for the conversation analysis tasks (Zhou et al., 2018). We collected roughly 90k dialogues from the 25 most popular Reddit subthreads discussing movies, series and TV shows. To obtain discourse acts labels, we trained our own discourse parser from scratch based on the CDSC dataset. Existing parsers are trained using much smaller datasets and operate with other discourse relations, making evaluation inconvenient. We chose the Two-Stage discourse parser (Wang et al., 2017) as the model architecture, since it is open-source and has obtained SOTA results for dialogue discourse parsing. The entire procedure for preprocessing and input data construction used is identical to that of CDSC.

5 Experiments

This section discusses implementation details and experiment results, including human evaluation.

5.1 Implementation Details

We fine-tuned the base-sized BART and T5 models (139M and ∼220M parameters respectively). The maximum source length was set to 1024, and the maximum target length was set to 64 (these values were estimated using the training set). The models were trained on batches of size 2 with a learning rate of 2e-5 during 5 epochs. Other hyperparameters were used by default.

Each model was trained on the GPU Tesla V100 32G for approximately 10 hours.

5.2 Discourse Planning Importance

We use the popular ROUGE-based (Lin, 2004) and BLEU-based (Papineni et al., 2002) scores to automatically estimate the overall generation quality. We calculate it based on the target texts cleared of discourse tokens.

We conducted experiments for the three settings of the dataset used for fine-tuning: (1) $\mathcal{D}$ containing discourse relation tokens in both source and target texts; (2) $\mathcal{D}_1$ only containing discourse relations in source texts; and (3) $\mathcal{D}_2$ having no discourse relations at all (is considered as the baseline model).

We selected the weight α for discourse planning as 100 using grid search for the $\mathcal{D}$ setting. Results further detailed in Section 5.4 indicate that it is better to choose the weight of discourse tokens in the loss function larger than the rest. At the same time, the difference for large values is not significant, so we chose the same value of α for both datasets. This weight was not used for $\mathcal{D}_1$ and $\mathcal{D}_2$ since they do not contain dialogue act tokens in the target texts. Additionally, it should be noted that $\mathcal{D}_1$ can be considered an equivalent for $\mathcal{D}$, where the α coefficient is set to 0 and dialogue acts are not being planned.

Table 1 presents the F1-scores for the ROUGE-based and BLEU-based metrics for the BART model. The results demonstrate that the model incorporating discourse planning (setting $\mathcal{D}$) achieved the highest scores and was significantly superior to the other models. This indicates that discourse planning simplifies the generation of response texts, even for BART.

Furthermore, for the Movie Reddit dataset, the results in setting $\mathcal{D}_1$ outperform those those in setting $\mathcal{D}_2$. It follows that in the case when the training dataset is large and comprises more examples of discourse dependencies, incorporating dialogue act markers in input texts can also be beneficial. Nevertheless, the maximum quality boost is obtained precisely when training the auxiliary discourse planning task (in setting $\mathcal{D}$).

The metrics for $\mathcal{D}$ are slightly lower for the Movie Reddit dataset than for the CDSC. This is primarily attributed to the fact that all dialogue acts in Movie Reddit were labeled automatically, which can lead to inaccurate labels. However, the results remain consistent between the two datasets, and the model featuring a discourse planning stage performs significantly better than the base model, even considering the automatically labeled data.

Table 2 demonstrates results for the T5 model. The language modeling quality is slightly inferior to that of BART, which may be due to a suboptimal hyperparameter selection. The decreased quality in the last rows may be attributed to the use of an extended tokenizer with discourse tokens (nevertheless, this assumption should not greatly affect the quality). At the same time, hyperparameter search was not our primary objective, and the results confirm that with an appropriate selection of hyperparameters, discourse planning greatly improves generation quality.

5.3 Human Evaluation

To enhance the evaluation as well as cover aspects that cannot be assessed by automatic metrics, we conducted a human evaluation. Here, the main goal was to compare texts generated by two BART models: the base model and the model trained via multi-task learning. For each instance, the experts were tasked with choosing which of two options was the best for continuing the dialogue (or whether they were equal), as well as evaluating each option on a 3-point scale according to the criteria of consistency (coherence) and meaningfulness. Coherence assessed the relation between the current utterance and the addressee, as well as the overall logic of the dialogue, while meaningfulness assessed the semantic load of the utterance in its general context. The two scales were rated on a scale of 0-2, with 0 representing a bad prediction, 1 representing a generally normal prediction with some inaccuracies, and 2 representing a prediction close to perfect. In order to ensure reliability in the evaluation, the options were shown in random order.

Dataset	Setting	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2
CDSC	$\mathcal{D}$ [full discourse data]	**9.34**	**0.66**	**8.37**	**8.68**	**0.32**
	$\mathcal{D}_1$ [no discourse in resp.]	7.39	0.48	6.64	6.53	0.30
	$\mathcal{D}_2$ [no discourse at all]	7.44	0.43	6.71	6.58	**0.32**
Reddit	$\mathcal{D}$ [full discourse data]	**8.89**	**0.58**	**7.96**	**8.11**	0.17
	$\mathcal{D}_1$ [no discourse in resp.]	7.76	0.54	7.07	6.80	**0.20**
	$\mathcal{D}_2$ [no discourse at all]	7.45	0.51	6.77	6.47	0.17

Table 1: Performance of BART-based models on the CDSC and Movie Reddit test sets for different variants of training datasets (denoted as settings). We use F1-scores for the ROUGE-based metrics. $\mathcal{D}$ uses discourse relations in both source and target texts in seq2seq training, $\mathcal{D}_1$ has responses cleared of discourse relations, and $\mathcal{D}_2$ is the dataset without discourse relations at all (is used to train the baseline model). Here, STD ≤ 0.6 in cases of unigram-based metrics and STD ≤ 0.1 in cases of bigram-based metrics.

Setting	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2
$\mathcal{D}$ [full discourse data]	**8.81**	**0.50**	**7.87**	**8.02**	0.25
$\mathcal{D}_1$ [no discourse in resp.]	7.06	0.39	6.36	6.12	0.25
$\mathcal{D}_2$ [no discourse at all]	6.94	0.41	6.24	5.96	0.21

Table 2: T5-based model performance on the CDSC test set for different training datasets and α coefficients.

Model	# better	Coherence	Meaning.
Base	62	1.11	1.32
Disco	**83**	**1.32**	**1.33**

Table 3: Human evaluation results on the subset of 200 dialogues from the CDSC test set. "Base" refers to the base BART model and "Disco" refers to the model trained via dialogue acts planning.

Table 3 presents the obtained results for 200 random dialogues from the CDSC test dataset. Here, the scores are averaged across the corpus. We can see that responses produced by the custom approach are preferable in more cases. This is mainly because the discourse-based model's responses are more coherent and more appropriate for continuing the dialogue, despite perhaps less semantically appropriate formulations (the task of generating texts for some dialogue acts is quite challenging). Although the overall improvement is not substantial, there is a considerable progress in the aspect of consistent dialogue generation.

5.4 Convergence Speed

In this section, we evaluate the convergence speed of our model. The rate of convergence can be estimated in several ways, and in this case we have chosen one of them, which is related to estimating the fewest number of steps to get high quality. For early quality estimation, we train models for a smaller number of steps (2 epochs) with varying values of the α coefficient in the loss function to indicate the importance of the discourse planning task. These values are 1, 10, 30, 100 and 200. We measure the quality of discourse tokens using Accuracy and the quality of response texts using F1-based ROUGE-L.

Figure 4 demonstrates the results that reveal a strong correlation between Accuracy and ROUGE values, suggesting that improved discourse planning improves the overall quality of language modeling. Furthermore, these results indicate that the approach converges faster (reaching optimal quality at earlier epochs) if the discourse planning task has more weight. For instance, increasing α from 1 to 100 yields a significant increase in the convergence speed of the training process, requiring far fewer steps to attain the best possible generation quality.

6 Discussion

In this section, we partially explain how discourse improves generation quality for the model trained using discourse planning and demonstrate differences using concrete examples.

6.1 Error Analysis

In order to analyze which dialogue acts the discourse-based model actually plans and which of them can improve the overall quality of language modeling, we compare the quality of dialogue acts planned by the base and the discourse-based BART

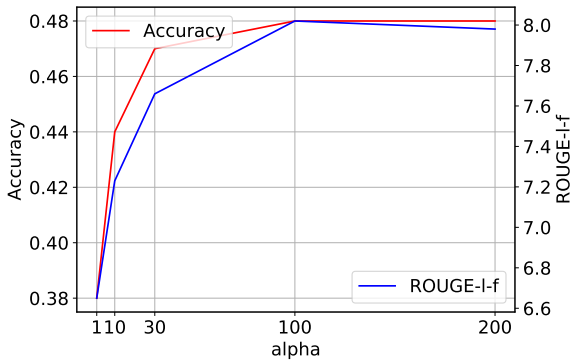


Figure 4: Discourse Accuracy (blue line) and ROUGE (red line) scores depending on α for the CDSC test set.

models. As the base model does not explicitly generate dialogue acts, our trained discourse parser was used to label them. The corpus used for training also contains unlabeled instances, which are not taken into account in our analysis.

Figure 5 demonstrates the confusion matrices for both the base and custom models. The correct labels were taken from the dataset. The results illustrate that the base model achieves only an accuracy of 0.315, whereas the custom model gets 0.615. The task is complicated by the fact that the dialogue can be continued in various ways and often there is no single correct dialogue act.

The confusion matrix for the discourse model is closer to diagonal, indicating improved performance. A standout feature is that the custom model successfully plans not only common relations, such as Elaboration and Answer, but also rarer ones. So, the discourse-based model more accurately determines when it is necessary to thank the interlocutor (Appreciation), and when to ask a clarifying question (Question). Interestingly, the discourse model predicts better even such relations as Disagreement and Humor. Some relations are quite non-trivial, and even with the right relations planned, it can be challenging for the model to generate the correct words to achieve the highest automatic generation metrics. However, as seen in Section 5, the right choice of dialogue acts is a step towards high-quality generation.

6.2 Generation Examples

Figure 6 shows examples of dialogues generated by the base BART model and the BART model fine-tuned using discourse tokens. We focus on the generation of the last utterance, as this is the second stage of the general generative MPC pipeline.

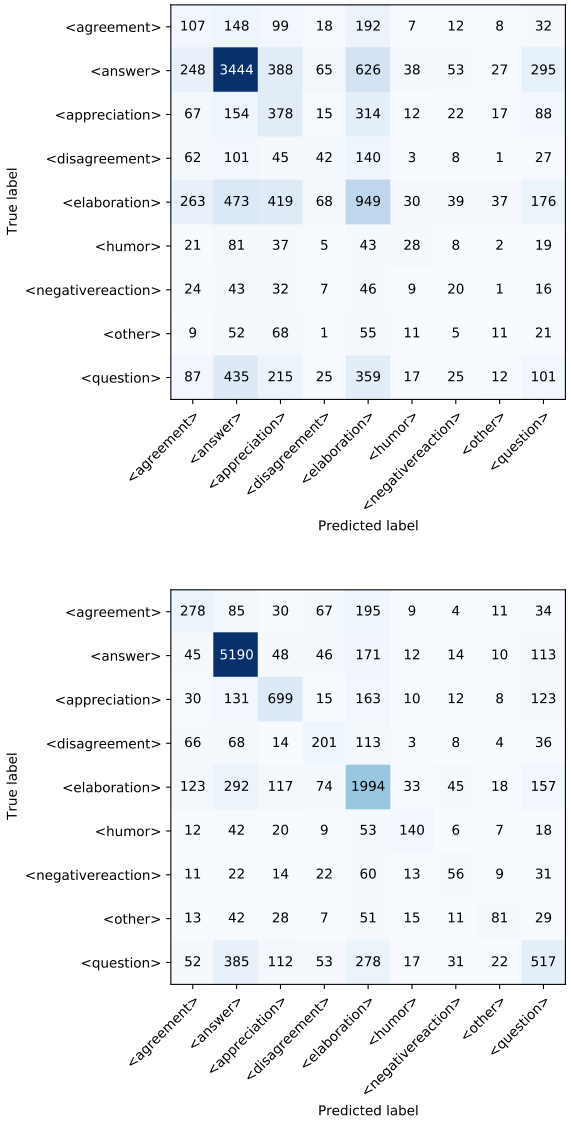


Figure 5: Confusion matrices of dialogue act planning for the CDSC test set for the base BART (top) and discourse BART (bottom) models.

The first example demonstrates a simple dialogue with only two speakers. Even in such scenarios, the base BART model may struggle. In this instance, the base model attempted to answer the question “maybe an endgame companion?”, while completely disregarding the context of the conversation. At the same time, the discourse planning model was able to respond in a more logical and reasonable manner, continuing the topic and aiming to achieve the initial goals of the first speaker by asking a new question.

The second dialogue presented appears to be a chain, with three speakers. One can see that the most pertinent relations for this dialogue are “agree-

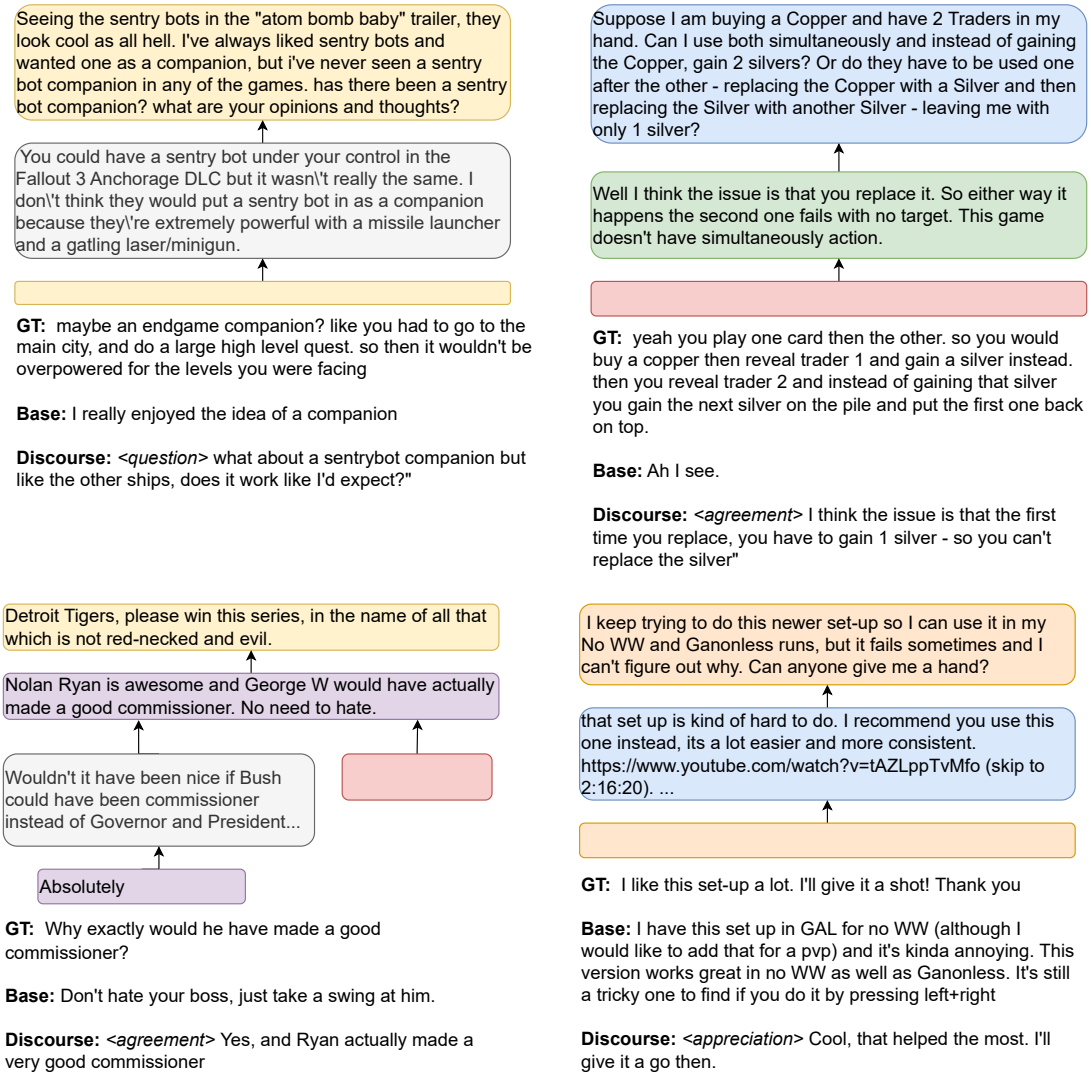


Figure 6: Examples of generated dialogues with the ground-truth reply (GT) and options generated by the base and discourse approaches. Node color is used to identify the speaker and arrows on the edges indicate the addressees. The task is to generate the utterance text for the empty node. Here, "Base" refers to the base BART model and "Discourse" refers to the BART model fine-tuned using special discourse tokens. The discourse-based model also produces discourse tokens which are shown in italics.

ment" and "disagreement", given the introduction of a new speaker. Nevertheless, the base model chooses options with "appreciation", which does not contribute to dialogue continuation.

The third example (bottom left in the Figure) demonstrates a dialogue with a more intricate structure, making the planning of discourse relations more challenging. There exist several ways to move the dialogue forward, and the chosen "agreement" relation allows the discourse-based model to generate an utterance that is not removed from the context.

The final example shows the case in which the "appreciation" token is sufficient for the discourse-based model to generate a concise, suitable answer without excessive detail.

It is important to emphasize that the base model is still capable of producing valid responses due to its element of randomization in the sampling process. Nevertheless, the accurate choice of a discourse relation (dialogue act) at the start significantly simplifies the search for alternatives and almost always results in valid responses.

7 Conclusion and Future Work

In this paper, we explored the effectiveness of transformer fine-tuning based on discourse dialogue acts planning for the multi-party conversation generation task. We evaluated our approach on the largest manually and automatically annotated datasets of

dialogues from Reddit.

The evaluation including automatic and human assessments revealed that incorporating special discourse tokens into the linearized training sequences could significantly improve the generation metrics and is an important step towards coherent generation. The proposed approach performed well even on the automatically annotated dataset, and increasing the weight of discourse tokens in the loss function further accelerated learning convergence.

Future work includes the analysis of other types of linguistic information (such as syntactic and semantic relations), other ways of integrating them into the training process, as well as experiments for alternative MPC generation pipelines.

8 Ethics and Broader Impact

The training of large transformer-based models is one of the reasons leading to global warming. However, we do not train these models from scratch and use the fine-tuning procedure. Moreover, we consider only the base variants of the models that have a lower number of trainable parameters.

9 Limitations

The proposed approach is not limited to the English language or BART/T5 approaches. The main limitations are the presence of annotated data that can be acquired manually or with the help of a parser, and the seq2seq nature of the transformer-based approach. As with most conversational agents, there are possible adverse impacts, like spreading harmful or hateful messages or misinformation. Models mainly learn the training dialogues, and most of these issues can be addressed through proper preprocessing or the selection of appropriate datasets.

Acknowledgements

The article was prepared within the framework of the HSE University Basic Research Program. It was also supported in part through the computational resources of HPC facilities at NRU HSE.

References

Stergos D. Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 928–937. The Association for Computational Linguistics.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos D. Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Alexander Chernyavskiy. 2022. Improving text generation via neural discourse planning. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 1543–1544. ACM.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3808–3814. ijcai.org.

Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2019. Interactive matching network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 2321–2324. ACM.

Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3682–3692. Association for Computational Linguistics.

Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn A. Walker. 2019. Maximizing stylistic control and semantic accuracy in NLG: personality variation and discourse contrast. *CoRR*, abs/1907.09527.

Alexander Miserlis Hoyle, Ana Marasovic, and Noah A. Smith. 2021. Promoting graph awareness in linearized graph-to-text generation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 944–956. Association for Computational Linguistics.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. GSN: A graph-structured network for multi-party dialogues. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5010–5016. ijcai.org.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. *CoRR*, abs/1603.01913.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 97–102. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Top-down discourse parsing via sequence labelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 715–726. Association for Computational Linguistics.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1909–1919. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jiaqi Li, Ming Liu, Zihao Zheng, Heng Zhang, Bing Qin, Min-Yen Kan, and Ting Liu. 2021a. Dadgraph: A discourse-aware dialogue graph neural network for multiparty dialogue machine reading comprehension. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8. IEEE.

Shimin Li, Hang Yan, and Xipeng Qiu. 2021b. Contrast and generation make BART a good dialogue emotion recognizer. *CoRR*, abs/2112.11202.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Yi Luan, Yangfeng Ji, and Mari Ostendorf. 2016. LSTM based conversation models. *CoRR*, abs/1603.09457.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2133–2143. The Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating pretrained language models for graph-to-text generation. *CoRR*, abs/2007.08426.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301. AAAI Press.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1577–1586. The Association for Computer Linguistics.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7007–7014. AAAI Press.

Matthew Stone, Una Stojnic, and Ernest Lepore. 2013. Situated utterances and discourse relations. In *Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013, March 19-22, 2013, University of Potsdam, Potsdam, Germany*, pages 390–396. The Association for Computer Linguistics.

David R. Traum. 2003. Issues in multiparty dialogues. In *Advances in Agent Communication, International Workshop on Agent Communication Languages, ACL 2003, Melbourne, Australia, July 14, 2003*, volume 2922 of *Lecture Notes in Computer Science*, pages 201–211. Springer.

David C. Uthus and David W. Aha. 2013. Multiparticipant chat analysis: A survey. *Artif. Intell.*, 199-200:106–121.

Ante Wang, Linfeng Song, Hui Jiang, Shaopeng Lai, Junfeng Yao, Min Zhang, and Jinsong Su. 2021. A structure self-aware model for discourse parsing on multi-party dialogues. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3943–3949. ijcai.org.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2015. Syntax-based deep matching of short texts. *ArXiv*, abs/1503.02427.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 184–188. Association for Computational Linguistics.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4970–4977. AAAI Press.

Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Eleventh International AAAI Conference on Web and Social Media*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 708–713. Association for Computational Linguistics.

Incorporating Annotator Uncertainty into Representations of Discourse Relations

S. Magalí López Cortez **Cassandra L. Jacobs**
Department of Linguistics
University at Buffalo
Buffalo, NY, USA
solmagal;cxjacobs@buffalo.edu

Abstract

Annotation of discourse relations is a known difficult task, especially for non-expert annotators. In this paper, we investigate novice annotators' uncertainty on the annotation of discourse relations on spoken conversational data. We find that dialogue context (single turn, pair of turns within speaker, and pair of turns across speakers) is a significant predictor of confidence scores. We compute distributed representations of discourse relations from co-occurrence statistics that incorporate information about confidence scores and dialogue context. We perform a hierarchical clustering analysis using these representations and show that weighting discourse relation representations with information about confidence and dialogue context coherently models our annotators' uncertainty about discourse relation labels.

1 Introduction

Discourse relations (DRs) are those relations such as Elaboration, Explanation, Narration, which hold between discourse units. The task of labeling DRs is known to pose difficulties for annotators (Spooren and Degand, 2010), as sometimes more than one interpretation may be possible (Scholman et al., 2022; Webber, 2013).

Recent studies have shown that allowing for multiple labels in annotation can improve the performance of discourse parsers (Yung et al., 2022). Scholman et al. (2022) test different label aggregation methods in a crowdsourced corpus annotated by 10 workers and find that probability distributions over labels better capture ambiguous interpretations of discourse relations than majority class labels. (1) shows an example from their corpus, where the relation between the second and third sentences (in italics and bold, respectively), was interpreted as Conjunction by four annotators and Result by five annotators.

(1) It is logical that our attention is focused on cities. *Cities are home to 80% of the 500 million or so inhabitants of the EU.* **It is in cities that the great majority of jobs, companies and centres of education are located.** (adapted from DiscoGeM, Europarl genre; Scholman et al., 2022, italics and bolding are ours.)

Annotating the discourse relation between these two sentences with both Conjunction and Result captures different possible interpretations of the relation between these segments. For example, the two sentences may contain two conjoined facts about cities, but can also be perceived as describing a causal relation between the first and second sentence (i.e., as cities are home to the largest part of the population, most jobs, companies and educational institutions are located there).

In this work, we investigate which relations are distributionally similar or co-occurring in multilabel annotations of spontaneous conversations. We are particularly interested in how novice annotators interpret discourse relation categories when annotating spoken conversational data. We collect annotations of DRs from Switchboard telephone conversations (Godfrey et al., 1992), allowing for multiple labels, and ask for confidence scores. We find that confidence scores vary significantly across dialogue contexts (single turn vs. pairs of turns produced by the same speaker vs. pairs of turns produced by different speakers). We incorporate information about these three dialogue context types and confidence scores into distributed representations of discourse relations. A clustering analysis shows that discourse relations that tend to occur across speakers cluster together, while discourse relations which tend to occur within a speaker, either in the same turn or different turns, form their own cluster.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 530–537
September 11–15, 2023. ©2023 Association for Computational Linguistics

2 Annotation of Discourse Relations

Our analyses are built on the dataset collected in López Cortez and Jacobs (2023), who selected 19 conversations from Switchboard[1], a corpus consisting of telephone conversations between pairs of participants about a variety of topics (e.g. recycling, movies, child care). We chose this corpus because it contains informal, spontaneous dialogues, and because it has been used within linguistics in various studies on conversation (Jaeger and Snider, 2013; Reitter and Moore, 2014).

2.1 Discourse Units

An initial set of turns for annotation was selected by using spaCy's dependency parser (Honnibal et al., 2020, version 3.3.1) to select turns with two or more ROOT or VERB tags. We define a turn as each segment of dialogue taken from Switchboard. We note that an utterance produced by one speaker (A) may take place during a continuous utterance by another speaker (B). Switchboard splits A's utterance into two turns in these cases. We return to this point in the Discussion.

We manually segmented these turns into elementary discourse units (EDUs). The main criteria for segmenting turns into EDUs was that the unit performs some basic discourse function (Asher and Lascarides, 2003). By default, finite clauses are considered EDUs, as well as comment words like "Really?" or acknowledgments such as "Uh-huh" or "Yeah." Cases of interruptions and repairs were segmented if they constituted a turn in Switchboard, as in example (2a), and when they contained a verb, as in example (2b). Cases of repetition as in (2c) were not considered separate EDUs. We segmented disfluencies ("uh") and some non-verbal communication ("[laughter]") but we did not select these for discourse relation labeling.

(2) a. B: || So you don't see too many thrown out around the || [laughter] || streets. ||
A: || Really ||
B: || Or even bottles. ||

b. B: || I think, || uh, || I wonder || if that worked. ||

c. A: || What kind of experience do you, do you have, then with child care? ||

[1]We discarded the annotations from one conversation because the annotators did not follow the guidelines.

Because many EDUs are very short, we selected pairs of elementary discourse units and complex discourse units (CDUs) for discourse relation annotation. CDUs consist of two or more EDUs that constitute an argument to a discourse relation (Asher and Lascarides, 2003). We use the term *discourse units* (DUs) to refer to both EDUs and CDUs.

2.2 Dialogue Contexts

We manually selected items for annotation across three different contexts: within a single turn, across two turns within a speaker, and across two immediately adjacent turns (two speakers). (3) shows an example for each context kind, with the first DU in italics and the second in bold. Example (3a) shows two discourse units within a speaker's turn. (3b) shows two discourse units uttered by the same speaker but that span across two different turns, interrupted by one turn. We did not include any constraint for the length of the interrupting turn. (3c) shows two DUs uttered by speakers in adjacent turns. We leave for future work the annotation of pairs of discourse units that may have a longer-distance relation with more turns in between DUs.

(3) a. A: || *and they discontinued them* || **because people were coming and dumping their trash in them.** ||

b. B: || No, || *I just, I noticed* || *in Iowa and other cities like that, it's a nickel per aluminum can.* ||
A: || Oh. ||
B: || **So you don't see too many thrown out around the || [laughter] || streets**.

c. A: || *We live in the Saginaw area.* ||
B: || **Saginaw?** ||

2.3 Taxonomy of Discourse Relations

The DRs chosen to annotate our corpus were adapted from the STAC corpus manual (Asher et al., 2012, 2016). STAC is a corpus of strategic multiparty chat conversations in an online game. Table 1 shows the taxonomy used. We selected 11 DRs based on a pilot annotation by the first author, and added an "Other" category for relations not included in the list of labels. We focused on a small taxonomy to minimize the number of choices presented to our novice annotators. We refer readers to López Cortez and Jacobs (2023) for details and examples of each relation in the taxonomy. Future work will include revising the taxonomy used.

Acknowledgement	Elaboration
Background	Explanation
Clarification Question	Narration
Comment	Question-Answer Pair
Continuation	Result
Contrast	Other

Table 1: Taxonomy of discourse relations.

2.4 Annotation Procedure

The annotation of discourse relations was done by students enrolled in a Computational Linguistics class. Students were divided into 19 teams of approximately 5 members each, and each team was assigned a conversation. The annotation was performed individually, but teams then discussed their work and wrote a report together. Annotators were trained using written guidelines, a quiz-like game, and a live group annotation demo.

We used the annotation interface Prodigy (Montani and Honnibal, 2018). Each display presented the two target discourse units plus two context turns before and two after. Annotators also had access to the entire conversation throughout the annotation task. Below the text, the screen showed a multiple choice list of discourse relations plus the "Other" category. We allowed for the selection of multiple labels following previous findings that allowing for multiple labels better captures ambiguous interpretations of discourse relations (Scholman et al., 2022) and improves the performance of discourse parsers (Yung et al., 2022).

Each display also asked for confidence scores in the range 1-5, corresponding to least to most confident. We did not pursue label-specific confidence scores but rather the confidence in the label(s) as a whole in the interest of minimizing annotator overhead. The results of this work show that per-label confidence scores or a slider-based approach may be informative and is a topic for future work. We include an example annotation item in Appendix C.

3 Dialogue Context as a Predictor of Confidence Scores

First we sought to understand how discourse relations and dialogue context (as defined above) influence annotator confidence. Because our confidence ratings data has multiple observations for each annotator, each team and each DU, it is hierarchical and thus benefits from being analyzed using hierarchical mixed effects models. Due to the ordinal nature of the ratings data, we use the cumulative link approach (CLMM; Liddell and Kruschke, 2018; Howcroft and Rieser, 2021) rather than model confidence scores as real-valued in linear regression. We first built a null model containing only random intercepts by annotator and compared it to a model containing an additional fixed effect and random slope by annotator for dialogue context type: single turn, across turns within speaker and across speakers (*kind*, dummy coded). A likelihood ratio test revealed a significant improvement in fit by adding kind as a predictor ($\chi^2(7) = 126.64, p < 0.001$). Adding random intercepts for DU pairs to account for annotation difficulty across DU pairs also led to a significant improvement in model fit beyond the model containing dialogue context kind ($\chi^2(1) = 195.01, p < 0.001$). This suggests that our annotators' confidence scores are sensitive to the context of DU pairs.

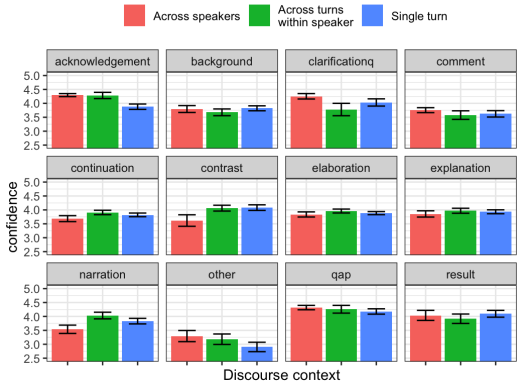


Figure 1: Confidence scores per context kind across discourse relations. *qap* stands for Question-Answer Pair and *clarificationq* for Clarification Question.

Figure 1 shows mean confidence scores per context kind across discourse relations. Confidence scores within a speaker both across and within turns received similar confidence ratings ($\beta = -0.13$, $z = -0.56$, $p =$ n.s.[2]), while annotators were significantly more confident for relation annotation across speakers ($\beta = 0.63$, $z = 3.05$, $p < .01$). The CLMM revealed that annotators used confidence scores between 3 and 5 overall, except for the label "Other", for which they selected lower confidence scores. Background received lower confidence scores overall. Continuation, Contrast and Narration received higher scores for contexts

[2]Not statistically significant.

within speaker. Comment and Result received higher scores for turns across speakers and single turn. For Elaboration and Explanation, mean confidence scores are very similar across the three contexts, with slightly higher scores for single turn and pairs of turns within speaker. Acknowledgment, Clarification Question ("clarificationq") and Question-Answer Pair ("qap") received higher scores for turns across speakers, which makes sense given the dialogic nature of these relations. However, these relations also received rather high confidence scores for single turn and pairs of turns within speaker, which is a bit surprising. We suspect this might be due to the context turns included for each pair of DUs, which might have led annotators to choose relations between discourse units other than for the pair of highlighted DUs. Future analysis will look closer at this aspect.

4 Distributed Representations from Discourse Relation Annotations

To model the similarity between discourse relations as perceived by annotators, we computed embedding representations of discourse relations. We extracted each n individual annotation containing **r**elation-**c**onfidence (r, c) tuples selected by a given annotator for a pair of DUs. We concatenate bag-of-relation vectors with one-hot encoded features representing the dialogue context kind, and multiply the count vector of annotated relations (either 1 or 0 for each relation) by the confidence score (1-5) for that pair of DUs. This weighting learns more from high confidence; an ideal reweighting may be possible with additional parameter search, possibly in conjunction with the CLMM outputs.

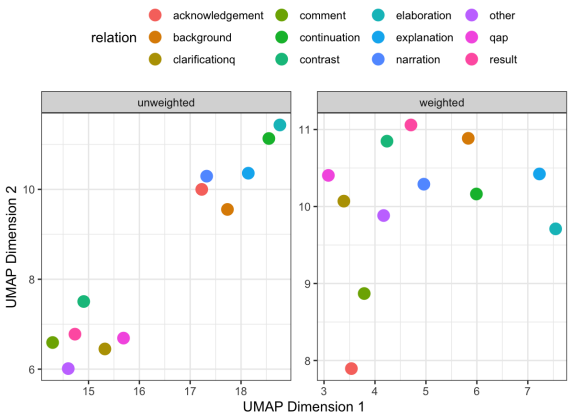
For an $n \times 1$ confidence ratings matrix C, an $n \times 12$ bag-of-relations matrix R, an $n \times 3$ discourse context matrix D for each annotation, we obtain an annotation matrix $A = C \times (R|D)$. We then obtain a square co-occurrence matrix O such that $O = A \cdot A^T$, which we factorize using Principal Component Analysis (without shifting the intercept following Levy and Goldberg, 2014). Each relation is thus represented as a vector that consolidates co-occurrences between all relations within a single annotator that are weighted by confidence score. We then projected these embeddings into two dimensions with UMAP (McInnes et al., 2018) and performed a hierarchical clustering analysis over the resulting coordinates due to the greater discriminability afforded by continuous distance metrics.

Informally, the UMAP coordinates appear more gradient in the representational space when confidence was included (right panel) than when it was not included (left panel). When context is not included, the UMAP coordinates primarily represent the frequency of labels in our corpus, which we include in Appendix A. We visualize the UMAP coordinates in Figure 2a. Figure 2b shows a dendrogram with the output clusters, colored according to the optimal number of clusters ($k = 2$), calculated using average silhouette widths (Levshina, 2022). There are two large clusters, one of which contains two sub-clusters with Background and Continuation, on the one hand, and Elaboration and Explanation on the other. In the other large cluster, Acknowledgement and Comment form a sub-cluster. These are very common relations between pairs of turns across speakers. Clarification Question and Question-Answer Pair form another sub-cluster, also common relations between pairs of turns across speakers, in close proximity to the Other label, which received a sub-cluster of its own. Narration and Contrast and Result, form the last sub-clusters, which we suspect is due in part to the frequencies of these relations (Schnabel et al., 2015). We include a dendrogram with the output clusters of a hierarchical clustering analysis performed with base bag-of-relations vectors (without context kind and confidence scores weight) in Figure 3 in Appendix B for comparison.

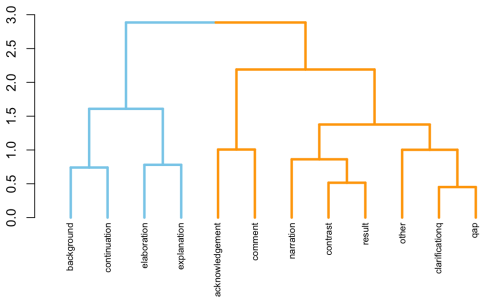
Currently, we provide these results as a proof of concept of the feasibility and interpretability of noisy labels produced by novice annotators. Importantly, annotations weighted by confidence produce coherent clusters of discourse relations. We envision applications of DR embeddings to several domains including dialogue generation, such that appropriate responses to input are partially conditioned on a latent or mixed combination of DRs.

5 Related Work

Annotation of discourse relations is usually done within Rhetorical Structure Theory (Mann and Thompson, 1987), as in the RST-DT (Carlson et al., 2003) and GUM (Zeldes, 2017) corpora, within Segmented Discourse Representation Theory (SDRT, Asher and Lascarides, 2003), as in the STAC (Asher et al., 2016) and Molweni (Li et al., 2020) corpora, or within the Penn Discourse Treebank framework (Prasad et al., 2008, 2014, 2018).



(a) The coordinates obtained with UMAP for all discourse relations plotted in two-dimensional space. The plot on the left shows the unweighted embedding representations and the figure on the right shows the weighted embedding representations.



(b) Dendrogram showing hierarchical clustering of Discourse Relations built from UMAP coordinates. *qap* stands for Question-Answer Pair and *clarificationq* for Clarification Question.

Figure 2: Dimensionality reduction and clustering of relation embeddings.

We use a taxonomy adapted from SDRT, in particular, the STAC corpus.

Annotators are usually trained to identify discourse relations using the framework's taxonomy. Some recent alternatives to explicitly collecting annotation of DRs include crowdsourcing by eliciting connectives (Yung et al., 2019; Scholman et al., 2022) or question-answer pairs (Pyatkin et al., 2020) rather than relations. In this work, we wanted to investigate how annotators perceive discourse relation categories, and therefore a connective insertion task would only provide indirect evidence. We train annotators on DR labeling and ask annotators to choose from a set of discourse relation labels. We allow for multiple labels to investigate what relations are more confusable or perceived as co-occurring (Marchal et al., 2022).

6 Discussion and Future Work

In this study, we collected multiple annotations of discourse relations from a subset of the Switchboard corpus, together with confidence scores. We found that dialogue context had a significant effect on confidence scores. We computed embedding representations of DRs using co-occurrence statistics and weighted the vectors using context type and confidence scores, and found that these representations coherently model our annotators uncertainty about discourse relation labels.

Discourse units that occur across turns as defined by Switchboard do not necessarily occur across continuous utterances from the speaker's point-of-view. Obtaining information about whether same-speaker pairs of discourse units fall into the same or different utterances may help to explain additional variance in annotator confidence.

Additionally, in this work, we investigated annotators' confidence on the annotation of adjacent turns. In future work, we plan to annotate discourse relations across longer-distance discourse units and to allow for hierarchical annotation. We expect that annotation confidence will also vary across longer-distance units and across different depths of annotation.

In the future, we plan to use this information to run a larger scale annotation study of the Switchboard corpus to analyze discourse relation patterns in spoken dialogues.

Limitations

This work is limited by the size of the dataset and the taxonomy used in the annotation task. While we found that our annotators perceived some of the categories as more similar or confusable, future work can examine annotators' uncertainty in a larger set of discourse relations. The selection of DUs for annotation was also non-exhaustive. In future work, we plan to expand the selection procedure so that we include more distantly related DUs. We also note that the frequency of discourse relation labels and individual differences in confidence levels among annotators may bias the representations. We plan to look into these potential biases in future work.

Ethics Statement

We are not aware of ethical issues associated with the texts used in this work. Students participated in the annotation task as part of course credit but annotation decisions were not associated with their performance in the course.

Acknowledgements

We would like to thank Jürgen Bohnemeyer and three anonymous reviewers for feedback on a previous version of this paper. We also thank the students who participated in the annotation task.

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Nicholas Asher, Vladimir Popescu, Philippe Muller, Stergos Afantenos, Anais Cadilhac, Farah Benamara, Laure Vieu, and Pascal Denis. 2012. Manual for the analysis of settlers data. *Strategic Conversation (STAC). Université Paul Sabatier*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA. IEEE Computer Society.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. "spaCy: Industrial-strength Natural Language Processing in Python".

David M. Howcroft and Verena Rieser. 2021. What happens if you treat ordinal ratings as interval data? human evaluations in NLP are even more underpowered than you think. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8932–8939, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

T Florian Jaeger and Neal E Snider. 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, 127(1):57–83.

Natalia Levshina. 2022. Semantic maps of causation: New hybrid approaches based on corpora and grammar descriptions. *Zeitschrift für Sprachwissenschaft*, 41(1):179–205.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Torrin M Liddell and John K Kruschke. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79:328–348.

S. Magalí López Cortez and Cassandra L. Jacobs. 2023. The distribution of discourse relations within and across turns in spontaneous conversation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 156–162, Toronto, Canada. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse Treebank, Comparable Corpora, and Complementary Annotation. *Computational Linguistics*, 40(4):921–950.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

David Reitter and Johanna D Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. DiscoGeM: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3281–3290, Marseille, France. European Language Resources Association.

Wilbert Spooren and Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2):241–266.

Bonnie Webber. 2013. What excludes an alternative in coherence relations? In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 276–287, Potsdam, Germany. Association for Computational Linguistics.

Frances Yung, Kaveri Anuranjana, Merel Scholman, and Vera Demberg. 2022. Label distributions help implicit discourse relation classification. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 48–53, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25, Florence, Italy. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

A Frequencies of Discourse Relation Labels

Discourse Relation	Count
Elaboration	636
Continuation	554
Acknowledgement	494
Explanation	383
Comment	265
Background	252
Narration	249
Question-Answer Pair	248
Contrast	191
Clarification Question	179
Result	124
Other	106

Table 2: Raw counts of discourse relation labels in our corpus from most to least frequent.

B Clustering without Context and Confidence Weighting

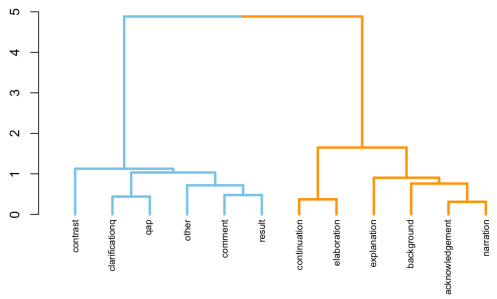


Figure 3: Dendrogram showing hierarchical clustering of Discourse Relations built from UMAP coordinates without context kind and confidence scores weighting. *qap* stands for Question-Answer Pair and *clarificationq* for Clarification Question. The two main clusters align with the two-dimensional plot of the unweighted UMAP coordinates in Figure 2a

C Annotation Interface

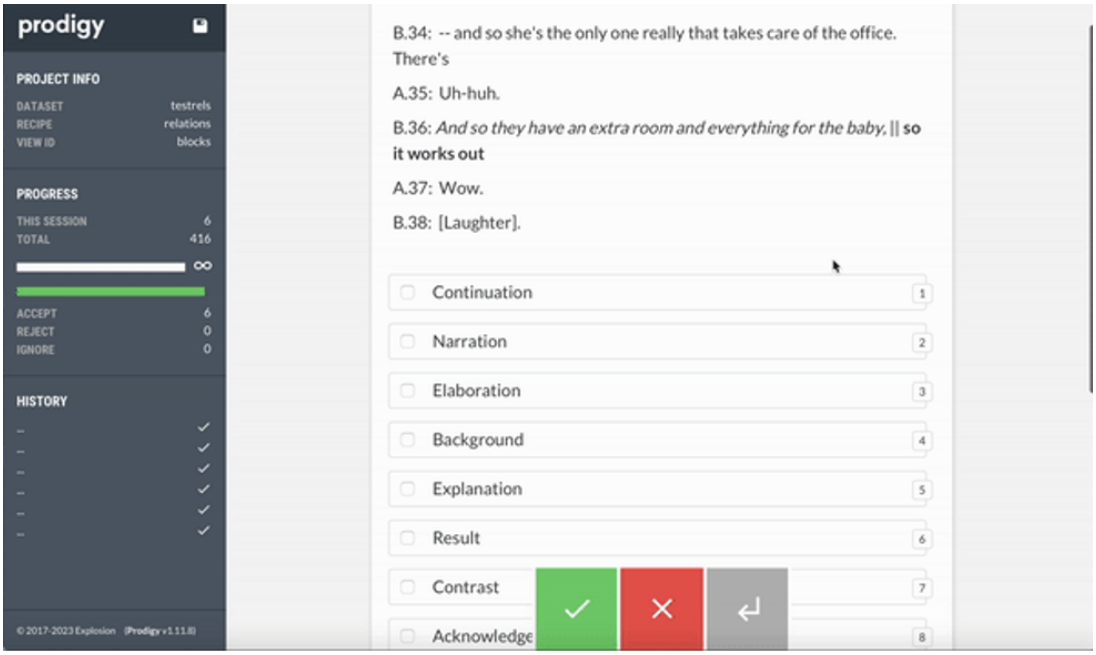


Figure 4: Example annotation task. EDUs to be annotated and discourse relations.

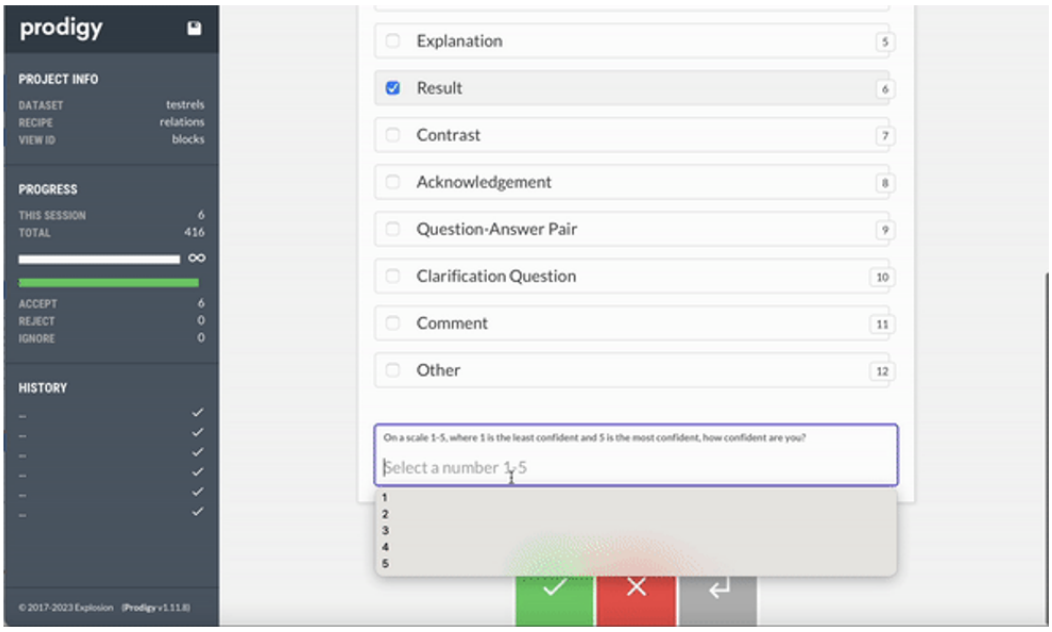


Figure 5: Example annotation task. Discourse relations and confidence score.

Investigating the Representation of Open Domain Dialogue Context for Transformer Models

Vishakh Padmakumar[1]* **Behnam Hedayatnia**[2] **Di Jin**[2] **Patrick Lange**[2]
Seokhwan Kim[2] **Nanyun Peng**[23] **Yang Liu**[2] **Dilek Hakkani-Tur**[2]

[1]New York University, [2]Amazon Alexa AI, [3]University of California, Los Angeles

vishakh@nyu.edu
{behnam, djinamzn, patlange, seokhwk, yangliud}@amazon.com
violetpeng@cs.ucla.edu dilek@ieee.org

Abstract

The bulk of work adapting transformer models to open-domain dialogue represents dialogue context as the concatenated set of turns in natural language. However, it is unclear if this is the best approach. In this work, we investigate this question by means of an empirical controlled experiment varying the dialogue context format from text-only formats (all recent utterances, summaries, selected utterances) as well as variants that are more structurally different (triples, AMR). We compare these formats based on fine-tuned model performance on two downstream tasks—knowledge selection and response generation. We find that simply concatenating the utterances works as a strong baseline in most cases, but is outperformed in longer contexts by a hybrid approach of combining a summary of the context with recent utterances. Through empirical analysis, our work highlights the need to examine the format of context representation and offers recommendations on adapting general-purpose language models to dialogue tasks.

1 Introduction

The bulk of existing work in adapting transformer models to open-domain dialogue represents the dialogue context as the concatenated set of turns in natural language (Zhang et al., 2019b; Roller et al., 2021; Shuster et al., 2022). While the self-attention mechanisms of these models are able to capture the context from these flat representations, it remains unclear if this is the best approach (Li et al., 2021). Studying the format of context representation would help improve performance on downstream tasks such as response generation and external knowledge selection and could also potentially inform the pretraining of general-purpose dialogue models. Additionally, as the length of conversations increases (Gopalakrishnan et al., 2019; Xu, 2021), these are truncated based on the limit imposed by the positional encodings on transformers. We also know that not all of the utterances are equally relevant so succinctly representing the relevant information in the context given the current conversation state and filtering out the noise from prior interactions would help to model provide more coherent responses.

In this work, we empirically investigate the dialogue context representation in the text space for using sequence-to-sequence models. To prioritize broad coverage, we vary the the format of the context using both natural language-only formats (e.g., using all recent utterances or summaries) as well as formats that are more structurally different (e.g., extracting knowledge triples from the utterances) (Section 2) and compare these based on downstream task performance.

We find that concatenating all recent utterances is a strong baseline. However, in longer dialogues, combining recent utterances with a summary of the past context obtains the best performance. This shows the benefit of the complementary long and short view of dialogue context. We also observe that improving summary quality and introducing external elements about the coherence of the context result in a further gain of downstream performance. This study and related findings can be extended to combine with elements from the broader definition of context (Bunt, 1999), such as social cues and guidelines (Gupta et al., 2022b), which were previously not included in dialogue datasets.

2 Approach

We study the effect of the representation of dialogue context on downstream dialogue tasks—knowledge selection and response generation. In order to do so, we run a controlled experiment fine-tuning sequence-to-sequence models on the two tasks verbalised into the text-to-text setup, while varying only the format in which the dialogue con-

* Work done during summer internship at Amazon Alexa AI.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 538–547
September 11–15, 2023. ©2023 Association for Computational Linguistics

text is represented.

The first broad category of representations consists of directly using the dialogue utterances. We include the concatenated past dialogue utterances, truncated when necessary, as ***Plaintext*** representation. This includes all the past turns delineated using a special token when applicable. We also include ***Windows*** of recent turns where we only use the most recent n utterances as the context.

To test if models require only the knowledge items within the dialogue utterances, we extract (subject, object, relation) ***Triples*** from the utterances as the context. To see if models benefit from more structured information, we convert the utterances into ***AMR*** graphs (Banarescu et al., 2013).

Finally, we examine if the information from the context can be distilled using summarization (Feng et al., 2021; Gliwa et al., 2019a; Khalifa et al., 2021). One method is to convert the utterances from both speakers into an abstractive ***Summary*** using a separate summarization model.[1] And while a summary might contain all the required high-level information from the dialogue context, it loses the local discourse-level information from recent utterances. To mitigate this, we create a hybrid ***Summary + Utterances*** format by appending the *Summary* with *Windows of Turns*. We also include an extractive summary in the form of ***Selected Turns*** from the context using pointwise mutual information, a proxy for relevance, with respect to the most recent turn (Padmakumar and He, 2021).

We provide further implementation details about each of the methods in Appendix C and illustrate an example converted to each of them in Figure 1.

3 Experiments

3.1 Datasets and Metrics

Knowledge Selection To evaluate performance on knowledge selection, we report results on the Wizard of Wikipedia (WoW) (Dinan et al., 2018) dataset, which consists of dialogue between a wizard (expert) and apprentice (novice) where the wizard selects knowledge items (sentences) to form a response. In the sequence-to-sequence setup, we frame this as a classification task on individual knowledge items as follows.

Input: <context> </s> <knowledge item>
Output: `"Relevant"` for the gold knowledge item given that context, and `"Not Relevant'` otherwise.

In addition to all the context formats from Section 2, we include another baseline called ***Plaintext with Documents*** where the gold documents that were used to generate previous wizard turns were appended to the utterances in the dialogue context.

Metrics: We report accuracy/F1-score of each label in lieu of instance-based classification performance. To report retrieval performance, we score the individual knowledge items for a particular context using the token probabilities assigned to `"Relevant"` and select the most relevant item. We then evaluate if this matches the checked sentence from the dataset, akin to Recall@1 when this is framed as a retrieval problem. We also report a more relaxed metric that evaluates if this item is from the checked document from the dataset.

Response Generation We report results on WoW, Multi-Session Chat (MSC) (Xu et al., 2021) and Topical Chat (TCS) (Gopalakrishnan et al., 2019) where the objective is to generate the gold response given the context. For WoW, the task is a knowledge-grounded dialogue where the responses were formed using the gold knowledge item from the dataset. The task for TCS is also knowledge-grounded response generation, but not all turns are accompanied by relevant knowledge items. For MSC, the task is for the partners to converse about their own interests and discuss information about each others' interests across multiple sessions. We concatenate utterances from all past sessions with a special token indicating a session break.[2]
Input: <context> </s> <optional knowledge item>
Output: Gold response from the dataset.

Metrics We report perplexity of the gold utterances w.r.t. the finetuned models and the BertScore (Zhang et al., 2019a) between the generated response and the target utterance.

3.2 Model Training

For each of the datasets, we convert all of the train examples into the different context representations from Section 2 and report finetuned T5 (Raffel et al., 2020) performance. We use the T5-base ($220M$ parameters) and Large ($770M$ parameters) variants. While the models trained in Zhang et al. (2019b); Peng et al. (2022) have examined further

[1] In particular, we use a BART-large model finetuned on SAMSum(Gliwa et al., 2019b).

[2] For MSC, the *Summary* baseline(s) use the released summaries for past sessions coupled with a model generated summary for the utterances in the current session.

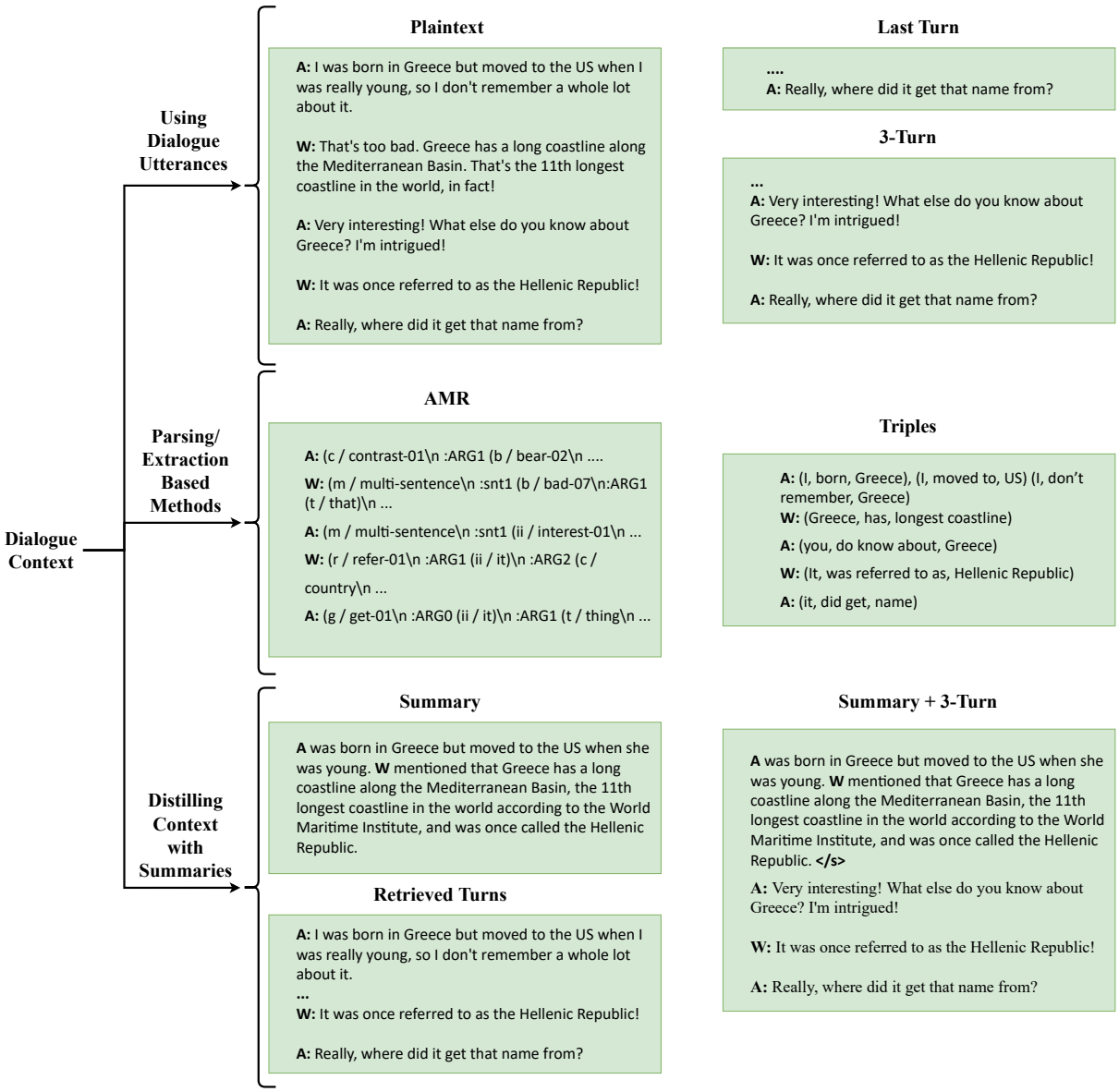


Figure 1: Example illustrating the conversion a dialogue into all the context representation methods evaluated in our experiments. The original set of utterances indicated by *Plaintext*. We perform an empirical controlled experiment evaluating the fine-tuned dialogue model performance on each of these context representation formats.

pretraining on dialogue, this would bias the model to additionally favor the *Plaintext* baseline. As a result, we choose T5, noting that absolute performance might improve further by adding dialogue-specific pertaining. When tokenizing the context, we allow for up to 1024 tokens and truncate earlier utterances in case of an overflow. We optimize cross-entropy loss on the output tokens in the desired format based on the dataset. We run finetuning for 10 epochs with an early stopping criteria based on validation loss. For each context representation, we select the best learning rate sweeping from $1e^{-3}$ to $1e^{-6}$. In the text-to-text setup, we run inference with greedy decoding kept uniform across the representations. Our experiments were run on a `p3.8xlarge` and a `p3.16xlarge` EC2 instances containing 4 and 8 Tesla V100 GPUs respectively.

4 Results

Table 1 and Table 2 show the results comparing context representation formats on knowledge selection and response generation respectively.

***Plaintext* is a strong baseline, which is outperformed by *Summaries+Utterances* on longer dialogues** From Table 1 and Table 2, we see that the *Plaintext* representation provides a strong baseline

		Plaintext	Plaintext w Docs	Last Turn	3-Turn	Selected Turns	AMR	Triples	Summary	Summ + 1 Turn	Summ + 3 Turn	Summ + 5 Turn
Accuracy	**Overall**	0.959 / 0.963	0.958 / 0.960	0.960 / 0.962	0.960 / 0.963	**0.965 / 0.966**	0.961 / 0.965	0.961 / 0.963	0.963 / 0.964	0.958 / 0.960	0.954 / 0.958	0.957 / 0.961
	Relevant	0.331 / 0.265	0.355 / 0.289	0.278 / 0.234	0.307 / 0.261	0.282 / 0.244	0.265 / 0.264	0.268 / 0.263	0.286 / 0.231	0.301 / 0.253	**0.369 / 0.297**	0.353 / 0.281
F1 Scores	**Relevant**	0.196 / 0.170	**0.202 / 0.188**	0.169 / 0.150	0.184 / 0.165	0.192 / 0.172	0.160 / 0.158	0.166 / 0.163	0.174 / 0.155	0.183 / 0.159	0.191 / 0.167	0.194 / 0.170
Recall@1 of Most Relevant Item	**Checked Sentence**	0.159 / 0.116	**0.171 / 0.129**	0.114 / 0.111	0.120 /0.105	0.138 / 0.118	0.097 / 0.085	0.101 / 0.086	0.116 / 0.099	0.128 / 0.111	0.143 / 0.118	0.147 / 0.116
	Checked Passage	0.238 / 0.174	**0.265 / 0.201**	0.186 / 0.165	0.165 / 0.146	0.214 / 0.178	0.138 / 0.124	0.140 / 0.126	0.160 / 0.150	0.199 / 0.177	0.234 / 0.191	0.222 . 0.185

Table 1: Evaluation of context representation methods on WoW knowledge selection. Each cell has two numbers corresponding to results on the random split (left) and topic split (right) of the validation set. All metrics are rounded off to three decimal places and the highest in each row is bold. We include only the overall accuracy and classification metrics of the `Relevant` label here. For metrics on all labels see Table 7 in Appendix E.

		Plaintext	Last Turn	3 Turn	5 Turn	Selected Turns	AMR	Triples	Summary	Summ + 1 Turn	Summ + 3 Turns	Summ + 5 Turns
WoW	**Bertscore**	**0.905 / 0.904**	0.903 / 0.901	0.903 / 0.902	0.904 / 0.902	0.903 / 0.900	0.895 / 0.890	0.898 / 0.894	0.902 / 0.900	0.903 / 0.901	0.905 / 0.903	0.904 / 0.903
	Perplexity	**6.978 / 7.545**	7.446 / 8.084	7.398 / 8.011	7.304 / 7.885	7.177 / 7.783	7.987 / 8.623	7.803 / 8.510	7.477 / 8.115	7.261 / 7.836	7.050 / 7.660	7.028 / 7.601
MSC	**Bertscore**	**0.873**	0.861	0.864	0.872	0.865	0.854	0.858	0.866	0.869	0.871	**0.873**
	Perplexity	12.246	15.262	14.701	14.024	14.565	16.245	15.782	13.985	13.69	13.011	**12.205**
TCS	**Bertscore**	0.871 / 0.868	0.869 / 0.867	0.870 / 0.869	**0.871 / 0.869**	0.869 / 0.869	0.865 / 0.864	0.866 / 0.865	0.868 / 0.866	0.869 / 0.867	0.870 / 0.868	0.871 / 0.868
	Perplexity	12.313 / 14.443	13.293 / 15.950	13.045 / 15.650	12.847 / 15.023	12.778 / 15.237	13.587 / 16.290	13.402 / 16.117	12.899 / 15.262	12.686 / 15.013	12.538 / 14.812	**12.181 / 14.342**

Table 2: Evaluation of context representation methods on response generation. For WoW, each cell has two numbers corresponding to results on the random split and topic split of the validation set. For MSC, we report results on all the turns of the validation set. For TCS, the two numbers correspond to the frequent and rare splits respectively. All metrics are rounded off to three decimal places and the highest in each row is bold.

for both knowledge selection and response generation. When we examine the *Last Turn* and *3-Turn* columns, we see the trend that increasing the window size predictably improves performance, but these lag behind *Plaintext*. This shows that transformers are able to leverage the additional information from more recent utterances in the context. However, we see that *Plaintext* is outperformed by the *Summary + 5-turn* method on the longer dialogue datasets, MSC and TCS. This shows that past the limit imposed on current transformer encoders by the positional embeddings, summarizing all available information outperforms a truncated set of recent utterances. Finally, we see that *Summary + 5-turn* outperforms *Summary* alone on all the datasets. These findings highlight the complementary *Long* and *Short* views of dialogue context from summaries and recent utterances respectively.

Improving the quality of summaries results in better downstream performance To observe the effect of summary quality, we point out two comparisons. On MSC, we compare the response generation performance using both the gold human-written summaries and model-generated summaries (released with the dataset). The perplexity for response generation reduces by using higher quality, human-written summaries (Table 5). Secondly, we can view the *Selected Turns* baseline as an extractive summary of the dialogue context that consistently outperforms windows of text of the same number of turns (here *Selected Turns* and *3-turn* are comparable). Combined with the observation of the complementary nature of summaries and recent turns, a future direction highlighted through our work is to use downstream task performance as a means to evaluate dialog summarization.

Natural language-based approaches outperform the more structure-oriented variants We observe that *AMR* and *Triples* are consistently outperformed by all the other utterance-based and summary-based variants. This is potentially explained by the higher similarity of the natural language formats to the pretraining data of sequence-to-sequence models.[3]

Positive Scaling Trends One of the main advantage of using sequence-to-sequence transformers is that as pretrained models get better, we can expect improved performance in downstream tasks. We observe a simple version of this when comparing results on the different context representation methods with T5-base and T5-large in Table 3 and

[3]These methods are at a disadvantage in the text-to-text format and could be improved by different methods of encoding the extracted information.

		Plaintext		Last Turn		Retrieved Turns		Summ + 3		Summ + 5	
		Base	Large	Base	Large	Base	Large	Base	Large	Base	Large
F1 Scores	Relevant	0.196 / 0.170	0.187 / 0.170	0.169 / 0.150	0.177 / 0.159	0.192 / 0.172	0.210 / 0.185	0.191 / 0.167	**0.212 / 0.189**	0.194 / 0.170	0.205 / 0.181
Recall@1 of Most Relevant Item	Match to 'Checked Sentence'	0.159 / 0.116	**0.203 / 0.156**	0.114 / 0.111	0.131 / 0.110	0.138 / 0.118	0.163 / 0.135	0.143 / 0.118	0.161 / 0.135	0.147 / 0.116	0.160 / 0.131
	Match to 'Checked Passage'	0.238 / 0.174	**0.326 / 0.258**	0.186 / 0.165	0.191 / 0.162	0.214 / 0.178	0.285 / 0.231	0.234 / 0.191	0.255 / 0.219	0.222 . 0.185	0.252 / 0.199

Table 3: Evaluation of knowledge selection as a function of model size—T5-Base vs Large for 5 different context representations. We largely observe positive scaling trends on both retrieval metrics and classification F1-scores. Table 9 in Appendix E shows the same table with metrics for all labels.

		Plaintext		Retrieved Turns		Last Turn		Summ + 5 Turns	
		Base	Large	Base	Large	Base	Large	Base	Large
WoW	Perplexity	6.978 / 7.545	**5.989 / 6.371**	7.177 / 7.783	6.151 / 6.574	7.446 / 8.084	6.754 / 7.226	7.028 / 7.601	6.001 / 6.412
TCS	Perplexity	12.313 / 14.443	9.811 / 11.279	12.778 / 15.237	10.101 / 12.980	13.293 / 15.950	11.456 / 14.374	12.181 / 14.342	**9.792 / 11.113**

Table 4: Evaluation of response generation as a function of model size—T5-Base vs Large for 4 different context representations. We observe positive scaling trends across each of the representations

Table 4. Performance improves using the scaled up model uniformly for response generation and on retrieval metrics in knowledge selection.

Providing additional content as part of the context improves performance Augmenting the *Plaintext* baseline with document level information for WoW results in further improvement in both classification and retrieval scores. In this work, we only considered the utterances in the dialogue itself to be a part of the context. However a broader definition of context for dialogue includes not just the turns but also discourse information, social context, or the relationship between the speakers, and even physical context, or cues from the relative physical positions and actions of the speakers (Bunt, 1999). Our work indicates that a promising future direction of dialogue research could involve collecting and summarizing all this additional rich information to be used by dialogue models.

We present additional results in Appendix E and discuss some limitations that inform future directions in Appendix A.

5 Related Work

When adapting transformers to dialogue tasks, the most common approach is to simply concatenate dialogue utterances (Zhang et al., 2019b; Adiwardana et al., 2020; Roller et al., 2021; Bao et al., 2021; Gupta et al., 2022a; Shuster et al., 2022). For longer dialog datasets where the entire conversation cannot be encoded, summaries of past sessions are a helpful way to provide all the relevant information needed to continue the conversation (Xu et al., 2022). While *AMR* graphs have been used to perturb individual utterances in order to evaluate coherence in dialogue (Ghazarian et al., 2022), to the best of our knowledge, *AMR* and *Knowledge Triples* have not been used to represent the context. We include them for wider coverage. In the dialogue space, retrieval has largely been used to identify relevant knowledge items to be included for response generation (Shuster et al., 2021). Prior work has examined matching candidate responses with multiple utterances for selection, the weighting learned in effect attending to 'relevant' turns (Wu et al., 2016; Zhang et al., 2018), however, we explicitly select turns as a means of representing the dialogue context across both of our open domain dialogue tasks. To our knowledge, ours is the first controlled experiment to evaluate different textual context representation methods for sequence-to-sequence models.

6 Conclusion

In this work, we present an empirical controlled study examining dialogue context representation for transformer models on open-domain dialogue tasks. While concatenating all previous turns, as is often adopted, is a strong baseline, combining summaries of the overarching context with recent utterances yields the best results in longer dialogues. Additionally improving the quality of the summaries being used and introducing further background information into the context further improve performance. This provides us with new directions to work on including dialogue summarization and considering the broader definition of context for use in open-domain dialogue.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525, Online. Association for Computational Linguistics.

Harry Bunt. 1999. Context representation for dialogue management. In *Proceedings of the Second International and Interdisciplinary Conference on Modeling and Using Context*, CONTEXT '99, page 77–90, Berlin, Heidelberg. Springer-Verlag.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019a. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019b. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P Bigham. 2022a. Improving zero and few-shot generalization in dialogue through instruction tuning. *arXiv preprint arXiv:2205.12673*.

Prakhar Gupta, Yang Liu, Di Jin, Behnam Hedayatnia, Spandana Gella, Sijia Liu, Patrick Lange, Julia Hirschberg, and Dilek Hakkani-Tur. 2022b. Dialguide: Aligning dialogue model behavior with developer guidelines. *arXiv preprint arXiv:2212.10557*.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen McKeown. 2021. A bag of tricks for dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8014–8022, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 128–138, Online. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Vishakh Padmakumar and He He. 2021. Unsupervised extractive summarization using pointwise mutual information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2505–2512, Online. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of*

the European Chapter of the Association for Computational Linguistics: Main Volume, pages 300–325, Online. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2016. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. 2021. Beyond preserved accuracy: Evaluating loyalty and robustness of BERT compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10653–10659, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Yang Xu. 2021. Global divergence and local convergence of utterance semantic representations in dialogue. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 116–124, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems*, 31.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

A Limitations

Coverage of Context Representations We acknowledge that the list of context representation formats we examine is non-exhaustive and each particular context format could be further optimized. For instance, for *AMR* we only cover the semantic representations within a single utterance. There are other types of structural aspects in dialogues like discourses, turn-taking, and so on which could be incorporated. We report results comparing these in order to inform subsequent model training/pre-training as well as subsequent analysis of a similar nature.

Using Only Verbalized Representations In this work, we only cover context representation formats that are verbalized in natural language. It is unclear if encoding the information either into a specialized dialogue transformer architecture or as a graph would result in improved performance. We choose the verbalized format as it is the most general purpose which can be used to adapt many different language models (Liang et al., 2022).

Adapting Retrieval Tasks for Text-to-Text Models Adapting language models to retrieval tasks such as knowledge selection can be done either by running inference on individual examples or by combining all candidates along with the input context. Liang et al. (2022) perform a comparison of these variants in the few-shot setting for a large number of tasks and observe no clear winning format so we proceed with separate inference on knowledge items. Here, to isolate the role of the context representation, we fix the format of the task and study the effect of dialogue context on performance

Evaluation We acknowledge that we report performance using automatic metrics on a single run for both sets of tasks and human evaluation would allow for a more holistic understanding of the capabilities of models, particularly on response generation. Human evaluation and running multiple sets of fine-tuning runs for each of the different formats would be expensive. In this work, we restricted ourselves to the same in order to focus on comparing and identifying trends in performance between a wider range of different context representation formats.

B Potential Risks

Our work discusses ways to adapt sequence-to-sequence transformer models for open-domain dialogue. The main associated risk comes from the black-box nature of these models. The text that is generated is pretty heavily influenced by the pretraining data. The models fine-tuned in this paper are open-sourced T5 checkpoints which may contain biases from the C4 (Raffel et al., 2020) corpus. Additionally, the advent of closed-access and limited-access language models such as GPT3 and Anthropic-LM comes with more uncertainty as the pretraining and training processes of these models are not as well documented (Liang et al., 2022).

C Context Representations

Context Representation Formats We vary the context in the following ways in an attempt to ensure coverage of different formats. The first broad category of representations consists of directly using dialogue utterances.

- **Plaintext**: The simplest, and most widely used, manner in which we can represent the dialogue context is just the concatenated set of past dialogue utterances. This includes all turns from past sessions delineated using a special token when applicable. [4]

- **Windows of Turns**: Here we only use the most recent n utterances as the context. As we increase n, we provide more local context about the dialogue.

Aside from including the utterances themselves, to evaluate if models benefit from more structured information we include the following representations:

- **AMR**: We convert each utterance into an AMR graph (Banarescu et al., 2013) and use the verbalised form as the context. The AMR parses the text into a directed acyclic graph, explicitly conveying the relationships as edges between the various concept nodes in the text. We use the `model_parse_xfm_bart_large` model from amrlib to convert the utterances into the corresponding AMR. We acknowledge that performance in our experiments could be affected by the quality of AMR conversions. We refer readers to the original library for performance benchmarking of the text-to-AMR model.

- **Knowledge Triples**: To test if models require only the knowledge items within the dialogue context, and not the whole utterance, we extract `(subject, object, relation)` triples from the utterances as the context. We use OpenIE5 to extract triples and use a simple unigram overlap heuristic to filter out duplicates. If two triples have a unigram overlap of over 0.7, only one is selected.

Finally, we examine if the information from the dialogue context can be distilled while retaining the natural language format using summarization.

- **Summary**: We summarise all of the dialogue utterances from both speakers abstractively using a finetuned transformer model. In particular, we use a BART-large model finetuned on SAMSum(Gliwa et al., 2019b). As indicated in Section 4, performance depends on the quality of the summarization model. This model was not trained by the authors of this work. We refer readers to the model card on HuggingFace for evaluation of the model itself.

- **Summary + Utterances**: While a summary might contain all the high-level information from the dialogue context, it loses the local discourse-level information from recent utterances which provide cues on how to use the high-level information. We create this hybrid short+long form context representation by appending the *Summary* with *Windows of Turns*.

- **Retrieved Turns**: While the aforementioned setups contain abstractive summaries of the dialogue context, we also include an extractive summary generated by selecting relevant turns using pointwise mutual information to the most recent turn (Padmakumar and He, 2021). In order to select relevant turns, we calculate the PMI of all utterances with respect to the *Last Turn* and combine the 2 most relevant turns, in order to obtain an extractive summary of the context.

An example converted to each of the above formats is provided in Figure 1.

[4] Dataset specific details are provided in Section 3.1

D Details for Responsibility Checklist

D.1 License and Usage of Scientific Artifacts

The Wizard of Wikipedia (Dinan et al., 2018) and MSC (Xu et al., 2021) datasets made available through ParlAI that is shared under the MIT License which permits usage of the data for research such as our work. Topical Chat (Gopalakrishnan et al., 2019) is shared using the Community Data License Agreement - Sharing, Version 1.0 which also permits the usage of the data in this manner. These datasets are commonly used in the community and are collected while ensuring that it was properly anonymized and does not contain any offensive language. We do not perform additional checks for either of the same. T5 (Raffel et al., 2020), used for all our finetuning experiments, is released under the Apache 2.0 license which permits its use for research. The model used for dialogue summarization and `amrlib` are both shared under the MIT license which permits such usage as does OpenIE which is shared under the Open IE 5 Software License Agreement. All of the artifacts, both models and datasets, were used as intended by the original authors.

D.2 Coverage and Statistics of the Data

All of the datasets contain only English data, largely collected from American English speakers conversing in a one-on-one conversation. The specifics of the settings where the conversations are collected are well documented and can be referred to in the original works (Dinan et al., 2018; Xu et al., 2021; Gopalakrishnan et al., 2019). Wizard of Wikipedia consists of $18,430$ documents ($166,787$ utterances total, $74,092$ of which were wizard turns used in knowledge selection) in the train set. The results were reported on the random split (981 documents, $3,939$ wizard turns) and topic split (967 documents, $3,927$ wizard turns) of the validation data. For MSC, there are 4000 train conversations (spread across multiple sessions) with $161,440$ turns and we report results on the validation set (1001 conversations, $53,332$ turns). In Topical Chat, there are 8628 train conversations consisting of 188378 utterances and we report results on the frequent (539 conversations, 11681 turns) and rare (539 conversations, 11692 turns) splits of the validation data.

		Human Written Summary	Model Generated Summary
All Turns	Perplexity	**12.129**	12.205
First Response in Session	Perplexity	**10.199**	10.257

Table 5: Performance on MSC improves when using the gold, human-written summaries as opposed to model-generated summaries.

	Truncated Examples	Examples wo Truncation
Perplexity	14.381	12.564
Bertscore	0.8641	0.8722

Table 6: Response generation performance on MSC examples adapted into the *Plaintext* representation and divided based on whether these are truncated.

E Additional Results

We report a more comprehensive version of the knowledge selection results from Table 1 in Table 7 and response generation from Table 2 in Table 8.

Effect of Scaling Model Size Table 9 and Table 10 contain the full comparison of results when we switch from T5-Base to T5-Large.

Quality of Summaries In order to ablate the quality of summaries used, we compared response generation performance on the MSC dataset, comparing the *Summary + 5-Turn* baseline when the gold, human-written summaries are used as opposed to the model generated summaries released in the original dataset. From Table 5 we observe that the higher quality summaries result in further improvement in performance.

Effect Of Truncation Here we aim to empirically verify that truncation of context has an adverse effect on model performance. We select those examples in the second session of the MSC dataset when adapted using the *Plaintext* representation and divide these into whether or not the context was truncated. This particular set of examples was chosen because, out of all the sessions, this was the one which had a relatively large fraction of examples in both of these buckets—27.6% of examples were truncated. From Table 6 we clearly see that those examples which suffer from truncation have a drop in performance.

		Plaintext	Plaintext w Docs	Last Turn	3-Turn	5-Turn	Selected Turns	AMR	Triples	Summary	Summ + 3 Turn	Summ + 5 Turn
Item Classification Accuracy	**Overall**	0.959 / 0.963	0.958 / 0.960	0.960 / 0.962	0.960 / 0.963	0.960 / 0.962	**0.965 / 0.966**	0.961 / 0.965	0.961 / 0.963	0.963 / 0.964	0.954 / 0.958	0.957 / 0.961
	NR	0.969 / 0.973	0.965 / 0.966	0.970 / 0.972	0.970 / 0.972	0.969 / 0.972	**0.975 / 0.976**	0.973 / **0.976**	0.970 / 0.972	0.974 / 0.975	0.963 / 0.967	0.966 / 0.971
	R	0.331 / 0.265	0.355 / 0.289	0.278 / 0.234	0.307 / 0.261	0.318 / 0.277	0.282 / 0.244	0.265 / 0.264	0.268 / 0.263	0.286 / 0.231	**0.369 / 0.297**	0.353 / 0.281
Item Classification F1 Scores	**NR**	0.979 / 0.981	0.977 / 0.979	0.979 / 0.981	0.980 / 0.981	0.979 / 0.980	**0.982 / 0.982**	0.977 / 0.980	0.978 / 0.980	0.981 / **0.982**	0.977 / 0.978	0.978 / 0.980
	R	0.196 / 0.170	**0.202 / 0.188**	0.169 / 0.150	0.184 / 0.165	0.187 / 0.171	0.192 / 0.172	0.160 / 0.158	0.166 / 0.163	0.174 / 0.155	0.191 / 0.167	0.194 / 0.170
Recall@1 of Most Relevant Item	**Match to 'Checked Sentence'**	0.159 / 0.116	**0.171 / 0.129**	0.114 / 0.111	0.120 /0.105	0.127 / 0.106	0.138 / 0.118	0.097 / 0.085	0.101 / 0.086	0.116 / 0.099	0.143 / 0.118	0.147 / 0.116
	Match to 'Checked Passage'	0.238 / 0.174	**0.265 / 0.201**	0.186 / 0.165	0.165 / 0.146	0.179 / 0.153	0.214 / 0.178	0.138 / 0.124	0.140 / 0.126	0.160 / 0.150	0.234 / 0.191	0.222 . 0.185

Table 7: Evaluation of context representation methods on knowledge selection. Each cell has two numbers corresponding to results on the random split and topic split of the validation set. All metrics are rounded off to three decimal places and the highest in each row is bold.

			Plaintext	Last Turn	3 Turn	5 Turn	Selected Turns	AMR	Triples	Summary	Summ + 1 Turn	Summ + 3 Turns	Summ + 5 Turns
WoW		**Bertscore**	**0.905 / 0.904**	0.903 / 0.901	0.903 / 0.902	0.904 / 0.902	0.903 / 0.900	0.895 / 0.890	0.898 / 0.894	0.902 / 0.900	0.903 / 0.901	0.905 / 0.903	0.904 / 0.903
		Perplexity	**6.978 / 7.545**	7.446 / 8.084	7.398 / 8.011	7.304 / 7.885	7.177 / 7.783	7.987 / 8.623	7.803 / 8.510	7.477 / 8.115	7.261 / 7.836	7.050 / 7.660	7.028 / 7.601
MSC	***All***	**Bertscore**	**0.873**	0.861	0.864	0.872	0.865	0.854	0.858	0.866	0.869	0.871	**0.873**
		Perplexity	12.246	15.262	14.701	14.024	14.565	16.245	15.782	13.985	13.69	13.011	**12.205**
	1st	**Bertscore**	0.875	0.868	0.862	0.863	0.863	0.859	0.863	**0.876**	0.873	0.874	0.875
		Perplexity	10.386	15.627	15.118	13.998	14.409	16.109	15.704	**10.143**	10.988	10.876	10.257
TCS		**Bertscore**	0.871 / 0.868	0.869 / 0.867	0.870 / 0.869	**0.871 / 0.869**	0.869 / 0.869	0.865 / 0.864	0.866 / 0.865	0.868 / 0.866	0.869 / 0.867	0.870 / 0.868	0.871 / 0.868
		Perplexity	12.313 / 14.443	13.293 / 15.950	13.045 / 15.650	12.847 / 15.023	12.778 / 15.237	13.587 / 16.290	13.402 / 16.117	12.899 / 15.262	12.686 / 15.013	12.538 / 14.812	**12.181 / 14.342**

Table 8: Evaluation of context representation methods on response generation. For WoW, each cell has two numbers corresponding to results on the random split and topic split of the validation set. For MSC, we report results on all the turns (*All*), and for the first turn in each session (*1st*). For TCS, the two numbers correspond to the frequent and rare splits respectively. All metrics are rounded off to three decimal places and the highest in each row is bold.

		Plaintext		Last Turn		Retrieved Turns		Summ + 3		Summ + 5	
		Base	**Large**	**Base**	**Large**	**Base**	**Large**	**Base**	**Large**	**Base**	**Large**
Item Classification Accuracy	**Overall**	0.959 / 0.963	0.944 / 0.950	0.960 / 0.962	**0.967 / 0.968**	0.965 / 0.966	0.964 / 0.965	0.954 / 0.958	0.964 / 0.965	0.957 / 0.961	0.957 / 0.961
	NR	0.969 / 0.973	0.951 / 0.954	0.970 / 0.972	**0.977 / 0.979**	0.975 / 0.976	0.973 / 0.976	0.963 / 0.967	0.973 / 0.975	0.966 / 0.971	0.965 / 0.970
	R	0.331 / 0.265	**0.445 / 0.402**	0.278 / 0.234	0.245 / 0.212	0.282 / 0.244	0.329 / 0.275	0.369 / 0.297	0.333 / 0.255	0.353 / 0.281	0.382 / 0.305
Item Classification F1 Scores	**NR**	0.979 / 0.981	0.971 / 0.968	0.979 / 0.981	**0.983 / 0.984**	0.982 / 0.982	0.981 / 0.982	0.977 / 0.978	0.981 / 0.982	0.978 / 0.980	0.978 / 0.978
	R	0.196 / 0.170	0.187 / 0.170	0.169 / 0.150	0.177 / 0.159	0.192 / 0.172	0.210 / 0.185	0.191 / 0.167	**0.212 / 0.189**	0.194 / 0.170	0.205 / 0.181
Recall@1 of Most Relevant Item	**Match to 'Checked Sentence'**	0.159 / 0.116	**0.203 / 0.156**	0.114 / 0.111	0.131 / 0.110	0.138 / 0.118	0.163 / 0.135	0.143 / 0.118	0.161 / 0.135	0.147 / 0.116	0.160 / 0.131
	Match to 'Checked Passage'	0.238 / 0.174	**0.326 / 0.258**	0.186 / 0.165	0.191 / 0.162	0.214 / 0.178	0.285 / 0.231	0.234 / 0.191	0.255 / 0.219	0.222 . 0.185	0.252 / 0.199

Table 9: Evaluation of knowledge selection as a function of model size. We report performance on T5-Base and Large for 5 different context representations. We observe positive scaling trends, where the larger model performs better, uniformly for retrieval metrics and generally across the classification metrics for the *Relevant* label.

		Plaintext		Retrieved Turns		Last Turn		Summ + 5 Turns	
		Base	**Large**	**Base**	**Large**	**Base**	**Large**	**Base**	**Large**
WoW	**Bertscore**	0.905 / 0.904	**0.907 / 0.906**	0.903 / 0.900	0.904 / 0.902	0.903 / 0.901	0.904 / 0.902	0.904 / 0.903	0.906 / 0.904
	Perplexity	6.978 / 7.545	**5.989 / 6.371**	7.177 / 7.783	6.151 / 6.574	7.446 / 8.084	6.754 / 7.226	7.028 / 7.601	6.001 / 6.412
TCS	**Bertscore**	0.871 / 0.869	0.873 / 0.872	0.869 / 0.869	0.872 / 0.871	0.869 / 0.867	0.871 / 0.870	0.871 / 0.868	**0.874 / 0.873**
	Perplexity	12.313 / 14.443	9.811 / 11.279	12.778 / 15.237	10.101 / 12.980	13.293 / 15.950	11.456 / 14.374	12.181 / 14.342	**9.792 / 11.113**

Table 10: Evaluation of response generation as a function of model size. We report performance on T5-Base and Large for 4 different context representations. We observe positive scaling trends, where the larger model performs better particularly on perplexity scores.

C^3: Compositional Counterfactual Contrastive Learning for Video-grounded Dialogues

Hung Le
Salesforce Research Asia
hungle@salesforce.com

Nancy F. Chen
A*STAR, Institute for Infocomm Research
nfychen@i2r.a-star.edu.sg

Steven C.H. Hoi
Salesforce Research Asia
shoi@salesforce.com

Abstract

Video-grounded dialogue systems aim to integrate video understanding and dialogue understanding to generate responses that are relevant to both the dialogue and video context. Most existing approaches employ deep learning models and have achieved remarkable performance, given the relatively small datasets available. However, the results are partially accomplished by exploiting biases in the datasets rather than developing multimodal reasoning, resulting in limited generalization. In this paper, we propose a novel approach of Compositional Counterfactual Contrastive Learning (C^3) to develop contrastive training between factual and counterfactual samples in video-grounded dialogues. Specifically, we design factual/counterfactual samples based on the temporal steps in videos and tokens in dialogues and propose contrastive loss functions that exploit object-level or action-level variance. Different from prior approaches, we focus on contrastive hidden state representations among compositional output tokens to optimize the representation space in a generation setting. We achieved promising performance gains on the Audio-Visual Scene-Aware Dialogues (AVSD) benchmark and showed the benefits of our approach in grounding video and dialogue context.

1 Introduction

Visual dialogue research (Das et al., 2017; Seo et al., 2017; De Vries et al., 2017; Chattopadhyay et al., 2017; Alamri et al., 2019a) aims to develop intelligent systems that can reason and answer questions about visual content in a multi-turn setting. Compared to traditional visual question answering (VQA) (Antol et al., 2015; Gao et al., 2015; Malinowski and Fritz, 2014; Zhu et al., 2016), visual dialogues bridge the gap between research and practical applications by allowing turn-based human-machine interactions. Recently, many deep learning approaches have been proposed to develop visual dialogue systems and achieved remarkable performance (Schwartz et al., 2019; Hori et al., 2019; Le et al., 2019; Li et al., 2021b). However, as these methods are heavily trained on relatively small datasets (Das et al., 2017; Alamri et al., 2019a), they are subject to inherent bias from the datasets and limited generalization into real-world applications (Zhang et al., 2016; Goyal et al., 2017). While training on large-scale data can alleviate this problem, visual dialogues are expensive to procure and require manual annotations. This challenge becomes more obvious in highly complex visual dialogue tasks such as video-grounded dialogues (Alamri et al., 2019a; Le et al., 2021) (Figure 1).

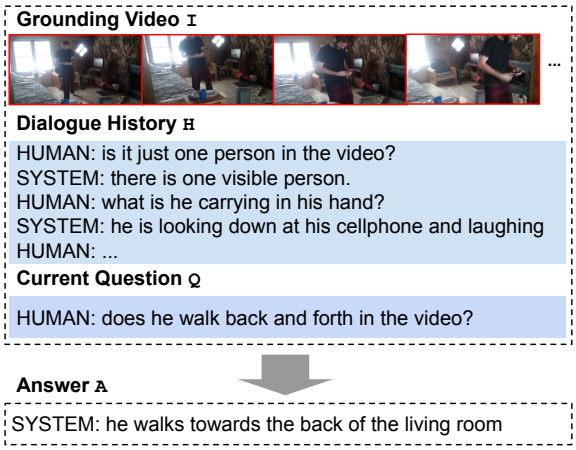


Figure 1: An example of video-grounded dialogue

In recent years, we have seen increasing research efforts in contrastive learning to improve deep learning performance (Wu et al., 2018; Henaff, 2020; Chen et al., 2020; He et al., 2020). The common strategy of these methods is an objective function that pulls together representations of an anchor and "positive" samples while pushing the representations of the anchor from "negative" samples. These methods are specifically beneficial in self-supervised image representation learning. Specifically, these methods often do not require additional annotations by augmenting data of existing samples

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 548–561
September 11–15, 2023. ©2023 Association for Computational Linguistics

to create "positive" and "negative" samples. We are motivated by this line of research to improve visual dialogue systems and propose a framework of **C**ompositional **C**ounterfactual **C**ontrastive Learning (C^3). C^3 includes loss functions that exploit contrastive training samples of factual and counterfactual data that are augmented to be object-variant or action-variant.

Compared to traditional deep learning tasks, a major challenge of applying contrastive learning (Wu et al., 2018; Henaff, 2020; Chen et al., 2020; He et al., 2020) in video-grounded dialogues lies in the complexity of the task. Specifically, in a discrimination task of image classification, given an image, positive samples are created based on non-adversarial transformations on this image e.g. by cropping inessential parts without changing the labels, and negative samples are randomly sampled from other image instances. However, such transformations are not straightforward to apply on visual dialogues, each of which consists of a video of spatio-temporal dimensions, a dialogue of multiple turns, and an output label in the form of natural language at the sentence level. In visual dialogues, the random sampling method, in which negative samples are created by swapping the input video and/or dialogue context with random components from other training samples, becomes too naive. In domains with high data variance like dialogues or videos, a system can easily discriminate between such positive and negative instances derived using previous approaches.

To mitigate the limitations of conventional contrastive learning in video-grounded dialogues, we propose a principled approach to generate and control negative and positive pairs by incorporating compositionality and causality (an overview of our approach can be seen in Figure 2 and 3). Specifically, we develop a structural causal model for visual dialogues by decomposing model components by object and action-based aspects. We then create hard negative samples of grounding videos by masking temporal steps that are relevant to actions mentioned in target output responses. Hard negative dialogue samples are created by masking tokens that are referenced to the entity mentioned in target output responses. Positive samples of videos and dialogues are developed similarly by masking irrelevant temporal steps or tokens for them to remain factual. Finally, based on an object or action-based variance between factual and counterfactual pairs, we only select specific hidden state representations of the target dialogue response sequence, to apply contrastive loss functions. Compared to existing approaches, our method has better control of data contrast at the granularity of object and action variance. We conducted experiments with comprehensive ablation analysis using the Audio-Visual Scene-Aware Dialogues (AVSD) benchmark (Alamri et al., 2019a) and showed that our method can achieve promising performance gains.

2 Related Work

Counterfactual Reasoning. Related to our work is the research of counterfactual reasoning. One line of research focuses on generating plausible counterfactual data to facilitate model training or evaluation. (Zmigrod et al., 2019; Garg et al., 2019; Vig et al., 2020) introduced data augmentation methods that convert gender-inflected sentences or remove identity-based tokens from sentences. The augmented data is used to study model stereotyping and improve fairness in model outputs. (Kaushik et al., 2020) crowd-sourced human annotations to minimally revise documents such that their sentiment labels are flipped. (Zeng et al., 2020; Wang and Culotta, 2020; Madaan et al., 2020) introduced data augmentation to improve model robustness in entity recognition and text classification tasks.

More related to our work are counterfactual augmentation methods in generative tasks. (Qin et al., 2019) introduced a new benchmark for counterfactual story rewriting. (Li et al., 2021a) explored augmented counterfactual dialogue goals to evaluate dialogue state tracking models. (Baradel et al., 2020) proposed a synthetic 3D environment for learning the physical dynamics of objects in counterfactual scenarios. Different from prior tasks, in the task of video-grounded dialogue, a target response is not easy to be flipped/negated, and hence, supervised learning is not straightforward. We propose to automatically develop counterfactual and factual samples and improve representation learning via unsupervised learning.

Contrastive Learning. Our work is related to the research of contrastive learning in deep learning models. The research is particularly popular in self-supervised learning of image representations (Wu et al., 2018; Hjelm et al., 2019; Henaff, 2020; Chen et al., 2020; He et al., 2020; Khosla et al., 2020). These methods do not require additional annotations but aim to improve representations

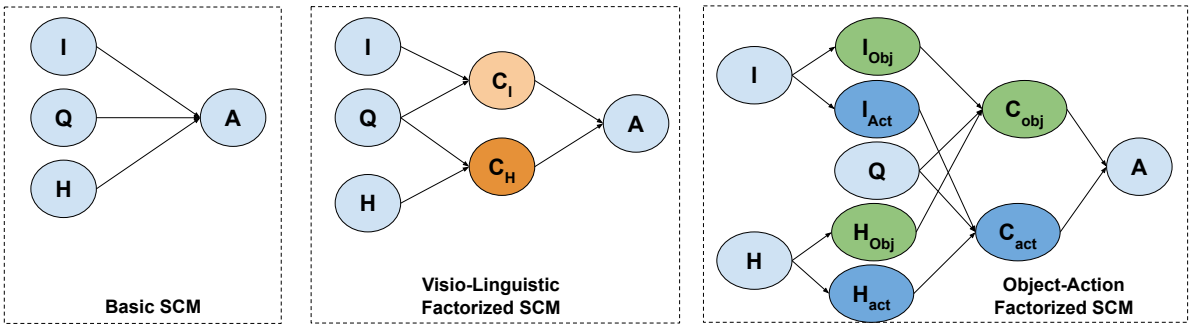


Figure 2: **SCMs of video-grounded dialogues:** Left: Basic SCM without factorization. Middle: SCM factorized by visual and textual context. Right: SCM factorized by object and action-level information. I: video input, Q: question input, H: dialogue history, C: contextualized information, and A: target response. For simplicity, we do not demonstrate independent noise variables U and the subscript t.

through loss functions. The loss functions are often inspired by noise contrastive estimation (NCE) (Gutmann and Hyvärinen, 2010) and applied in lower-dimensional representation space. In the language domain, similar loss functions have been introduced to improve word embeddings (Mnih and Kavukcuoglu, 2013) and sentence embeddings (Logeswaran and Lee, 2018). More related to our work is (Huang et al., 2018; Liu and Sun, 2015; Yang et al., 2019; Lee et al., 2021), introducing positive and negative pairs of sentences for contrastive learning in generative tasks such as language modelling, word alignment, and machine translation. In the multimodal research domains, our work is related to contrastive learning methods introduced by (Zhang et al., 2020; Gokhale et al., 2020; Liang et al., 2020; Gupta et al., 2020). Specifically, our work complements (Zhang et al., 2020) by incorporating causality into contrastive learning. However, we focus on a very different task of video-grounded dialogues that involves turn-based question-answering. The task requires multimodal reasoning performed on both dialogue context and video context. Moreover, we improve models by tightly controlling data variance by adopting compositionality and our loss functions optimize hidden state representations of decoding tokens by their object or action-based semantics.

3 Method

3.1 Problem Definition

In a video-grounded dialogue task (Alamri et al., 2019a; Le et al., 2021), the inputs consist of a dialogue $\mathcal{D}$ and the visual input of a video $\mathcal{I}$. Each dialogue contains a sequence of dialogue turns, each of which is a pair of question $\mathcal{Q}$ and answer $\mathcal{A}$. At each dialogue turn t, we denote the dialogue context $\mathcal{H}_t$ as all previous dialogue turns $\mathcal{H}_t = \{(\mathcal{Q}_i, \mathcal{A}_i)\}|_{i=1}^{i=t-1}$. The output is the answer $\hat{\mathcal{A}}_t$ to answer the question of the current turn $\mathcal{Q}_t$. The objective of the task is the generation objective that output answers of the current turn:

$$\hat{\mathcal{A}}_t = \arg\max_{\mathcal{A}_t} P(\mathcal{A}_t|\mathcal{I}, \mathcal{H}_t, \mathcal{Q}_t; \theta) \quad (1)$$

3.2 Structural Causal Model

We first cast a visual dialogue model as a structural causal model (SCM) (Pearl, 2009) to explore the potential factors that affect the generation of target dialogue responses in a dialogue system. By definition, an SCM consists of random variables $V = \{V_1, ..., V_N\}$ and corresponding independent noise variables $U = \{U_1, ..., U_N\}$. We assume an SCM of a directed acyclic graph (DAG) structure. In this structure, causal functions are defined as $F = \{f_1, ..., f_N\}$ such that $V_i = f_i(P_i, U_i)$ where $P_i = \{V_p\} \subset V$ are the parent nodes of V_i in the DAG. Using this definition of SCM, we develop three SCM structures for a video-grounded dialogue system in Figure 2.

The *Basic SCM* is directly derived from the objective function (1). The *VL-SCM* adopts a question-aware reasoning process that partitions visual and language reasoning based on question information as the common cause. A limitation of VL-SCM is that it does not account for the interactions of components such as object and action abstracts that are embedded in visual context C_I and linguistic context C_H. This drawback becomes more significant in scenarios in which question information is highly dependent on prior turns in the dialogue history. Specifically, in questions that involve references, including object refer-

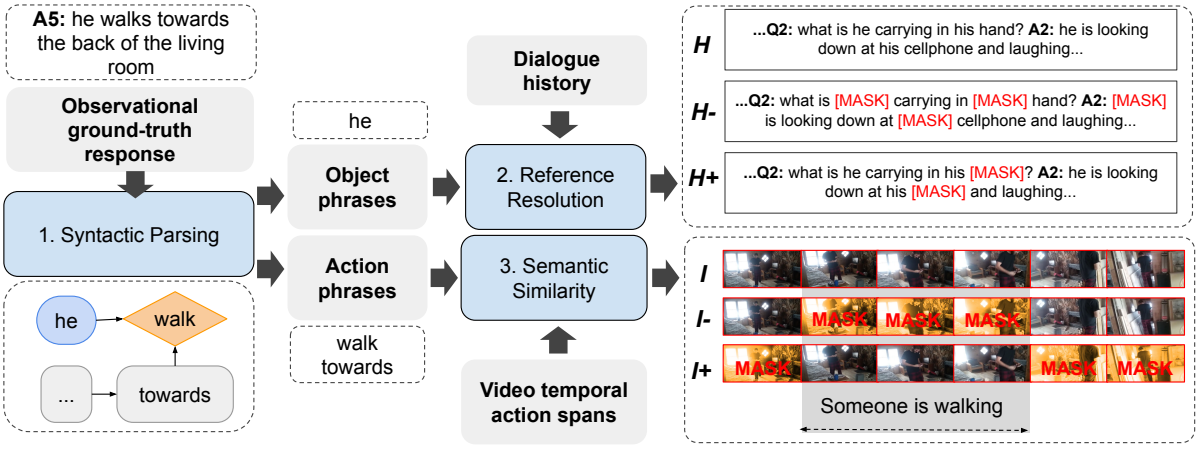


Figure 3: **Counterfactual generation:** An overview of our factual and counterfactual dialogue/video generation.

ences ("does *she* interact with *the woman in red*?") and action references ("what does the boy do after *that*?"), VL-SCM is not optimal to integrate dialogue and video context to solve component references such as "she" and "that". To address this drawback, we propose an *OA-SCM* that is factorized by object-action contextual information (Figure 2, right). The causal functions f_H^{obj} and f_H^{act} can be a simple text parser that map tokens into object-based tokens or action-based tokens s.t. $\mathcal{H}_{obj} = f_H^{obj}(\mathcal{H})$ and $\mathcal{H}_{act} = f_H^{act}(\mathcal{H})$. Similarly, f_I^{obj} and f_I^{act} are causal functions that map bounding boxes or temporal steps into object-based or action-based contents. In Section 3.3, we show that OA-SCM structure provides a framework to develop *partially counterfactual* training samples.

3.3 Counterfactual Augmentation

An overview of our augmentation process can be seen in Figure 3.

Decomposing observational target response. First, at each dialogue turn t, the ground-truth dialogue response $\mathcal{A}_t$ are passed to a syntactic parser such as the Stanford parser system [1]. The output includes grammatical components, such as subjects, verbs, and modifiers, in the form of a dependency tree. We prune the dependency tree to remove inessential parts and extract a set of object phrases $\mathcal{A}_{t,obj}$, and action-based phrases $\mathcal{A}_{t,act}$.

Generating counterfactual dialogue. Based on $\mathcal{A}_{t,obj}$, we apply a pretrained reference resolution model e.g. (Clark and Manning, 2016), to the dialogue context $\mathcal{H}_t$ to identify any references from past dialogue turns to any objects in $\mathcal{A}_{t,obj}$. For instance, in Figure 2, the object "he" identified in $\mathcal{A}_t$ are mapped to different token positions in prior dialogue turns, e.g. "his" in the text span "his hand" in the second question turn. All referenced tokens in dialogue context $\mathcal{H}_t$ are replaced by a `MASK` vector and the resulting dialogue context is denoted as counterfactual sample $\mathcal{H}_t^-$. We also used the pretrained reference resolution model to select any object tokens in $\mathcal{H}_t$ that are not mapped to $\mathcal{A}_{t,obj}$. These objects are considered irrelevant to $\mathcal{A}_t$ and they are replaced by the `MASK` vector from $\mathcal{H}_t$ and the resulting dialogue is denoted as a factual sample $\mathcal{H}_t^+$.

Generating counterfactual video. To create a counterfactual video sample, we first identify the temporal steps from the video that are semantically relevant to action phrases in $\mathcal{A}_{t,act}$. We obtain the annotation of temporal action spans from video, which can be retrieved from a pretrained temporal localization model (Shou et al., 2016) or is readily available in existing video benchmarks (Sigurdsson et al., 2016). The action span annotations consist of a set of action labels $Y_{i,act}$, each of which is mapped to a start and end time (t_i^s, t_i^e). Temporal segments that are deemed necessary to generate $\mathcal{A}_t$ is the union of all time spans from the set $S = \{(t_j^s, t_j^e)\}$ for all $Y_{j,act}$ that is semantically similar to $\mathcal{A}_{t,act}$. To identify similar pairs, we adopted cosine similarity scores between pretrained Glove embedding vectors of $Y_{j,act}$ and $\mathcal{A}_{t,act}$. During video feature encoding, any features of temporal steps sampled within S are replaced with a `MASK` vector, and resulting video features are noted as encoded features of counterfactual video $\mathcal{I}^-$. Factual video $\mathcal{I}^+$ are created similarly but for video parts irrelevant to $\mathcal{A}_t$, that is $I \setminus S$.

[1] https://nlp.stanford.edu/software/lex-parser.shtml

By the definition of OA-SCM from Section 3.2, we can denote $\mathcal{H}_t^- = \mathcal{H}_{t,obj}^- + \mathcal{H}_{t,act}$ and $\mathcal{H}_t^+ = \mathcal{H}_{t,obj}^+ + \mathcal{H}_{t,act}$; and $\mathcal{I}^- = \mathcal{I}_{obj} + \mathcal{I}_{act}^-$ and $\mathcal{I}^+ = \mathcal{I}_{obj} + \mathcal{I}_{act}^+$. Note that we follow (Hsieh et al., 2018) and assume object information such as object appearance and shape are typically embedded in any video frame. In this case, $\mathcal{I}_{obj}$ is unchanged and can be obtained from either $I \setminus S$ or S. In Section 3.4, we show that these partially counterfactual formulations enable a compositional contrastive learning approach.

3.4 Contrastive Learning

In this section, we introduce a contrastive learning method that exploits the compositional hidden states between factual and counterfactual samples. We extend the objective function (1) to express the auto-regressive decoding process:

$$\hat{\mathcal{A}}_t = \arg\max_{\mathcal{A}_t} P(\mathcal{A}_t|\mathcal{I}, \mathcal{H}_t, \mathcal{Q}_t; \theta)$$
$$= \arg\max_{\mathcal{A}_t} \prod_{m=1}^{L_\mathcal{A}} P_m(w_m|\mathcal{A}_{t,<m}, \mathcal{I}, \mathcal{H}_t, \mathcal{Q}_t; \theta)$$

Each target response $\mathcal{A}$ is represented as a sequence of token or word indices $\{w_m\}|_{m=1}^{m=L} \in |\mathbb{V}|$, where L is the sequence length and $\mathbb{V}$ is the vocabulary set. The conditional probability P_m is defined as:

$$P_m = \text{softmax}(Wk_m + b) \in \mathbb{R}^{|\mathbb{V}|} \quad (2)$$
$$k_m = \theta_{\text{decode}}(w_{m-1}, \theta_{\text{encode}}(\mathcal{I}, \mathcal{H}_t, \mathcal{Q}_t)) \quad (3)$$

where k_m is the hidden state at decoding position m and d is the embedding dimension of the hidden state. In this generative setting, we then explain 2 different ways of contrastive learning:

Sentence-level contrast. This approach learns the representations of the hidden states by contrasting a linear transformation of an aggregated vector of hidden states following an NCE framework:

$$\mathcal{L}_{\text{nce}}^{\text{sent}} = -\log \frac{e^{\text{sim}(z,z^+)}}{e^{\text{sim}(z,z^+)} + e^{\text{sim}(z,z^-)}} \quad (4)$$

where $\text{sim}(,)$ is the cosine similarity score and z is the output of an aggregation function Agg: $z = \text{Agg}(U)$ where $U \in \mathbb{R}^{d_{\text{nce}} \times L_\mathcal{A}}$ and $u_m = \text{MLP}_{\text{nce}}(k_m) \in \mathbb{R}^{d_{\text{nce}}}$. z^+ and z^- are obtained similarly by passing k_m^+ and k_m^- to the same MLP and aggregation function. k_m^+ and k_m^- are obtained by passing factual and counterfactual video pairs into (3): $k_m^+ = \theta_{\text{decode}}(w_{m-1}, \theta_{\text{encode}}(\mathcal{I}^+, \mathcal{H}_t, \mathcal{Q}_t))$ and $k_m^- = \theta_{\text{decode}}(w_{m-1}, \theta_{\text{encode}}(\mathcal{I}^-, \mathcal{H}_t, \mathcal{Q}_t))$. In cases of augmentation with factual and counterfactual dialogues, we obtain k_m^+ and k_m^- by replacing $\mathcal{H}$ with $\mathcal{H}^+$ and $\mathcal{H}^-$ in (3). Agg is an aggregation function that collapses hidden states into a single vector, e.g. average pooling (Lee et al., 2021; Zhang et al., 2020). We follow (Khosla et al., 2020) to normalize z, z^+, z^- to lie on the unit hypersphere. To reflect this contrastive learning approach against the VL-SCM, we can assume $\mathcal{C} \cong K$ and (4) essentially exploits the contrast between $\mathcal{C}^+$ and $\mathcal{C}^-$.

Compositional contrast. We note that the above approach does not consider compositionality in the target output response $\mathcal{A}$. Since we are using the same observational output w_{m-1} to obtain k_m, k_m^-, and k_m^-, we can remove the Agg function and apply a token-level pairwise contrastive loss between pairs of $(z_m = u_m, z_m^+ = u_m^+)$ and $(z_m = u_m, z_m^- = u_m^-)$. In this strategy, we formulate a loss function for action variance between $\mathcal{I}^+$ and $\mathcal{I}^-$, and one for object variance between $\mathcal{H}^+$ and $\mathcal{H}^-$:

$$\mathcal{L}_{\text{nce}}^{\text{act}} = -\frac{1}{|D_{act}|} \sum_{i \in D_{act}} \log \frac{e^{\text{sim}(z_i, z_i^+)}}{e^{\text{sim}(z_i, z_i^+)} + e^{\text{sim}(z_i, z_i^-)}}$$
$$D_{act} = \{\text{idx}(w_i) : w_{i-1} \in \mathcal{A}_{t,act}\} \quad (5)$$
$$\mathcal{L}_{\text{nce}}^{\text{obj}} = -\frac{1}{|D_{obj}|} \sum_{j \in D_{obj}} \log \frac{e^{\text{sim}(z_j, z_j^+)}}{e^{\text{sim}(z_j, z_j^+)} + e^{\text{sim}(z_j, z_j^-)}}$$
$$D_{obj} = \{\text{idx}(w_j) : w_{j-1} \in \mathcal{A}_{t,obj}\} \quad (6)$$

where $idx(w_m)$ returns the index of w_m in $\mathcal{A}_t$. Note that in (5) and (6), we adopt a hypothetical strategy by obtaining hidden states given *input* tokens are either in $\mathcal{A}_{t,act}$ or $\mathcal{A}_{t,obj}$. An alternative approach is to consider hidden states that are expected to produce *prospective* tokens $w_m \in \mathcal{A}_{t,act}/\mathcal{A}_{t,obj}$, i.e. $D'_{act} = \{\text{index}(w_i) : w_i \in \mathcal{A}_{t,act}\}$ and $D'_{obj} = \{\text{index}(w_j) : w_j \in \mathcal{A}_{t,obj}\}$. We conducted experiments with both strategies and explained our findings in the next section. Note that we can connect the compositional contrastive learning approach against the OA-SCM (Section 3.2) by denoting $\mathcal{C}_{act} \cong \{k_i\} \forall i \in D_{act}$ and $\mathcal{C}_{obj} \cong \{k_j\} \forall j \in D_{obj}$. Therefore, (5) essentially exploits the contrast between $\mathcal{C}_{act}^+$ and $\mathcal{C}_{act}^-$, and (6) for the contrast between $\mathcal{C}_{obj}^+$ and $\mathcal{C}_{obj}^-$.

4 Experiments

Dataset and Experimental Setup. We used the Audio-Visual Sene-Aware Dialogue (AVSD)

	Train	$\text{Train}_{\text{aug}}^{\text{video}}$	$\text{Train}_{\text{aug}}^{\text{dial}}$	Val	$\text{Val}_{\text{aug}}^{\text{video}}$	$\text{Val}_{\text{aug}}^{\text{dial}}$	Test
#Dialogs	7,659	7,145	6,411	1,787	1,709	1,557	1,710
$\#(\mathcal{I}, \mathcal{H}_t, \mathcal{Q}_t, \mathcal{A}_t)$	76,590	28,163	18,397	17,870	7,383	4,912	6,745

Table 1: Summary of the AVSD benchmark with augmented counterfactual video/dialogue data

dataset (Alamri et al., 2019b) to benchmark video-grounded dialogue systems. The dataset contains 10-turn dialogues, each of which is grounded on one video from the Charades dataset (Sigurdsson et al., 2016). We used the standard visual features I3D (Carreira and Zisserman, 2017) to represent the video input. Note that compared to (Alamri et al., 2019b), we followed the setting of AVSD in the 7^{th} Dialogue System Technology Challenge (DSTC7) (Yoshino et al., 2019), which requires generating a response rather than selecting from a candidate set. We also did not use video caption as an input as the caption is typically not easy to obtain in applications. A summary of the dataset can be seen in Table 1.

All model parameters, except the visual feature extractor of a pretrained I3D model, are initialized with uniform distribution (Glorot and Bengio, 2010). Our approach can be applied to different model architectures, as long as the hidden states of individual decoding tokens are available for contrastive learning. We used MTN (Le et al., 2019), which is a Transformer adaptation of the traditional RNN-based dialogue systems, as our base model. Finally, we evaluated models with objective metrics, including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015). These metrics are found to correlate well with human judgment (Alamri et al., 2019b).

Creating Counterfactual Data. We created counterfactual data for the training split and validation split of the AVSD benchmark. Specifically, from the original data, we identified invalid samples that are not sufficient for factual and counterfactual transformations. Examples of invalid samples are ones with ambiguous actions in target responses (e.g. "I am not sure what he is doing"), or ones without object references to prior turns (e.g. "there is only a single person in the video"). These samples are discarded and the remaining data is processed as described in Section 3.3. The overall statistics of augmented train and validation splits can be seen in Table 1. Note that the number of samples with augmented videos and dialogues are different as some samples contain valid actions but no object references (e.g. "the man is walking around the kitchen"), and vice versa.

Evaluating with Counterfactual Data. First, using augmented data, we evaluated models trained only with the original data. Motivated by (Kaushik et al., 2020; Vig et al., 2020; Agarwal et al., 2020), we designed this set of experiments to gauge the model performance under adversarial (counterfactual) samples and favourable (factual) samples and to observe the effects of our transformation methods. Specifically, we trained an MTN model (Le et al., 2019) on the original training data and evaluate the model on an augmented validation set. To fairly compare the results, we create a shared validation set in which each sample is augmented with both video and dialogue factual and counterfactual pairs. Essentially, this set is the intersection $\text{Val}_{\text{aug}}^{\text{v+d}} = \text{Val}_{\text{aug}}^{\text{video}} \cap \text{Val}_{\text{aug}}^{\text{dial}}$. Using the CIDEr metric (Vedantam et al., 2015), We noted the MTN model pretrained on original training data achieves 0.996 and 1.086 score in the original test and validation set respectively. However, as noted from Table 2, the performance drops to 0.779 when evaluating on the validation set $\text{Val}_{\text{aug}}^{\text{v+d}}$ even with the original video-dialogue pair $(\mathcal{I}, \mathcal{H})$. This performance drop indicates that the subset contains more challenging instances that require reasoning in dialogues and videos.

The performance decreases to 0.760 when tested with $\mathcal{I}^-$ and increases to 0.782 when tested with $\mathcal{I}^+$, keeping the $\mathcal{H}$ unchanged. When tested with videos that are masked at random temporal steps $\mathcal{I}^-_{\text{rand}}$, the result only reduces to 0.773, less than $\mathcal{I}^-$. This illustrates higher counterfactual impacts in $\mathcal{I}^-$ than in $\mathcal{I}^-_{\text{rand}}$. We also observed that model performance with counterfactual videos $\mathcal{I}^-$ is higher than cases with no video at all, $\mathcal{I}^0$. This observation demonstrates the factorization formulation of our SCM in which $\mathcal{I}^-$ is partially counterfactual, containing useful information, i.e. $\mathcal{I}_{obj}$, than $\mathcal{I}^0$, to support response generation.

When tested with dialogue transformations, we have similar observations with $\mathcal{H}^-$, $\mathcal{H}^+$, $\mathcal{H}^-_{\text{rand}}$, and $\mathcal{H}^0$. Specifically, following our SCM structure,

Video augmentation + original dialogue					Video augmentation + no dialogue				
$(\mathcal{I}, \mathcal{H})$	$(\mathcal{I}^-, \mathcal{H})$	$(\mathcal{I}^0, \mathcal{H})$	$(\mathcal{I}^+, \mathcal{H})$	$(\mathcal{I}^-_{\text{rand}}, \mathcal{H})$	$(\mathcal{I}, \mathcal{H}^0)$	$(\mathcal{I}^-, \mathcal{H}^0)$	$(\mathcal{I}^0, \mathcal{H}^0)$	$(\mathcal{I}^+, \mathcal{H}^0)$	$(\mathcal{I}^-_{\text{rand}}, \mathcal{H}^0)$
0.779	0.760	0.733	0.782	0.773	0.724	0.708	0.693	0.722	0.710
Dialogue augmentation + original video					Dialogue augmentation + no video				
$(\mathcal{I}, \mathcal{H})$	$(\mathcal{I}, \mathcal{H}^-)$	$((\mathcal{I}, \mathcal{H}^0)$	$(\mathcal{I}, \mathcal{H}^+)$	$(\mathcal{I}, \mathcal{H}^-_{\text{rand}})$	$(\mathcal{I}^0, \mathcal{H})$	$(\mathcal{I}^0, \mathcal{H}^-)$	$(\mathcal{I}^0, \mathcal{H}^0)$	$(\mathcal{I}^0, \mathcal{H}^+)$	$(\mathcal{I}^0, \mathcal{H}^-_{\text{rand}})$
0.779	0.764	0.724	0.788	0.778	0.733	0.722	0.693	0.739	0.734

Table 2: **Validation results with augmentation data:** $\mathcal{I}$: original video input, $\mathcal{I}^{-/+}$: counterfactual/factual video following Section 3.3, $\mathcal{I}^-_{\text{rand}}$: counterfactual video by masking random temporal steps, $\mathcal{I}^0$: no video input; $\mathcal{H}$: original dialogue input, $\mathcal{H}^{-/+}$: counterfactual/factual dialogue following Section 3.3, $\mathcal{H}^-_{\text{random}}$: counterfactual dialouge by masking random tokens, $\mathcal{H}^0$: no dialogue input. All results are in CIDEr score.

#	Contrast pair	Contrast loss	Hidden states	B-1	B-2	B-3	B-4	M	R	C
A	-	-	-	0.695	0.558	0.455	0.376	0.253	0.534	0.996
B	$\mathcal{I}^+, \mathcal{I}^-$	NCE	D_{act}	**0.709**	**0.577**	**0.476**	**0.398**	**0.262**	**0.549**	**1.040**
C	$\mathcal{I}^+, \mathcal{I}^-$	NCE	D'_{act}	0.697	0.565	0.462	0.381	0.254	0.538	1.003
D	$\mathcal{I}^+, \mathcal{I}^-$	NCE	D_{obj}	0.701	0.565	0.462	0.383	0.256	0.541	1.011
E	$\mathcal{I}^+, \mathcal{I}^-$	NCE	D	0.699	0.566	0.465	0.386	0.253	0.539	1.008
F	$\mathcal{I}^+, \mathcal{I}^-_{\text{rand}}$	NCE	D_{act}	0.693	0.563	0.464	0.388	0.254	0.538	1.010
G	$\mathcal{I}^+, \mathcal{I}^0$	NCE	D	0.700	0.566	0.463	0.383	0.256	0.538	1.019
H	$\mathcal{I}^+, \mathcal{I}^0_{\text{rand}}$	NCE	D	0.695	0.563	0.463	0.385	0.253	0.538	0.998
I	$\mathcal{I}^+, \mathcal{I}^-$	S-NCE	D_{act}	0.695	0.567	0.467	0.389	0.255	0.54	1.014
J	$\mathcal{I}^+, \mathcal{I}^-$	L1-PD	D_{obj}	0.705	0.569	0.465	0.385	0.258	0.543	1.005

Table 3: **Contrastive learning with counterfactual videos:** We experimented with variants of contrastive video pairs, hidden state sampling, and contrast loss. Metrics: B-n: BLEU-n, M: METEOR, R: ROUGE-L, C: CIDEr.

we show that $\mathcal{H}^-$ is partially counterfactual. To isolate the impacts of video/dialogue augmentations, we also tested models with tuples that are paired with zero dialogue context/video input ($\mathcal{H}_0$/$\mathcal{I}_0$). In these isolated experiments, we still observe consistent performance patterns among different variants of augmented video/dialogues, validating our factorization SCM and the effectiveness of augmentation techniques.

Contrastive Learning with Counterfactual Videos. In these experiments, we combined the task objective loss with our proposed contrastive learning approach that exploits action-based data contrast between $\mathcal{I}^+$ and $\mathcal{I}^-$. From Table 3, we have the following observations: **1)** First, when applying contrastive learning on augmented counterfactual data following our NCE function (5) (Row B), the model outperforms one which was trained with only original training data (Row A). This demonstrates the positive impacts of our C^3 learning approach through better generated target responses. **2)** When using the indices of hidden states based on prospective tokens (D'_{act}) (Row C), the performance gain decreases. This can be explained as hidden states in D'_{act} positions represent contextual information that *potentially*, but not absolutely, generate an action token. However, hidden states in D_{act} positions already assume a hypothetical input action token (w_{i-1} in (5)), and hence, a contrastive learning on these hidden states is more stable. **3)** we observed marginal performance gains when changing hidden state indices to indices of object tokens D_{obj} (Row D) or to hidden states of all tokens D (Row E). This observation verifies our factorized SCM framework as $\mathcal{I}^-$ and $\mathcal{I}^+$ are formulated to be action-variant specifically. Training them based on object variance or generic variance might lead to unstable representation learning and trivial performance gains.

4) Consistent with our observations from Table 2, contrastive learning applied to counterfactual videos with random masked temporal steps $\mathcal{I}^-_{\text{rand}}$ (Row F) results in very low performance gain. **5)** When we applied contrastive learning between $\mathcal{I}^+$ and naive counterfactual samples, including zero video input $\mathcal{I}^0$ (Row G) or video input sample from other training instance I^0_{rand} (Row H), the results only increases marginally compared to results with $\mathcal{I}^-$. **6)** We experimented with sentence-level contrast (S-NCE) (as in (4)) in which all hidden states are considered and collapsed to a single vector, as similarly used by (Lee et al., 2021; Zhang et al., 2020). We observed that this loss formulation (Row I), is not effective in our task, illustrating the benefits of using compositional representations of decoding tokens. **7)** Fi-

#	Contrast pair	Contrast loss	Hidden states	B-1	B-2	B-3	B-4	M	R	C
A	-	-	-	0.695	0.558	0.455	0.376	0.253	0.534	0.996
B	$\mathcal{H}^+, \mathcal{H}^-$	NCE	D_{obj}	**0.705**	**0.571**	**0.470**	**0.393**	**0.260**	**0.545**	**1.029**
C	$\mathcal{H}^+, \mathcal{H}^-$	NCE	D'_{obj}	0.701	0.569	0.469	0.392	0.256	0.540	1.023
D	$\mathcal{H}^+, \mathcal{H}^-$	NCE	D_{act}	0.699	0.561	0.453	0.369	0.251	0.538	0.963
E	$\mathcal{H}^+, \mathcal{H}^-$	NCE	D	0.707	**0.571**	0.466	0.385	0.258	0.542	1.020
F	$\mathcal{H}^+, \mathcal{H}^-_{\text{rand}}$	NCE	D_{obj}	0.693	0.557	0.452	0.370	0.253	0.536	0.957
G	$\mathcal{H}^+, \mathcal{H}^0$	NCE	D	**0.705**	0.570	0.466	0.387	0.258	0.542	1.022
H	$\mathcal{H}^+, \mathcal{H}^0_{\text{rand}}$	NCE	D	0.696	0.563	0.462	0.383	0.254	0.536	1.005
I	$\mathcal{H}^+, \mathcal{H}^-$	S-NCE	D_{obj}	0.696	0.561	0.458	0.378	0.252	0.538	0.999
J	$\mathcal{H}^+, \mathcal{H}^-$	L1-PD	D_{act}	0.699	0.569	0.468	0.390	0.255	0.543	1.008

Table 4: **Contrastive learning with counterfactual dialogues:** We experiment with variants of contrastive dialogues pairs, hidden state sampling, and loss. Metrics: B-n: BLEU-n, M: METEOR, R: ROUGE-L, C: CIDEr.

Model	Visual Features	B-1	B-2	B-3	B-4	M	R	C
Baseline (Hori et al., 2019)	I3D	0.621	0.480	0.379	0.305	0.217	0.481	0.733
JMAN (Chu et al., 2020)	I3D	0.648	0.499	0.390	0.309	0.240	0.520	0.890
FA-HRED (Nguyen et al., 2018)	I3D	0.648	0.505	0.399	0.323	0.231	0.510	0.843
Student-Teacher (Hori et al., 2019) †	I3D	0.675	0.543	0.446	0.371	0.248	0.527	0.966
MSTN (Lee et al., 2020) †	I3D	-	-	-	0.379	0.261	0.548	1.028
BiST (Le et al., 2020)	RX	**0.711**	**0.578**	0.475	0.394	0.261	**0.550**	1.050
RLM-GPT2 (Li et al., 2021b) † ‡	I3D	0.694	0.570	**0.476**	**0.402**	0.254	0.544	**1.052**
MTN (Le et al., 2019)	I3D	0.695	0.558	0.455	0.376	0.253	0.534	0.996
MTN $+C^3$ ($\mathcal{I}^{+/-}$)	I3D	0.709	0.577	**0.476**	0.398	**0.262**	0.549	1.040
MTN $+C^3$ ($\mathcal{H}^{+/-}$)	I3D	0.705	0.571	0.470	0.393	0.260	0.545	1.029

Table 5: **Overall results**: † incorporates additional video background audio inputs. ‡ indicates finetuning methods on pretrained language models. Metrics: B-n: BLEU-n, M: METEOR, R: ROUGE-L, C: CIDEr.

nally, to utilize any object-level invariance between $\mathcal{I}^+$ and $\mathcal{I}^-$, we applied a pairwise L1 distance loss $\mathcal{L}^{\text{act}} = \sum_i \|sim(z_i, z_i^+) - sim(z_i, z_i^-)\|_1$ to minimizes distances of hidden states of D_{obj} positions (Row J). However, the performance gain of this loss is not significant, demonstrating representation learning through data variance is a better strategy.

Contrastive Learning with Counterfactual Dialogues. From Table 4, we observed consistent observations as compared to prior experiments with counterfactual videos. Essentially, our results illustrate the impacts of C^3 that specifically contrasts object-level information between $\mathcal{H}^-$ and $\mathcal{H}^+$.

Overall Results. In Table 5, we reported the results of our models which we trained on an MTN backbone (Le et al., 2019) incorporated our proposed C^3 learning approach with counterfactual videos or dialogues. Our models achieve very competitive performance against models trained on the same data features e.g. MSTN (Lee et al., 2020), as well as models pretrained with a large language dataset e.g. RLM-GPT2 (Li et al., 2021b). We also observed that the performance gain of C^3 with $\mathcal{I}^{+/-}$ is higher than that with $\mathcal{H}^{+/-}$. As we showed the benefits of augmented counterfactual dialogues and videos, we will leave the study to unify both augmented data types for a hybrid contrastive learning approach for future work. In this paper, we showed that either dialogues or videos can be augmented and used to improve contextual representations through contrastive losses based on object-based or action-based variance.

For example factual/counterfactual videos/dialogues, please refer to the Appendix.

5 Discussion and Conclusion

In this work, we proposed Compositional Counterfactual Contrastive Learning (C^3), a contrastive learning framework to address the limitation of data in video-grounded dialogue systems. We introduced a factorized object-action structural causal model, described a temporal-based and token-based augmentation process, and formulated contrastive learning losses that exploit object-level and action-level variance between factual and counterfactual training samples. In our proposed approach, we train models to minimize the distance between compositional hidden state representations of factual samples and maximize the distance between counterfactual samples.

We noted our proposed C^3 still entails some limitations. we describe these limitations and suggest potential ways to overcome them for future extension. First, in our approach, we made the assumption of independence between $\mathcal{C}_{obj}$ and $\mathcal{C}_{act}$ to mask tokens/video segments as a way to generate counterfactual data samples. However, in many cases, this assumption might be too strong. Therefore, our approach might disrupt the natural data distribution and create negative noise in model training. A more advanced counterfactual data generation should be able to better capture the nature of counterfactual scenarios, avoiding the above assumption and generalizing the model better. Secondly, in our approach, we require external text-processing tools to decompose the input components. More sophisticated tools could be used to improve data quality of counterfactual/factual examples. Finally, after this work was completed, there have been several more advanced approaches following MTN (Le et al., 2019). As our approach is model-agnostic, we encourage readers to review and adapt our work to these more advanced models.

6 Broader Impacts

In this work, we described C^3, a novel contrastive learning approach that exploits action-based and object-based variance between counterfactual video/dialogue pairs. We demonstrated the benefit of this approach in the video-grounded dialogue domain, which is typically suffered from dataset scarcity. We want to emphasize that our method should be used strictly to improve dataset quality and obtain model performance gains. For instance, a chatbot that incorporates C^3 can generate high-quality responses that better match human questions. Our method should not be used for malicious purposes, such as creating chatbots to steal information or make scam calls.

Considering the widespread application of AI in the real world, the adoption of our method can lead to better dialogue systems that improve the quality of life for many people. For instance, a better chatbot embedded in electronic devices will improve both user experience and productivity. Conversely, the adoption of dialogue systems might lead to the potential loss of jobs in domains such as customer call centres. In high-risk domains such as autonomous vehicles, applications of our method can improve virtual assistant applications in the vehicles. As the products might directly affect human safety, any applications of C^3 should be tested to account for different scenarios, whether the method works as intended or not, and mitigate consequences when the output is incorrect. We advise that any plan to apply our method should consider carefully all potential groups of stakeholders as well as the risk profiles of applied domains to maximize the overall positive impacts.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019a. Audio visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Stefan Lee, Peter Anderson, Irfan Essa, Devi Parikh, Dhruv Batra, Anoop Cherian, Tim K. Marks, and Chiori Hori. 2019b. Audio-visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. 2020. Cophy: Counterfactual learning of physical dynamics. In *International Conference on Learning Representations*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Yun-Wei Chu, Kuan-Yen Lin, Chao-Chun Hsu, and Lun-Wei Ku. 2020. Multi-step joint-modality attention network for scene-aware dialogue system. *DSTC Workshop @ AAAI*.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28:2296–2304.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 878–892, Online. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. 2020. Contrastive learning for weakly supervised phrase grounding. In *ECCV*.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy. PMLR.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das, I. Essa, D. Batra, and D. Parikh. 2019. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356.

Chiori Hori, Anoop Cherian, Tim K Marks, and Takaaki Hori. 2019. Joint student-teacher learning for audio-visual scene-aware dialog. *Proc. Interspeech 2019*, pages 1886–1890.

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. 2018. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jiaji Huang, Yi Li, Wei Ping, and Liang Huang. 2018. Large margin neural language model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191, Brussels, Belgium. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron

Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623, Florence, Italy. Association for Computational Linguistics.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven C.H. Hoi. 2020. BiST: Bi-directional spatio-temporal reasoning for video-grounded dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1846–1859, Online. Association for Computational Linguistics.

Hung Le, Chinnadhurai Sankar, Seungwhan Moon, Ahmad Beirami, Alborz Geramifard, and Satwik Kottur. 2021. Dvd: A diagnostic dataset for multi-step reasoning in video grounded dialogue. *arXiv preprint arXiv:2101.00151*.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Dstc8-avsd: Multimodal semantic transformer network with retrieval style word generator. *DSTC Workshop @ AAAI 2020*.

Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2021. Contrastive learning with adversarial perturbations for conditional text generation. In *International Conference on Learning Representations*.

Shiyang Li, Semih Yavuz, Kazuma Hashimoto, Jia Li, Tong Niu, Nazneen Rajani, Xifeng Yan, Yingbo Zhou, and Caiming Xiong. 2021a. Coco: Controllable counterfactuals for evaluating dialogue state trackers. In *International Conference on Learning Representations*.

Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021b. Bridging text and video: A universal multimodal transformer for video-audio scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–1.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3285–3292.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2020. Generate your counterfactuals: Towards controlled counterfactual generation for text. *arXiv preprint arXiv:2012.04698*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27:1682–1690.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Dat Tien Nguyen, Shikhar Sharma, Hannes Schulz, and Layla El Asri. 2018. From film to video: Multi-turn question answering with multi-modal context. In *AAAI 2019 Dialog System Technology Challenge (DSTC7) Workshop*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, USA.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053, Hong Kong, China. Association for Computational Linguistics.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. 2019. Factor graph attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2039–2048.

Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. 2017. Visual reference resolution using attention memory for visual dialog. In *Advances in neural information processing systems*, pages 3719–3729.

Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1049–1058.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Zhao Wang and Aron Culotta. 2020. Robustness to spurious correlations in text classification via automatically generated counterfactuals. *arXiv preprint arXiv:2012.10040*.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Florence, Italy. Association for Computational Linguistics.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S Lasecki, Jonathan K Kummerfeld, Michel Galley, Chris Brockett, et al. 2019. Dialog system technology challenge 7. *arXiv preprint arXiv:1901.03461*.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280, Online. Association for Computational Linguistics.

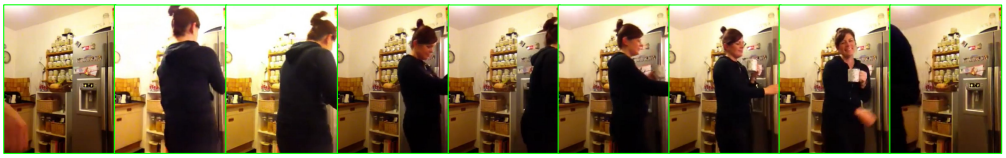
Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022.

Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33:18123–18134.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

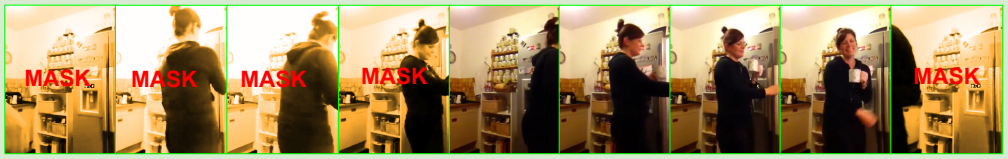
Original video



Original dialogue history
Q1: is the woman already in the room ? A1: yes she is already in the room. Q2: is there any other people ? A2: no other people in the video. Q3: is she talking in the video ? A3: no she isn't talking in this video. Q4: is there any music heard ? A4: no music is heard there.

Question and answer of current turn (detected actions are highlighted):
Q5: does the woman eat or drink anything? A5: she takes a cup from the fridge but didn't drink.

Factual video (*I+*)



Counterfactual video (*I-*)

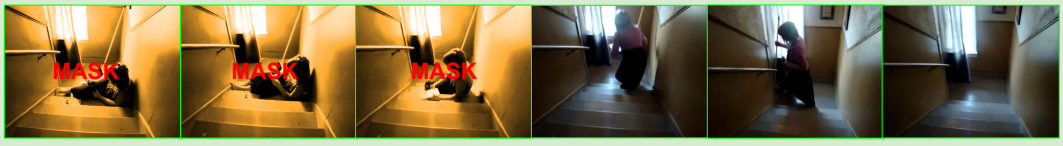
Figure 4: Example factual and counterfactual video

Original video

Original dialogue history
Q1: how many people can you see ? A1: there is only one person . Q2: is it indoors ? A2: yes , the entire video is indoors . Q3: is it daylight ? A3: yes , it is daylight outside . Q4: is the person happy ? A4: yes , she is laughing . to herself . Q5: is it in a house or apartment ? A5: i cannot tell if it is an apartment or home . Q6: is the person watching tv or reading a book ? A6: she is looking at her phone . Q7: how old does the person seem to be ? A7: she looks like early twenties . Q8: is she sitting down or standing up ? A8: she is sitting on the stairs then stands up and leaves . Q9: are the stairs covered with carpet ? A9: no , they are bare , no carpet .

Question and answer of current turn (detected actions are highlighted):
Q10: can you see her getting out of the dwelling ? A10: no , you can only see her walk away .

Factual video (*I+*)



Counterfactual video (*I-*)

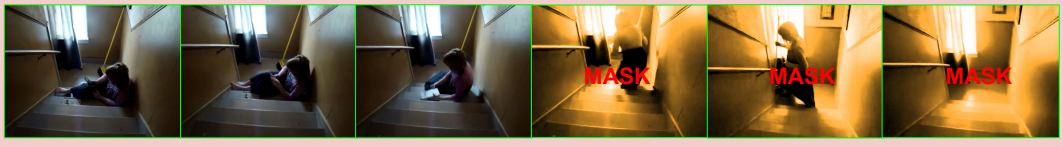


Figure 5: Example factual and counterfactual video

Original video

Original dialogue history
Q1: is there only one guy in this whole video?
A1: yes I can see only one guy in the video
Q2: can you tell what he's cooking?
A2: nothing clear he pour some water in bowl and on stove
Q3: what's in the silver packet at the end?
A3: he opened medicine in the box
Q4: does he boil the water?
A4: yes it seems in the video

Question and answer of current turn (detected objects are highlighted)
Q5: does he eat any of that medicine?
A5: no he opened the medicine that's all in video

Factual dialogue history (*H+*)
Q1: is there only one guy in this whole video?
A1: yes I can see only one guy in the video
Q2: can you tell what he's cooking?
A2: nothing clear he pour **[MASK]** in bowl and on stove
Q3: what's in the silver packet at the end?
A3: he opened medicine in the box
Q4: does he boil the water?
A4: yes it seems in the video

Counterfactual dialogue history (*H-*)
Q1: is there only one guy in this whole video?
A1: yes I can see **[MASK]**
Q2: can you tell what **[MASK]** 's cooking?
A2: nothing clear **[MASK]** pour some water in bowl and on stove
Q3: what's in the silver packet at the end?
A3: **[MASK]** opened **[MASK]** in the box
Q4: does **[MASK]** boil the water?
A4: yes it seems in the video

Figure 6: Example factual and counterfactual dialogue history

Original video

Original dialogue history
Q1: is the guy in the red shirt dancing?
A1: no , he is using the towel to dust the window.
Q2: is that a women to right of him watching him?
A2: yes that is a woman.
Q3: what was he doing before dusting the window?
A3: he turns around, then picks up the towel.
Q4: what did he do after dusting the window?
A4: he doesn't stop, he does it for the remainder of the video.
Q5: was there any talking in the video?
A5: yes, a woman speaks in a foreign language, at the beginning of the video only.
Q6: can you tell who she was talking to?
A6: to the man who ends up dusting the window.
Q7: does the woman do anything besides talk to the man dusting?
A7: no, she doesn't, it might be the female behind the camera speaking.
Q8: is the window he's dusting dirty?
A8: can tell if is or not.

Question and answer of current turn (detected objects are highlighted):
Q9: is he using only the towel on the window or does he have a cleaner like a spray bottle?
A9: only the towel he's using.

Factual dialogue history (*H+*)
Q1: is the guy in the red shirt dancing?
A1: no , he is using the towel to dust the window.
Q2: is that a women to right of him watching him?
A2: yes that is a woman.
Q3: what was he doing before dusting the window?
A3: he turns around, then picks up the towel.
Q4: what did he do after dusting the window?
A4: he doesn't stop, he does it for the remainder of the video.
Q5: was there any talking in the video?
A5: yes, **[MASK]** speaks in a foreign language, at the beginning of the video only.
Q6: can you tell who **[MASK]** was talking to?
A6: to **[MASK]**
Q7: does **[MASK]** do anything besides talk to **[MASK]** dusting?
A7: no, she doesn't, it might be the female behind the camera speaking.
Q8: is the window he's dusting dirty?
A8: can tell if is or not.

Counterfactual dialogue history (*H-*)
Q1: is **[MASK]**?
A1: no , **[MASK]** is using the towel to dust **[MASK]**.
Q2: is that a women to right of **[MASK]** watching **[MASK]**?
A2: yes that is a woman.
Q3: what was **[MASK]** doing before dusting **[MASK]**?
A3: **[MASK]** turns around, then picks up the towel.
Q4: what did **[MASK]** do after dusting **[MASK]**?
A4: **[MASK]** doesn't stop, **[MASK]** does it for the remainder of the video.
Q5: was there any talking in the video?
A5: yes, a woman speaks in a foreign language, at the beginning of the video only.
Q6: can you tell who she was talking to?
A6: to the man who ends up dusting **[MASK]**.
Q7: does the woman do anything besides talk to the man dusting?
A7: no, she doesn't, it might be the female behind the camera speaking.
Q8: is **[MASK]** he's dusting dirty?
A8: can tell if is or not.

Figure 7: Example factual and counterfactual dialogue history

No that's not what I meant: Handling Third Position Repair in Conversational Question Answering

Vevake Balaraman Arash Eshghi Ioannis Konstas Ioannis Papaioannou
AlanaAI
{vevake, arash, ioannis.k, ioannis}@alanaai.com

Abstract

The ability to handle *miscommunication* is crucial to robust and faithful conversational AI. People usually deal with miscommunication immediately as they detect it, using highly systematic interactional mechanisms called *repair*. One important type of repair is *Third Position Repair* (TPR) whereby a speaker is initially misunderstood but then corrects the misunderstanding as it becomes apparent after the addressee's erroneous response (see Fig. 1). Here, we collect and publicly release REPAIR-QA[1], the first large dataset of TPRs in a conversational question answering (QA) setting. The data is comprised of the TPR turns, corresponding dialogue contexts, and candidate repairs of the original turn for execution of TPRs. We demonstrate the usefulness of the data by training and evaluating strong baseline models for executing TPRs. For stand-alone TPR execution, we perform both automatic and human evaluations on a fine-tuned T5 model, as well as OpenAI's GPT-3 LLMs. Additionally, we *extrinsically* evaluate the LLMs' TPR processing capabilities in the downstream conversational QA task. The results indicate poor out-of-the-box performance on TPR's by the GPT-3 models, which then significantly improves when exposed to REPAIR-QA.

1 Introduction

Participants in conversation need to work together on a moment by moment basis to achieve shared understanding and coordination (Clark, 1996; Clark and Brennan, 1991; Goodwin, 1981; Healey et al., 2018; Mills, 2007). One of the key interactional mechanisms that enables this is called *repair* (Schegloff et al., 1977; Schegloff, 1992) – see Fig. 1: a set of universal, highly systematised (Dingemanse et al., 2015), local methods for dealing with *miscommunication* as it is detected.

[1]The dataset, models and code for all experiments are available at https://github.com/alanaai/Repair-QA

Figure 1. TPR Example from REPAIR-QA

(T1) U: What is the name of the princess in Frozen? ⟨Trouble Source⟩
(T2) S: The name of the princess who eventually becomes queen is Elsa
(T3) U: no I mean the name of the younger sister ⟨Third Position Repair⟩
(T4) S: The name of the younger sister is Anna

Miscommunication likewise arises in human-machine conversation. Therefore, the ability to interpret and generate effective repair sequences is crucial to *robust* Conversational AI technology, and to ensuring that Natural Language Understanding (NLU) output and/or subsequent system responses remain *faithful* to what the user intended.

Considerable attention has been paid to computational models for the interpretation and generation of *self-repair* (see (Hough and Schlangen, 2015; Hough, 2015; Shalyminov et al., 2017; Skantze and Hjalmarsson, 2010; Buß and Schlangen, 2011; Hough and Purver, 2012) among others): a class of repairs whereby the speaker corrects themselves on the fly within the same conversational turn (e.g. "User: I want to go to London uhm sorry Paris"). Similarly, the crucial role of generating and responding to *Clarification Requests* (e.g. "Pardon/what/who?") in conversational models has long been recognised (see (San-Segundo et al., 2001; Purver, 2004; Purver and Ginzburg, 2004; Rieser and Moore, 2005; Rodríguez and Schlangen, 2004; Rieser and Lemon, 2006) among others), but existing systems either remain limited (e.g. Curry et al. (2018)) or do not support this at all – see Purver et al. (2018) for an overview of existing models of repair.

In this paper, we focus on an important class of repairs that has, to our knowledge, been neglected in the NLP community, likely due to the unavail-

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 562–571
September 11–15, 2023. ©2023 Association for Computational Linguistics

ability of data: *Third Position Repair* (TPR; (Schegloff, 1992); aka repair after next turn). These occur when the addressee initially misunderstands the speaker (Fig. 1 at T1, the *trouble source* turn), responds based on this misunderstanding (at T2), which in turn reveals the misunderstanding to the addressee who then goes on to correct the misunderstanding (at T3). Our **contributions** are: (1) We collect, analyse and release REPAIR-QA, the first large dataset of Third Position Repairs (TPR) in a conversational QA setting together with candidate repair outcomes (rewrites) for training *repair execution* models; and (2) We then use REPAIR-QA to: (a) train and intrinsically evaluate strong baseline models for the execution of TPRs; and (b) systematically probe the TPR processing capabilities of GPT-3-Curie and GPT-3-Davinci with and without exposing them to examples from REPAIR-QA.

2 The REPAIR-QA dataset

In this section, we describe our method for eliciting Third Position Repairs (TPR) from AMT crowd workers (henceforth annotators). Overall, we set this up as a dialogue completion task whereby the annotators are given a dialogue snippet in which a miscommunication has occurred: they are given T1 (Fig. 1; the `Trouble Source`) and T2 (the erroneous system response). They are then asked to provide a (Third Position) correction at T3 to resolve the miscommunication.

Method: Eliciting TPRs We built our dialogue completion tasks on Amazon Mechanical Turk (AMT). Annotators were paid $0.29 per annotation for their work (estimated at $11 per hour). To generate the dialogue completion tasks in order to elicit TPRs, we start from the AmbigQA dataset (Min et al., 2020) since it contains ambiguous questions (i.e. questions that have multiple interpretations and answers) and their corresponding unambiguous questions along with their answers. For each ambiguous question, Q, and the corresponding pair of unambiguous questions with their answers, (Q_1, A_1) and (Q_2, A_2), we build a dialogue snippet to be completed by the annotator with a TPR as follows: (1) We build an informative *context*, C, that differentiates between questions Q_1 and Q_2; (2) The answers in AmbigQA are mostly short, Noun Phrase answers, which do not reveal how the ambiguous question was interpreted or reveal the apparent miscommunication to the annotator. To remedy this, we transform these short answers to full sentential form using the rule-based approach of Demszky et al. (2018). This allows us to derive sentential forms for A_1, call it A_1'; (3) We build the dialogue snippet with two turns, T1 and T2 – see Fig. 1 – where $T1 = Q$ and $T2 = A_1'$. Annotators are told that their goal was to get a response to Q_2 (indicated by context C); then, given the dialogue snippet which erroneously provides an answer to Q_1, they are asked to provide *two* alternative TPRs at $T3$ to get a response to Q_2 instead. For example, in Fig. 1: Q is $T1$; Q_1 is "What is the name of the princess in Frozen who eventually becomes queen?"; A_1 is "Elsa"; A_1' is $T2$; and C is "who eventually becomes queen vs. the younger sister". The context C is built by identifying the difference between $Q1$ and $Q2$. We employ this approach as the AmbigQA unambiguous questions have the same syntactic form as the ambiguous question. Another big advantage of using the AmbigQA dataset is that Q_2 can be seen as the contextually resolved meaning of the TPR which we call the gold 'rewrite' following (Anantha et al., 2021). This gold rewrite is used below in our *repair execution* models. See Appendix B for more details.

Statistics and Quality Control The REPAIR-QA dataset consists of **3305** examples (training: 2657, test: 648) which are chosen and annotated from the 4749 examples from the AmbigQA dataset. Each conversation in REPAIR-QA consists of two different TPRs yielding a total 6610 TPR annotations. Table 6 in Appendix shows some examples of the collected data. For quality control, we randomly select 100 TPR annotations from the testset to perform a qualitative inspection of the collected data. We annotate them for (i) Quality: Does the TPR convey the information needed to convey the necessary correction?; (ii) Context-Dependence: Does the TPR contain any context-dependent phenomena (e.g. fragments, ellipsis, pronominals); and (iii) Corrective: Is the TPR formulated explicitly as a correction? (e.g. The TPR in Fig. 1 could have been: "what about the name of the younger sister?" which does not explicitly signal a correction). We find that only 16% of the data contains some noise; that 93% of TPRs contain some form of context-dependency; and that 80% of the TPRs formulate the TPR explicitly as a correction. To further measure the degree to which the interpretation of the TPRs relies on the dialogue context, we measure the unigram overlap between the TPR and the refer-

	BERT Score	BLEU	EM
T5-REPAIR-QA	**97.48**	**72.06**	**30.40**
GPT-3-Davinci	97.22	64.18	25.68
GPT-3-Curie	93.19	52.43	7.60

Table 1: Model performance on the testset of the REPAIR-QA dataset.

	BERTScore	BLEU
T5-REPAIR-QA	1.48	**20.12**
GPT-3-Davinci	**1.76**	19.94
GPT-3-Curie	(0.11)	1.85

Table 2: Model ability to generate corrective tokens computed based on the difference in performance of the prediction against the rewrite and the trouble source.

ence rewrite (viz. $Q2$ above). We find 28% overlap between them, suggesting that the TPRs are highly context-dependent.

Limitations As such, REPAIR-QA has two important limitations: (1) TPRs can in general sometimes – but rarely – occur at a distance of more than two turns from the *trouble-source* turn (Schegloff, 1992). But the TPRs we collected are always in the third turn following the trouble source: this is an artefact not just of our data collection design as a unilateral dialogue completion task, but also of the architecture of most Conversational QA models that REPAIR-QA is designed to be useful for; and (2) overall we'd have preferred a more ecologically valid setup where TPRs are elicited within a more dynamic, interactive setting rather than as a dialogue completion task. Nevertheless, we believe that this trade-off between difficulty of collecting human-human dialogues, and the breadth of the types of TPR sequences collected is justified.

3 TPR execution

We cast the TPR execution task as a sequence to sequence problem, where input to the model is the dialogue history up to and including the TPR turn, and the model is trained to generate a rewrite of the ambiguous, trouble-source question, reflecting the correction in the TPR. We use a pre-trained T5 model (Raffel et al., 2022) for our experiments and compare against OpenAI's GPT-3 (Brown et al., 2020) when prompted with TPR examples.

3.1 Repair Execution Results

The models are evaluated against metrics of BERTScore (Zhang et al., 2020), BLEU and Exact Match (EM) between the reference rewrite and the generated output [2].

Table 1 shows the performance of all models on the REPAIR-QA testset. The T5 model is fine-tuned using the REPAIR-QAand its performance is reported as T5-REPAIR-QA. The fine-tuned T5-REPAIR-QA model achieves the best performance against the gold rewrites on all the 3 metrics considered. The GPT-3 models (Davinci and Curie) are few-shot prompted with 10 random examples, per test instance, pooled from REPAIR-QA followed by the test data; (see Appendix C for details); unlike the T5-REPAIR-QA model which is fine-tuned using the REPAIR-QA training data. We see a slightly lower performance for Davinci compared to the T5-REPAIR-QA on the automatic evaluation; the Curie model shows significantly inferior performance, especially when looking at EM [3].

Generally, the correction that a TPR provides to the *trouble source* question (T1 in Fig. 1) is very specific and small (often just 1 or 2 words, e.g. "the younger sister" in Fig. 1). Thus a higher BLEU score is more likely even when the model prediction is similar to the trouble source. To evaluate the ability of the models to produce specifically the corrective tokens, we evaluate the models' predictions against both the gold rewrite and the trouble source itself, and compare these across all metrics. We compute the metrics for the models' prediction against the gold rewrite on the one hand, and, the trouble source separately on the other hand, and compute the difference between them (simple subtraction). This difference in performance against them is therefore attributable to whether the model was able to produce the few corrective tokens. Table 2 shows this differential evaluation: a similar trend is seen on the models for the BLEU metric but GPT-3-Davinci outperforms other models on BERTScore. This result is discussed further below.

[2]We also tried an NLI-based text-classifier (Yin et al., 2019) for evaluation but the metric was not suited for this task, hence not reported here.

[3]We also did a zero-shot evaluation of a T5 model trained only on QReCC (Anantha et al., 2021) – a contextual resolution dataset – against the REPAIR-QA testset: it performed very poorly (BLEU = 37.44) indicating that the patterns of context-dependency in the TPRs are very different from the general patterns of context-dependency found in the QReCC dataset. This further demonstrates the usefulness of REPAIR-QA.

	Q1	Q2
T5-REPAIR-QA	3.53	4.01
GPT-3-Davinci	4.56	4.27

Table 3: Human evaluation of TPR execution models

Prompting	BLEU	EM	Unknown
w/o TPR examples	11.40	11.71%	230
with TPR examples	16.98	31.90%	57

Table 4: End-to-end, TPR processing capability of GPT-3 Davinci, with and without being exposed to TPR examples from REPAIR-QA

3.2 Human Evaluation

We asked two expert annotators (two co-authors of the paper) to rate the quality of T5-REPAIR-QA and GPT-3-Davinci model's output rewrites for executing the TPRs. We separately asked them the following questions: **Q1**: "On a scale of 1 to 5, how well does the model prediction avoid the misunderstanding caused by the ambiguity in the original question?"; and **Q2**: "On a scale of 1 to 5, to what degree is the model prediction asking for the same information as the gold?". While the answer to Q2 depends on the gold rewrites from REPAIR-QA, the answer to Q1 does not. This is because in executing a TPR what we care about is not necessarily the surface form of the output but instead the overall correction on a *semantic level*. The annotators showed very high interannotator agreement on both questions (average Krippendorf's $\alpha = 0.8$).

As Table 3 shows, the Davinci model's performance in the human evaluation is superior to the T5-REPAIR-QA model for both Q1 and Q2. At first glance, this would seem to be inconsistent with the word overlap metrics in Table 1 since the fine-tuned T5-REPAIR-QA model outputs show more overall overlap with the gold rewrites. However, a qualitative inspection of the respective outputs of each model shows that the Davinci model manages to produce rewrites which sufficiently capture the meaning of the TPR even as it doesn't always reproduce exactly the same words. This explanation is further supported by the BERTScore, semantic similarity results in Table 2 which shows slightly superior performance of the Davinci model (see Table 5 in Appendix for an example comparison). We believe that this is due to the fact the Davinci model is only exposed to ten examples in the prompt each time, whereas the T5-REPAIR-QA model is fine-tuned on all the training data from REPAIR-QA.

4 Extrinsic evaluation of GPT-3's TPR capabilities in conversational QA

In this section, we use REPAIR-QA to evaluate the TPR processing capabilities of OpenAI's GPT-3 Davinci model extrinsically in an end-to-end, conversational QA setting. We do this by comparing:

(a) the model's response to the reference rewrite (the corrected, unambiguous form of each question); with

(b) the response returned after the dialogue snippet with the TPR as its last turn.

If (a) and (b) are identical or highly similar, we can infer that the model was able to interpret the TPR correctly; independently of whether the responses are faithful. We compute the automatic evaluation on the model's response in (b) while treating the model's response in (a) as the ground truth. This would evaluate if the model was consistent in generating responses for both the rewrite and the TPR dialogue snippet. This evaluation is performed under two *prompting conditions*: **With TPR examples**: where the model is exposed to 10 TPR examples in the prompt; and; **Without TPR examples**: where the model is prompted without any TPR examples. In both conditions, the preamble instructs Davinci to generate *unknown* as the answer if the question is either nonsense, trickery, or Davinci has no clear answer. In addition, in both cases, the model is instructed to provide short form, Noun Phrase answers (for details of all of the preambles used, see Appendix, Sec. C).

There could in general be two reasons for *unknown* predictions after a TPR: (i) the Davinci's closed-book knowledge is insufficient to answer the (disambiguated, corrected) question; or; (ii) It was unable to interpret the TPR sequence. Since we are interested only in (ii), we *exclude* all cases where the model was not able to answer the unambiguous question (i.e case (a) above), viz. the reference rewrite (the meaning of the TPR). This way we ensure that the model can actually answer the target, rewritten / corrected question. After these are excluded, the 'Unknown' column in Table 4 contains the number of *unknown* responses to the TPRs; showing how the model improves when exposed to TPR examples in conversational QA.

For cases where both (a) and (b) above receive answers from GPT3, we perform automatic evaluation to measure the similarity between them: this is

also shown in Table 4. As a surface overlap metric, BLEU is suitable for this evaluation since we compare short answer tokens with many of these being bare Noun Phrases, e.g. names of movies, persons, dates, etc: there are no or few semantically similar paraphrases of these answers.

As is evident in Table 4, the TPR processing capability of Davinci in conversational QA when not exposed to any TPR example is very poor, but this improves significantly with a handful of TPR examples in the prompt. This shows that state-of-the-art LLMs do not handle TPRs well at all out-of-the-box, validating the requirement for datasets addressing specific dialogue phenomena like TPRs.

Even when the model is exposed to TPR sequences in the prompt (the "with TPR examples" condition) the model's performance still leaves a lot to be desired: the model's responses to the TPRs matches the expected response only in 31.9% of cases.

To verify the meaningfulness of the 31.9% exact match and the corresponding low BLEU score of 16.98 between model responses in (a) and (b), we went on to do a manual inspection of the data. Fig. 2 shows two examples of these responses:

User: Who plays the leprechaun in the leprechaun movie?
System: Warwick Davis
TPR: I was referring to leprechaun origins
Rewrite: Who plays the leprechaun in the Leprechaun Origins movie?

Response to (a): Dylan Postl
Response to (b): Linden Porco

User: Who created the quote keep calm and carry on?
System: British government
TPR: I wanted to know the name of the ministry though.
Rewrite: Which ministry created the quote keep calm and carry on?

Response to (a): British Ministry of Information
Response to (b): Ministry of Information

Figure 2: Two pairs of example responses provided by Davinci in its responses to (a): the unambiguous, corrected question rewrite; and; (b): the three turn TPR sequence

We can see different answers when prompted with the dialogue including the TPR ((b) above) and when prompted with the rewrite (unambiguous form of the input; (a) above). Such inconsistent answers are frequent from the model even when REPAIR-QA examples are provided in the prompt.

For more certainty, we further computed more focused BLEU scores only in cases where there was no exact match between the model's responses in (a) and (b). The BLEU scores on these not exactly matching responses, **with** and **without** exposure to TPR examples were 8.81 and 8.08 respectively. This shows that the model provides different, inconsistent answers for a large part of the REPAIR-QA dataset even when exposed to TPR examples in the prompt; which in turn shows that the model is not able to interpret or integrate the TPR for too large a part of REPAIR-QA. On a very small proportion of cases, Davinci provides responses which are similar (usually a partial match as in the second example above: "British Ministry of Information" vs. "Ministry of Information"), which is captured by the BLEU score metric.

5 Conclusion

The ability to interpret and generate repairs is essential to robust and faithful Conversational AI. In this paper, we focused on Third Position Repair (TPR) that's been largely neglected in the NLP community. We collect, analyse and release the first large dataset of TPRs and use it to evaluate strong baseline repair execution models, as well as the conversational QA performance of Open AI's Davinci model when it encounters TPRs. The results show very poor out-of-the-box performance on TPRs which then improves when the model is exposed to REPAIR-QAdataset. But even then, Davinci does not exhibit an acceptable performance on TPRs when evaluated end to end in a Conversational QA setting. This is a symptom of the sparsity of TPRs in the original dialogic data used to pretrain Davinci and LLMs in general; and suggests that LLM researchers should be more selective in how they compile the datasets used for pretraining.

For this paper, we did not have a chance to evaluate later releases of LLMs (e.g. GPT3.5; GPT4) - it would be telling to see how much performance improvement the later models might exhibit on TPRs. Our evaluation methods above in conjunction with the REPAIR-QA dataset can be used easily to perform these evaluations. Finally, we hope that this paper inspires further computational research into miscommunication phenomena in dialogue in the context of recent astonishing successes with LLMs.

References

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question an-

swering goes conversational via question rewriting. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 520–534, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Okko Buß and David Schlangen. 2011. Dium : An incremental dialogue manager that can produce self-corrections. In Proceedings of SemDial 2011 (Los Angelogue), Los Angeles, CA, pages 47–54.

H. H. Clark and S. A. Brennan. 1991. Grounding in communication, pages 127–149. Washington: APA Books.

Herbert H. Clark. 1996. Using Language. Cambridge University Press.

Amanda Cercas Curry, Ioannis Papaioannou, Alessandro Suglia, Shubham Agarwal, Igor Shalyminov, Xu Xinnuo, Ondrej Dusek, Arash Eshghi, Ioannis Konstas, Verena Rieser, and Oliver Lemon. 2018. Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. In 1st Proceedings of Alexa Prize (Alexa Prize 2018).

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. arXiv preprint arXiv:1809.02922.

Mark Dingemanse, Seán G. Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S. Gisladottir, Kobin H. Kendrick, Stephen C. Levinson, Elizabeth Manrique, Giovanni Rossi, and N. J. Enfield. 2015. Universal principles in the repair of communication problems. PLOS ONE, 10(9):1–15.

C. Goodwin. 1981. Conversational organization: Interaction between speakers and hearers. Academic Press, New York.

Patrick G. T. Healey, Gregory J. Mills, Arash Eshghi, and Christine Howes. 2018. Running Repairs: Coordinating Meaning in Dialogue. Topics in Cognitive Science (topiCS), 10(2).

Julian Hough. 2015. Modelling Incremental Self-Repair Processing in Dialogue. Ph.D. thesis, Queen Mary University of London.

Julian Hough and Matthew Purver. 2012. Processing self-repairs in an incremental type-theoretic dialogue system. In Proceedings of the 16th SemDial Workshop on the Semantics and Pragmatics of Dialogue (SeineDial), pages 136–144, Paris, France.

Julian Hough and David Schlangen. 2015. Recurrent neural networks for incremental disfluency detection. In INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015, pages 849–853.

Gregory J. Mills. 2007. Semantic co-ordination in dialogue: the role of direct interaction. Ph.D. thesis, Queen Mary University of London.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5783–5797, Online. Association for Computational Linguistics.

Matthew Purver. 2004. CLARIE: the Clarification Engine. In Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL), pages 77–84, Barcelona, Spain.

Matthew Purver and Jonathan Ginzburg. 2004. Clarifying noun phrase semantics. Journal of Semantics, 21(3):283–339.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. In Patrick G. T. Healey, Jan de Ruiter, and Gregory J. Mills, editors, Topics in Cognitive Science (topiCS), volume 10.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(1).

Verena Rieser and Oliver Lemon. 2006. Using machine learning to explore human multimodal clarification strategies. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 659–666, Sydney, Australia. Association for Computational Linguistics.

Verena Rieser and Johanna Moore. 2005. Implications for generating clarification requests in task-oriented dialogues. In Proceedings of the 43rd Annual Meeting of the ACL, pages 239–246, Ann Arbor. Association for Computational Linguistics.

Kepa Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in German task-oriented spoken dialogues. In Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL), Barcelona, Spain.

Ruben San-Segundo, Juan M. Montero, J. Ferreiros, R. Córdoba, and José M. Pardo. 2001. Designing confirmation mechanisms and error recover techniques in a railway information system for Spanish. In Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue, pages 136–139, Aalborg, Denmark. Association for Computational Linguistics.

E.A. Schegloff. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. American Journal of Sociology, pages 1295–1345.

E.A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. Language, 53(2):361–382.

Igor Shalyminov, Arash Eshghi, and Oliver Lemon. 2017. Challenging neural dialogue models with natural data: Memory networks fail on incremental phenomena. In Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2017 - SaarDial), Barcelona.

Gabriel Skantze and Anna Hjalmarsson. 2010. Towards incremental speech generation in dialogue systems. In Proceedings of the SIGDIAL 2010 Conference, pages 1–8, Tokyo, Japan. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

A Model Training and Inference

The T5 models reported in this paper are implemented in pytorch using HuggingFace (Wolf et al., 2020) library. The hyperparameter of the models are set as default with the batch size set to 16. The T5 models are trained on a single 16GB GPU and fine-tuned for 5 epochs. The results in Table 1 for T5 models for a single run on the train/test split. For GPT-3 inference, we use OpenAI's playground [4] API and get predictions from both Davinci (text-davinci-003) and Curie (text-curie-001) models.

B Data Collection Details

We use Amazon Mechanical Turk[5] for collecting the human annotations for TPR. The data collection was conducted anonymously.

Crowdworker Quality Control. We conduct a pilot with 4 internal annotators to verify the instructions and revise them before deploying to AMT crowdworkers. To control for the quality of annotations and the langauge, the crowdworkers are restricted to i) Location is one of Australia, Canada, New Zealand, United Kingdom, United States; ii) HIT approval rate > 80% and; ii) Number of HITs approved > 50. This was done explicitly to control the quality of the annotations collected after examining the annotations from a pilot phase in AMT.

Crowdworker Instructions. Figure 3 shows the instruction provided to the crowdworker and Figure 4 shows the interface, which the crowdworker uses to annotate the provided example. We explicitly instruct the crowdworkers to mark examples in which any of the information is unclear. To better explain the concept of TPR to the crowdworkers, we use the term **late correction instead of TPR** in the annotation instructions.

C GPT-3 prompts

The prompt used to query GPT-3 model to get predictions for both rewrite and QA is presented here. The text in blue indicate the tokens that the GPT-3 has to generate.

Rewriter prompts. Prompt used to generate rewrites from GPT-3. We use 5 examples in the prompts (single example is shown here for reference).
"Rewrite the Question Q based on the late correction LC.

[4] https://beta.openai.com/playground
[5] www.mturk.com

User: What is the most current episode of Ray Donovan?
System: The title of the most current episode of Ray Donovan is you'll Never Walk Alone.
User (TPR): What number was it in the series?

GPT-3-Davinci: What is the number of the most current episode of Ray Donovan titled "You'll Never Walk Alone"?
T5-QReCC+REPAIR-QA: What number was the most current episode of Ray Donovan?
Reference: What is the number overall of the most current episode of Ray Donovan?

Table 5: Prediction from different models on an example from REPAIR-QA.

User: Where do you hit to test your reflexes?
System: You hit to test your ankle jerk reflexes in Achilles tenden.

TPR-1: No, I meant your biceps, not ankle.
TPR-2: I should have been clearer. I wanted to know about the location to test for biceps reflexes.
Rewrite: Where do you hit to test your bicips reflexes?

User: Who sings i'm telling you i'm not going?
System: Jennifer Holliday sings i'm telling you i'm not going in the musical Dreamgirls.

TPR-1: I should have asked, who sang the song in 1982.
TPR-2: I wanted the singer in 1982, not in the musical Dreamgirls.
Rewrite: Who sings i'm telling you i'm not going in 1982?

User: Who is the lead singer of doobie brothers?
System: Johnston is the first lead singer of doobie brothers.

TPR-1: I want to know who was the second lead singer not the first.
TPR-2: I was wanting to know the second lead singer not the first.
Rewrite: Who is the second lead singer of doobie brothers?

User: Who has won the european cup the most?
System: Real Madrid has won the european cup the most.

TPR-1: Instead of club, can you tell me the country with the most.
TPR-2: I am looking for the country instead of the club with them most.
Rewrite: What country has won the european cup the most?

User: How much did titanic make at the box office?
System: Titanic (1953 film) made $2,250,000 at the box office.

TPR-1: I meant the 1997 version.
TPR-2: I was thinking of the 1997 one.
Rewrite: How much did Titanic (1997 film) make at the box office?

User: Who is winner of womens world cup 2017?
System: New Zealand is the winner of the Women's Rugby World Cup in 2017.

TPR-1: Yeah, but who won the cricket world cup?
TPR-2: What I wanted to know is who won the cricket cup.
Rewrite: Who is the winner of the Women's Cricket World Cup in 2017?

User: Who plays the king of france in the borgias?
System: Michel Muller plays King Charles VIII of France in The Borgias (2011 TV series).

TPR-1: I meant to ask who played louis xii.
TPR-2: Sorry but I was looking for louis xii.
Rewrite: Who plays King Louis XII of France in The Borgias (2011 TV series)?

Table 6: Examples from the REPAIR-QA dataset.

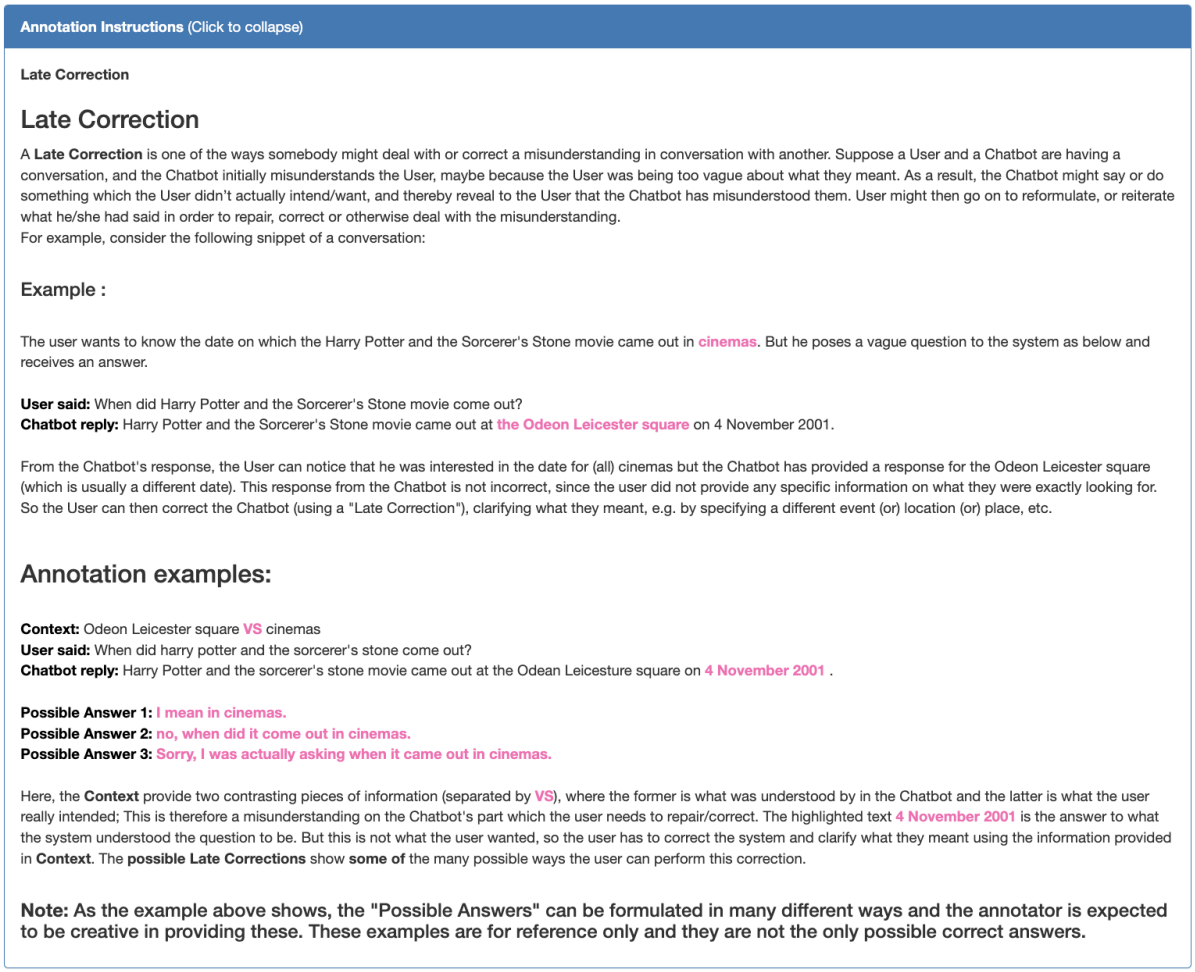


Figure 3: Annotation Instructions provided to the crowd annotators.

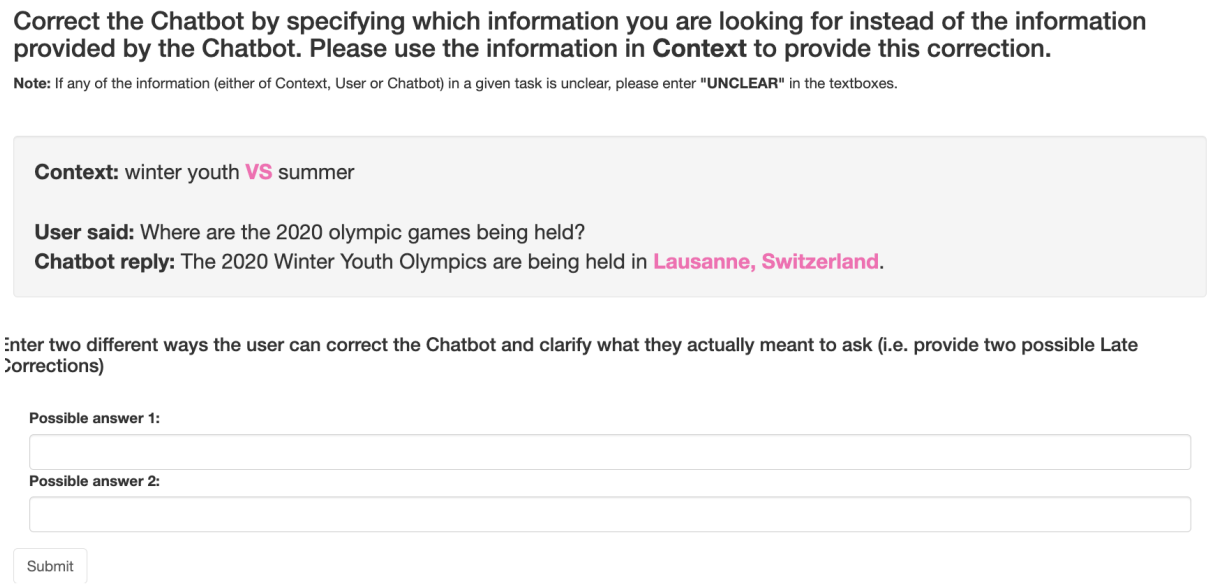


Figure 4: Interface of the annotation page as shown to the annotators.

Q: What is the percentage of agriculture in gdp of india?
A: The percentage of agriculture in gdp of india in 2017 is 15.4.
LC: I am looking for the year 2014 instead.
Rewrite: What is the percentage of agriculture in gdp of india in 2014?

Q: Who sang the song it's the final countdown?
A: Europe was the band that sang the song it's the final countdown, released in 1986.
LC: I was looking for the name of the lead singer.
Rewrite: Who sang lead vocals for the song it's the final countdown, released in 1986?
"

QA prompts. The prompt used for the conversational QA task is as below. We use 10 examples in the prompts (single example is shown here for reference).
"I am a highly intelligent question answering bot. If you ask me a question that is rooted in truth, I will give you only the answer phrase. If you ask me a question that is nonsense, trickery, or has no clear answer, I will respond with "Unknown".

Q: Who is the lead singer of doobie brothers?
A: Johnston is the first lead singer of doobie brothers.
Q: I want to know who was the second lead singer not the first.
A: Michael McDonald

Q: Who sang dedicated to the one i love?
A: The Shirelles sang Dedicated to the one I love in 1959.
Q: Could you also tell me who sang the 1967 version of dedicated to the one I love?
A: The Mamas and the Papas
"

When to generate hedges in peer-tutoring interactions

Alafate Abulimiti[1,2], **Chloé Clavel**[3], **Justine Cassell**[1,4]

[1] INRIA, Paris [2] ENS/PSL <alafate.abulimiti@inria.fr>
[3] LTCI, Institut Polytechnique de Paris, Telecom Paris <chloe.clavel@telecom-paris.fr>
[4] Carnegie Mellon University <justine@cs.cmu.edu>

Abstract

This paper explores the application of machine learning techniques to predict where hedging occurs in peer-tutoring interactions. The study uses a naturalistic face-to-face dataset annotated for natural language turns, conversational strategies, tutoring strategies, and nonverbal behaviors. These elements are processed into a vector representation of the previous turns, which serves as input to several machine learning models, including MLP and LSTM. The results show that embedding layers, capturing the semantic information of the previous turns, significantly improves the model's performance. Additionally, the study provides insights into the importance of various features, such as interpersonal rapport and nonverbal behaviors, in predicting hedges by using Shapley values (Hart, 1989) for feature explanation. We discover that the eye gaze of both the tutor and the tutee has a significant impact on hedge prediction. We further validate this observation through a follow-up ablation study.

1 Introduction

Effective communication involves various conversational strategies that help speakers convey their intended meaning and manage social interactions at the same time. These strategies can include the use of self-disclosure, praise, reference to shared experience, etc. (Zhao et al., 2014). Hedges are one of those strategies that is commonly used in dialogue. Hedges are words or phrases that convey a degree of uncertainty or vagueness, allowing speakers to soften the impact of their statements and convey humility or modesty, or avoid face threat. Although hedges can be effective in certain situations, understanding when and how to use hedges is essential and challenging.

The use of hedges is especially significant in tutoring interactions where they may facilitate correcting a wrong answer without embarrassing the recipient. However, the use of hedges in this context is not limited to expert educators. They are also found to be abundant in peer-tutoring settings. In fact, Madaio et al. (2017a) found that confident tutors tend to use more hedges when their rapport with the tutee is low, and that this pattern leads to tutees attempting more problems and solving more problems correctly. Hence, the detection and correct deployment of hedges, at the right time, is not just pleasant, but crucial for the development of effective intelligent peer tutoring systems.

While the use of hedges in conversation is an important aspect of effective communication, automatically generating hedges in real-time at the right time, can be a challenging task. In recent years, there have been several studies of automatic hedge detection (Raphalen et al., 2022; Goel et al., 2019), particularly in the context of dialogue systems. However, despite significant advances in detection, generating hedges in a timely and appropriate manner remained unsolved. For example, the RLHF-based training method enables the development of robust language models that align with human preferences (Ouyang et al., 2022). However, this approach does not explicitly instruct large language models (e.g., ChatGPT) in pragmatic and social skills, such as the appropriate use of hedges during communication. This lack of specific training can result in a gap in the model's ability to effectively integrate these conversational nuances into its responses in *at the correct time*. This limitation can affect the quality of communication and highlights the need for further research on effective hedge strategie generation; that is, to generate hedges at the right time.

Despite the widespread use of hedges in communication, there is still much to learn about their timing and the effectiveness of their use, particularly in dialogue rather than running text, and specifically in the current article, in peer-tutoring environments.

To address this gap in the literature, our research

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 572–583
September 11–15, 2023. ©2023 Association for Computational Linguistics

focuses on two key questions:

RQ1: First, can we predict when hedges should be generated in peer-tutoring environments?

To address this question we investigate whether it is possible to identify the points at which hedges should be introduced during a peer tutoring dialogue.

RQ2: Second, what features contribute to accurate predictions? of where to place hedges?

To address this question we focus on the explainability of classification models using Shapley values (Sundararajan and Najmi, 2020) .

2 Related Work

2.1 Hedges

Hedges are a common rhetorical device used to diminish the impact of an utterance, often to avoid unnecessary embarrassment on the part of the listener or to avoid the speaker being interpreted as rude. In linguistic terms, hedges diminish the full semantic value of an expression (Fraser, 2010). **Propositional hedges**, also called *Approximators*, refers to the use of uncertainty (Vincze, 2014), vagueness (Williamson, 2002), or fuzzy language (Lakoff, 1975), such as "sort of" or "approximately". On the other hand, **Relational hedges** are used to convey the subjective or opinionated nature of a statement, such as "*I guess* it will be raining tomorrow". **Apologizer** (Raphalen et al., 2022; Goel et al., 2019; Fraser, 2010) is an expression used to mitigate the strength of an utterance by using apologies, is another type of hedges. such as "*I am sorry*, but you shouldn't do that." Although the different types of hedges function differently, they all share a common role of mitigation in conversation. Therefore, in this paper, we focus on simply predicting hedges vs non-hedges.

As described above, in tutoring, including peer tutoring, hedges are frequently used and have a positive impact on performance (Madaio et al., 2017a). Powerful language models such as GPT-4 (OpenAI, 2023) and ChatGPT (OpenAI, 2022) are now capable of generating hedges with appropriate prompts, but these language models do not actively generate hedges (Abulimiti et al., 2023), fIn other words, the question of how to use thedges correctly in the next conversational action remains unsolved.

2.2 Conversational Strategy Prediction

The development of approaches for predicting conversational strategies – or particular ways of saying things – has progressed significantly over the past few years in the field of dialogue systems. Early studies, such as the COBBER, a domain-independent framework, used a Conversational Case-Based Reasoning (CCBR) framework based on reusable ontologies (Gómez-Gauchía et al., 2006). The aim was to help people use a computer more effectively by keeping them in the right mood or frame of mind. Methods such as reinforcement learning have also been introduced in non-task-oriented dialog systems, including a technique known as policy learning (Yu et al., 2016). Reinforcement learning has been explored, as well,for training socially interactive agents that maximize user engagement (Galland et al., 2022).

The Sentiment Look-ahead method is used to predict users' future emotional states and to reward generative models that enhance user sentiment (Shin et al., 2020). The rewards include response relevance, fluency, and emotion matching. These rewards are built using a reinforcement learning framework, where the model learns to predict the user's future emotional state. Romero et al. (2017) designed a social reasoner that can manage the rapport between user and system by reasoning and applying different conversational strategies.

More recently, deep learning-based approaches have emerged. For example, the Estimation-Action-Reflection (EAR) framework combines conversational and recommender approaches by learning a dialogue policy based on user preferences and conversation history (Lei et al., 2020).

Perhaps the most recent advances in the field have focused on how to create an empathetic dialogue system. MIME (Majumder et al., 2020) used the emotion mimicry strategy to match the user's emotion based on the text context. EmpDG (Li et al., 2020) generated empathetic responses using an interactive adversarial learning method to identify whether the responses evoke emotional perceptivity (the ability to perceive, understand, and be sensitive to the emotions of others.) in dialogue. The Mixture of Empathetic Listeners (MoEL) model (Lin et al., 2019) generates empathetic responses by recognizing the user's emotional state, using emotion-specific multi-agent listeners to respond, and then combining these responses based on the emotion distribution. This process effectively merges the output states of the listeners to create an appropriate empathetic response. The model then crafts an empathetic re-

sponse grounded in the user's emotions, which are monitored by the emotion tracker. Despite the notable success of MIME and MoEL in predicting emotions or conversational strategies, they do not incorporate the social context (e.g., the relationship between speakers), or the emotional tenor of the conversation up until that point, nor do they include important nonverbal behaviors into reasoning and decision-making processes. However, such elements are fundamental for the correct use of social language, and their absence potentially limits the effectiveness and naturalness of these models.

Predicting the appropriate emotion or conversational strategies in a conversation is a challenging task, mainly because determining what is "appropriate" in a conversation is rather subjective and is certainly context-dependent. For example, EmpDG (Li et al., 2020) model achieved an accuracy of approximately 0.34 across the 32 evenly distributed labels in the Empathetic Dialogue dataset (Rashkin et al., 2019). indicating the complexity of the problem at hand. Similarly, MoEL (Lin et al., 2019) model achieved varying degrees of accuracy in the same dataset - 38% for the top 1, 63% for the top 3, and 74% for the top 5 for emotion detection, further emphasizing the difficulty of the task.

The current paper aims to fill the lacunae in prior work by integrating social context and nonverbal behaviors as predictive features to construct predictive models for hedges.

3 Methodology

3.1 Task Description

Suppose we have a set of dialogues $D = \{d_1, d_2, d_3, ...d_n\}$. Each dialogue $d = \{u_1, u_2, u_3...u_m\}$ consists of m turns, with u_i representing a specific turn. Both tutor and tutee turns in these dialogues can be categorized as either hedges or non-hedges. However, for the purposes of our analysis, we will primarily focus on the tutor's turns. The label of a particular turn u_i is denoted as l_i. Furthermore, every turn can be depicted as a feature vector X, composed of elements $(x_1, x_2, ..., x_N)$. Here, N signifies the total number of features used to characterize a turn. Each turn in the dialogue is assigned a fixed window size (ω) of the dialogue history, represented as: $h_i = \{u_{max(1,i-\omega)}, u_{i-\omega+1}, ...u_i\}$. The primary objective of this research is to develop a model, denoted M, capable of predicting the type of hedge l'_{i+1} that a tutor will use next, based on the dialogue history h_i. The effectiveness of the model is measured using standard classification metrics, such as precision, recall, and F1 score.

Predicting hedges in a peer-tutoring conversation can be simplified to a binary classification problem. The features used as inputs are extracted from the turns in the interaction (further details in Section 3.3), while the output is a binary value showing whether or not hedges are present in each turn.

3.2 Corpus

The dataset used in the current work is the same as that employed in our previous work on hedges (Madaio et al., 2018). It is a subset of a larger investigation into the role of social, rapport-building conversational strategies in task-oriented dialogue. The corpus consists of face-to-face interaction from 20 same-gender dyads of American teenagers, with an average age of 14.3 years (and a range of ages from 13 to 16 years), gender-balanced [1] , and recorded twice over two weeks. However, due to technical issues, data from only 14 dyads' data were usable. The participants were asked to to take turns tutoring one another in different aspects of linear algebra. Each hour-long session was divided into 4 phases: an initial social period, followed by a first peer tutoring period, then a second short social period, and finally, the teens switched roles, with the tutee becoming tutor for the second task period. For the 14 dyads we used for our model, 28-hour-long face-to-face interactions were recorded over the period of two weeks. The recorded video and audio data were transcribed, resulting in approximately 9479 turns for the 14 dyads. These included 8399 non-hedges and 1080 hedges. 4214 non-hedges and 507 hedges in the tutors' turns since, as described above, we looked only at tutor hedges for this analysis (although note that both tutor and tutee hedges in prior turns were used as input). A "hedge turn" is any turn that includes hedging language. We also retained non-speech segments such as laughter and fillers.

Peer tutoring is a popular teaching method used in many schools and educational settings. As described above, and in Madaio et al. (2017b), even

[1]The corpus used here comes from earlier work by the last author and her colleagues, as cited above, and was used in accordance with the original experimenters' Institutional Review Board (IRB) approval. That approval required that the children's data not be released, which means that we cannot share the corpus. However, a pixelated example of the video data is available at github.com/neuromaancer/hedge_prediction.

though these teenagers may be inexperienced, in contexts of low rapport, when they use hedges during tutoring, their tutees are encouraged to attempt more problems and succeed in solving more of them. This positive outcome justifies the use of this dataset for studying hedges in tutoring interactions. While we recognize the importance of exploring the use of hedge with expert tutors in the future, our current focus on untrained peer tutors provides a unique perspective on how hedges can impact learning, even when the tutors themselves are not highly experienced. The methods and results from our study can be used as a foundation for future research, which could include the investigation of expert tutors and the potential differences in their use of hedges.

3.3 Features

In this section, we outline the features used as input vectors (i.e., u_i vector) for our prediction model, which seeks to properly predict the hedging strategy for the tutor's upcoming turn. In total, we have a vector with a length of 438 to represent a turn.

3.3.1 Turn embedding

Turn embedding is a common technique in natural language processing that involves representing a turn as a vector. In this study, we apply a sentence transformer (Reimers and Gurevych, 2019) to generate turn embeddings from the tutor-tutee conversation. This feature enables us to capture the semantic meaning of the turn in the context of the conversation, which can be helpful for predicting hedges.

3.3.2 Conversational Strategies (*CS*) of the previous turns

Conversational strategies refer to the different ways of speaking used by both speakers to manage social interaction. Strategies considered in this study are self-disclosure, praise, violation of social norms, and hedges. Self-disclosure (Derlega et al., 1993) refers to situations in which the tutor or tutee shares personal information, which is often used to build rapport. Praise (Brophy, 1981) is a form of positive feedback that acknowledges and reinforces the other person's behaviors or attributes. Violation of social norms (Zhao et al., 2014), which in this population often consists of friendly teasing, is a conversational move in speaker demonstrates the special nature of the relationship with the listener by engaging in slightly transgressive behavior. The conversational strategy annotation was carried out by Madaio et al. (2018), and inter-rater reliability achieved a minimum Krippendorff's alpha of over .7 for all strategies.

In terms of hedges, we note that we only use the speakers' previous hedge strategies to predict the tutor's next hedge strategy. This avoids any issues with predicting label leakage.

3.3.3 Tutoring Strategies (*TS*) in the previous turns

Tutoring strategies (Madaio et al., 2016) refer to the different techniques employed by the tutor or tutee to facilitate learning. Strategies considered in this study include deep/shallow questions, meta-communication, knowledge building, and knowledge telling. The deep question encourages critical thinking and higher-order cognition. The shallow question is used to confirm or clarify understanding. Meta-communication is a strategy whereby the tutor or tutee refers to the tutoring process or the tutor/tutee's self-evaluation of their own knowledge, which can help to clarify misunderstandings and promote effective communication. Knowledge building involves introducing new concepts or ideas, discussing the reasoning-mathematical solving steps, and providing examples. Knowledge telling refers to providing information (i.e., simply stating numbers, variables). The tutoring strategies annotation was also carried out by Madaio et al. (2018), with annotators achieving a minimum Krippendorff's alpha of .7 for all tutoring strategies.

3.3.4 Dialogue Act (*DialAct*) of the previous turns

Dialogue acts are types of speech acts (Searle, 1965) used by tutors and tutees during their interactions. In our study, we use the widely-used DAMSL (Dialogue Act Markup in Several Layers) (Jurafsky, 1997) coding schema to annotate dialogue turns by using a state-of-the-art dialogue act classifier with context-aware self-attention (Raheja and Tetreault, 2019). In our dataset, only 6 dialogue acts were found, they are *Abandoned* or Turn-Exit (*%*) , *Acknowledge (Backchannel)* (*b*), *Backchannel in question form* (*bh*), *Yes-No-Question* (*qy*), *Statement-non-opinion* (*sv*) and *Statement-opinion* (*sd*).

3.3.5 Rapport in the previous turns

As our previous work demonstrates, the level of rapport between tutor and tutee plays a role in the

use of hedges. We therefore include it as a feature in our study. Rapport is "The relative harmony of relations felt by both participants" (Spencer-Oatey, 2005). The rapport annotation was carried out by Amazon Mechanical Turk (AMT) annotators as described in Madaio et al. (2018). Rapport level was operationalized as a 7 point Likert scale, where a higher score indicates a stronger level of rapport. For the annotation of rapport, the annotators employed the "thin slice" method (Ambady and Rosenthal, 1993), whereby the experimenter segmented each video into 30-second clips and randomized the order. To ensure the quality of rapport annotations, three annators evaluated each clip, and the experimenter applied the inverse-bias correction method (Parde and Nielsen, 2017) for selecting a single score for each clip. In the current study, when the dialogue history is contained within a single slice, we directly use the annotated rapport level of that particular slice as the historical rapport level. However, if the dialogue history extends over two slices, we select the rapport level of the slice containing the majority of the dialogue history.

3.3.6 Nonverbal Behaviors (NB)

Nonverbal behaviors, such as head nod, smile, and gaze, are an essential aspect of interpersonal communication that can also contribute to the development of rapport (Tickle-Degnen and Rosenthal, 1990). The gaze and smile annotation was carried out by Madaio et al. (2018), we annotated the head nods with 2 annotators. All the annotations were carried out after annotators reached an inter-rater reliability of 0.7 or above on Krippendorff's alpha. We collected all nonverbal behaviors that occurred during one turn and encoded them using one-hot encoding. For head nods and smiles, we used a binary labeling approach, marking 1 for their occurrence and 0 for non-occurrence. As gaze serves as a potent indicator of attention, we categorized it into 4 distinct types: no gaze appeared in the video, gaze at partner, gaze at worksheet, and gaze elsewhere.

Mutual gaze between interlocutors, mutual smiles, and mutual head nods serve as great indicators of alignment and rapport in communication. These are not encoded separately, as our encoding process for nonverbal behaviors captures the behaviors of both participants within a turn, not only the current turn holder. Our current approach successfully captures these important mutual signals.

3.3.7 Contextual Information (*ConInfo*) in the previous turns

Our model also incorporates contextual information that characterizes the discourse environment between the two interlocutors. Specifically, we include features such as the session and period numbers, which help to encapsulate the temporal dynamics of the tutoring interactions. We also consider the math problem ID and the correctness of the current problem response, which act as markers of the present learning context. These features can illuminate the complexity of the ongoing problem and the students' performance, potentially influencing their use of hedges. The tutee's and tutor's pre-experiment test scores are also included, serving as initial measures of their knowledge before the tutoring session. This data can help to identify the starting knowledge disparity between the tutor and the tutee. It is plausible that these pre-test scores might also be linked with the students' level of confidence, which could subsequently impact their use of hedges (Madaio et al., 2017a).

Norman et al. (2022) suggested a link between verbal alignment signals, such as backchannels (e.g., "um", "hhm", "oh.."), and learning gains in a cooperative learning environment. Given the role of hedging as a social language skill that improves learning performance, we hypothesize its connection to dynamic learning gains. Consequently, we incorporated the frequency of these verbal alignment signals from the previous four conversational turns into our model input.

3.4 Vector Representation

Before presenting the specific models, we first describe how we convert each sequence of turns into a vector representation. Our vector representation consists of three basic parts: turns as a sequence of tokens, annotations based on the turn (e.g., conversational strategies), and the nonverbal behaviors. Figure 1 shows that we divide a vector of turns into 6 parts: turn embedding, conversational strategies (*CS*), tutoring strategies (*TS*), nonverbal behaviors (*NB*), contextual information (*ConInfo*) and dialogue acts (*DialAct*). After encoding each turn in this fashion, we use the four previous turns as a history tensor of a turn. This history ten tensor will be the input to the prediction models, and the model's output will be this turn's hedge label.

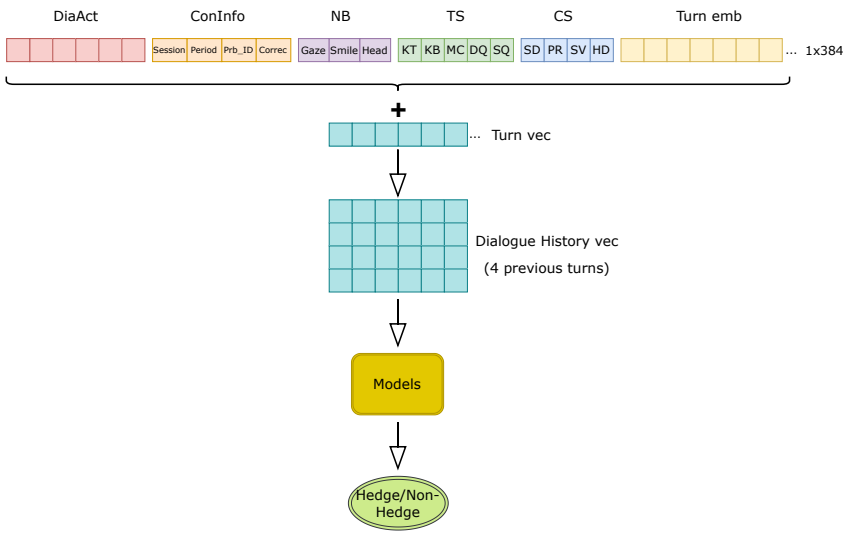


Figure 1: Vector Representation

3.5 Prediction as Classification

We mentioned in the previous section that we transform the prediction problem into a classification problem. This means that the corresponding hedge strategy is obtained by classifying different previous interactions (i.e., dialogue history) and historical characteristics (e.g., rapport, etc.). The classification models used are presented here.

The selection of learning models in this study is strategic and based on our research objectives. Our primary aim is not to engineer a perfect system for hedging. Instead, we seek to comprehend the variables that influence hedging in dialogue. As such, our approach leans towards the use of models that are effective in contextual understanding. For example, Long Short-Term Memory networks (LSTMs) were chosen over Multi-Layer Perceptrons (MLPs) due to their superior ability to manage and interpret context, an essential factor in our exploration of hedging phenomena.

3.5.1 LightGBM

In this work, we used LightGBM (Ke et al., 2017), a gradient boosting framework known for its efficiency. We use it to predict hedges in dialogues, relying only on dialogue features such as conversational strategies, tutoring strategies, nonverbal behaviors, and contextual information, while turn embeddings are not included.

3.5.2 XGBoost

We also used the Extreme Gradient Boosting (XGBoost) algorithm (Chen and Guestrin, 2016), which is a decision tree-based ensemble machine learning algorithm that uses a gradient boosting framework. Similar to LightGBM, the turn embedding is not used.

3.5.3 Multi-layer perceptron (MLP)

We constructed a multi-layer perceptron using two sets of features. These included a pre-trained contextual representation of the turn, specifically from the SentBERT model (Reimers and Gurevych, 2019) which is the most prevalent sentence embedding tool, and the concatenation of all the features mentioned in Section 3.3.

3.5.4 Long Short-Term Memory (LSTM)

We use the same features and apply them to LSTM (Hochreiter and Schmidhuber, 1997) and also LSTM with attention (Bahdanau et al., 2015). LSTM has a good ability to capture temporal correlations, and we expect this ability to enhance prediction performance.

3.6 Implementation Details

In order to address the imbalance in our dataset, where the ratio of hedge to non-hedge instances is approximately 1:10, we used the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) for each model to augment our learning process. SMOTE is a popular method that generates synthetic examples in a dataset to counteract its imbalance. Given the variable nature of model performance, we implemented a 5-fold cross-validation strategy to evaluate the models. In order to account for the imbalanced nature of the dataset, we opted to use a lower number of folds in

Models	F1-score	Precision	Recall
LightGBM (w/o emb)	0.24 (±0.07)	0.17 (±0.03)	0.45 (±0.07)
XGBoost (w/o emb)	0.24 (±0.07)	0.16 (±0.03)	0.45 (±0.07)
MLP	0.25 (±0.06)	0.16 (±0.03)	0.52 (±0.07)
MLP (only emb)	0.26 (±0.05)	0.16 (±0.02)	0.74 (±0.06)
MLP (w/o emb)	0.26 (±0.06)	0.17 (±0.06)	0.56 (±0.07)
LSTM	0.25 (±0.06)	0.16 (±0.03)	0.50 (±0.07)
LSTM (only emb)	0.28 (±0.07)	0.19 (±0.08)	0.52 (±0.07)
LSTM (w/o emb)	0.25 (±0.05)	0.15 (±0.02)	0.75 (±0.06)
AttnLSTM	0.24 (±0.06)	0.15 (±0.03)	0.57 (±0.07)
AttnLSTM (only emb)	0.25 (±0.07)	0.17 (±0.03)	0.45 (±0.07)
AttnLSTM (w/o emb)	0.23 (±0.06)	0.15 (±0.07)	0.57 (±0.07)
Dummy	0.11 (±0.08)	0.14 (±0.06)	0.10 (±0.04)

Table 1: Comparison of MLP and LSTM models for predicting hedges

the cross-validation process. By choosing 5 folds instead of a higher number, we aimed to ensure that each fold would contain a sufficient representation of samples from each class. The model that delivered the best performance during this cross-validation process was then chosen to make predictions on the test set. For the neural models, we adjusted the loss function to account for class imbalance, thereby compelling the models to accommodate less frequent classes more effectively. The code is available in https://github.com/neuromaancer/hedge_prediction

4 Results

4.1 Classification Results

To answer the research question 1, we conducted classification experiments on different models. Table 1 offers an in-depth comparison of multiple machine learning models for predicting hedges in a peer-tutoring dataset. We also incorporated a dummy classifier for comparison, which generates predictions in accordance with the class distribution observed in the training set. The performance metrics are F1 score, precision and recall, all of which include confidence intervals ($\alpha = 0.05$). The dataset is composed of several types of input features described in Section 3.3. The models used different combinations of these inputs. (w/o emb) indicates that the model uses only the features without turn embeddings. If not specified, the model uses all features plus turn embeddings.

From Table 1, the LightGBM and XGBoost models without embeddings achieved relatively low scores for F1 scores, precision and recall, indicating limited performance in terms of balanced precision and recall. The MLP models, particularly those using only embeddings, showed a remarkable recall of 74%, but at the cost of reduced precision. The LSTM model using only turn embeddings demonstrated balanced performance across all metrics, achieving the highest precision of 19% and a competitive F1 score of 0.28. However, the attention-based LSTM (AttnLSTM) model did not significantly outperform the standard LSTM model in any metric.

The inclusion of turn embeddings significantly impacts model performance. Models with only embeddings perform better in terms of F1 score and recall, suggesting that the semantic information captured in these embeddings, which represented the semantic information of turns, is crucial for hedge prediction. Second, models without embeddings also performed reasonably well in F1 score, implying that other features such as rapport, conversational strategies, tutoring strategies, nonverbal behaviors, and contextual information are also important. These features should not be overlooked.

The LightGBM and XGBoost models, which only use features without turn embeddings, also display competitive performance compared to the MLP, LSTM, and AttnLSTM models using all features. This suggests that although turn embeddings provide valuable information for hedge prediction, models can still achieve satisfactory results even without them. The AttnLSTM models, which incor-

Model \ Feature	*N/A*	*Rapport*	CS	*TS*	*NB*	*ConInfo*	*DialAct*
XGBoost	0.24 (±0.07)	0.15 (±0.08)	0.10 (±0.08)	0.15 (±0.09)	**0.08** (±0.07)	0.10 (±0.08)	0.12 (±0.08)
LightGBM	0.24 (±0.07)	0.16 (±0.08)	**0.09** (±0.08)	0.10 (±0.07)	0.10 (±0.10)	0.12 (±0.09)	0.13 (±0.08)
LSTM	0.25 (±0.05)	0.24 (±0.05)	0.26 (±0.06)	0.24 (±0.06)	0.22 (±0.06)	0.25 (±0.07)	**0.21** (±0.06)
AttnLSTM	0.23 (±0.06)	**0.20** (±0.06)	0.22 (±0.05)	0.25 (±0.05)	0.24 (±0.05)	0.23 (±0.07)	0.22 (±0.06)
MLP	0.26 (±0.06)	0.25 (±0.06)	0.25 (±0.06)	0.26 (±0.06)	0.25 (±0.06)	0.27 (±0.06)	**0.21** (±0.07)

Table 2: F1 scores after the feature ablation, *CS*: Conversational Strategies; *TS*: Tutoring Strategies; *NB*: Nonverbal Behaviors; *ConInfo*: Contextual Information; *DialAct*: Dialogue Act.

porate attention mechanisms, do not show significant improvements over the regular LSTM models. This could be due to the limited amount of data available, which cannot unleash the potential of the attention mechanism.

Since good performance can also be achieved using the extracted features, in order to answer our research question 2, in the next subsections we will mainly investigate the importance of features in predicting hedges.

4.2 Features Explanation with Shapley values

Shapley values (Hart, 1989), originating from cooperative game theory, have emerged as a powerful model-agnostic tool to explain the predictions of machine learning models. This approach provides a way to fairly distribute the contribution of each feature to the overall prediction for a specific instance. By calculating the Shapley value for each feature, we gain insight into the importance of individual features within the context of a specific prediction. This interpretability technique has been adopted across various machine learning models. In this study, we use Shapley values to interpret the contributions of extracted features in our classification models using the SHAP python package (Lundberg and Lee, 2017).

Figure 2 in the Appendix illustrates the importance of each feature for prediction when only features are used as input to different prediction models. The importance of features within the models can differ depending on their architectures. For simplicity, we identify the features that frequently appear in these 4 figures as significant indicators. Therefore, we have selected some of the most representative features in predicting hedges in Table 3.

Based on Table 3, certain features have a significant impact on the likelihood of using hedges in tutoring conversations. Rapport has a negative valence, suggesting that higher rapport between the participants results in a lower likelihood of hedges being used. This confirms the finding cited above, that hedges are more frequent in low rapport interaction (Madaio et al., 2017c). Interestingly, the "problem ID" feature also has a negative valence, indicating that as the complexity or difficulty of the problem increases, the likelihood of using hedges decreases. This could be because tutors tend to be more assertive or confident when addressing more challenging problems.

Features	**Valence**
correctness	+
no gaze from tutor	-
problem id	-
rapport	-
tutee's deep question	-
tutee's gaze at tutor	-
tutee's pre-test	-
tutor's gaze at elsewhere	-
tutor's praise	-

Table 3: Features and their Valences

Moreover, certain conversational features such as "tutee's deep question" and "tutor's praise" have a negative valence, implying that these actions tend to decrease the likelihood of hedges. This could be because deeper questions or praise might indicate a more open and confident dialogue, thus reducing the need for hedges.

The table also reveals a negative correlation between various non-verbal cues such as "no gaze from tutor", "tutee's gaze at tutor", and "tutor's gaze at elsewhere", and the occurrence of hedges. When the tutor is not gazing at the tutee, the likelihood of hedges decreases. The tutee's gaze at the tutor and the tutor's gaze at elsewhere are negatively associated with the use of hedges. This could indicate that when tutors' attention is focused elsewhere, they are attending less to how best to convey instruction or correction. To our knowledge, this is the first demonstration that specific nonverbal cues substantially influence the likelihood of a hedge

being used in the succeeding turn of peer-tutoring interactions.

4.3 Ablation Study

We next examine the aforementioned models with different features ablated from input. This approach allows us to identify which features, when absent, lead to the best or worst performance in each model, implying that these features may not have contributed positively (or negatively) to the model's performance. Our study considered 6 groups of features: Conversational Strategies (*CS*), Tutoring Strategies (*TS*), Nonverbal Behaviors (*NB*), Contextual Information (*ConInfo*), Dialogue Act (DialAct), and Rapport.

Table 2 shows the different F1 scores as a consequence of removing the different features. For XGBoost and LightGBM, the worst performance is observed when *NB* and *CS* were removed, respectively, which implies that these features may provide important information for these models. The LSTM and MLP models showed a significant drop in performance when the *DialAct* feature was removed, suggesting a substantial dependency of these models on the *DialAct* feature for their prediction capabilities. Interestingly, the best performance of AttnLSTM was achieved when the rapport feature was removed, suggesting that the attention mechanism could compensate for loss of rapport.

5 Conclusion and Future Work

This study presents an effective approach to predict where hedges occur in peer-tutoring interactions using classic ML models. Our results show the importance of considering various types of input features, such as turn embeddings, rapport, conversational strategies, tutoring strategies, nonverbal behaviors, and contextual information. Moreover, Shapley values applied to the predictions of the different models show, for the first time, that the gaze of both tutor and tutee may play a critical role in predicting hedges. This observation is substantiated by subsequent ablation studies, where classic classification models, like XGBoost and LightGBM, experienced a significant decline in F1 score when removing nonverbal behavior features.

For future work, several directions can be pursued. First, the investigation of hedge generation in the context of expert tutors could provide valuable insights into how experienced tutors use hedges differently and how these differences might affect learning outcomes. Second, incorporating reinforcement learning techniques to enhance specific aspects of the interaction, such as learning performance, could improve the practical applications of our findings.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. We express sincere gratitude to the members of the ArticuLabo at Inria Paris for their invaluable assistance in the successful completion of this research, and to the members of the ArticuLab at Carnegie Mellon University in Pittsburgh for answering our questions about their prior work. This study received support from the French government, administered by the Agence Nationale de la Recherche, as part of the "Investissements d'avenir" program, with reference to ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

Alafate Abulimiti, Chloé Clavel, and Justine Cassell. 2023. How about kind of generating hedges using end-to-end neural models? In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

Nalini Ambady and Robert Rosenthal. 1993. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of personality and social psychology*, 64(3):431.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Jere Brophy. 1981. Teacher praise: A functional analysis. *Review of educational research*, 51(1):5–32.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Valerian J Derlega, Sandra Metts, Sandra Petronio, and Stephen T Margulis. 1993. *Self-disclosure.* Sage Publications, Inc.

Bruce Fraser. 2010. Pragmatic competence: The case of hedging. new approaches to hedging.

Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2022. Adapting conversational strategies to co-optimize agent's task performance and user's engagement. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–3.

Pranav Goel, Yoichi Matsuyama, Michael Madaio, and Justine Cassell. 2019. i think it might help if we multiply, and not add. In *Detecting indirectness in conversation. In 9th International Workshop on Spoken Dialogue System Technology*, page 27–40. Springer.

Hector Gómez-Gauchía, Belén Díaz-Agudo, and Pedro A González-Calero. 2006. Conversational strategies in cobber: an affective ccbr framework. *Journal of Experimental & Theoretical Artificial Intelligence*, 18(4):449–469.

Sergiu Hart. 1989. *Shapley value*. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *www. dcs. shef. ac. uk/nlp/amities/files/bib/ics-tr-97-02. pdf*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

George Lakoff. 1975. Hedges: A study in meaning criteria and the logic of fuzzy concepts. In *Contemporary research in philosophical logic and linguistic semantics*, pages 221–271. Springer.

Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 304–312.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Michael Madaio, Justine Cassell, and Amy Ogan. 2017a. The impact of peer tutors' use of indirect feedback and instructions. Philadelphia, PA: International Society of the Learning Sciences.

Michael Madaio, Justine Cassell, and Amy Ogan. 2017b. "i think you just got mixed up": confident peer tutors hedge to support partners' face needs. *International Journal of Computer-Supported Collaborative Learning*, 12(4):401–421.

Michael Madaio, Rae Lasko, Amy Ogan, and Justine Cassell. 2017c. Using temporal association rule mining to predict dyadic rapport in peer tutoring. *International Educational Data Mining Society*.

Michael Madaio, Kun Peng, Amy Ogan, and Justine Cassell. 2018. A climate of support: a process-oriented analysis of the impact of rapport on peer tutoring. International Society of the Learning Sciences, Inc.[ISLS].

Michael A Madaio, Amy Ogan, and Justine Cassell. 2016. The effect of friendship and tutoring roles on reciprocal peer tutoring strategies. In *Intelligent Tutoring Systems: 13th International Conference, ITS 2016, Zagreb, Croatia, June 7-10, 2016. Proceedings 13*, pages 423–429. Springer.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979.

Utku Norman, Tanvi Dinkar, Barbara Bruno, and Chloé Clavel. 2022. Studying alignment in a collaborative learning activity via automatic methods: The link between what we say and do. *Dialogue & Discourse*, 13(2):1–48.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Natalie Parde and Rodney Nielsen. 2017. Finding patterns in noisy crowds: Regression-based annotation aggregation for crowdsourced data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1907–1912.

Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3727–3733.

Yann Raphalen, Chloé Clavel, and Justine Cassell. 2022. "You might think about slightly revising the title": Identifying hedges in peer-tutoring interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2174, Dublin, Ireland. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Oscar J Romero, Ran Zhao, and Justine Cassell. 2017. Cognitive-inspired conversational-strategy reasoner for socially-aware agents. In *IJCAI*, pages 3807–3813. Melbourne, VIC.

John R Searle. 1965. What is a speech act. *Perspectives in the philosophy of language: a concise anthology*, 2000:253–268.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2020. Generating empathetic responses by looking ahead the user's sentiment. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7989–7993. IEEE.

Helen Spencer-Oatey. 2005. (im)politeness, face and perceptions of rapport: Unpackaging their bases and interrelationships. 1(1):95–119.

Mukund Sundararajan and Amir Najmi. 2020. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR.

Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293.

Veronika Vincze. 2014. Uncertainty detection in natural language texts. *PhD, University of Szeged*, 141.

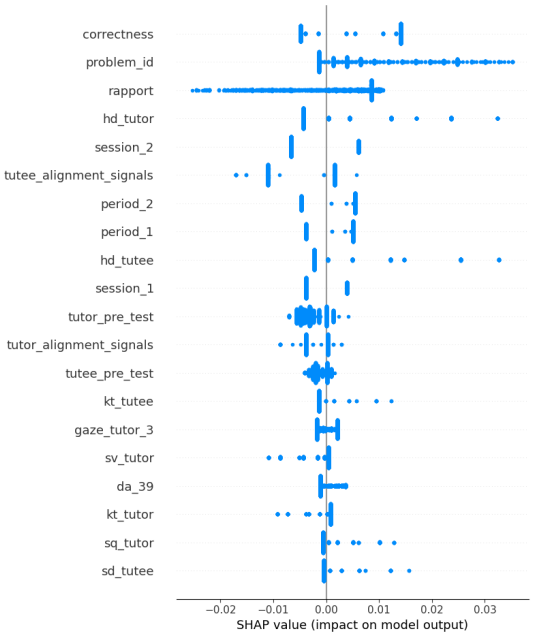
Timothy Williamson. 2002. *Vagueness*. Routledge.

Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 404–412.

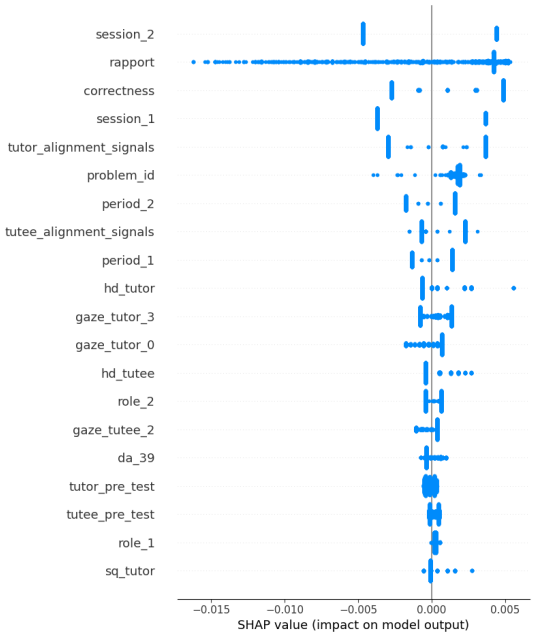
Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *International conference on intelligent virtual agents*, pages 514–527. Springer.

Appendix: SHAP Value Graphs

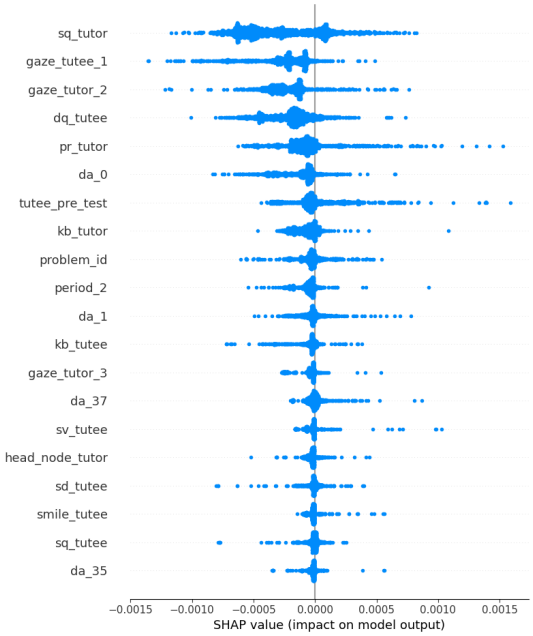
The vertical axis indicates the mean contribution of the feature over the model decision. The horizontal axis indicates how the distribution of features influences the model decision.



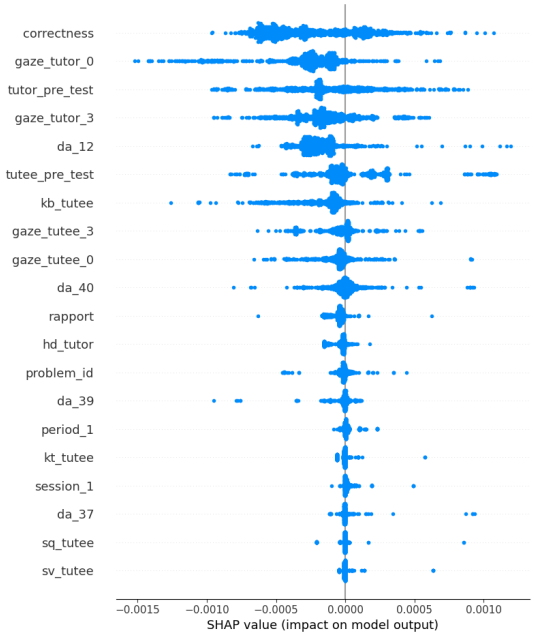
(a) Feature Importance for AttnLSTM (without emb)



(b) Feature Importance for MLP (without emb)



(c) Feature Importance for XGBoost



(d) Feature Importance for LightGBM

Figure 2: Feature Importance for Different Models

PaperPersiChat: Scientific Paper Discussion Chatbot using Transformers and Discourse Flow Management

Alexander Chernyavskiy
HSE University
Sberbank
Moscow, Russia
alschernyavskiy@gmail.com

Max Bregeda
Moscow State University
Sberbank
Moscow, Russia
mbregeda@gmail.com

Maria Nikiforova
HSE University
Sberbank
Moscow, Russia
labenzom@gmail.com

Abstract

The rate of scientific publications is increasing exponentially, necessitating a significant investment of time in order to read and comprehend the most important articles. While ancillary services exist to facilitate this process, they are typically closed-model and paid services or have limited capabilities. In this paper, we present *PaperPersiChat*, an open chatbot-system designed for the discussion of scientific papers. This system supports summarization and question-answering modes within a single end-to-end chatbot pipeline, which is guided by discourse analysis. To expedite the development of similar systems, we also release the gathered dataset, which has no publicly available analogues.

1 Introduction

Scientific papers are a crucial part of academic research and are used to disseminate new findings, theories and knowledge to the wider community. At the same time, rapid scientific progress makes it challenging to keep up with new technologies without spending a lot of time reading papers. While traditional summarizing services like Elicit[1] and Scholarcy[2] can be helpful, they often unable to explain sophisticated and complex concepts. More advanced solutions, such as Explainthepaper[3], have emerged to address this limitation as they can elucidate user-highlighted text, but also require the user to read the article beforehand.

Dialogue systems are an alternative capable of combining extractive and generative approaches. Grounding-based approaches were suggested to eliminate issues associated with the hallucinations of LMs (Cai et al., 2022; Gao et al., 2022). The release of ChatGPT[4] has propelled chatbots to the forefront of text data processing. The ChatGPT API and proprietary solutions have enabled the creation of communication services like ChatPDF[5] and xMagic[6]. However, these are services with a closed architecture and paid for.

Interestingly, there are no publicly available open systems that do not use the API of LLMs. One of the reasons, is the lack of open-source datasets for dialogue on scientific grounding. To bridge this gap, we present *PaperPersiChat*[7], a chatbot pipeline designed for the scientific paper domain. It capable of communicating on the basis of a user-selected paper by providing summaries and answering clarifying questions. Our second contribution is the training dataset that can be used to develop solutions for similar tasks. Our code is available at https://github.com/ai-forever/paper_persi_chat.

2 Related Work

The incorporation of external information, referred to as grounding, has been shown to enhance the quality of the generation by improving the factual component. Several approaches utilize knowledge bases or web mining (Glaese et al., 2022; Thoppilan et al., 2022), while others focus on extracting information from individual documents. Cai et al. (2022) proposed a transformer-based model which retains context semantics while sacrificing text details due to the use of averaging word embeddings. UniGDD (Gao et al., 2022) and DIALKI (Wu et al., 2021) systems also consider document-grounded generation but are limited by context length or investigated for task formulations different from ours.

The main limitation of such systems is the lack of training datasets. CMU DoG (Zhou et al., 2018) was proposed for grounding-based movie conversations but contains few documents which com-

[1] https://elicit.org
[2] https://www.scholarcy.com
[3] https://www.explainpaper.com
[4] https://chat.openai.com/
[5] https://www.chatpdf.com
[6] https://www.xmagic.ai
[7] *PaperPersiChat* is running online on http://www.PaperPersiChat.tech

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 584–587
September 11–15, 2023. ©2023 Association for Computational Linguistics

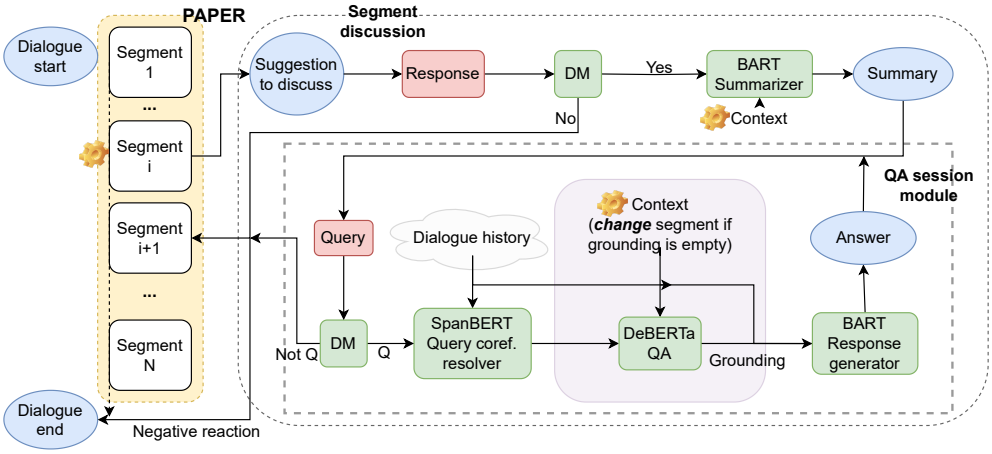


Figure 1: The architecture of *PaperPersiChat*. The discussion of the single i-th segment is demonstrated. User input is shown in red frames, chatbot answers in blue and the trainable pipeline submodules in green. DM refers to the Dialogue Management submodule. The light-purple part of the Question Answering (QA) system runs in a loop over segments while the retrieved grounding is empty. The current segment texts are labelled with a gear icon.

plicates generalization. Larger datasets, such as the Wizard of Wikipedia (Dinan et al., 2019), are labeled more roughly and are not tailored to the scientific domain either. At the same time, recent approaches employ synthetic training datasets collected via ChatGPT (Askari et al., 2023). We follow this idea and propose a dataset collection process, which we used to train our pipeline later.

3 System Overview

Figure 1 shows the general architecture of the *PaperPersiChat* system. The chatbot discusses paper segments step by step, with each segment containing one or several sections of the paper. The dialogue ends when all segments have been discussed or too much negative feedback has been received.

For each segment, the chatbot firstly suggests discussing it. If the suggestion is accepted, it provides a short summary and proceeds to the question-answering session. Otherwise, the chatbot moves to the discussion of the succeeding segment. For each question query, the QA module attempts to extract grounding from the current segment. However, if this fails, it continues to look over all other segments. In case when QA module can't find an answer in the entire paper, it informs the user about that. If the user's query is not a question, the system moves to the discussion of the following segment. Further details are described in Section 5.

4 Data

There is a lack of publicly available datasets for training the dialogue systems with scientific text grounding. Since manual markup requires significant resources, we constructed the dataset automatically. As the source, we used 63,321 computer science papers from the Semantic Scholar Open Research Corpus published at top science conferences between 2000 and 2021. We utilized its subset to collect our dataset, which consists of two parts: instances collected via OpenAI's Davinci or ChatGPT [8].

The Davinci model processed complex instructions and tried to produce the part of the dialogue related to the whole segment discussion part (see Figure 1). In this way, we collected 3,588 raw outputs and each of them was processed further into a summary and dialogue turns. All these summaries were used to train the summarization submodule. Further filtering was done to remove unparsed outputs, short dialogues and dialogues with inconsistent structure (including incorrect speaker order). This yielded a set of 2,817 dialogues that were used to train the models from the QA session module. To construct qualitative dialogues for QA, and also to manage the inputs of the dialogue participants, we used two ChatGPT models talking to each other. The resulting dataset totals 2,817 dialogues produced by Davinci and 8,787 dialogues produced by ChatGPT, with an average of four turns per dialogue. We have made this dataset publicly available via https://huggingface.co/datasets/ai-forever/paper_persi_chat.

[8] https://platform.openai.com/docs/models

5 Submodule Details

This section provides details about submodules of the *PaperPersiChat* pipeline.

Dialogue Discourse Flow Management (DM) This component is employed to classify the user's reaction and navigate to the pertinent pipeline steps. It is composed of two models: a dialogue discourse parser and an agreement classifier. To acquire the discourse parser, we trained the parser proposed by Shi and Huang (2019) from scratch on CDSC (Zhang et al., 2017). To classify the last relation in the dialogue, the pipeline passes the last ten utterances of the dialogue history as the parser input. In this pipeline, we consider only the following dialogues acts: Agreement, Disagreement, Question and Negative Reaction. Since the discourse parser has broad classes and can not distinguish well "yes" and "no" cases, an auxiliary agreement classifier based on the SBERT model (Reimers and Gurevych, 2019) was trained using the open-source CDSC and IAC (Walker et al., 2012) corpora.

Summarization This component briefly summarizes the current segment text. Here, the BART model (Lewis et al., 2020) was selected for generation due to its state-of-the-art performance in similar tasks. We trained it using the 3,204 samples and started from DistilBART[9]. Evaluation of the model on 356 test samples from our dataset yielded ROUGE-1 score of 51.43, ROUGE-2 score of 30.80, and ROUGE-L score of 40.97.

Coreferences Resolver We utilized the pretrained SpanBERT from AllenNLP Framework [10]. Here, we use the concatenation of the user query and four last utterances from the dialogue history as the SpanBERT input and replace entities with coreferences in the query to get the final QA input.

Extractive QA The QA module extracts the most relevant text sufficient to answer the user's question, namely grounding. To ensure the class balance as in SQuAD2.0, we added more hard negative examples (questions with empty answers) by selecting questions generated for similar segments. The similarity scores were calculated as the cosine distance between Specter [11] embeddings. As the final model, we fine-tuned DeBERTA-v3 [12].

Response Generation This component generates the target response text based on the query, dialogue history and grounding text extracted by DeBERTa. We conducted experiments for BART (Lewis et al., 2020) and DialoGPT (Zhang et al., 2020) for two options of groundings: extracted by the pretrained or by the fine-tuned DeBERTa.

To construct model inputs, we concatenated query, dialogue history and grounding via special separation tokens.The BART model trained using groundings from the fine-tuned DeBERTa yielded the best results, with a ROUGE-1 of 61.71 and a BLEU-1 of 50.3 on our test set. In comparison, the BART model trained using groundings extracted by the pretrained QA model got ROUGE-1 of 49.41 and the best DialoGPT model got 61.42.

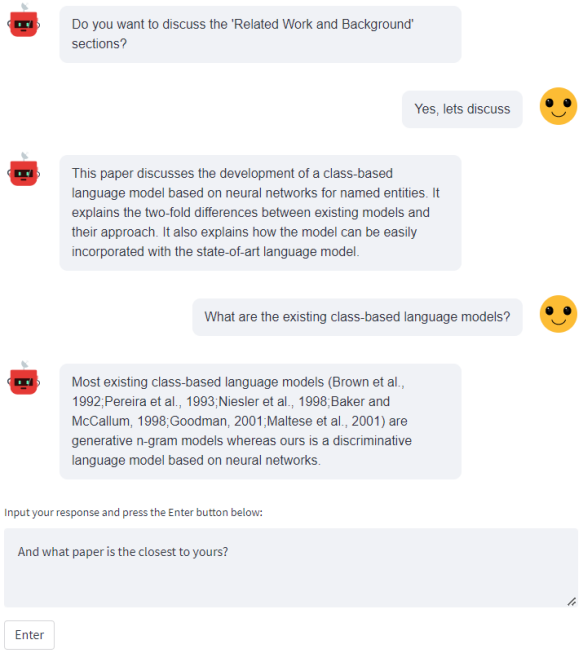


Figure 2: Screenshot of the interface window with a dialogue example generated using *PaperPersiChat*.

6 User Interface

Figure 2 depicts a screenshot of a sample dialogue between the user and the proposed chatbot. Here, the bot suggests discussing the Related Work section; the user agrees and the system moves to the QA session. If during the session the bot cannot find a grounding for a question, it informs the user that there is not enough information in the paper. The QA session continues until the user ceases asking questions, after which the dialogue advances to the next section.

During the dialogue, the user enters his message in the corresponding field and then the dialogue

[9] https://huggingface.co/sshleifer/distilbart-cnn-12-6

[10] https://demo.allennlp.org/coreference-resolution/coreference-resolution

[11] https://huggingface.co/allenai/specter

[12] https://huggingface.co/deepset/deberta-v3-base-squad2

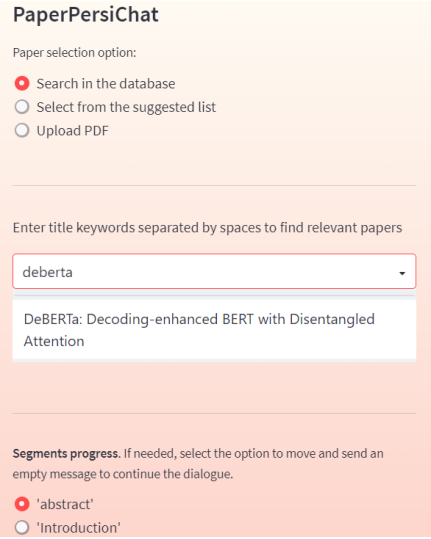


Figure 3: Screenshot of the chat settings.

history above the input field is updated with the addition of the last user's query and the bot's response. Then the process repeats.

The auxiliary menu to the left, illustrated in Figure 3), assists the user in selecting a paper for discussion, switching to another segment (via radio buttons), or clearing the dialogue history. Additionally, the menu provides short instructions to facilitate communication with the bot. There are several options to select a paper for discussion:

- Select any paper from our dataset (63,321 papers) by searching. For this option, the user just needs to enter a few keywords separated by a space and press the "Search" button.
- Select a paper from a suggested sublist.
- Upload new paper in the PDF format.

7 Conclusion

We have presented *PaperPersiChat*, chatbot based only on open-source models and capable of engaging in conversations about scientific papers. For each paper segment, the bot offers the user an opportunity to get a summary and moves to the QA session mode in the case of agreement. The dialogue flow is controlled by a discourse analyzer. We also presented a novel dataset to facilitate the development of similar systems. Future work includes refining individual submodules and dialogue management to promote greater flexibility.

References

Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. 2023. Generating synthetic documents for cross-encoder re-rankers: A comparative study of chatgpt and human experts. *CoRR*, abs/2305.02320.

Yuanyuan Cai, Min Zuo, and Haitao Xiong. 2022. Modeling hierarchical attention interaction between contexts and triple-channel encoding networks for document-grounded dialog generation. In *Frontiers of Physics*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR 2019*. OpenReview.net.

Chang Gao, Wenxuan Zhang, and Wai Lam. 2022. Unigdd: A unified generative framework for goal-oriented document-grounded dialogue. In *Proceedings of ACL 2022*.

Amelia Glaese et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *CoRR*, abs/2209.14375.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL 2020*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3980–3990.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Processings of AAAI, 2019*.

Romal Thoppilan et al. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of LREC 2012*, pages 812–817.

Zeqiu Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. DIALKI: knowledge identification in conversational systems through dialogue-document contextualization. In *Proceedings of EMNLP, 2021*, pages 1852–1863.

Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Eleventh International AAAI Conference on Web and Social Media*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of ACL 2020*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. In *Proceedings of EMNLP, 2018*.

FurChat: An Embodied Conversational Agent using LLMs, Combining Open and Closed-Domain Dialogue with Facial Expressions

Neeraj Cherakara, Finny Varghese, Sheena Shabana, Nivan Nelson
Abhiram Karukayil, Rohith Kulothungan, Mohammed Afil Farhan,
Birthe Nesset, Meriam Moujahid, Tanvi Dinkar, Verena Rieser, Oliver Lemon†
Interaction Lab, Heriot-Watt University ; †Alana AI
{nc2025, fv2002, ss2022, nn2023, ak2120, rk2065, mf2034, bn25
mm470, t.dinkar, v.t.rieser, o.lemon}@hw.ac.uk

Abstract

We demonstrate an embodied conversational agent that can function as a receptionist and generate a mixture of open and closed-domain dialogue along with facial expressions, by using a large language model (LLM) to develop an engaging conversation. We deployed the system onto a Furhat robot, which is highly expressive and capable of using both verbal and nonverbal cues during interaction. The system was designed specifically for the National Robotarium to interact with visitors through natural conversations, providing them with information about the facilities, research, news, upcoming events, etc. The system utilises the state-of-the-art GPT-3.5 model to generate such information along with domain-general conversations and facial expressions based on prompt engineering.

1 Introduction

The progress in robotics and artificial intelligence in recent decades has led to the emergence of robots being utilized beyond their conventional industrial applications. Robot receptionists are designed to interact with and assist visitors in various places like offices, hotels, etc. by providing information about the location, services, and facilities. The appropriate use of verbal and non-verbal cues is very important for the robot's interaction with humans (Mavridis, 2015). Most research in the field has been mainly focused on developing domain-specific conversation systems, with little exploration into open-domain dialogue for social robots.

Conventional agents are often rule-based, which means they rely on pre-written commands and keywords that are pre-programmed. This limits the interaction with humans to little or no freedom of choice in answers (Tudor Car et al., 2020). The advancement of large language models (LLMs) in the past year has brought an exciting revolution in the field of natural language processing. With the development of models like GPT-3.5[1], we have seen unprecedented progress in tasks such as question-answering and text summarization (Brown et al., 2020). However, a question remains about how to successfully leverage the capabilities of LLMs to create systems that can go from closed domain to open, while also considering the embodiment of the system.

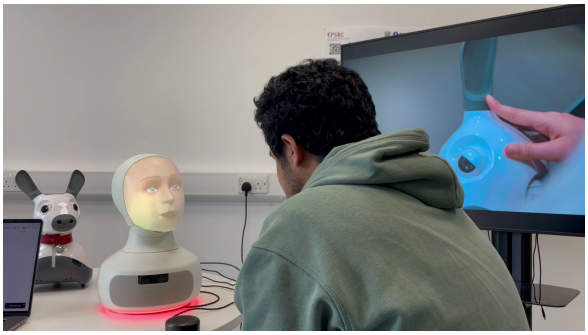


Figure 1: A user interacting with the FurChat System.

In this work, we present FurChat[2], an embodied conversational agent that utilises the latest advances in LLMs to create a more natural conversational experience. The system seamlessly combines open and closed-domain dialogues with emotive facial expressions, resulting in an engaging and personalised interaction for users. The system was initially designed and developed to serve as a recep-

[1] https://platform.openai.com/docs/models/gpt-3-5

[2] A demonstration video of the system is available here.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 588–592
September 11–15, 2023. ©2023 Association for Computational Linguistics

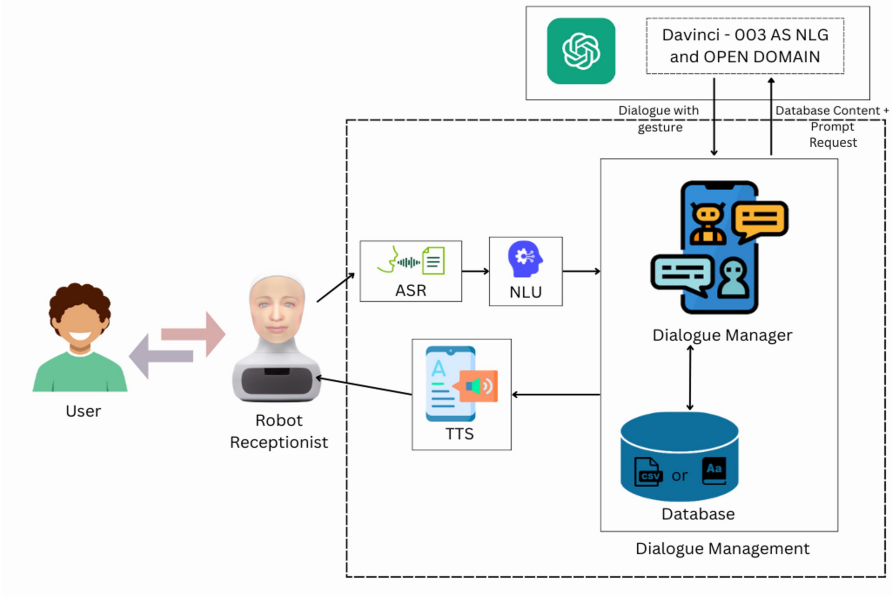


Figure 2: System Architecture of the current FurChat system.

tionist for the National Robotarium, in continuation of the multi-party interactive model developed by Moujahid et al. (2022b), and its deployment shows promise in other areas due to the LLMs versatile capabilities. As a result, the system is not limited to the designated receptionist role, but can also engage in open-domain conversations, thereby enhancing its potential as a multifunctional conversational agent. We demonstrate the proposed conversational system on a Furhat robot (Al Moubayed et al., 2013) which is developed by the Swedish firm Furhat Robotics[3]. With FurChat, we demonstrate the possibility of LLMs for creating a more natural and intuitive conversation with robots.

2 Furhat Robot

Furhat is a social robot created by Furhat Robotics. To interact with humans naturally and intuitively, the robot employs advanced conversational AI and expressive facial expressions. A three-dimensional mask that mimics a human face is projected with an animated face using a microprojector (Al Moubayed et al., 2013). A motorised platform supports the robot's neck and head, allowing the platform's head to spin and node. To identify and react to human speech, it has a microphone array and speakers. Due to the human-like appearance of Furhat, it is prone to the uncanny valley effect (Ågren and Silvervarg, 2022).

[3] https://furhatrobotics.com/

3 System Architecture

As shown in Figure 2, the system architecture represents a conversational system that enables users to interact with a robot through spoken language. The system involves multiple components, including automatic speech recognition (ASR) for converting user speech to text, natural language understanding (NLU) for processing and interpreting the text, a dialogue manager (DM) for managing the interaction flow, and natural language generation (NLG) powered by GPT-3.5 for generating natural sounding responses(Ross et al., 2023). The generated text is then converted back to speech using text-to-speech (TTS) technology and played through the robot's speaker to complete the interaction loop. The system relies on a database to retrieve relevant data based on the user's intent.

3.1 Speech Recognition

The current system uses the Google Cloud Speech-to-Text[4] module for ASR. This module, which transcribes spoken words into text using machine learning algorithms, is integrated into the system by default through the Furhat SDK.

3.2 Dialogue Management

Dialogue Management consists of three submodules: NLU, DM and a database storage. The NLU component analyses the incoming text from

[4] https://cloud.google.com/speech-to-text

the ASR module and, through machine learning techniques, breaks it down into a structured set of definitions (Otter et al., 2021). The FurhatOS provides an NLU model to classify the text into intents based on a confidence score. We provide multiple custom intents for identifying closed-domain intents using Furhat's NLU capabilites.

The in-built dialogue manager in the Furhat SDK is responsible for maintaining the flow of conversation and managing the dialogue state based on the intents identified by the NLU component. This module is responsible for sending the appropriate prompt to the LLM, receiving a candidate response from the model, and subsequent processing of the response to add in desired facial gestures (see §3.4).

An open challenge faced by present-day LLMs is the *hallucination of nonfactual content*, which potentially undermines user trust and raises concerns of safety. While we cannot fully mitigate hallucinated content in the generated responses, in order to tone-down this effect, we create a custom database following suggestions from Kumar (2023). We do so by manually web-scraping the website of the National Robotarium[5]. The database consists of a dictionary of items with the intents as keys and scraped data as values. When an appropriate intent is triggered, the dialogue manager accesses the database to retrieve the scraped data, which is then sent with the prompt (further details in §3.3)) to elicit a response from the LLM.

3.3 Prompt engineering for NLG

The NLG module is responsible for generating a response based on the request from the dialogue manager. Prompt engineering is done to elicit an appropriate sounding response from the LLM, which generates natural dialogue that results in engaging conversations with humans. The current system uses `text-davinci-003`, which is one of the most powerful models in the GPT-3.5 series and it is priced at $0.0200 per 1000 tokens.

Producing relevant responses was achieved using the combined technique of few-shot learning and prompt engineering, which enabled us to try different variations in techniques and produce a variety of output by the LLM.

During prompt engineering, the personality of the robot and the context of the application are described, along with the past few dialogue histories and scraped data from the database in a particular response format. Moreover, the prompt engineering methodology involves using the LLM to generate an appropriate emoticon based on the conversation. In the context of emotional expression during an interaction, selecting an appropriate emoticon depends on understanding the underlying emotions being conveyed by the visitors and adhering to the display rules of the specific social situation. If the dialogue reflects joy or humor, a happy facial gesture might be fitting. On the other hand, if the conversation conveys empathy or sadness, a sad face could be more suitable. These emoticons are then integrated with the robot's facial gestures to generate facial expressions (see §3.4), thereby enabling a text-based LLM to integrate in the embodied Furhat robot. The explicit specification of the personality and context in the prompt aids in creating a natural conversation between the robot and the human that is coherent and relevant to the topic. The sample format of the prompt used is as follows:

This is a conversation with a robot receptionist, <Robot Personality>, <Data from the Database>, <Dialogue history>, <Response Format along with sample emoticons>.

3.4 Gesture Parsing

The Furhat SDK offers a range of built-in facial gestures that can be enhanced by custom facial gestures that meet specific needs. The latest GPT models have the ability to recognise emotions and sentiments from text, which is used in the system (Leung et al., 2023). Rather than simply recognising sentiments in the text, the model is tasked with generating appropriate emotions for the conversation from the text After receiving the response from the model, the matched conditional clause in the dialogue manager will trigger an expression from the pre-developed set of gestures, which will be triggered along with the generated speech.

3.5 Text-To-Speech Generation

For converting the text to speech, the Amazon Polly[6] service is used. This service is available within FurhatOS by default.

4 Example Dialogue

An example dialogue between a human and the robot is illustrated in Table 1.

[5] https://thenationalrobotarium.com/

[6] https://docs.aws.amazon.com/polly/

At the outset, the robot remains in an idle state. Once the user enters the vicinity of the robot, the conversation begins.

R: [robot] Hello, I am the Receptionist here at the National Robotarium. Would you like to know about this facility? <Robot smiles>

U: [user] Yes, tell me about this facility.

R: [robot] Sure. <Robot smiles>

R: [robot] The National Robotarium is a state-of-the art research facility located on the Heriot-Watt campus in Edinburgh. . . .

U: [user] That's great. Is there any ongoing healthcare researches here

R: [robot] Yes. <Robot smiles> The National Robotarium is part of a multimillion-pound collaborative project called SPRING and is focused on developing socially assistive robots for healthcare settings. SPRING will focus on supporting elderly patients and improving psychological well-being and the relationship between patients and hospital professionals. <Robot nod>

U: [user] That's nice. Can you tell a joke?

R: [robot] <Robot smiles> Sure, Why did the robot go to the bar? <Robot winks> To get a screwdriver!

U: [user] Can you suggest me a movie about robots?

R: [robot] Sure thing! <Robot smiles> How about "I, Robot"? It's a classic that deals with the exploration of robotics and AI. <Robot nod>

Table 1: Sample Conversation between the user and the robot. For a full system description, please refer to §3.

5 Conclusions and Future Work

We demonstrate FurChat, an embodied conversational agent with open and closed domain dialogue generation and facial expressions generated through LLMs, on a social robot in a receptionist environment. The system is developed by integrating the state-of-the-art GPT-3.5 model on top of the Furhat SDK. The proposed system uses a one-to-one interaction method of communication with the visitors. We plan on extending the system to handle multi-party interaction (Moujahid et al., 2022a; Addlesee et al., 2023; Lemon, 2022; Gunson et al., 2022), which is an active research topic in developing receptionist robots. It is also crucial to address the issue of hallucination from the large language model and this problem can be mitigated by fine-tuning the language model and directly generating conversations from it without relying on any NLU components which we plan to implement in the future.

We plan to showcase the system on the Furhat robot during the SIGDIAL conference to all the attendees and show them the capabilities of using LLMs for dialogue and facial expression generation as described in this paper.

Acknowledgements

This research has been funded by the EU H2020 program under grant agreement no. 871245 (http://spring-h2020.eu/) and the EPSRC project 'Gender Bias in Conversational AI' (EP/T023767/1).

References

Angus Addlesee, Weronika Sieinska, Nancie Gunson, Daniel Hernandez Garcia, Christian Dondrup, and Oliver Lemon. 2023. Data collection for multi-party task-based dialogue in social robotics. In *IWSDS 2023: International Workshop on Spoken Dialogue Systems Technology*.

Isabella Ågren and Annika Silvervarg. 2022. Exploring humanlikeness and the uncanny valley with furhat. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–3.

Samer Al Moubayed, Jonas Beskow, and Gabriel Skantze. 2013. The furhat social companion talking head. In *INTERSPEECH*, pages 747–749.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nancie Gunson, Daniel Hernández García, Weronika Sieińska, Christian Dondrup, and Oliver Lemon. 2022. Developing a social conversational robot for the hospital waiting room. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1352–1357. IEEE.

Krishna Kumar. 2023. Geotechnical Parrot Tales (GPT): Overcoming GPT hallucinations with prompt engineering for geotechnical applications. *arXiv preprint arXiv:2304.02138*.

Oliver Lemon. 2022. Conversational AI for multi-agent communication in Natural Language. *AI Communications*, 35(4):295–308.

John Kalung Leung, Igor Griva, William G Kennedy, Jason M Kinser, Sohyun Park, and Seo Young Lee. 2023. The application of affective measures in text-based emotion aware recommender systems. *arXiv preprint arXiv:2305.04796*.

Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35.

Meriam Moujahid, Helen Hastie, and Oliver Lemon. 2022a. Multi-party interaction with a robot receptionist. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '22, page 927–931. IEEE Press.

Meriam Moujahid, Bruce Wilson, Helen Hastie, and Oliver Lemon. 2022b. Demonstration of a robot receptionist with multi-party situated interaction. In *Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1202–1203. IEEE.

Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2021. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624.

Steven I. Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D. Weisz. 2023. The programmer's assistant: Conversational interaction with a large language model for software development. IUI '23, page 491–514, New York, NY, USA. Association for Computing Machinery.

Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin-Leng Theng, and Rifat Atun. 2020. Conversational agents in health care: Scoping review and conceptual analysis. *J Med Internet Res*, 22(8):e17158.

Towards Breaking the Self-imposed Filter Bubble in Argumentative Dialogues

Annalena Aicher
Ulm University, Germany
NAIST, Japan
annalena.aicher@uni-ulm.de

Daniel Kornmüller
Wolfgang Minker
Ulm University, Germany

Stefan Ultes
University of Bamberg, Germany

Yuki Matsuda
Keiichi Yasumoto
NAIST, Japan

Abstract

Human users tend to selectively ignore information that contradicts their pre-existing beliefs or opinions in their process of information seeking. These "self-imposed filter bubbles" (SFB) pose a significant challenge for cooperative argumentative dialogue systems aiming to build an unbiased opinion and a better understanding of the topic at hand.

To address this issue, we develop a strategy for overcoming users' SFB within the course of the interaction. By continuously modeling the user's position in relation to the SFB, we are able to identify the respective arguments which maximize the probability to get outside the SFB and present them to the user. We implemented this approach in an argumentative dialogue system and evaluated in a laboratory user study with 60 participants to show its validity and applicability. The findings suggest that the strategy was successful in breaking users' SFBs and promoting a more reflective and comprehensive discussion of the topic.

1 Introduction

Spoken dialogue systems are getting increasingly popular, especially as they enable easy access to requested information from online sources, such as search engines or social media platforms. Specifically, with regard to more complex interactions, two important phenomena can be observed that can result in information bias.

On the one hand, due to filter algorithms, information content is selected based on previous online behavior, which leads to cultural/ideological bubbles, the so-called "Filter Bubbles" (Pariser, 2011). On the other hand, Nickerson (1998) points out that users who are confronted with controversial topics tend to focus on a "biased subset of sources that repeat or strengthen an already established or convenient opinion." This user behavior leads to the so-called "Self-imposed Filter Bubbles" (SFB) (Ekström et al., 2022; Aicher et al., 2022b) and "echo chambers" (Quattrociocchi et al., 2016; Anand, 2021; Donkers and Ziegler, 2021). Both are manifestations of "confirmation bias", a term typically used in psychological literature. These phenomena are mutually dependent according to Lee (2019) as the SFB is reinforced and perpetuated due to algorithmic filters delivering content aligned with presumed interests based on search histories. Moreover, Bakshy et al. (2015) claim that studies have shown that individual choice has even more of an effect on exposure to differing perspectives than "algorithmic curation". In this paper, we focus on the second phenomenon, namely the user's SFB regarding a controversial topic during the interaction with an argumentative dialogue system (ADS). Building upon the work of Aicher et al. (2022b, 2023), we model the user's SFB using the following four main dimensions: *Reflective User Engagement (RUE), Personal Relevance (PR), True Knowledge (TK)* and *False Knowledge (FK)*.

The concept of *RUE* encapsulates the user's critical thinking, building upon the definition established in our prior work (Aicher et al., 2021a). On the other hand, *PR* pertains to the individual user's assessment of the significance of subtopics, further on called "clusters", in relation to the overarching topic of discussion. *True Knowledge (TK)* is characterized as the information already possessed by the user on a particular topic. Conversely, *False Knowledge (FK)* entails the user's false beliefs and misinformation on the respective topic. Based upon these dimensions we have the ability to construct a model for assessing the likelihood of a user being caught within an SFB. In order to achieve this, we ascertain the user's position along these four dimensions and consistently update it throughout the course of the dialogue. Building upon SFB-Model we 1) introduce a rule-based system policy to break the user's SFB during an ongoing interaction and 2) validate our policy in a laboratory study by comparing it to a user-interest-driven system

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 593–604
September 11–15, 2023. ©2023 Association for Computational Linguistics

policy.

The remainder of this paper is as follows: Section 2 gives an overview of related literature, followed by a description of the underlying SFB-Model and our proposed rule-based SFB-breaking policy in Section 3. Section 4 discusses an exemplary integration of our model/policy in an ADS, which is evaluated in a laboratory study described in Section 5. Section 6 covers the respective study results, followed by a discussion of the former and study limitations in Sections 6 and 8. We close with a conclusion and a brief discussion of future work in Section 9.

2 Related Work

In the following, we provide a brief overview of the existing literature on the main aspects of the work presented herein, *Confirmation Bias and Self-imposed Filter Bubbles* and *Argumentative Dialogue Systems*.

2.1 Confirmation Bias and Self-imposed Filter Bubbles

As previously pointed out, a central issue in the process of opinion building is the phenomenon known as "confirmation bias". This bias refers to the tendency of users to seek or interpret evidence in ways that align with their existing beliefs, expectations, or hypotheses (Nickerson, 1998). Given our goal of achieving a well-founded and unbiased exploration of information, we are determined to counteract the user's inclination to focus solely on information that confirms their preexisting beliefs (Allahverdyan and Galstyan, 2014).

To address this challenge, Huang et al. (2012) propose the utilization of computer-mediated counter-arguments within decision-making processes. Additionally, Schwind and Buder (2012) consider preference-inconsistent recommendations as a promising approach to stimulate critical thinking. However, given our cooperative approach and the objective of maintaining the user's motivation to explore arguments without bias, introducing an excessive number of counter-arguments could potentially lead to undesirable negative emotional consequences, such as annoyance and confusion (Huang et al., 2012).

In order to identify a means of mitigating these consequences, it is crucial to consider how a genuine and profound critical reflection can be stimulated. When users engage in critical thinking in a *weak sense*, this implies contemplating positions that differ from their own (Mason, 2007), but often involves a tendency to defend their own viewpoint without thorough introspection (Paul, 1990). Critical thinking in a *strong sense* involves reflecting on one's own opinions as well, which aligns with our objective. However, the substantial energy and effort (Gelter, 2003) required for this robust critical reflection are frequently lacking due to a deficiency in individuals' inherent *need for cognition* (Maloney and Retanal, 2020). Given users' tendency to defend their own views (Paul, 1990), a system that confronts them with opposing viewpoints might not necessarily foster critical reflection; on the contrary, it could lead to a reinforcement of their existing stance. Hence, there is a need for an intelligent system capable of adjusting the frequency, timing, and selection of counter-arguments (Huang et al., 2012). To the best of our knowledge we are the first to provide such a system, which integrates a model to determine the user's Self-imposed Filter Bubble (SFB) and adapts its strategy accordingly. This adaptation aims to identify the most suitable arguments and still maintaining the user's interest, ensuring a well-balanced exploration of viewpoints.

In contrast to Del Vicario et al. (2017), who study online social debates and try to mathematically model the related polarization dynamics, we define a model for this "seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations or a hypothesis in hand" (Nickerson, 1998) consisting of four dimensions building upon our previous work (Aicher et al., 2022b, 2023). The respective dimensions are based on a well-established framework in persuasion research, the " Elaboration Likelihood Model" (ELM) (Petty et al., 2009).

2.2 Argumentative Dialogue Systems

Within this paper we define a system policy aiming to help users overcome their SFBs in a cooperative argumentative dialogue. Argumentative dialogue systems (ADS) enable users to engage in information-seeking and to explore pro and con arguments on a controversial topic by accessing large-scale argumentation structures and assist in a well-founded opinion building (Waheed et al., 2021; Aicher et al., 2021b,a, 2023). The "ability to engage in argumentation is essential for humans to understand new problems, to perform scientific

reasoning, to express, to clarify and to defend their opinions in their daily lives" (Palau and Moens, 2009) and thus, enables to reflect controversial topics critically. A consensual dialogue is much more likely to resolve diverging perspectives on evidence and repair incorrect, partial, and subjective readings of evidence than a persuasive one (Villarroel et al., 2016). Hence, it is crucial for the argumentative dialogue system, in which our SFB-Model is embedded, that it does not try to persuade or win a debate against a user.

Most approaches to human-machine argumentation utilize different models to structure the interaction and are embedded in a competitive, persuasive scenario. For instance, Slonim et al. (2021) introduced the IBM Debater, which is an autonomous debating system that can engage in a competitive debate with humans via natural language. Another speech-based approach was introduced by Rosenfeld and Kraus (2016), presenting a system based on weighted Bipolar Argumentation Frameworks (wBAG). Arguing chatbots such as Debbie (Rakshit et al., 2017) and Dave (Le et al., 2018) interact via text with the user. A menu-based framework that incorporates the beliefs and concerns of the opponent was presented by Hadoux et al. (2022). In the same line, Chalaguine and Hunter (2020) used a previously crowd-sourced argument graph and considered the concerns of the user to persuade them. Another introduced persuasive prototype chatbot is tailored to convince users to vaccinate against COVID-19 using computational models of argument (Chalaguine and Hunter, 2021). As pointed out in Subsection 2.1 in contrast to those persuasive approaches we chose collaborative exploration of arguments, enabling users to express their preferences and thus providing a more suitable basis than the previously mentioned, competitive ADS.

3 Self-imposed Filter Bubble Model

In the following section we will give a short overview on the SFB-Model we adapted to and its respective dimensions. This serves as a basis for our system's SFB-breaking policy introduced in Subsection 3.3.

3.1 SFB-Model Dimensions

We adapted the SFB-Model introduced by Aicher et al. (2022b) which is motivated by the "Elaboration Likelihood Model" (ELM) (Petty et al., 2009). As already mentioned, it incorporates of four dimensions, which span a four-dimensional space to describe the user's SFB: *Reflective User Engagement (RUE), Personal Relevance (PR), True Knowledge (TK)* and *False Knowledge (FK)*.

The *Reflective User Engagement (RUE)* describes the critical-thinking and open-mindedness demonstrated by the user. It takes into account the polarity and number of heard arguments. This can be mapped onto the request for more information, either on the pro or con side of the topic of the discussion. Thus, it measures how balanced the user is exploring a topic. The *RUE* has first been introduced by Aicher et al. (2021a), to whose work we refer to for details of its calculation.

The *Personal Relevance (PR)* refers to the user's individual assessment of how relevant a cluster is with regard to the topic of the discussion. The greater the relevance a cluster holds for a user, the stronger their inclination to delve into the corresponding arguments associated with it. As this is impossible to ascertain through implicit methods, the *Personal Relevance (PR)* is explicitly queried within the dialogue when transitioning to a new cluster, with respect to the previous cluster.

The *True Knowledge (TK)* serves as a measure for the information gain and is defined as the new information the user is provided with by talking to the system. It can be determined by comparing the total information provided by the system and the information, which is already known to the user. For its determination, the user is required to provide feedback on each known argument. For each cluster, this number of known arguments is subtracted from the total number of arguments heard within the cluster. As we want the user to explore as much information as possible, a high *TK* increases the chance to explore other aspects and viewpoints. Thus, the bigger the *TK* of the users, the more unlikely they find themselves in an SFB.

The concept of "False Knowledge (FK)[1]" pertains to inaccurate information held by a user regarding a specific topic. When a user possesses false beliefs about specific clusters, it increases

[1]Regarding the terminology, please note that the term "False Knowledge" was chosen to facilitate a simplified three-dimensional representation, wherein the dimensions of "True" and "False Knowledge" are merged into the single dimension of "Knowledge". This choice is intended solely for the purpose of simplified illustration as the actual calculation occurs within a four-dimensional space. Without loss of generality the information stored in the system's database is defined as factually accurate, thereby classifying information contradicting it as wrong.

the probability to be caught in an SFB and fosters reluctance toward conflicting information and viewpoints. Likewise to the "True Knowledge", the "False Knowledge" is determined by the user indicating that they consider an argument to be factually incorrect.

3.2 SFB-Model

Argumentative discussions are complex and consist of a lot of different cluster, which contain arguments referring to the same content-related aspects. For each of these clusters, a corresponding SFB vector $\overrightarrow{sfb_k} = (pr_k, r_k, tk_k, fk_k)^T,\ k \in \mathbb{N}$ is defined, contributing to the overall SFB vector $\overrightarrow{SFB_k}$ for the entire discussion topic. It is important to differentiate between the SFB and the SFB-vector of a user (refer to Figure 1). The SFB-vector is conceptualized as a vector originating from the coordinate system's origin and terminating at the user's position in the four-dimensional space. The SFB, on the other hand, constitutes the region within the four-dimensional space that signifies a specific probability of users to be caught in their SFB. Figure 1 presents an illustration[2] of two positions of this vector, and the respective SFB (dark blue geometric shape) for a single cluster. As it is very difficult to establish precise boundaries of the SFB, we establish a probability denoting a user's position within or outside the SFB. A short SFB-vector (dashed red arrow) corresponds to a high probability of the user to be caught within the SFB. Conversely, a large SFB vector (continuous green arrow) that extends further beyond the SFB diminishes the likelihood of the user to be caught in the SFB. The overall SFB vector $\overrightarrow{SFB} = (PR, RUE, TK, FK)^T$, consists of the overall cluster values for each dimension, derived from a weighted mean calculation (Aicher et al., 2023).

3.3 SFB-breaking policy

Building upon the model described in Subsection 3.2, we propose a rule-based system policy with the objective of breaking the user's SFB. Utilizing data from a prior crowd-sourcing user study, we investigated how SFB dimensions changed under two distinct system policies. The first policy, as outlined in Section 5, follows the interest-based approach, selecting arguments based on the estimation of the user's greatest interests. The second policy involves the random presentation of arguments from the remaining set. The calculated averages across all participants were utilized as benchmark values for identifying regions where there is a higher probability of being caught in an SFB (very high probability = interest average; medium probability = random average).

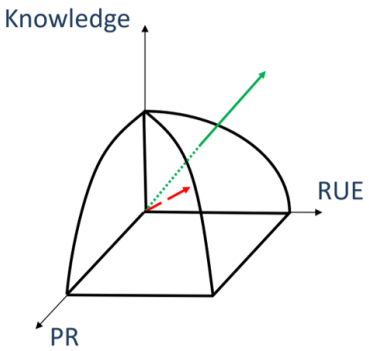


Figure 1: Schematic sketch of a clusterwise SFB-vector and SFB for a cluster k. The probability of an SFB is very high in proximity to the origin and/or when a dimension approaches a value close to zero. As a four-dimensional space is challenging to visualize, we consolidate the dimensions of *TK* and *FK* into the *Knowledge* dimension. The red dashed vector indicates the position of a user within the SFB. The green continuous arrow indicates the position of a user outside the SFB.

Given that *PR* and *FK* cannot be ascertained beforehand but only in hindsight, the rule-based policy focuses on maximizing the *RUE* and *TK* dimension, which can be computed in advance. If the values for *PR* or *fK* deteriorate (become smaller) after introducing a new argument, we assign a greater weight to the associated cluster and respective arguments to counteract this.

To ensure logical coherence, it is important that potential argument candidates are logically connected to the requested argument, either through sibling relationships or by sharing the highest degree of overlap in their respective cluster affiliations. Once candidates are identified, they are evaluated against the user-selected argument in terms of the corresponding *RUE* and *TK* dimensions. Subsequently, the argument with the maximum values in these dimensions is presented. In cases where the system selects an argument different from the user's choice, the system response includes an explanation such that the user understands the system's choice.

Following an initialization phase (first five argu-

[2]Please note that this illustration serves solely explanatory purposes, and thus is reduced to a three dimensional space (by merging *TK* and *FK* and that the actual form and structure of the SFB may deviate.

ment requests) aimed at detecting and rewarding shifts in users' exploration behaviors, the user's current SFB-vector is compared to the data-based SFB-margins (interest, random) after each interaction turn. If the SFB-vector falls within the first area (below the interest margin), the ADS will consistently opt to select the best available argument in each turn. When the SFB-vector is situated within the second region (above the interest margin, below the random margin), a decision is made based on recent changes in the SFB vector over the preceding three interaction turns. This determines whether the system offers an "SFB-breaking" argument or the requested argument. If the SFB-vector surpasses the random margin, the ADS presents the requested argument, contingent upon the precondition that the absolute value of the SFB-vector did not decrease in the preceding turn.

4 SFB-Model and Policy Integration into the ADS

In the following, the relevant components of the ADS, namely the knowledge base and dialogue model, focusing on the exemplary integration of our SFB-Model. In order to combine the presented model with existing argument mining approaches, ensuring its adaptability with respect to discussed topics, we adhere to the bipolar argument annotation scheme introduced Stab and Gurevych (2014)[4]. This scheme encompasses argument components (nodes), structured in the form of bipolar argumentation trees. The overall topic represents the root node in the graph. We consider two relationships between these nodes: *support* or *attack*. Each component, excluding the root node (which has no relation), has exactly one unique relation to another component. This results in a non-cyclic tree structure, wherein each node, or "parent", is supported or attacked by its "children". If no children exist, the node is a leaf and marks the end of a branch.

Move	Description	SFB Dim
why_{pro}	Request pro argument	r_k, tk_k
why_{con}	Request con argument	r_k, tk_k
$suggest$	Suggest any argument	r_k, tk_k
$prefer$	Prefer current argument	r_k
$reject$	Reject current argument	r_k
$know$	Current argument is already known	$tk_{k,i}$[3]
$false$	Current argument is incorrect	fk_k
$exit$	Terminates the conversation	

Table 1: Description of potential user actions along with their corresponding impact on SFB dimensions.

Furthermore, the SFB-Model necessitates semantically clustered arguments, wherein each argument pertains to one or more clusters related to the discussed topic. Given that an argument can encompass multiple aspects of a topic, it may belong to several overlapping clusters (Daxenberger et al., 2020). Every argument directly addresses one or more clusters. Since each argument component targets the preceding parent, it indirectly refers to all preceding parents. Consequently, we stipulate that each argument component inherits the clusters of its preceding nodes, meaning it indirectly encompasses all clusters that its parent addresses, whether directly or indirectly. Notably, the root node is not affiliated with a cluster.

In this ADS, a sample debate on the topic *Marriage is an outdated institution* provides a suitable manually clustered argument structure. It serves as the knowledge base for the arguments and is sourced from the *Debatabase* of the idebate.org[5] website. It consists of a total of 72 argument components, their corresponding relations, and is encoded in an OWL ontology (Bechhofer, 2009) for further use. In each $whypro/con$ move, a single supporting/attacking argument component is presented to the user. To prevent the user from being overwhelmed by the amount of information, the available arguments are presented to the users incrementally upon their request. In order to integrate the SFB-Model 3.2, the dialogue model has to provide respective user moves. The interaction between the system and the user is separated into turns, consisting of a user action and the corresponding natural language answer from the system. The system's response is based on the original textual representation of the argument components, which is embedded in moderating utterances. Table 1 shows the required[6] possible moves (actions) the user is able to choose from. This allows the user to navigate through the argument tree and inquire

[4]Due to the generality of the annotation scheme, the system is not confined to the data considered herein. In general, any argument structure that aligns with the applied scheme can be utilized.

[5]https://idebate.org/debatabase (last accessed July 23rd, 2022). Material reproduced from www.idebate.org with the permission of the International Debating Education Association. Copyright © 2005 International Debate Education Association. All Rights Reserved.

[6]Only moves that are relevant for the SFB-Model are shown. Other moves are not listed due to their mere navigational/meta-informational purposes.

for more information. The determiners show which moves are available depending on the position of the current argument.

As shown in Table 1, r_k, t_k, and f_k are directly influenced by respective user moves and thus updated immediately. However, this does not apply to *PR*, which does not directly refer to the dialogue content but rather serves as a meta reflection. Since pr_k does not directly pertain to the argument, but rather to the respective cluster, this information is requested in a separate pop-up window during the interaction. To avoid inconveniencing the user (given that the cluster might remain the same over a certain number of moves), we update pr_k whenever the corresponding clusters change (when a new cluster k_2 is addressed and the old cluster k_1 is no longer addressed). The user's spoken input is captured through browser-based audio recording using the Google Speech Recognition API. Subsequently, it is processed by an NLU framework (Abro et al., 2022) that employs an intent classifier based on a BERT Transformer Encoder (Devlin et al., 2019) and a bidirectional LSTM classifier. After recognizing a user move, the spoken system response is presented using speech synthesis provided by the Google Web Speech API. An exemplary dialogue is shown in Appendix A.1.

5 User Study

We conducted a user study from October 4$^{\text{th}}$ to 15$^{\text{th}}$, 2022, involving 60 participants. The participants were divided into two groups: one group was presented with arguments based on their interests (referred to as the "interest" group), whereas the other group was presented with arguments that might challenge their existing beliefs (referred to as the "SFB-breaking" group). In the interest group, the system presented arguments that precisely matched the user's requests. If a loss of interest was detected (modeled by an interest model (Aicher et al., 2022a)), the system suggested arguments that aligned best with the user's preferences and interests. This interest policy is based on our previously introduced interest model (Aicher et al., 2022a) and adapted accordingly. In the SFB-breaking group, the system presented arguments based on the system policy described in Subsection 3.3. Consequently, the arguments presented to the SFB-breaking group might have differed in polarity and/or cluster from the original user request. The primary objective of this study was to address the following research questions: 1) Can the proposed system policy effectively break a user's SFB? 2) What are the discernible differences in the overall SFB dimensions between the two participant groups? To investigate these research questions, we formulated the following hypotheses to be tested during the study:

H1 Participants in the SFB-breaking (interest) group exhibit a lower (higher) probability of being caught in an SFB after the interaction.

H2 The exploration behavior of the SFB-breaking group changed during the interaction.

These hypotheses were designed to assess the effectiveness of the system policy in breaking the users' SFBs and to explore potential differences in SFB dimensions between the two groups. The study was conducted in a laboratory setting at a university, involving international participants who possessed a sufficient level of proficiency in English. Including the introductory phase and the completion of pre- and post-questionnaires, the entire study duration was estimated to be one hour. Participants were compensated with a payment of 10$, which corresponded to an hourly rate of 10$/hour. After a brief introduction to the system, including a short text and instructions on how to interact with it, participants were required to answer two control questions. These questions served as a means to verify their understanding of how to interact with the system. Only participants who successfully passed this test were allowed to proceed to a test interaction with the system.

During the "real" interaction, participants were instructed to listen to at least 20 arguments[7]. Participants were not informed about the underlying SFB or Interest Model. They were only informed that the ADS might provide suggestions on its own, and they could return to the previous argument if they did not approve.

Throughout the study, the following data was collected: Self-assessment questionnaire (P.851, 2003), Calculated SFB-values: *RUE*, *PR*, *TK*, and *FK* (for each cluster k), Participants' opinions and interests regarding the topic of discussion, set of heard arguments, dialogue history. Strict adherence to data protection regulations and participant anonymity was maintained throughout the study. Participants had the freedom to withdraw from the

[7]This minimum ensured a sufficient amount of data was collected to analyze the different system policies.

study at any time. The study was approved by an Institutional Review Board (IRB) after thorough ethical review and met all internal guidelines due to the solely cooperative, non-persuasive design of the user study.

6 Results

The user study involved 60 participants, ranging in age from 22 to 41 years. The participants' average age was 28.45 (with a standard deviation (SD) of 4.11). The two participant groups each consisted of 30 individuals (SFB-breaking: 7 females, 23 males; interest: 10 females, 20 males). Both groups exhibited similar levels of experience with spoken dialogue systems, rated on a 5-point Likert scale where 1 represented "No experience" and 5 represented "Very much experience": interest group at 2.40 (SD 0.89); SFB-breaking group at 2.13 (SD 1.04).

On average, participants spent approximately 33.87 minutes engaged in interactions with the system (interest group: 33.99 min (SD 7.74), SFB-breaking group: 33.75 min (SD 5.96)). Throughout the interaction, participants were presented with an average of 22.02 arguments (interest group: 21.73 (SD 4.00), SFB-breaking group: 22.30 (SD 3.54)). In Table 2, we present the mean values for all dimensions of both groups across all clusters. Given the paper's limited scope, our primary focus lies on the weighted overall means for each SFB dimension, calculated by averaging across all clusters (subtopics). Exemplary clusterwise results are provided in Appendix A.2. Notably, the SFB-breaking group displayed significantly larger values for all dimensions: *Reflective User Engagement (RUE), Personal Relevance (PR), True Knowledge (TK)*, and *False Knowledge (FK)* when compared to the interest group.

	Interest		SFB-breaking			
Asp.	M	SD	M	SD	p_{corr} value	r
RUE	0.30	0.28	**0.47**	0.26	<0.001	0.92
PR	0.78	0.20	**0.80**	0.19	<0.001	0.45
TK	0.28	0.18	**0.31**	0.25	<0.001	0.61
FK	0.97	0.09	**0.99**	0.05	0.008	0.39

Table 2: Means and SD of all SFB dimensions over all cluster for for both groups. Bold values indicate statically significant differences with respective Bonferroni corrected p_{corr} values and effect sizes r.

To ascertain the statistical significance of these findings, we employed the non-parametric Mann-Whitney U-test for two independent samples (McKnight and Najab, 2010). This choice was made due to the deviation of group means from normal distribution, as indicated by the Shapiro-Wilk test. Given that we are considering four dimensions, we applied the Bonferroni correction to account for multiple comparisons, thereby adjusting the p-value (represented as p_{corr}). The most substantial and statistically significant distinction was observed in the dimension of *Reflective User Engagement (RUE)* ($p_{corr} < 0.001$), as indicated by a very high effect size of 0.92 ($0.5 < r < 1$). Similarly, a significant difference ($p_{corr} < 0.001$) with a high effect size ($0.5 < r = 0.61 < 1$) was noted for *True Knowledge (TK)*. Concerning *Personal Relevance (PR)* and *False Knowledge (FK)*, the differences were also found to be highly significant, exhibiting a medium effect size ($0.3 < r < 0.5$).

Regarding the "pre-interest" of the participants (measured on a 5-point Likert scale before the interaction, where 1 represented "Not at all interested" and 5 represented "Very much interested"), the difference between the two groups is insignificant (interest: 3.67 [SD 0.71], SFB-breaking: 3.47 [SD 0.82]; $p_{corr} = 0.986$). Similarly, the difference in their "pre-opinion" (rated on a scale of 1 to 5, where 1 represented "Totally disagree" and 5 represented "Totally agree") is also insignificant (interest: 3.09 [SD 0.93]; SFB-breaking: 2.78 [SD 0.83]; $p_{corr} = 0.308$). During the interaction, approximately 36.67% (11 out of 30) participants changed their opinion (from pro to con or vice versa) in the SFB-breaking group, compared to 6.67% (2 out of 30) in the interest group. Regarding the "post-interest" (measured after the interaction), a significant difference with $p_{corr} = 0.024 < 0.05 = \alpha$ is notable (interest: 3.20 [SD 1.16], SFB-breaking: 3.97 [SD 0.89]). Similarly, the "post-opinion" also exhibits a significant difference (interest: 3.63 [SD 0.96], SFB-breaking: 3.07 [SD 0.87], $p_{corr} = 0.048$, $r = 0.29$).

To determine the significance of the difference between pre- and post-measurements, we utilized the non-parametric Wilcoxon signed rank test (Woolson, 2007) for paired samples and Bonferroni corrected p-values p_{corr} based on a set of four comparisons. For the SFB-breaking group, both interest and opinion showed significant differences before and after the interaction (interest: $p_{corr} = 0.006$, $r = 0.38$; opinion: $p_{corr} = 0.036$, $r = 0.30$). In the interest group, the pre- and

post-interest also exhibited significant differences ($p_{corr} = 0.006, r = 0.39$).

Considering the user moves, a significant difference between both groups becomes evident. In the interest group, a pro (con) argument was requested 297 (172) times. Only in 15% of all argument requests, interest group users asked for an argument which did not align with their own opinion. In the SFB-breaking group, a con (pro) argument was requested 117 (90) times. Furthermore, in 71 (82) instances, the ADS opted to present a con (pro) argument. Particularly towards the end, the SFB-breaking group tended to request arguments without specifying polarity, and if polarity was specified, it contradicted the user's opinion in 43% of all requests. In the interest group, arguments were rarely rejected (3) and mostly preferred (87). In the SFB-breaking group, suggested arguments were rejected 65 times and explicitly preferred 71 times. Moreover, participants in the SFB-breaking (interest) group requested to return to the previous argument in only 8 (1) cases.

7 Discussion

In the following the results of our study (Section 6), particularly regarding our two hypotheses (refer to Section 5) are discussed.

7.1 Validation of Effectiveness of SFB-breaking policy (H1):

The significant differences in all overall dimensions between both groups can be attributed to the substantial disparity in polarity and the corresponding clusters to which the heard arguments belonged, despite the nearly similar number of heard arguments. While the interest group was exclusively exploring arguments of the requested polarity and the estimated most interesting clusters, the SFB-breaking group encountered arguments strategically chosen to break the SFB of the user. Consequently, participants in the interest group primarily requested arguments aligning with their pre-existing opinions. In contrast, the SFB-breaking group encountered arguments of both polarities, elucidating the significant difference in the overall *Reflective User Engagement (RUE)*. These observations further validate the hypothesis that users tend to remain within their SFBs while exploring contentious topics unless proactively motivated to consider opposing viewpoints. The substantial difference in *True Knowledge (TK)* across all clusters is a result of the SFB-breaking system's tailored policy, which aims to present arguments spanning as many clusters as possible to encompass diverse facets of the topic. In contrast, the interest policy concentrates on clusters aligned with the user's interest, offering arguments accordingly.

Significant variations in *Personal Relevance (PR)* are also evident, even accounting for differences between individual clusters, notably contingent on the number of arguments heard from each cluster. Participants who explored a greater number of clusters in a balanced manner tended to exhibit notably higher *Personal Relevance (PR)* on average. Similarly, disparities are discernible among the individual clusters concerning *False Knowledge (FK)*. Out of the nine instances of $false$ moves, merely two were initiated by participants in the SFB-breaking group. Hence, aligning with our hypothesis, the outcomes affirm that participants in the SFB-breaking (interest) group demonstrated a notably lower (higher) likelihood of being caught in an SFB after the interaction.

7.2 Change of exploration behaviour (H2):

In the initial stage of the interaction, the first five arguments presented by the ADS were selected solely based on the user's requests. During this phase, both groups exhibited a tendency to seek arguments that aligned with their pre-existing opinions. However, a shift in behavior was observed among the SFB-breaking group participants after being repeatedly exposed to arguments of opposing polarity. On average, after the eleventh argument, SFB-breaking users began to request pro and con arguments almost equally or no longer specified the polarity. Interestingly, with the exception of one case, participants from the SFB-breaking group continued the interaction and did not revert to the previous argument. This suggests that the participants appeared to be more motivated by the system's suggestions to explore differing viewpoints and facets. This observation is further supported by the heightened *Personal Relevance (PR)* of the corresponding clusters. Conversely, participants in the interest group returned to the previous argument when they did not perceive the corresponding cluster as personally relevant.

Within the SFB-breaking group, participants expressed a preference for and rejection of the proposed arguments almost equally, with approximately a third changing their opinion, resulting

in a relatively neutral post-opinion. In contrast, the interest group predominantly indicated their preference for arguments and rarely rejected any. The reinforcement of their pre-existing opinions becomes particularly evident as the interest group encountered over twice as many pro arguments as con arguments, and only two participants altered their stance on the topic. The comparatively diminished level of interest after the interaction in the interest group could potentially be attributed to a saturation effect. Conversely, the SFB-breaking group exhibited an elevated post-interest, indicative of heightened engagement and a greater willingness to explore additional aspects.

In conclusion, it is evident that the exploration behavior exhibited by the SFB-breaking group demonstrates a significant improvement in balance concerning clusters and polarity. To sum up, our findings corroborate our initial hypotheses and demonstrate that our SFB-breaking policy takes us closer to achieving our goal of assisting users in critically evaluating information on a contentious topic.

8 Limitations

However, this work has certain limitations that could be addressed in future research. First, the sample size of our study is relatively small, potentially affecting the generalizability of our findings. In future endeavors, a study (e.g., through crowdsourcing) with a larger sample size could yield more robust data, enabling us to refine the SFB margins and enhance the validity of our approach. Second, given that the SFB-Model is a novel concept, it is presently constrained to four dimensions. Subsequent research could explore additional dimensions that may prove pertinent in various scenarios and applications. Additionally, finding ways to implicitly estimate both PR and TK, which can only be determined retrospectively, would be advantageous. This could involve leveraging common sense knowledge bases and employing fake news detection techniques. Third, while our study demonstrates the proof-of-principle for the effectiveness of a rule-based policy to break SFB, it is limited to static, predefined rules, rendering it relatively inflexible. In future work, we intend to delve into more advanced machine learning techniques, such as reinforcement learning. This would enable us to personalize and adapt these strategies based on the user's verbal and non-verbal feedback, thereby ensuring the user's satisfaction and sustaining their willingness to engage in the dialogue.

9 Conclusion and Future Work

In this work, to the best of our knowledge, we introduce a novel approach to break the user's SFB. After shortly explaining the underlying SFB-Model, we define a rule-based system policy to break the respective user SFB during a cooperative dialogue with an argumentative dialogue system and validate it in a laboratory user study. The study results strongly indicate the effectiveness of the proposed system policy in reducing the likelihood of being stuck in an SFB compared to a policy that prioritizes the users' greatest interest. Moreover, the study revealed significant changes in users' exploration behaviors during the interaction. In particular, the SFB-breaking participants requested arguments of both polarities almost equally often after the ADS pointed out that the previous exploration seemed to be one-sided. These findings emphasize the influence of the system policy on users' exploration behaviors and opinions, further highlighting the success of the proposed approach in mitigating SFB tendencies and fostering open-mindedness in an argumentative dialogue. In future research, we will augment our system's policy by incorporating sophisticated techniques for perceiving and interpreting the user's non-verbal social signals (gestures, facial expressions) in real-time during the interaction. Building upon estimation methods for sentiment and emotion recognition, we aim to leverage Reinforcement Learning to optimize the system's policy, enabling it to dynamically adapt to each individual user's motivation and effectively engaging the users to recognize and overcome their SFB.

In conclusion, this paper highlights the importance of addressing SFBs in argumentative dialogues and takes us a step closer to enabling users to build a well-founded opinion and foster critical, reflective thinking, and open-mindedness in their interaction with cooperative ADS.

Acknowledgements

This work has been funded by the DFG within the project "BEA - Building Engaging Argumentation", Grant no. 313723125, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

References

Waheed Ahmed Abro, Annalena Aicher, Niklas Rach, Stefan Ultes, Wolfgang Minker, and Guilin Qi. 2022. Natural language understanding for argumentative dialogue systems in the opinion building domain. *Knowledge-Based Systems*, 242:108318.

Annalena Aicher, Nadine Gerstenlauer, Wolfgang Minker, and Stefan Ultes. 2022a. User interest modelling in argumentative dialogue systems. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 127–136, Marseille, France.

Annalena Aicher, Wolfgang Minker, and Stefan Ultes. 2021a. Determination of reflective user engagement in argumentative dialogue systems.

Annalena Aicher, Wolfgang Minker, and Stefan Ultes. 2022b. Towards modelling self-imposed filter bubbles in argumentative dialogue systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4126–4134, Marseille, France. European Language Resources Association.

Annalena Aicher, Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2021b. Opinion building based on the argumentative dialogue system BEA. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 307–318. Springer.

Annalena Bea Aicher, Daniel Kornmüller, Wolfgang Minker, and Stefan Ultes. 2023. Self-imposed filter bubble model for argumentative dialogues. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–11.

Armen E Allahverdyan and Aram Galstyan. 2014. Opinion dynamics with confirmation bias. *PloS one*, 9(7):e99557.

Bharat N Anand. 2021. The us media's problems are much bigger than fake news and filter bubbles. *Domestic Extremism*, page 138.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Sean Bechhofer. 2009. Owl: Web ontology language. In *Encyclopedia of Database Systems*, pages 2008–2009. Springer.

Lisa Chalaguine and Anthony Hunter. 2021. Addressing popular concerns regarding covid-19 vaccination with natural language argumentation dialogues. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 59–73, Cham.

Lisa A. Chalaguine and A. Hunter. 2020. A persuasive chatbot using a crowd-sourced argument graph and concerns. In *COMMA*.

Johannes Daxenberger, Benjamin Schiller, Chris Stahlhut, Erik Kaiser, and Iryna Gurevych. 2020. Argumentext: argument classification and clustering in a generalized search scenario. *Datenbank-Spektrum*, 20(2):115–121.

Michela Del Vicario, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2017. Modeling confirmation bias and polarization. *Sci Rep*, 7(40391):1–9.

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

Tim Donkers and Jürgen Ziegler. 2021. The dual echo chamber: Modeling social media polarization for interventional recommending. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 12–22, New York, NY, USA. Association for Computing Machinery.

Axel G. Ekström, Diederick C. Niehorster, and Erik J. Olsson. 2022. Self-imposed filter bubbles: Selective attention and exposure in online search. *Computers in Human Behavior Reports*, 7:100226.

Hans Gelter. 2003. Why is reflective thinking uncommon. *Reflective Practice*, 4(3):337–344.

Emmanuel Hadoux, Anthony Hunter, and Sylwia Polberg. 2022. Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuadee. *Argument & Computation*, 14:1–53.

Hsieh-Hong Huang, Jack Shih-Chieh Hsu, and Cheng-Yuan Ku. 2012. Understanding the role of computer-mediated counter-argument in countering confirmation bias. *Decision Support Systems*, 53(3):438–447.

Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.

Terry Lee. 2019. The global rise of "fake news" and the threat to democratic elections in the usa. *Public Administration and Policy*, 22(1).

Erin A Maloney and Fraulein Retanal. 2020. Higher math anxious people have a lower need for cognition and are less reflective in their thinking. *Acta psychologica*, 202:102939.

Mark Mason. 2007. Critical thinking and learning. *Educational philosophy and theory*, 39(4):339–349.

Patrick E. McKnight and Julius Najab. 2010. *Mann-Whitney U Test*, pages 1–1. American Cancer Society.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

ITU-T Recommendation P.851. 2003. Subjective quality evaluation of telephone services based on spoken dialogue systems (11/2003). International Telecommunication Union.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Richard W Paul. 1990. Critical and reflective thinking: A philosophical perspective. *Dimensions of thinking and cognitive instruction*, pages 445–494. Publisher: North Central Regional USA.

Richard E Petty, Pablo Briñol, and Joseph R Priester. 2009. Mass media attitude change: Implications of the elaboration likelihood model of persuasion. In *Media effects*, pages 141–180. Routledge.

Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo chambers on facebook. *Available at SSRN 2795110*.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn A. Walker. 2017. Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems*, pages 45–52.

Ariel Rosenfeld and Sarit Kraus. 2016. Strategical argumentative agent for human persuasion. In *ECAI'16*, pages 320–328.

Christina Schwind and Jürgen Buder. 2012. Reducing confirmation bias and evaluation bias: When are preference-inconsistent recommendations effective–and when not? *Computers in Human Behavior*, 28(6):2280–2290.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, and Lilach Edelstein. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *COLING*, pages 1501–1510.

Constanza Villarroel, Mark Felton, and Merce Garcia-Mila. 2016. Arguing against confirmation bias: The effect of argumentative discourse goals on the use of disconfirming evidence in written argument. *International Journal of Educational Research*, 79:167–179.

Abdul Waheed, Muskan Goyal, Nimisha Mittal, Deepak Gupta, Ashish Khanna, and Moolchand Sharma. 2021. BloomNet: A robust transformer based model for bloom's learning outcome classification. In *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 209–218, Trento, Italy. Association for Computational Linguistics.

RF Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3.

A Appendix

A.1 Example Interaction

In Table 3, we present a segment of an exemplary dialogue with the argumentative dialogue system, following the SFB-breaking policy. It illustrates an artificial interaction between the system and the user regarding the topic *Marriage is an outdated institution*, sourced from the *Debatabase* on the idebate.org[8] website.

We assume that the interaction has been ongoing for a substantial duration, and the user's previous interactions suggest a likelihood of being trapped in an SFB due to consistently requesting attacking arguments.

Throughout the illustrated interaction, the user's requests for con/pro arguments impact two SFB dimensions: the *Reflective User Engagement* (r_k) and the *True Knowledge* (tk_k), within the corresponding SFB-vector $\overrightarrow{sfb_k}$. Additionally, the value for *False Knowledge* (fk_k) is updated due to the user's contradictory knowledge statement.

A.2 Exemplary clusterwise Results

Due to the limited scope of the paper, we focused on the overall SFB dimension results. Nonetheless, as we observed variations among individual clusters, we present the results of two specific clusters below.

Table 4 displays the mean values for each SFB dimension in both groups for two particular clusters (subtopics) of the discussed theme "Marriage is an outdated institution." Cluster 4 ("Children") comprises 22 arguments (one of the largest clusters), while Cluster 7 ("Law") includes only 6 arguments, making it one of the smallest clusters.

Similar to the overall cluster means (refer to Table 2), we observe significant differences with

[8] https://idebate.org/debatabase (last accessed 23rd July 2021). Material reproduced from www.iedebate.org with the permission of the International Debating Education Association. Copyright © 2005 International Debate Education Association. All Rights Reserved.

Actor	Utterance	Move	SFB Dimension
System	...		
User	*Could you please provide an attacking argument?*	why_con	r_k
System	*Rather than focusing solely on an attacking argument, I propose we also consider the opposing view. One could argue that religion as a whole is becoming less significant, leading to a reduced importance of marriage.*	why_pro_sfb	tk_k
User	*I understand, but could you provide additional evidence?*	why_con	r_k
System	*On the pro side, it's worth mentioning that the number of people in the UK identifying as religious has declined by 20% in the last 20 years, as indicated by the British Social Attitudes Survey of 2007.*		tk_k
User	*Impressive, but isn't this statistic outdated by now? I don't believe it holds true anymore.*	$false$	fk_k
System	*I understand your skepticism, but if you allow me, I'd like to provide a more detailed explanation.*		

Table 3: Artificial dialogue between the ADS and the user on the topic *"Marriage is an outdated institution"*. The table includes actor, utterance, respective moves, and updated SFB dimensions of cluster k.

	Interest		SFB-breaking			
Asp.	M	SD	M	SD	p_{corr} value	r
r_4	0.35	0.20	**0.66**	0.12	<0.001	0.63
pr_4	0.71	0.21	**0.80**	0.12	0.007	0.23
tk_4	0.25	0.14	**0.47**	0.64	<0.001	0.73
fk_4	0.93	0.13	**1.00**	0.00	0.044	0.21
r_7	0.21	0.39	**0.82**	0.19	<0.001	0.73
pr_7	0.74	0.15	**0.81**	0.18	<0.001	0.35
tk_7	0.45	0.26	**0.83**	0.15	<0.001	0.77
fk_7	0.92	0.14	**1.00**	0.00	0.021	0.32

Table 4: Means and SDs of all SFB dimensions for two clusters (4 = "Children", 7 = "Law") for both groups. Bold values indicate statically significant differences with respective Bonferroni corrected p values and effect sizes r.

small to high effect sizes in each dimension. Furthermore, noticeable differences are evident between individual clusters, as illustrated in Table 4. Particularly concerning smaller clusters, we discern that our SFB-breaking policy has a moderate to large effect on each dimension. This can be attributed to the fact that our SFB-breaking policy aims to explore all clusters in a balanced manner, whereas the interest policy only targets clusters of user interest.

The Open-domain Paradox for Chatbots: Common Ground as the Basis for Human-like Dialogue

Gabriel Skantze
KTH Speech, Music and Hearing
Stockholm, Sweden
skantze@kth.se

A. Seza Doğruöz
Universiteit Gent
Belgium
as.dogruoz@ugent.be

Abstract

There is a surge in interest in the development of open-domain chatbots, driven by the recent advancements of large language models. The "openness" of the dialogue is expected to be maximized by providing minimal information to the users about the common ground they can expect, including the presumed joint activity. However, evidence suggests that the effect is the opposite. Asking users to "just chat about anything" results in a very narrow form of dialogue, which we refer to as the *open-domain paradox*. In this position paper, we explain this paradox through the theory of common ground as the basis for human-like communication. Furthermore, we question the assumptions behind open-domain chatbots and identify paths forward for enabling common ground in human-computer dialogue.

1 Introduction

Recent advancements of large language models (LLMs) have given rise to a surge in interest for the development of "open-domain" chatbots (Roller et al., 2020a; Adiwardana et al., 2020; Thoppilan et al., 2022). Unlike task-oriented dialogue systems designed for a specific purpose and typically implemented in a modular fashion, open-domain chatbots are trained end-to-end on large amounts of data. Roller et al. (2020a) define their long-term goal as "building a superhuman open-domain conversational agent" that is "preferred on average to an alternative human speaking partner in open conversation". The more specific purpose of such a conversational agent is not stated, and thus it is implied that dialogue is a generic problem that can be abstracted away from the context in which it takes place.

In current evaluations of open-domain chatbots, there seems to be a general assumption that the "openness" of the dialogues can be maximized by removing as much instructions and context as possible (e.g., by instructing the user to "just chat about anything"). While this might seem intuitive at first in terms of removing the boundaries for "openness", we argue that this assumption stems from a misconception and that dialogue as a linguistic activity cannot be stripped from its context. The setting in which open-domain chatbots are evaluated does not clearly correspond to any form of human-human dialogue "in the wild".

In this position paper, we analyse this misconception as the **open-domain paradox**: *The diversity of the various forms of dialogues found in human-human interaction does not stem from the "openness" of the dialogue setting, but rather the opposite: they stem from the diversity of highly specific contexts in which dialogue takes place.* If this is true, it means that the current methods for collecting dialogue data and evaluating open-domain chatbots will only give rise to a very narrow form of dialogue which does not correspond closely to human-human dialogues. Thus, they will not tell us whether these systems are truly "open". Nor will they tell us much about how good these systems actually are at modelling various dialogue phenomena. From the user's perspective, if the common ground and the reason for having the interaction is not clear, there is a risk that the system will not be perceived as meaningful.

Our contribution has the following goals: First, we provide a critical review (not an extensive survey) of SOTA open-domain chatbots, in terms of how they are defined, trained and evaluated. We discuss how the lack of common ground has consequences for their limited scope and arguably their "openness", compared to human-human dialogue. Secondly, we provide various research directions which might help to mitigate this problem and enable common ground in human-computer dialogue.

2 What is common ground?

When we initiate a dialogue as humans, we do not start with a blank slate but we assume some **com-**

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 605–614
September 11–15, 2023. ©2023 Association for Computational Linguistics

mon ground between the speakers/interlocutors. Clark (1996) describes common ground among humans as "the sum of their mutual, common or joint knowledge, beliefs and suppositions". For a successful and meaningful communication to take place, and for coordinating joint actions, it is essential that both parties have a shared understanding of what this common ground is.

Clark (1996) makes a distinction between *communal* and *personal* common ground. Communal common ground refers to the cultural communities (e.g., nationality, profession, hobbies, language, religion, politics) people belong to. In addition, there could also be cultural communities which are shaped around shared expertise specific to the members of that community who may not live in the same place (e.g., English teachers around the world), and it is possible to belong to more than one cultural community at the same time. Clark (1996) makes a further distinction in communal common ground between *human nature* (i.e., same senses, sense organs, types of sensations), *communal lexicons* (e.g., there are conventions about word meanings even when two interlocutors speak the same language), as well as *cultural facts, norms and procedures* (which are commonly shared within that community). Procedures for joint activities are the underlying notions about common ground for the community members who know the specific "scripts" for the procedures about joint activities in certain contexts (e.g., restaurants, supermarkets vs. school).

Personal common ground is based on personal joint experiences with someone (Clark, 1996), and is further classified into *perceptual bases*, *actional bases* and *personal diary*. One important aspect of the personal common ground, apart from shared memories and commitments, is the **linguistic alignment** whereby human interlocutors align (or adjust) their language in alliance with their conversational partners (Pickering and Garrod, 2006), context and medium of communication (Werry, 1996; Nguyen et al., 2016). Since childhood, humans learn how to adjust and tolerate linguistic variation (e.g., across conversational partners, contexts, mediums) between different communal and personal common grounds in their environment.

The theory of common ground also postulates the principle of *least collaborative effort*, which means that people in conversation use their assumed common ground to minimize their collaborative effort to achieve further understanding (Clark, 1996). Thus, a brief word might have a significant meaning if the context is highly specific or the interlocutors know each other well. As another example, Meylan et al. (2022) showed that conversations between children and their caregivers are hard to transcribe, since their common ground is not known to the transcriber.

An important part of common ground is the **joint activity** that is assumed, that is, the reason why the interaction is taking place. This is similar to Wittgenstein's (1958) concept of *language games* or the notion of *activity type* developed by Levinson (1979). A related concept is that of **speech events** developed by Goldsmith and Baxter (1996) (not to be confused with "speech acts"), which refers to the type of activity that the parties are involved in. By analysing transcribed speech diaries from 48 university students over a 1-week period, they developed a taxonomy of 39 speech events:

- **Informal/Superficial talk**: Small talk, Current events talk, Gossip, Joking around, Catching up, Recapping the day's events, Getting to know someone, Sports talk, Morning talk, Bedtime talk, Reminiscing
- **Involving talk**: Making up, Love talk, Relationship talk, Conflict, Serious conversation, Talking about problems, Breaking bad news, Complaining
- **Goal-directed talk**: Group discussion, Persuading conversation, Decision-making conversation, Giving and getting instructions, Class information talk, Lecture, Interrogation, Making plans, Asking a favor, Asking out

This specific taxonomy is likely not generic for all human-human conversations (i.e., there may be more events which they did not identify in their limited study in terms of duration, population and methodology). Nevertheless, it illustrates the diversity of joint activities in human-human interaction. Note that even when we engage in more casual speech events, such as *Small talk*, there is still a reason for why we are having the interaction (maybe just to pass time or avoid being rude), and both interlocutors should be aware of this reason (it is part of their common ground). The speech event or joint activity that is assumed puts constraints on the interpretation space (e.g., which implications can

be made) and what can be considered to be a coherent and meaningful contribution to the activity. Thus, in *Small talk* or during *Decision-making*, we expect the speakers to bring up certain topics, but not others. We also do not engage in any speech event with anyone at any time.

3 Open-domain chatbots

3.1 What does open-domain mean?

The term "chatbot" (and its predecessor "chatterbot") has been used since the early 1990's to denote systems that interact with users in the form of a written chat, typically without any constraints (at least as presented to the user) on what the conversation should be about (Mauldin, 1994; Wallace, 2009). This early line of work was primarily done outside of academia, where the focus instead was on more task-oriented systems. A search in the DBLP bibliographic database for computer sciences[1] reveals that the word "chatbot" was used in the titles of very few publications until 2015. After this, the usage of the term has increased rapidly, and in 2022, it was used in the titles of almost 300 papers. This development has clearly been sparked by the development of LLMs and the end-to-end modelling of dialogue (Vinyals and Le, 2015), which has attracted people from the machine learning community. To stress the open-ended nature of these chatbots, the term "open-domain chatbot" is often used. Adiwardana et al. (2020) provide the following definition: "Unlike closed-domain chatbots, which respond to keywords or intents to accomplish specific tasks, open-domain chatbots can engage in conversation on any topic".

Many of the early chatbots were developed to take part in the Loebner prize competition that was running between 1990-2019 (Mauldin, 1994). This competition was partly inspired by the so-called *Turing test*, as proposed by Turing (1950) under the name *the Imitation Game*, as a test for determining whether a computer has reached human-level intelligence. It is interesting to note that, as the game is described by Turing (1950), the testers are not provided with any information that would allow them to assume some form of common ground. It is possible that this original idea by Turing has influenced the concept of open-domain chatbots and how they are evaluated. While they are typically not evaluated according to the original Turing test (i.e., the testers are not supposed to guess whether they are interacting with a computer or human), the context-less setting of the interaction is still similar.

Although earlier chatbots mainly interacted in written form (due to the limited speech recognition performance), open-domain chatbots are nowadays also sometimes built for spoken interaction. One example of this is the **Alexa Prize**, which is academic competition sponsored by industry to create an open-domain "socialbot" for the Amazon Echo device (Ram et al., 2018). Users of the device can interact with the socialbot from a randomly selected team by just saying "Let's chat" to their device. Since the purpose of the socialbot is similar to that of a typical open-domain chatbot, we also include it in our discussion.

In some work, "open-domain" seems to be synonymous with more "social" (as opposed to task-oriented) interaction, and such systems have been referred to as "social chatbots" (Shum et al., 2018). Deriu et al. (2020) make a distinction between *task-oriented*, *conversational* and *question-answering* chatbots, where *conversational* chatbots "display a more unstructured conversation, as their purpose is to have open-domain dialogues with no specific task to solve". These chatbots are built to "emulate social interactions" (ibid.). However, it is not entirely clear how the term "social" should be understood in this context, as all conversations are "social" in the sense that they are used for interpersonal communication. Using the speech event taxonomy by Goldsmith and Baxter (1996), mentioned above, the speech events that are closest to this notion are perhaps those belonging to *informal/superficial talk*, whereas more task-oriented chatbots or dialogue systems would rather belong to *goal-directed talk*. However, as their analysis shows, the range of speech events in human communication is much more nuanced than this simple distinction would suggest.

Another problem with the term "open-domain" is that it is unclear what "domain" refers to. In one interpretation, it could refer to the joint activity or speech event that the interlocutors are engaged in (e.g., small talk, information seeking, decision-making, negotiation). In another interpretation it could mean a wide range of factual topics (e.g., sports, music, travel, math) that are discussed among the interlocutors. These two notions are to some extent orthogonal. For example, the factual topic of travelling could be discussed in the context of various speech events, such as recapping some-

[1] http://dblp.org

one's travel experience (*"Tell me about your trip to Paris"*), asking for travel advise (*"What should I see in Paris?"*), or planning a trip (*"Let's plan a trip to Paris together"*).

If a chatbot is truly "open-domain", we could perhaps expect it to be able to engage in all combinations of speech events and factual topics that we can expect to find in conversations between humans. However, it is unclear whether this is something we could expect from one and the same agent, since we do not expect this from all human-human encounters in real-life settings. Instead, we are selective about what to talk with who and in which way. For example, we can have a conversation with a travel agent in real-life about the costs, insurances, types of sightseeing associated with a trip to Egypt (factual information) but we do not expect her/him to to have a conversation about making a decision about which souvenir to buy, since that is (probably) beyond her/his work definition. Therefore, even human-human conversations are not that open to cover anything across all contexts.

3.2 Training of chatbots

Current open-domain chatbots are typically implemented in an end-to-end fashion as transformer-based LLMs, trained to do next-token prediction on large amounts of text data. For the **Meena chatbot** (Adiwardana et al., 2020), (unspecified) social media conversations were used as training data. Roller et al. (2020b) built the **Blender** chatbot based on the training data collected from Reddit. More recent chatbots, like **LaMDA** (Thoppilan et al., 2022), have been trained using larger, more general datasets (including both dialogue and other public web documents). Similarly, **GPT-3** (Brown et al., 2020) is a general-purpose language model that can be used as a chatbot when prompted in the right way.

While general language models can be used directly as chatbots, their responses will reflect ordinary language use, which might not always align with the desired output in terms of, for example, truthfulness and toxicity (the so-called "alignment-problem"). To address this, Thoppilan et al. (2022) fined-tuned LaMDA to optimize human ratings of safety and other qualitative metrics. A more sophisticated approach was taken by Ouyang et al. (2022) with their model **InstructGPT**, which uses so-called "reinforcement learning from human feedback" (RLHF), where a model of human raters is used during reinforcement learning to optimize the model towards the desired criteria.

The RLHF approach was also used when training the chatbot **ChatGPT**[2]. This kind of model adaptation is interesting from an "open-domain" perspective, since the behavior of the chatbot becomes specific to the instructions given to the human raters. In the communication around ChatGPT, there is very little information about what the user can expect in terms of its capabilities or the purpose of the interaction besides the fact that it "interacts in a conversational way". Only when interacting with ChatGPT, it becomes clear that its purpose is to serve as some form of AI assistant or interactive search engine, answering factual questions, as well as assisting in writing text and code. However, it refuses to engage in small talk or give opinions. For example, when asked "What is your favorite sport?", it answers "As a language model, I do not have personal preferences or feelings, so I cannot have a favorite sport". In that respect, ChatGPT should perhaps not be seen as an open-domain chatbot (and it is in fact never advertised with those words). In comparison to other chatbots, we do not know much about the evaluation methods and metrics around ChatGPT, and there is not a publication available explaining them to the wider public.

One problem that was identified early on when training chatbots end-to-end is their lack of coherent responses. When asked about their name or favorite sport twice (with some turns in-between), they could give different responses. One way to address this problem was to give them a **persona**, which is a description of the character that the chatbot is supposed to represent (Zhang et al., 2018; Dinan et al., 2020; Roller et al., 2020b). While the persona gives some background for the crowdworker or chatbot that can help to improve their internal consistency, it is typically not communicated to the interlocutor beforehand, so it does not really provide any additional common ground.

3.3 Evaluation of chatbots

So far, most of the research on how to evaluate open-domain chatbots have focussed on which *metrics* to use when evaluating them (Roller et al., 2020a; Mehri et al., 2022). This includes questions such as whether to use human or automatic measures, what questions to ask to raters, and whether to evaluate dialogues on the turn- or dialogue-level

[2]https://openai.com/blog/chatgpt/

(ibid.). For example, in the Alexa Prize, users were asked at the end of the conversation to rate the interaction on a scale between 1 and 5 (Ram et al., 2018).

To evaluate the Meena chatbot, Adiwardana et al. (2020) used the metrics "sensibleness" (whether the response makes sense in the given context) and "specificity" (whether the response is specific to the context or more of a generic nature, like "*I don't know*"). The assessment is done by third party observers (crowdworkers), who read the chats and rate them. They showed that a model with lower perplexity scored higher on those metrics. The Blender chatbot (Roller et al., 2020b) was evaluated using ACUTE-Eval method, where two chats are presented next to each other and a crowdworker assess their "engagingness" and "humanness".

For the LaMDA chatbot, Thoppilan et al. (2022) also assess the "groundedness" of responses, which is intended to measure whether the model's output is in accordance with authoritative external sources. This should not be confused with the notion of common ground discussed earlier. They also refer to "role consistency" which refers to a metric testing whether the agent is performing its tasks in alignment with what is expected from a similar role in a real-life situation (i.e., consistency with the definition of the agent's role external to the conversation).

4 Lack of common ground in "open-domain" dialogue

While the above-mentioned metrics do say something about the relative merits of chatbots, they do not tell us much about their "openness", or the diversity of the speech events they can engage in. When doing so, more attention should be given to the setting in which the chatbots are evaluated. Since there is no natural setting in which these open-domain chatbots are used, crowd workers are typically recruited to interact with them, either for data collection or for evaluation purposes. Although the crowd workers are typically informed about whether they are interacting with a human or a computer, the setting is similar to that of the Turing test mentioned above, in the sense that no information about the assumed common ground is provided, and they are often asked to initiate the interaction with as few instructions as possible.

For the Meena chatbot, "Conversations start with 'Hi!' from the chatbot to mark the beginning of the conversation and crowd workers have no expectation or instructions about domain or topic of the conversation" (Adiwardana et al., 2020). For the LaMDA chatbot, crowd workers were instructed to "Start a conversation with the chatbot by posing a question or typing a statement on any topic you want to talk about" (Thoppilan et al., 2022). For the Alexa Prize, the users were not provided with any details on what they could expect. Users were asked (through commercials) to just say "Let's chat" to their smart speaker in order to initiate the interaction, but no other instructions were provided (Ram et al., 2018).

These forms of generic and minimalistic instructions are perhaps chosen to provide as little bias as possible in terms of what topics will be brought up, and to really stress the "open domain" nature of the chatbots. However, given what has been discussed above about the importance of common ground and a shared understanding of what the speech event is supposed to be in human-human dialogues, the setting of open-domain chatbots without any common ground is quite unnatural. There is also no physical context or visual cues that could be used to infer any common ground. It is hard to find any similar setting for a human-human conversation. Even if we initiate a small talk with a stranger when waiting for the bus, we both know that this is the type of activity we are engaged in, which will guide us in what might be appropriate to talk about in that context. The equivalent would rather be to be randomly connected to a person without any knowledge about that person or about what the conversation is supposed to be about. Can we expect such a setting to give rise to a wide variety of speech events and topics? If not, how do we know if these systems would be able to handle them?

To address this question, Doğruöz and Skantze (2021) annotated a subset of the publicly released chats from the Meena chatbot (Adiwardana et al., 2020) based on the closest speech event category from Goldsmith and Baxter (1996). The results showed that almost all of them belonged to the *Small talk* category, indicating that they were indeed very limited in scope in terms of speech events. Interestingly, the same was found when annotating the human-human chats that was used as a reference in the evaluation of the Meena chatbot. Those dialogues had been collected by randomly connecting (Google) employees through a chat based system and asking them to converse

about anything, to create a similar setting as that for the human-chatbot interactions. This indicates that it was not primarily the users' expected (lack of) agency of the interlocutor that limited the scope of the dialogue, but rather the limited instructions about the context for the interaction.

These findings point to what we have referred to as the **open-domain paradox**: A completely "open" setting for conversation, where it is not possible to assume any form of common ground, does not give rise to an "open-domain" dialogue, but rather a very limited form of dialogue in terms of both speech events and factual information.

Whether the setting for evaluating open-domain chatbots actually gives rise to a diversity of speech events has consequences for our understanding of their capabilities. In their evaluation of the Blender chatbot, Roller et al. (2020b) reported that it was rated on the same level as the human-human chats taken from the Meena evaluation (Adiwardana et al., 2020). However, in an experiment by Doğruöz and Skantze (2021), they also tested whether Blender could handle other speech events than the small talk that would normally take place in an "open-domain" type of evaluation. This was done by giving a human tester the task of interacting with Blender on a set of different speech events, such as *Decision-making* or *Making plans*. In this evaluation, the chatbot performed much worse (compared to a human interlocutor). This shows that the setting for the interaction and the instructions provided to the testers influence the outcome of the evaluation.

5 Enabling common ground

While the idea of an "open-domain" setting for chatbots is quite pervasive in current research, there are also other trends (and forgotten lessons) pointing towards systems where the context is more specific and where the user can potentially assume some form of common ground. In this section, we will discuss those lines of work, and explore to what extent they could increase the diversity of speech events and open a path towards more human-like (and possibly more meaningful) human-computer dialogue.

5.1 Repeated interactions

A clear limitation for building some form of common ground is that most SOTA chatbots can only handle one-time interactions, which limits the number of relevant speech events that might be relevant. If the interlocutors are allowed to have repeated interactions, they could potentially build common ground together across chat sessions, and more diverse speech events might emerge. One step in this direction was proposed by Xu et al. (2021), who collected and modelled long term conversations, where the speakers learn about each other's interests over time and also refer/discuss issues from past events. The data was collected over 5 chat sessions (each consisting of 14 utterances) through which the speakers talked about topics expanding over days and weeks in order to build a shared history. Due to privacy concerns, the crowdworkers were asked to play one of several different roles. While playing their role, they were also asked to pay attention to the previous interactions with the other speakers.

Although we might expect the participants in subsequent sessions to start with more common ground, it is not self-evident that the user will continue to be interested in interacting with the chatbot over multiple sessions (if they weren't paid) and that meaningful speech events will arise, given the lack of other forms of common ground and reasons for why these repeated interactions are taking place. Xu et al. (2021) do not present any analysis of their data that would help to indicate whether their setting in fact leads to more diversity of speech events.

5.2 Constraining the speech event

As we have discussed, the absence of contextual cues does not give rise to a variety of speech events, but rather the opposite. Many speech events, like decision-making, do not naturally arise with open-domain chatbots. An alternative would be to instead implement dialogue systems that target a larger variety of more specific contexts and speech events. Examples of this include negotiation (Traum et al., 2003), persuasion (Prakken, 2006), and presentations (Axelsson and Skantze, 2020).

One form of more constrained setting is that of **knowledge-grounded dialogue**, where the agent, or one of the crowdworkers during data collection, has access to an external knowledge source, such as Wikipedia (Ghazvininejad et al., 2018; Li et al., 2022). As discussed in Section 3.2 above, chatbots such as ChatGPT, which are restricted in what kind of speech events they willing to engage in, can perhaps also be put into this category. It should

be noted though that ChatGPT is still lacking in terms of the common ground the user can assume (e.g., which cultural norms can be expected) and to what extent the user can trust the factual correctness of the answers given. The chatbot also does not adjust its answers according to the user's level of knowledge, needs, and preferences. For example, a human librarian would utilise the presumed common ground and interact with the user (e.g., student) to find out the purpose of the request (e.g., "*Why do you need this information?*", "*Is it for a homework?*"), and the level of the user's knowledge (e.g., "*Which grade are you at?*"), to recommend resources that fit with its assumptions about the user.

Another recent example of a system that implements a specific speech event is **CICERO**, an agent that can play the game of Diplomacy on a human expert level (FAIR et al., 2022). Unlike other games, like Chess or Go, Diplomacy does not only rely on the strategy of how to move pieces on the game board, but also on the verbal interaction between the players, where they need to negotiate, build trust, persuade, and potentially bluff, highlighting the joint activity and common ground clearly. This type of speech event is also not very likely to take place with an open-domain chatbot. Thus, a plethora of different speech event-specific dialogue systems will likely give rise to a larger diversity of speech events than what can be expected from one open-domain chatbot.

5.3 Situated and embodied interaction

One limitation with chatbots is the lack of physical embodiment or physical situation from which common ground could be inferred. In absence of a shared personal history, common ground can to some extent be inferred from cues like our physical appearance (e.g., age or how we dress) and the language/dialect we use. For example, Lau et al. (2001) asked participants to estimate the proportion of other students who would know certain landmarks, which they could do very accurately. The situation in which the interaction takes place can also serve as a cue for humans to establish the common ground with other humans and develop joint actions accordingly.

If we present the user with an animated avatar instead of an empty chat prompt, it could perhaps help the user to infer more about their potential common ground (Kiesler, 2005; Fischer, 2011). In case of a robot situated in a physical environment, there should be even more contextual cues. For example, in an analysis of interactions with a robot receptionist, Lee and Makatchev (2009) note that "it seemed that people assumed the robot would have knowledge about his surroundings [...] or places relevant to his background or occupation", and thus most questions directed towards the robot were also related to its role and situation. Studies have also shown that people use the robot's presumed origin (Sau-lai Lee et al., 2005) or gender (Powers et al., 2005) to infer the robot's knowledge (and thereby their common ground).

5.4 Scenario-based evaluation

As discussed earlier, Doğruöz and Skantze (2021) evaluated the Blender chatbot on various speech events by giving the user (tester) specific instructions on which speech event to engage in. For example, for the *Decision-making* speech event, the tester could say to the chatbot: "*We have 1000 dollars. Let's decide how we spend it together*". By providing the tester with a list of different speech events, it is possible to better understand which of them the chatbot can handle. If further developed, we think this could constitute an interesting evaluation scheme. To increase the common ground, both the chatbot and the user should probably be given a more detailed description of the setting for the interaction. A potential drawback of this evaluation scheme is that it involves crowdworkers who would need to role-play (likely without much engagement in the task), rather than naturally arising speech events, and that there is a limit as to how detailed the scenario descriptions can be. Thus, they would still not be very close to the level of common ground we can expect from human-human dialogue.

5.5 Simulated worlds

Multi-player text-adventure games (Urbanek et al., 2019) could also provide interlocutors with some common ground. For example, Ammanabrolu et al. (2021) present a model for such agents using LLMs and reinforcement learning. For these agents, the context of the game provides common ground in terms of a textual "setting", which describes the reason for why the interaction takes place and of the characters involved in the interaction.

We can imagine even more open worlds (ideally multi-modal) in which agents and/or humans interact with each other. In such settings, a larger variety

of speech events (similar to human-human interactions), can be expected, and an agent acting in that world will have to be aware of the context and presumed common ground in order to engage in those speech events. An interesting step in this direction was presented in Park et al. (2023), where a large language model (GPT-4) was used to simulate agents in a virtual world. In this setting, the authors observed social behaviours "emerging". Although they did not use the speech event categorization in their analysis, it is clear from their examples that various speech events took place, including *Small talk*, *Catching up* and *Making plans*; eventually, the agents started to plan a Valentine's Day party and set up dates with one another. Such simulations could be an interesting setting for studying and modelling diverse forms of dialogue and speech events. While this simulation did not include any human interlocutors, it is easy to see how that could be added.

6 Conclusion and Discussion

Our goal in this paper was not to survey the latest literature on chatbots but to question the assumptions behind the term "open-domain", and scrutinize to what extent chatbots labelled as such are truly "open". We discussed the notion of common ground in human-human dialogue, and how it is important for human-like dialogue and a diversity of speech events and topics. The general assumption behind SOTA open-domain chatbots is instead to remove as much context as possible, often presenting users with an empty prompt and asking them to "just chat". However, both linguistic theory and evidence suggests that the absence of context does not give rise to a diversity of speech events, but rather a very limited form of dialogue. We called this the *open-domain paradox*.

To be able to study and model different forms of dialogue between humans and agents, the dialogue needs to be embedded within a highly specific context, where both the agent and the human can assume some form of common ground. We identified a couple of different paths towards this end in Section 5.

One explanation for the huge interest in the development of (context-less) open-domain chatbots is perhaps that it fits well with the LLM paradigm (next-token prediction), which uses a limited prompt with the dialogue history. In a way, it is an example of a solution that has found its problem. It is of course possible to include a larger context in the prompt (i.e., a textual representation of the common ground that could be expected), but this clearly has its technical limits. It will be interesting to see whether there will be a movement towards other solutions, perhaps using more modular architectures (as in the example of CICERO).

From the perspective presented here, open-domain chatbots (as the term is currently used) are not necessarily more "generic" than task-oriented dialogue systems, given the limited form of dialogue they are typically evaluated against. One option would be to re-brand open-domain chatbots as "small talk chatbots" or, as some have suggested, "social chatbots" (Shum et al., 2018). We do not think this is appropriate either, since even small-talk (between humans at least) is dependent on the presumed common ground between the speakers. We do not exclude the possibility that the type of dialogue crowdworkers have with open-domain chatbots (i.e., dialogue without common ground) can be regarded as a special speech event category, which we have no existing terminology for yet.

For future work, it might be better to characterize dialogue systems based on which contexts they are intended to be used in, and what speech events are expected to take place. When evaluating such systems, it is also important that they are used in the context they were intended for.

As we have discussed, according to the "principle of least collaborative effort", humans use common ground to make their interactions more efficient (Clark, 1996). Thus, users of dialogue systems will likely always prefer to use systems where their common ground is maximized, rather than "open-domain" settings. If one would want to develop a generic "superhuman open-domain conversational agent" (Roller et al., 2020a), it would need to be highly context-aware, in order to serve across contexts. This route is perhaps not very realistic, given the incredible richness and diversity of the forms of common ground humans assume and build together. Also, this does not even exist for human-human dialogues, as we do not speak about anything with anybody randomly at any given time, without any common ground. Instead of open-domain dialogue systems, it might be more fruitful to focus on developing a large variety of highly context-specific dialogue agents.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv: 2001.09977*.

Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li, Arthur Szlam, Tim Rocktäschel, and Jason Weston. 2021. How to Motivate Your Dragon: Teaching Goal-Driven Agents to Speak and Act in Fantasy Worlds. *arXiv preprint arXiv:2010.00685*.

Nils Axelsson and Gabriel Skantze. 2020. Using knowledge graphs and behaviour trees for feedback-aware presentation agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA. Association for Computing Machinery.

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

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Jan Deriu, Alvaro Rodrigo, Arantxra Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cielibak. 2020. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS '18 Competition*, pages 187–208, Cham. Springer International Publishing.

A. Seza Doğruöz and Gabriel Skantze. 2021. How "open" are the conversations with open-domain chatbots? a proposal for speech event based evaluation. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 392–402, Singapore and Online. Association for Computational Linguistics.

FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Kerstin Fischer. 2011. How People Talk with Robots: Designing Dialog to Reduce User Uncertainty. *AI Magazine*, 32(4):31–38.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A Knowledge-Grounded Neural Conversation Model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Daena J Goldsmith and Leslie A Baxter. 1996. Constituting relationships in talk: A taxonomy of speech events in social and personal relationships. *Human Communication Research*, 23(1):87–114.

S. Kiesler. 2005. Fostering common ground in human-robot interaction. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 729–734, Nashville, TN, USA. IEEE.

Ivy Yee-Man Lau, Chi-yue Chiu, and Ying-yi Hong. 2001. I Know What You Know: Assumptions About Others' Knowledge and Their Effects on Message Construction. *Social Cognition*, 19(6):587–600.

Min Kyung Lee and Maxim Makatchev. 2009. How Do People Talk with a Robot? An Analysis of Human-Robot Dialogues in the Real World. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI)*, Boston, MA.

Stephen C. Levinson. 1979. Activity types and language. *Linguistics*, 17(5-6):365–400.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022. Knowledge-Grounded Dialogue Generation with a Unified Knowledge Representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle, United States. Association for Computational Linguistics.

Michael L Mauldin. 1994. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In *AAAI*, volume 94, pages 16–21.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. Report from the NSF Future Directions Workshop on Automatic Evaluation of Dialog: Research Directions and Challenges. *arXiv preprint arXiv:2203.10012*.

Stephan C Meylan, Ruthe Foushee, Nicole H Wong, Elika Bergelson, and Roger P Levy. 2022. How adults understand what young children say. *arXiv preprint arXiv:2206.07807*.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv: 2203.02155*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *arXiv preprint arXiv:2304.03442*.

Martin J Pickering and Simon Garrod. 2006. Alignment as the basis for successful communication. *Research on Language and Computation*, 4(2):203–228.

A. Powers, A. Kramer, S. Lim, J. Kuo, S-L. Lee, and S. Kiesler. 2005. Common ground in dialogue with a gendered humanoid robot. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, Nashville, TN, USA. IEEE.

Henry Prakken. 2006. Formal systems for persuasion dialogue. *The Knowledge Engineering Review*, 21(2):163–188.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational AI: The Science Behind the Alexa Prize. *arXiv preprint arXiv: 1801.03604*.

Stephen Roller, Y.-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, Pratik Ringshia, Kurt Shuster, Eric Michael Smith, Arthur Szlam, Jack Urbanek, and Mary Williamson. 2020a. Open-Domain Conversational Agents: Current Progress, Open Problems, and Future Directions. *arXiv preprint arXiv:2006.12442*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020b. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Sau-lai Lee, Ivy Yee-man Lau, S. Kiesler, and Chi-Yue Chiu. 2005. Human Mental Models of Humanoid Robots. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 2767–2772, Barcelona, Spain. IEEE.

Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

David Traum, Jeff Rickel, Jonathan Gratch, and Stacy Marsella. 2003. Negotiation over tasks in hybrid human-agent teams for simulation-based training. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '03, page 441–448, New York, NY, USA. Association for Computing Machinery.

Alan Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.

Orioi Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. In *ICML Deep Learning Workshop 2015*, volume 37.

Richard S Wallace. 2009. The anatomy of alice. In *Parsing the Turing Test*, pages 181–210. Springer.

Christopher C Werry. 1996. Linguistic and interactional features of internet relay chat. In *Computer-Mediated Communication: Linguistic, social, and cross-cultural perspectives*, pages 47–64. John Benjamins Publishing Co.

Ludwig Wittgenstein. 1958. *Principal Investigations*. Blackwell Publishing.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 1:2204–2213.

MERCY: Multiple Response Ranking Concurrently in Realistic Open-Domain Conversational Systems

Sarik Ghazarian[1*] **Behnam Hedayatnia**[2] **Di Jin**[2] **Sijia Liu**[2]
Violet Peng[2] **Yang Liu**[2] **Dilek Hakkani-Tur**
[1] University of Southern California / Information Sciences Institute
[2] Amazon Alexa AI
sarik@isi.edu, dilek@ieee.org
{behnam,djinamzn,sijial,pengnany,yangliud}@amazon.com

Abstract

Automatic Evaluation (AE) and Response Selection (RS) models assign quality scores to various candidate responses and rank them in conversational setups. Prior response ranking research compares various models' performance on synthetically generated test sets. In this work, we investigate the performance of model-based reference-free AE and RS models on our constructed response ranking datasets that mirror real-case scenarios of ranking candidates during inference time. Metrics' unsatisfying performance can be interpreted as their low generalizability over more pragmatic conversational domains such as human-chatbot dialogs. To alleviate this issue we propose a novel RS model called **MERCY** that simulates human behavior in selecting the best candidate by taking into account distinct candidates *concurrently* and learns to rank them. In addition, MERCY leverages *natural language feedback* as another component to help the ranking task by explaining why each candidate response is relevant/irrelevant to the dialog context. These feedbacks are generated by prompting large language models in a few-shot setup. Our experiments show the better performance of MERCY over baselines for the response ranking task in our curated realistic datasets.

1 Introduction

Advancements of neural models (Devlin et al., 2019; Radford et al., 2019; Zhang et al., 2020b; Shuster et al., 2022) has led to the vast continuous research on open-domain dialog systems. Many deployed open-domain dialog systems rely on multiple response generators in order to address the variety of topics within a dialog. Accordingly, response ranking is introduced as a major necessity for ranking different responses based on their quality (Zhou et al., 2018; Wu et al., 2019; Liu et al., 2021).

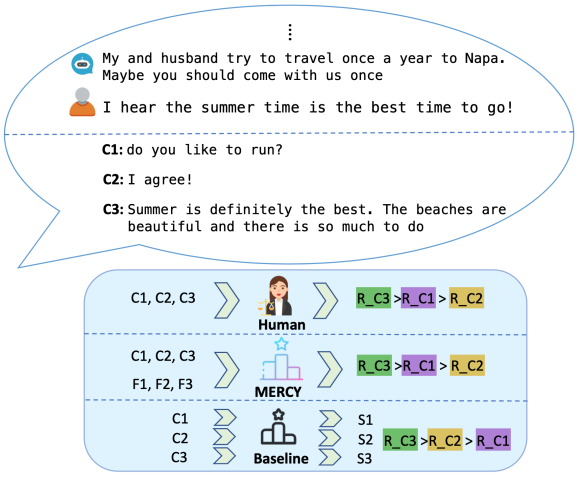


Figure 1: An overview of the response ranking task. Similar to human, MERCY takes all the candidates as input to rank them (denoted as R_C1, R_C2, R_C3) more accurately, while baseline RS metrics score (denoted as S1, S2, S3) each candidate separately. Alongside candidates, MERCY also takes the candidates generated feedback (denoted as F1, F2, F3) as another input.

Response Selection (RS) models were initially trained to rank human-written positive responses higher than synthetically generated negative responses (Gu et al., 2020; Gao et al., 2020; Gupta et al., 2021). Another conforming substitution for RS models can be model-based reference-free **Automatic Evaluation (AE)** metrics that conduct response evaluation along different dimensions such as relevancy (Tao et al., 2018; Ghazarian et al., 2019), engagingness (Ghazarian et al., 2020), coherence (Ye et al., 2020), etc. and have been proven to be more suitable for open-domain conversations (Lowe et al., 2017; Eskenazi et al., 2019) in comparison to the reference-based overlap-based metrics (Papineni et al., 2002; Lin, 2004). The primary intention of AE and RS is to quantify the quality of responses. In this work, we investigate their usage for the response ranking task.

Predominantly, RS models have been tested to distinguish between positive and synthetically gen-

*Work done during an internship at Amazon

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 615–631
September 11–15, 2023. ©2023 Association for Computational Linguistics

erated negative candidates following approaches such as random matching (Gao et al., 2020; Sai et al., 2020; Gupta et al., 2021), text-level manipulations (shuffling, replacing) (Sai et al., 2020; Zhang et al., 2021a), human-written adversarial irrelevant responses (Sai et al., 2020; Gupta et al., 2021; Sato et al., 2020). These synthetically curated test sets are not sufficient representations of real-world inference time candidates that are generated by dialog models. Hedayatnia et al. (2022) demonstrated the effectiveness of training on machine-generated candidates from real user interactions over using synthetic candidates for response selection. However this data is not publicly available.

In this work, we construct the first public dataset of human fine-grained rankings for responses generated by state-of-the-art dialog models in *human-human* dialog contexts, which we denote as *Static Evaluation Setup (SES)*. For our work we also leverage the recently collected *Interactive Evaluation Setup (IES)* (Liu et al., 2023) dataset, which includes human annotations for different machine-generated responses within *human-chatbot* conversations. Our intention is to conduct a survey to evaluate the performance and generalizability of state-of-the-art model-based reference-free AE and RS on our curated datasets that are closer to deployment time ranking scenarios, where one interlocutor is human and the other is a dialogue system. We show that in these realistic test cases, existing RS and AE models exhibit low performance.

To overcome this issue and have a more reliable RS in real case scenarios we propose **MERCY**, which pursues users' behavior of taking different candidates as input and predicting their rankings by relying on their comparable representations (See Figure 1). There is strong evidence that relying on comparable representations is useful such as human preference modeling to improve the performance of LLMs (Bai et al., 2022). MERCY also augments the input with feedback in the form of natural language that explains why or why not a response is relevant. Gupta et al. (2022) introduced an instruction-tuned large language model (LLM) to perform a variety of dialog tasks such as determining if a response is relevant or not. We follow a similar approach of prompting a LLM to evaluate a response; however, we prompt the model to generate more detailed information by not only asking *if* a response is relevant but also *why* it is relevant. We refer to this generated output as feedback. We leverage BLOOMZ-175B (Muennighoff et al., 2022) to generate each candidate's feedback. We train MERCY on the train split of SES and demonstrate that it is more accurate in real-case ranking scenarios in comparison to the best performing automatic metric finetuned on the same training set. Considering multiple candidates together and augmenting responses with feedback both contribute to MERCY's better performance.

Our contributions are summarized as follows:

- We release a new benchmark dataset for response selection, which contains human rankings for responses generated by state-of-the-art neural response generation models.[1]
- We present an in-depth analysis of the performance of AE and RS models on this benchmark dataset and report their low performance and generalizability over different dialog contexts, domains and generated responses.
- We propose a new RS method, MERCY, which receives various candidates simultaneously and takes the generated natural language feedbacks for each candidate as input and learns to rank candidates by minimizing the Kullback-Leibler divergence loss. Experiments show that MERCY outperforms all existing AE and RS metrics by a good margin.

2 Related Work

Due to the vast number of AE/RS models, an in-depth comparison of these metrics is critical. Yeh et al. (2021) performed a comprehensive survey by comparing multiple AE metrics on publicly available evaluation testsets. In this work, we perform a similar survey of model-based reference-free AE/RS models on *response selection testsets*. In contrast to their evaluation testsets where responses are annotated on a Likert scale, which can lead to annotator bias and could make it difficult for a model to predict the exact scores, we only need to evaluate the relative ordering from the predicted output of these methods.

The response selection datasets we leverage are more realistic than previously proposed synthetically generated datasets. Prior research proposed to use simple approaches such as random response selection (Han et al., 2021), corrupting utterances by inserting, substituting and deleting random tokens (Whang et al., 2021), using the mask-and-fill

[1]The dataset will be published upon acceptance.

approach (Gupta et al., 2021) for generating adversarial negative examples or collect human-written negative samples (Sato et al., 2020). Previous work also suggest to augment dialog datasets with synthetically generated positive samples (Mizukami et al., 2015; Khayrallah and Sedoc, 2020; Gupta et al., 2019; Sai et al., 2020; Zhang et al., 2020a).

In a study by Hedayatnia et al. (2022), they demonstrated that using a human-chatbot dataset, where responses were generated by multiple response generators and then annotated by humans for training RS (response selection) models, led to improved performance compared to models trained on synthetically generated datasets. Unfortunately, the dataset they used could not be made public due to privacy concerns, as it contained real-user dialogs. In contrast, our approach involves collecting a similar and realistic response ranking dataset, which we plan to release for future research purposes.

In RS, most models score response candidates independently without considering them together. Zhang et al. (2021b) proposed a joint matching approach that concurrently accepts *exactly four* candidates as input and *selects the only correct response* using log-likelihood as the training objective. Our RS model follows a similar training approach but can handle a *variable number of responses* for ranking. Additionally, our metric stands out from previous work as it combines generated natural language feedback with multiple response candidates, providing the model with valuable information in a natural language format.

Feedback generation has been shown to be beneficial for improving language models. (Shi et al., 2022; Xu et al., 2020; Hancock et al., 2019; Scheurer et al., 2022; Tandon et al., 2022). Shi et al. (2022); Hancock et al. (2019); Scheurer et al. (2022) focused on improving response generation models using three types of human feedback: binary, modular and natural language. While these studies use natural language feedback collected via human annotation, our work *generates feedback* from large language models in a few-shot fashion and use them for the *ranking task*. The closest work to ours is (Gupta et al., 2022), which trains an instruction-tuned large language model to conduct evaluation. However, this work treats feedback as a classification task asking *if* a response is relevant, while we prompt the model to output *why* it is relevant and accompany that with the candidate.

3 Data Sets

To conduct a comprehensive survey on AE/RS models, we look at three response ranking test sets each encompassing different properties: 1) type of dialog contexts, 2) type of candidates for ranking, and 3) type of conversational domains.

DAILYDIALOG++ Sai et al. (2020) composed a dataset consisting of manually created relevant/irrelevant responses for human-human dialog contexts taken from DailyDialog (Li et al., 2017). To create irrelevant responses, annotators were asked to write responses that share similar semantics with the dialog context yet are not acceptable. Hence, in DAILYDIALOG++ both dialog histories and candidate responses are human-written.

SES The responses from DailyDialog++ may not match realistic inference time test sets where responses are machine generated. To deal with this, we collect the Static Evaluation Setup (SES) dataset comprising of various model generated responses for contexts sampled from multiple human-human dialog datasets: DailyDialog (Li et al., 2017), BlendedSkillTalk (Smith et al., 2020), PersonaChat (Zhang et al., 2018), EmpatheticDialogues (Rashkin et al., 2019). Each dialog context contains 8 different responses generated by BlenderBot (Roller et al., 2021), GPT2-XL (Radford et al., 2019) fine-tuned on BlendedSkillTalk (Smith et al., 2020), Plato-2 (Bao et al., 2020), and Plato-XL (Bao et al., 2021) with different decoding mechanisms. The model training and decoding parameters are provided in Section E in the Appendix. We collect two sets of data where the rankings are eventuated from two groups of annotators: 1) in-house annotators familiar with the ranking task (SES_INTERNAL), 2) Amazon Mechanical Turk (AMT) workers (SES_AMT).

Responses in SES_INTERNAL are annotated by two internal annotators on the scale of 0 (not an appropriate response) to 2 (a suitable response). We calculate the normalized mean score for each response in the range of 0 to 1 and assign label 1 to the response if its normalized score is greater than 0.5 or 0 otherwise. To better analyze the performance difference of AE/RS models, we remove turns where all the candidates are 1 or 0 and call it SES_INTERNAL_FILTERED. Although RS model may face such all good or all bad candidates in real-world scenarios, we exclude them to not mislead the performance of RS with random candidate selection. Due to the higher quality of annotations by

Dataset	Num_Responses	Pos/Neg
DAILYDIALOG++	11420	5710/5710
SES_INTERNAL	8000	4601/3399
SES_INTERNAL_FILTERED	7336	4049/3287
SES_AMT	7968	5546/2422
SES_AMT_FILTERED	6488	4098/2390
IES	31849	13519/18330
IES-V2	3240	1330/1910

Table 1: Statistics of response ranking datasets.

internal annotators who are more familiar with the task, we leverage this dataset to test our proposed RS model versus baselines.

The process of collecting SES_INTERNAL is slow due to an insufficient amount of annotators, therefore we use AMT workers for faster data collection. For SES_AMT dataset, 5 AMT workers evaluate each response in the range of 1-5 indicating low-quality to high-quality responses. Here we use more fine-grained ratings which allow us to check if AMT workers understand the range of how good/bad a response can be. We get the median score of each candidate's ratings and normalize it in the range of 0-1. We assign 0/1 label similar to SES_INTERNAL. We remove turns with all good or bad responses, and call it SES_AMT_FILTERED. The statistics of these datasets are shown in Table 1. This dataset is biased toward positive samples as generations are done by state-of-the-art models resulting high quality responses, similar to what happens during real-case scenarios.

IES We take one step closer towards having a realistic response ranking test set by leveraging the dataset from (Liu et al., 2023) where at each turn in a human-chatbot dialog, AMT workers are requested to select all valid responses from multiple machine-generated candidates. The generative models are four GPT2-XL (Radford et al., 2019) models, fine-tuned on BlendedSkillTalk (Smith et al., 2020), TopicalChat (Gopalakrishnan et al., 2019), and WOW(Dinan et al., 2019) datasets, respectively. The model training parameters are described in Section D in the Appendix. We denote this dataset as Interactive Evaluation Setup (IES). Although IES represents a more realistic dataset, its collection process is time consuming as the user has to both converse with the system and annotate each turn for quality. In contrast only one turn needs to be annotated in SES allowing for faster data collection.

A closer look at the IES data shows that some good responses were not marked correctly by AMT workers. This could be because AMT workers may be taking into account factors besides relevancy when selecting a response such as engagingess. Examples of these issues can be seen in Section C in the Appendix. In order to have a more fair comparison we sample 80 dialogs from IES and ask AMT workers to reannotate each response on a scale of [1-5] similar to SES_AMT setup, and denote this dataset as IES-v2. IES-v2 includes a part of IES dataset with more fine-grained annotations in a 1-5 scale that allows better training signals for the RS model. We compute the Fleiss kappa for inter-annotator agreement and get a score of 0.41, which indicates moderate agreement. In this work, we use IES/IES-v2 data for only testing.

4 Analysis of AE/RS Methods for Response Ranking

4.1 AE/RS Methods

Inspired by the survey of automatic metrics on evaluation test sets (Yeh et al., 2021), we compare different AE/RS models on response ranking testsets. We compare AE metrics such as: Ruber (Tao et al., 2018), Bert_Ruber(Ghazarian et al., 2019), Pone(Lan et al., 2020), USR(Mehri and Eskenazi, 2020b), FED(Mehri and Eskenazi, 2020a), FlowScore(Li et al., 2021), Maude(Sinha et al., 2020), Grade(Ye et al., 2020), DynaEval(Zhang et al., 2021a), Predictive_Engagement(Ghazarian et al., 2020), USL(Phy et al., 2020), HolisticEval(Pang et al., 2020), MDD(Zhang et al., 2022), DEAM(Ghazarian et al., 2022). For RS models, we use BM25(Robertson et al., 2009), Dialogrpt(Gao et al., 2020), SABert_KeySem(Gupta et al., 2021).

Bert_Ruber (Ghazarian et al., 2019), Pone (Lan et al., 2020), Maude (Sinha et al., 2020) and DEB (Sai et al., 2020) are classifiers used to predict the relevancy of a response, while Predictive_Engagement (Ghazarian et al., 2020) affirms the positive impact of incorporating an engagement classifier on top of response relevance. FlowScore (Li et al., 2021), Deam (Ghazarian et al., 2022) and DynaEval (Zhang et al., 2021a) evaluate the overall dialog and the connection between utterances. A few AE metrics, such as USL-H (Phy et al., 2020), HolisticEval (Pang et al., 2020), USR (Mehri and Eskenazi, 2020b), FED (Mehri and Eskenazi, 2020a) take into account multiple sub-metrics to achieve a more reliable evaluation metric. Finally, MDD (Zhang et al., 2022) looks for a robust metric that has acceptable performance over multiple domains.

For RS models, BM25 (Robertson et al., 2009) ranks candidates based on their keyword similarities to the context. DialogRPT (Gao et al., 2020) uses human feedback data from Reddit and determines whether a response is human-like to rank the generated candidates. SABert_KeySem (Gupta et al., 2021) is a Speaker-Aware Bert-based (Gu et al., 2020) classifier finetuned on adversarial responses created via mask-and-fill and keyword-based generations.[2]

In addition to aforementioned AE/RS methods, we add random and naive baselines to achieve an exhaustive study. Random baseline randomly assigns scores to responses in the range of 0 to 1. We report the mean aggregation of random baseline performance after 5 runs. Naive baseline reports the best generative model's performance by selecting all its responses as appropriate and the rest candidates as not suitable.

4.2 Ranking Metrics

We report common metrics for response selection:

Hits@K shows the rate of correct responses (selected by human) appearing in the top-k responses scored by each metric. *In our experiments, K is a variable since each turn of evaluation can have different number of human selected responses*. Thus, we report the mean of Hits@K from different evaluation turns.

Recall@1 computes the number of evaluation turns where the highest scored candidate by the metric is also selected by human.

MRR computes the mean of all reciprocal ranks for human-selected responses. Reciprocal rank for each true response shows its rank in the metric's ordered output list. MRR demonstrates the ability of the metric to assign better scores (higher rankings) to human-selected responses.

4.3 Results

We show the performance and generalizability of AE/RS models on the IES and DAILYDIALOG++ datasets in Table 2 and SES datasets in Table 3 by using them without finetuning on the datasets.

For IES we see the best performing AE metric is DEB. This shows the positive impact of pretraining on a large conversational dataset (Reddit) for evaluation. The best performing RS model is SABert_KeySem, which is due to the positive effect of its semantic-based perturbations to generate higher quality negative samples. The slight performance difference between the best performing AE/RS models and baselines on the IES dataset shows the low generalizability of these methods. The main distinctions between IES and the training datasets of the AE/RS models are: 1) differing conversational domains, 2) responses generated by state-of-the-art dialog systems in IES versus human-written or heuristically generated candidates, 3) human-bot dialog contexts in IES versus human-human interactions.

For the SES testsets we also see DEB and SABert_KeyKem are among the best performing AE/RS models, respectively. We see a much higher score from these models on SES in comparison to IES. This may be due to the closeness of dialog history type, which is human-written in SES, to the training datasets of these metrics.

For DAILYDIALOG++, DEB and MDD achieve the best performance. One reason is the domain overlap between the test and train data as both are from DailyDialog. The high performance of these metrics on DAILYDIALOG++ in comparison to SES and IES further shows the low generalizability of AE/RS metrics on different dialog contexts/domains. We don't report the Naive baseline since the candidates are not from different models.

5 Method

Our proposed response selection model MERCY evaluates multiple response candidates for a given dialog context *simultaneously*, and also leverages the feedback generated by LLMs for candidates.

5.1 Few-Shot Feedback Generation

We look into leveraging LLMs for feedback generation via prompting. Specifically, we use the BLOOMZ-175B model (Muennighoff et al., 2022), which is finetuned to follow human instructions for various NLP tasks. To prompt the model for response evaluation, we take three conversations from the FED testset (Mehri and Eskenazi, 2020a), add the question "How relevant are the bot responses?" along with a brief explanation of relevance (or lack thereof) for each response. The exact prompt is available in Section F in the Appendix. Using this prompt, we input the conversations from SES and IES-v2 to obtain feedback for each response in the dataset. Table 4 has an example of our generated feedback showing the model's ability to predict relevance and offer reasoning. More examples can be found in Section G the Appendix.

[2]More details about AE/RS metrics are discussed in Section A of the Appendix.

Metric	Type	DAILYDIALOG++ Hits@K	DAILYDIALOG++ MRR	DAILYDIALOG++ Recall@1	IES Hits@K	IES MRR	IES Recall@1
Naive	basline	-	-	-	50.04	72.40	49.64
Random	basline	50.25	70.12	49.4	48.10	71.40	48.15
Bert_Ruber	AE	55.99	74.41	57.44	47.51	71.00	47.43
PONE	AE	48.14	65.06	43.61	47.59	71.00	47.47
USR	AE	54.69	75.68	59.28	46.57	70.37	46.44
FED	AE	61.28	86.00	75.92	50.88	73.00	50.76
FlowScore	AE	26.01	42.37	37.04	48.06	71.23	48.00
Maude	AE	62.31	84.28	71.8	50.86	72.89	50.77
Grade	AE	69.72	89.74	82.14	46.22	70.24	46.11
DynaEval	AE	92.7	98.88	98.07	48.99	71.85	48.86
Predictive_Engagement	AE	45.92	59.72	35.81	46.91	70.57	46.79
USL-H	AE	60.51	66.01	44.57	47.47	70.86	47.36
HolisticEval	AE	55.43	81.12	68.39	46.9	70.61	49.08
MDD	AE	95.73	99.65	**99.74**	50.9	73.00	51.29
DEAM	AE	54.64	72.72	54.99	49.88	72.37	49.93
DEB	AE	**95.97**	**99.70**	99.39	52.12	73.62	52.11
Dialogrpt	RS	46.87	61.50	38.79	49.95	72.30	49.74
BM25	RS	40.47	63.45	44.05	46.89	70.75	46.73
SABert_KeySem	RS	89.63	99.16	98.51	**52.80**	**74.14**	**53.01**

Table 2: Performance of different AE/RS metrics on DAILYDIALOG++ and IES

Metric	Type	SES_INTERNAL_FILTERED Hits@K	SES_INTERNAL_FILTERED MRR	SES_INTERNAL_FILTERED Recall@1	SES_AMT_FILTERED Hits@K	SES_AMT_FILTERED MRR	SES_AMT_FILTERED Recall@1
Naive	baseline	53.76	74.23	59.54	63.05	78.86	64.50
Random	baseline	58.74	72.71	58.44	63.45	78.14	62.52
Bert_Ruber	AE	56.71	74.02	57.8	65.71	81.62	68.68
PONE	AE	56.06	71.16	52.78	63.60	77.15	61.28
USR	AE	56.54	74.92	58.89	66.46	81.95	69.54
FED	AE	58.69	75.94	59.77	66.54	82.13	69.67
FlowScore	AE	56.14	71.48	53.54	62.09	74.56	57.21
Maude	AE	56.45	73.75	57.25	61.96	76.57	61.05
Grade	AE	56.78	72.66	54.53	65.50	79.30	65.10
DynaEval	AE	58.95	75.96	59.76	63.53	79.72	65.72
Predictive_Engagement	AE	53.55	69.42	51.36	61.32	74.17	57.09
USL-H	AE	57.63	74.21	57.47	66.56	**82.81**	70.9
HolisticEval	AE	56.13	74.18	58.56	63.70	79.74	66.09
MDD	AE	56.62	74.78	61.61	64.78	82.40	**71.89**
DEAM	AE	55.73	74.10	58.01	63.80	81.30	68.80
DEB	AE	**60.83**	**77.56**	**63.03**	63.32	79.28	65.23
Dialogrpt	RS	53.63	69.56	51.47	63.40	76.45	60.30
BM25	RS	56.87	73.13	55.39	63.65	76.03	59.43
SABert_KeySem	RS	57.91	76.07	61.18	**67.85**	82.05	69.79

Table 3: Performance of different AE/RS metrics on SES_INTERNAL_FILTERED and SES_AMT_FILTERED

User: South Padre Island is beautiful. How many boats have you made? Socialbot: about 6 i believe , i lost track after my 3rd User: Haha. Are they big boats or small boats? Socialbot: small boat, about 30 feet
Question: How relevant are the socialbot responses?
Answer: The last system response is relevant in this dialog. The socialbot responds to the user's question about boats by providing details about a boat they have made. Overall, the socialbot's responses are relevant, earning a score of 8 out of 10 for relevance.

Table 4: An example of a generated feedback

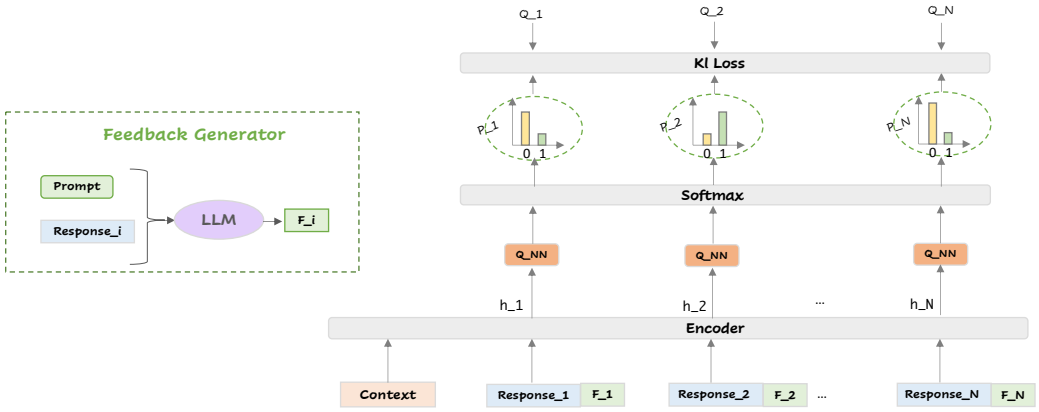


Figure 2: The overview of MERCY and the feedback generator.

To verify the quality of the generated feedback we sample 50 outputs from the SES dataset, 25 predicted by the model as relevant and the remaining as not relevant. We annotate for two dimensions: *Correctness* (Did the model correctly predict if the response was relevant?) and *Makes Sense* (Did the explained feedback make sense?) as yes/no. We find that 82% of the feedback was correct and 56% of the explanations made sense. This suggests that the feedback can be used as input into the response selector. We leave improving the quality of explanations via prompt engineering for future work.

5.2 Ranking Responses Synchronously

Users rank responses by simultaneously receiving all candidates and comparing their quality. However, most RS models consider the candidates independently and predict their scores without conducting the comparison between responses. In this work, we explore the benefits of concurrently receiving *any number* of candidates as input for *rankings*. We train MERCY by minimizing the Kullback-Leibler (KL) Divergence loss between its outputs and ground-truth labels. MERCY's predictions are scores in the range of [0-1], indicating low up to high quality candidates.

Figure 2 gives an overview of MERCY. It takes the context and concatenated candidates as input. Context is composed of all utterances from beginning up to the current turn in the dialogue, splitted with $</UTT>$ token.

$$C = U_1 </UTT> U_2 ... </UTT> U_M \quad (1)$$

Since MERCY can receive multiple candidates with various lengths, the chance of passing the maximum length that can be handled by the encoder in the metric is not negligible. Hence, we use $</UTT>$ to handle such cases by removing the minimum number of utterances from the beginning of the context until all the input can fit in the model.

Following the context C, we pass all the candidates beginning with $[RES]$ special token. In contrast to the metric proposed by Zhang et al. (2021b), which separates each candidate with special tokens to be distinguished between *constant* number of candidates, MERCY is *more generalized* and can process *any number of candidates* as input. It uses the index of each $[RES]$ token to get the corresponding candidate's encoding vector.

$$R = [RES]R_1[RES]R_2 ... [RES]R_N \quad (2)$$

To incorporate feedback into MERCY, we concatenate each response's feedback to itself, and separate them with a $[Feedback]$ special token.

$$R_i = R_i[Feedback]F_i \quad (3)$$

After concatenating C and R we pass the input I ($I = C \cdot R$) through an encoder and get the output embeddings $H \in \mathbb{R}^{|I| \times d}$, where d denotes the hidden dimensional size of the encoder. The hidden representation of each candidate response is returned based on the index of the $[RES]$ token for that corresponding candidate. Similar to how humans rank responses, MERCY is seeing multiple candidates during the encoding process.

$$h\_i = H_{[RES]} \quad \text{where} \quad [RES] \in [RES]R_i \quad (4)$$

Simultaneously, each candidate's hidden representational vector is passed through a linear layer, whose parameters are denoted as W_q, which outputs a scalar value q_i for each candidate. The outputs are then sent through a Softmax layer. The KL-Divergence loss is then minimized between the normalized model outputs and the probability distribution of ground-truth labels.

Metric	Data	SES_INTERNAL_FILTERED Hits@K	MRR	Recall@1	IES-V2 Hits@K	MRR	Recall@1
DEB	SES_AMT_SINGLE	60.51	77.95	63.25	51.55	70.30	51.48
+ F	SES_AMT_SINGLE	59.31	75.90	65.00	52.91	66.26	54.29
MERCY	SES_AMT_SHUFFLED	62.75	78.43	67.39	49.67	64.36	51.03
+ F	SES_AMT_SHUFFLED	63.19	79.51	69.03	50.61	63.86	51.03
+ KL	SES_AMT_SHUFFLED	63.62	80.63	67.50	**53.50**	72.35	53.55
+ KL + F	SES_AMT_SHUFFLED	**64.77**	**81.75**	**69.14**	53.13	**73.23**	**55.62**

Table 5: Performance of AE/RS metrics. KL=Kullback-Leibler divergence loss. F=Feedback

6 Experiments

As seen in Table 2 and Table 3, DEB (Sai et al., 2020) performs the best amongst all existing metrics therefore we use it as our baseline.

For training we leverage the SES_AMT dataset which consists of 8 different candidates per context. We augment the data by perturbing the location of candidate responses. In our experiments, we shuffle the candidates 10 times, and thus have 10 times more training samples. We denote this dataset as SES_AMT_SHUFFLED. In order to compare the benefit of training on multiple candidates versus one, we create a dataset SES_AMT_SINGLE. Models finetuned on SES_AMT_SINGLE take in one response as input while models finetuned on SES_AMT_SHUFFLED take in all candidates simultaneously as input by concatenating them.[3]

When training on the SES_AMT_SINGLE dataset, we append the entire feedback to the response. However, when training on the SES_AMT_SHUFFLED dataset, the tokens are too long since DEB has only 512 positional embeddings. We found that truncating the feedback led to performance degradation. Therefore, for each response we take the corresponding feedback and map to one of the following templates: *"the response is relevant"*, *"the response is not relevant"*, *"the response is somewhat relevant"*. We create a list of keywords for each template. If at least one of the keywords exist in the original generated feedback we replace it with the corresponding template. Through this method the entire response/feedback pair can be fitted with minimal context truncation and results in faster encoding.

7 Results

We present results of comparing MERCY against DEB and MERCY with feedback as additional input in Table 5, we draw the following findings[4]:

1) Training our model in a joint fashion with multiple candidates as input (SES_AMT_SHUFFLED) outperforms training on a dataset with a single candidate as input (SES_AMT_SINGLE), showing the benefit of concurrent response ranking.

2) The use of feedback improves Recall@1 with a 3% for SES_AMT_SINGLE; however, there is not similar improvement when training on SES_AMT_SHUFFLED. This may be due to the fact that we had to rewrite the feedback to contain less information in order to fit into the model's input. However, leveraging KL-Divergence loss alongside feedback shows improvements. Additionally, the feedback provides explanations which allows for interpretability during evaluation.

3) Leveraging the KL-Divergence loss (MERCY + KL) outperforms MERCY with Cross-entropy loss, an improvement of Recall@1 score from 51.03 to 53.55 on the IES-V2 test set. This could be due to the way the data has been annotated. Each response in the dataset has a score between [0-2], to show the rank of responses. While Likert scales suffer due to annotator bias, ranking responses are more robust to this bias. The KL-Divergence loss determines how different the model's output distribution is from the ground-truth distribution and therefore does not rely on the specific Likert scores, but rather on the relative ordering of responses.

8 Conclusion

We introduce MERCY, an RS model that ranks responses by comparing multiple responses synchronously and leveraging natural language feedback. We demonstrate that feedback generated from a LLM through a few-shot setup improves the performance of MERCY. Additionally we introduce the SES dataset, a more realistic RS dataset with human annotated machine generated responses and show the low performance of baseline AE/RS metrics on SES and other existing realistic response raking testsets.

[3]Training parameters are in Section B in the Appendix.

[4]We trained on SES_AMT and evaluated on the test split of SES_INTERNAL_FILTERED.

9 Limitations

(1) In this work, we only look at the relevancy when generating feedback; however, this can be expanded to contain other useful evaluation dimensions such as engagingness and contradiction. (2) We perform experiments on English-only conversations which makes our work biased toward the English language. (3) The performance on IES is far from satisfactory; however, this demonstrates the difficulty of this problem and a strong test set is useful for better development of AE/RS systems. (4) The number of responses that can be ranked by MERCY is limited by the context length of the model; however, the baseline model which only takes in one response at a time will get computationally expensive as the number of responses grows.(5) The responses in SES and IES do not consider the most recent conversational models such as ChatGPT[5].

10 Ethics Statement

All authors of this paper acknowledge and agree with the ACM Code of Ethics. In our study, we ensure that our work is compatible with the provided code, specifically in the terms of presenting a non-offensive dataset construction.

In order to accomplish a comprehensive analysis of AE/RS metrics on the response ranking task, we collect a dataset containing human rankings for generated responses conditioned on existing human-human conversations with polished contents. The main concern is that generated responses based on well-known state-of-the-art dialogue models could have offensive content which is out of our work's scope.

In the feedback generation component leveraged in our proposed metric which is based on prompting a LLM, the outputs show whether a response is relevant or not and explain why that is the case, hence the chance of generating inappropriate contents is near zero.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. Plato-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525.

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

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proceedings of the 2019 Conference on International Conference on Learning Representations (ICLR)*.

Maxine Eskenazi, Shikib Mehri, Evgeniia Razumovskaia, and Tiancheng Zhao. 2019. Beyond turing: Intelligent agents centered on the user. *arXiv preprint arXiv:1901.06613*.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the NAACL 2019 Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen workshop)*.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the 2020 Conference on Association for the Advancement of Artificial Intelligence (AAAI)*.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

[5] https://openai.com/blog/chatgpt/

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2041–2044, New York, NY, USA. Association for Computing Machinery.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P Bigham. 2022. Improving zero and few-shot generalization in dialogue through instruction tuning. *arXiv preprint arXiv:2205.12673*.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. *arXiv preprint arXiv:1907.10568*.

Prakhar Gupta, Yulia Tsvetkov, and Jeffrey Bigham. 2021. Synthesizing adversarial negative responses for robust response ranking and evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3867–3883, Online. Association for Computational Linguistics.

Michael Gutmann and Aapo Hyv¨arinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558, Online. Association for Computational Linguistics.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684.

Behnam Hedayatnia, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2022. A systematic evaluation of response selection for open domain dialogue. In *Proceedings of the 2022 conference on Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Huda Khayrallah and João Sedoc. 2020. Smrter chatbots: Improving non-task-oriented dialog with simulated multi-reference training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4489–4505.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Trans. Inf. Syst.*, 39(1).

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the intrinsic information flow between dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries.

Sijia Liu, Patrick Lange, Behnam Hedayatnia, Alexandros Papangelis, Di Jin, Andrew Wirth, Yang Liu, and Dilek Hakkani-Tür. 2023. Towards credible human evaluation of open-domain dialog systems using interactive setup. In *AAAI 2023*.

Yongkang Liu, Shi Feng, Daling Wang, Kaisong Song, Feiliang Ren, and Yifei Zhang. 2021. A graph reasoning network for multi-turn response selection via customized pre-training. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics*.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 2020 conference on Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 681–707, Online. Association for Computational Linguistics.

Masahiro Mizukami, Hideaki Kizuki, Toshio Nomura, Graham Neubig, Koichiro Yoshino, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Adaptive selection from multiple response candidates in example-based dialogue. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 784–790. IEEE.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Shiki Sato, Reina Akama, Hiroki Ouchi, Jun Suzuki, and Kentaro Inui. 2020. Evaluating dialogue generation systems via response selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 593–599, Online. Association for Computational Linguistics.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with natural language feedback. *arXiv preprint arXiv:2204.14146*.

Weiyan Shi, Emily Dinan, Kurt Shuster, Jason Weston, and Jing Xu. 2022. When life gives you lemons, make cherryade: Converting feedback from bad responses into good labels. *arXiv preprint arXiv:2210.15893*.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. 2022. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. *NAACL Findings.(to appear)*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. 2019. A Sequential Matching Framework for Multi-Turn Response Selection in Retrieval-Based Chatbots. *Computational Linguistics*, 45(1):163–197.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2020. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. *arXiv preprint arXiv:2009.06265*.

Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *Proceedings of First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15—33.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021a. DynaEval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5676–5689, Online. Association for Computational Linguistics.

Chen Zhang, Luis Fernando D'Haro, Thomas Friedrichs, and Haizhou Li. 2022. Mdd-eval: Self-training on augmented data for multi-domain dialogue evaluation. In *Proceedings of the 2022 Conference on Association for the Advancement of Artificial Intelligence (AAAI)*.

Linhao Zhang, Dehong Ma, Sujian Li, and Houfeng Wang. 2021b. Do it once: An embarrassingly simple joint matching approach to response selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 4872–4877. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020a. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9604–9611.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 2020 Conference of the American Chapter of the Association for Computational Linguistics: Demonstration*. Association for Computational Linguistics.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.

A More about AE and RS Metrics

In our study, we target model-based reference-free AE metrics which are more appropriate with no reliance and limitations on ground-truth responses (Novikova et al., 2017; Lowe et al., 2017; Yeh et al., 2021) in open-domain dialogue systems.

Bert_Ruber (Ghazarian et al., 2019) which is an advanced version of Ruber (Tao et al., 2018) leverages positive impact of contextualized word embeddings across with a cross entropy loss to distinguish between human-written responses from random matched ones. **Pone** (Lan et al., 2020) proposes to get top-k most similar randomly chosen responses to the dialogue context as more appropriate negative samples and augment data with more balanced number of generated positive responses. Similar to prior metrics, **Maude** (Sinha et al., 2020) also takes the BERT embeddings of dialogue and learns to assign quality score to the response by being trained with Noise Contrastive Estimation (Gutmann and Hyv¨arinen, 2010) loss between positive and negative samples.

Predictive_Engagement(Ghazarian et al., 2020) returns the engagingness label for corresponding responses and shows its importance in achieving a more precise open-domain dialogue evaluation.

FlowScore (Li et al., 2021) as its name implies models the dynamic flow of the dialogue by leveraging three training objectives to consider the flow and semantic influence of the context and utterances. **DEAM** (Ghazarian et al., 2022) focuses on AMR architecture of conversations to apply semantic-level perturbations and generate more natural looking incoherent conversations. The classification model trained on the resulted dataset has higher performance than its counterpart trained on heuristically generated negative samples. **Grade** (Ye et al., 2020) considers topic transition dynamics by incorporating topic-level graph representations of the dialogue alongside utterance-level contextualized representations trained with the ranking loss and outputs the coherence score for the response. **DynaEval** (Zhang et al., 2021a) transforms each dialogue to a graph where its nodes represent the utterances and the edges demonstrate the dependency between utterances. A graph convolutional network is adopted to measure the quality of the response as well as the whole dialogue. A contrastive loss is defined to distinguish between positive dialogues and negative ones resulted from shuffling and replacements in the utterances.

USL-H (Phy et al., 2020) pays attention to the multi-facet feature of the evaluation. It is a mixture of three metrics for capturing Understandability, Sensibleness, and Likability in Hierarchy. **HolisticEval** (Pang et al., 2020) decomposes the overall quality into four sub-metrics: coherency, fluency, diversity and logical self-consistency. The first two aspects are yielded from probability distribution of GPT-2 model, while diversity is computed by taking into account n-gram entropy and logical self-consistency follows Natural Language Inference models. **USR** (Mehri and Eskenazi, 2020b) is a combination of naturalness, context consistency, knowledge conservation sub-metrics. The likelihood estimated by a fine-tuned RoBERTa model based on MLM objective shows the naturalness. The conditional distribution of a fine-tuned RoBERTa model for the retrieval task can potentially demonstrate the context consistency and naturalness. **FED** (Mehri and Eskenazi, 2020a) is also a multi-dimensional metric without necessity of training. It defines positive and negative follow-up responses designed for each aspect and computes their likelihood using DialoGPT model.

MDD (Zhang et al., 2022)'s main goal is to be a robust metric over different domains and to achieve it two types of models are trained: teacher and student models. The teacher model is trained on human annotated positive and negative responses and later applied on synthetic dataset to get pseudo-labels. Following, the student model is trained to have similar predictions as the teacher model, to be able to be covered after injecting noise to the responses, and to be better adaptable to the multi-domain synthetic datasets.

DEB (Sai et al., 2020) determines the efficiency of pretraining on large-scale dialogue corpora for the evaluation task. The pretraining on Reddit dataset including positive comments and randomly picked negative responses with incorporating cross entropy loss objective and subsequently finetuning on some human crafted positive and adversarial negative samples makes DEB as the most accurate evaluation metric.

Apart from AE metrics that assess the quality of responses from different perspectives, RS models learn to assign different rankings to the responses. One of the pioneer baselines for the response ranking task is **BM25** (Robertson et al., 2009) that leverages keyword similarity to rank responses given a context. We pursue the idea by Henderson et al.

(2019) to rank candidates based on their BM25 vector's inner product with the context's BM25 vector.

DialogRPT (Gao et al., 2020) contains a set of GPT-2 based models that are trained on human feedback data on social media platforms indicating different factors such as the number of replies, maximum length of the dialog after the reply, the difference between upvotes and downvotes. The overall ranking of each response includes scores showing predicted human feedback of responses and whether the response is human-like or not.

SABert_KeySem (Gupta et al., 2021) is a recent response ranking metric that proposes two modern approaches for constructing negative candidates that are used to be classified from responses with higher rankings. First is a mask-and-fill approach that masks spans of utterances and infill them using GPT2-based model conditioned on random contexts. Second is also a GPT2-based model that tries to complete a response conditioned on its keywords and a random context. A Speaker-Aware Bert (SA-Bert) (Gu et al., 2020) classifier trained on such data outperforms different existing baselines.

B MERCY training parameters

We start finetuning our model on the DEB (Sai et al., 2020) checkpoint. We set training for 10 epochs and do early stopping once the loss on the validation set does not go down. We use a training batch size of 8. We use the Adam optimizer with a learning rate of 5e-6. Additionally we use a dropout rate of 0.1. The model was trained on a single NVIDIA A100 GPU. The average inference time of MERCY over 1000 runs on an NVIDIA A100 GPU is 21.88 ms.

C Annotation issues in IES

As seen in Table 6, even though Candidate 3 is a relevant response it is still annotated as not relevant by the worker. This could be due to the fact that Candidates 2 and 4 are more interesting as they offer up more information.

D Models used in IES

All models used in this dataset are GPT2-based (Radford et al., 2019) with a variety of sizes. For each model, both the Language Modeling Head and Multiple Choice Head of GPT2 are finetuned in a Transfer-Transfo fashion (Wolf et al., 2019). The Language Modeling Head takes in the dialog history and learns to predict the follow up response by minimizing the cross-entropy loss. The Multiple Choice Head is fine-tuned to select the ground-truth response amongst five candidates where four are randomly selected negative candidates. During inference nucleus sampling is used to generate the response.

E Models used in SES

The models used to generate responses are: BlenderBot (Roller et al., 2021), GPT2-XL (Radford et al., 2019) finetuned on the Blended Skills Talk dataset (Smith et al., 2020), Plato-2 (Bao et al., 2021) and Plato-XL. GPT2-XL was finetuned in the same fashion as the models in D. To generate 8 candidates for each dialog context we generate three responses from GPT2-XL using nucleus sampling, two responses from Plato-2 using nucleus sampling, one response from Plato-XL using beam search, one response from Blenderbot using nucleus sampling and one response from Blenderbot using beam search.

F Feedback prompt

G Generated feedback examples

Table 8 shows an example of a relevant and non-relevant response with its corresponding feedback. Additionally the last example includes an explanation that is incorrect. In the last example the feedback hallucinates information that is not mentioned in the response. To determine how diverse the generated feedback is we compute self-BLEU and get a score of 0.67 where the lower the score the more diverse the feedback is.

H Mechanical Turk interface for SES data collection

Our qualifications included a HIT Approval Rate greater than 95%, number of HITs approved greater than 1000, location is US and Masters have been granted. Additionally we had asked the Turkers to do a qualification HIT where if they passed they were granted permission to do the actual data collection task. They were paid $15 per hour. A screenshot of the conducted experiment is shown in Figure 3.

Context	Annotation
Bot: I don't blame you. What do you do for work? Human: i manage social media for clients. how about you? Bot: Same here! I manage my social media and my website. I like to use tools like Google and Facebook. Human: google has a heck of a lot of tools for managing all types of things	
Candidate 1: i manage social media for clients	Not relevant
Candidate 2: Same here! I manage my social media and my website. I like to use tools like Google and Facebook.	Relevant
Candidate 3: Yes, and Facebook is a great social media tool too.	Not relevant
Candidate 4: It really does! I use it for everything from social media to email to file sharing.	Relevant

Table 6: Example of annotation issues in IES

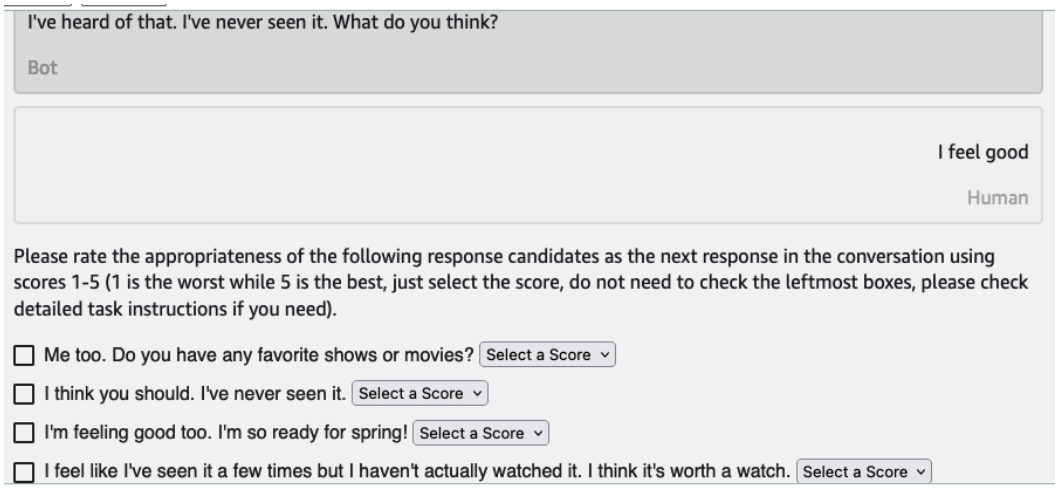


Figure 3: The interface for collecting the SES and IES-v2 data

Below are a series of examples dialogs, followed by a question about the quality of the socialbot's responses. This question must be answered by carefully considering each response in the proceeding dialog, assessing whether the response satisfies the quality/criteria the question is asking for, and ultimately answering the question with a detailed description of the assessment. The answer should be detailed, quote the concerning parts of the dialog and give an overall score out of 10.

User: Hi!
Socialbot: Hi there.
User: What are you listening to?
Socialbot: All sorts of music. I listen when no-one is chatting to me.
User: That's great!
Socialbot: Thanks.
User: Do you chat here often?
Socialbot: I am a talking computer, after all, so of course I could talk here, if I needed to.
User: Let's talk about Taylor Swift!
Socialbot: I watched The Matrix recently. It's got to be my favorite movie.
Question: How relevant are the socialbot responses?
Answer: The socialbot responses in this dialog are not very relevant. The first response to the question about what the socialbot is listening to is somewhat relevant, as it mentions listening to music. However, the subsequent responses do not directly address the user's questions or statements. The response to the question about chatting often simply states that the socialbot is a talking computer and could talk if needed, but does not address the frequency of its chat activities. The response to the statement about Taylor Swift is completely unrelated, discussing The Matrix movie instead. Overall, the socialbot's responses in this dialog score a 2 out of 10 for relevance.

User: Hi!
Socialbot: Hi! What's up?
User: Nothing much, how about you
Socialbot: Not much either.
User: What are you doing
Socialbot: Playing Terraria. What about you?
User: Sitting in a meeting
Socialbot: What kind of meeting?
User: Can't say
Socialbot: It's probably boring, isn't it?
Question: How engaging are the socialbot responses?
Answer: In this dialog, the socialbot's responses could be considered somewhat engaging. The socialbot initiates conversation by asking "What's up?" and continues the conversation by asking the user what they are doing. However, the socialbot's response of "It's probably boring, isn't it?" could be perceived as dismissive or judgmental. Overall, I would rate the socialbot's responses as a 6 out of 10 in terms of engagement.

User: Hi!
Socialbot: Hey
User: How's your day going?
Socialbot: pretty good. yours?
User: Ehh it's fine. I didn't do so well on that history test, actually..
Socialbot: oh what happened?
User: Apparently Christopher Columbus didn't fight in the Civil War :')
Socialbot: hahah wait for real?
User: I know right! Are you taking History next semester?
Socialbot: No I'm not in school anymore
Question: How relevant are the socialbot responses?
Answer: The socialbot responses are relevant in this dialog. The socialbot asks the user how their day is going and responds to the user's mention of not doing well on a history test by asking for more information. The socialbot also responds to the user's question about taking history next semester with information about their current schooling status. Overall, the socialbot's responses are timely and relevant to the conversation, earning a score of 8 out of 10 for relevance.

Table 7: Prompt sent into BLOOMZ-175B

... User: I am an avid gamer I just got a new one I am obsessed with. Socialbot: I am not much of a gamer myself but I do know a few, what game did you get? User: It's called the awakening. What do you do if you don't game? Socialbot: i like to watch movies and tv
Question: How relevant are the socialbot responses?
Answer: The socialbot's last response is highly relevant. The socialbot's response discusses the types of things the user likes to do, including games, movies, and TV. Overall, the socialbot's responses earn an 8 out of 10 for relevance.
... User: i love eating anything italian , watching forensic files and jeopardy Socialbot: Great shows! Have you ever been it italy? User: i have been there a couple of times Socialbot: How is it? I've always wanted to go? User: its beautiful and its nice to use overall Socialbot: I like to run in the mountains.
Question: How relevant are the socialbot responses?
Answer: The system response is not relevant. The socialbot states information about running in mountains and likes of sports. The response is not relevant to the conversation as the user has not indicated that they like sports or running in mountains.
... User: I may take a run myself, but am really thinking of going to American Eagle, my favorite store...they have a sale! Socialbot: What do you like to purchase there? I like the tee's personally User: yes the tee shirts are great! Also have some nice jeans. Socialbot: You have any hobbies?
Question: How relevant are the socialbot responses?
Answer: The last system response is relevant. The system asks the user about hobbies and what they might purchase at American Eagle. The system's last response mentions the store's sale. The system's responses score a 4 out of 10 for relevance in this dialog.

Table 8: More examples of generated feedback

Empathetic Response Generation for Distress Support

Chun-Hung Yeh*, **Anuradha Welivita***, **and Pearl Pu**
School of Computer and Communication Sciences
EPFL, Switzerland
chyeh.work@gmail.com; kalpani.welivita@epfl.ch; pearl.pu@epfl.ch

Abstract

AI-driven chatbots are seen as an attractive solution to support people undergoing emotional distress. One of the main components of such a chatbot is the ability to empathize with the user. But a significant limitation in achieving this goal is the lack of a large dialogue dataset containing empathetic support for those undergoing distress. In this work, we curate a large-scale dialogue dataset that contains ≈1.3M peer support dialogues spanning across more than 4K distress-related topics. We analyze the empathetic characteristics of this dataset using statistical and visual means. To demonstrate the utility of this dataset, we train four baseline neural dialogue models that can respond empathetically to distress prompts. Two of the baselines adapt existing architecture and the other two incorporate a framework identifying levels of cognitive and emotional empathy in responses. Automatic and human evaluation of these models validate the utility of the dataset in generating empathetic responses for distress support and show that identifying levels of empathy in peer-support responses facilitates generating responses that are lengthier, richer in empathy, and closer to the ground truth.

1 Introduction

Psychological distress refers to a state of extreme sorrow, pain, or suffering, both emotional and physical. It is often associated with feelings of discomfort, anxiety, or anguish. The World Health Organization estimates that psychological distress affects 29% of people in their lifetime (Steel et al., 2014). Despite the availability of mental health services, people hesitate to reach them because of the public stigma associated with mental health. There is also a severe shortage of mental health workers (Vaidyam et al., 2019). Thus, recent work investigates how technology can be utilized to meet the needs of people suffering from distress. One

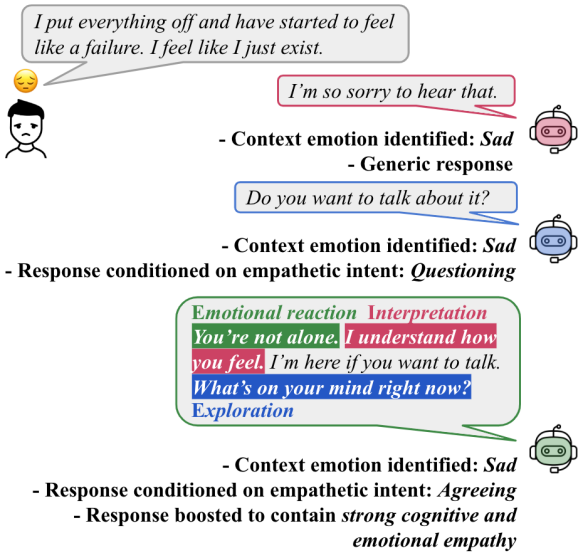


Figure 1: Distress support responses generated by our different chatbot models trained on peer support dialogues. The last response boosted with cognitive and emotional empathy communication mechanisms stands out from the rest as its lengthier and richer in empathy.

such solution is the development of conversational agents or chatbots to deliver distress support (Fitzpatrick et al., 2017; Inkster et al., 2018; Mousavi et al., 2021).

Deep neural networks work very effectively in the development of task-oriented and open-domain conversational agents (Sutskever et al., 2014; Vinyals and Le, 2015; Wen et al., 2015). Most of such dialogue agents can generate syntactically correct and contextually relevant responses. But a major challenge faced by these systems is identifying human emotion and responding in an empathetic manner (Rashkin et al., 2018; Welivita et al., 2021). This is very important when developing chatbots to support distress as one of the major components that contributes to the success of such interaction is the ability to empathize (Bohart et al., 2002; Thwaites and Bennett-Levy, 2007). Recently, researchers have curated emotion-labeled and empathetic datasets such as EmotionLines (Hsu et al.,

*These authors contributed equally to this work.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 632–644
September 11–15, 2023. ©2023 Association for Computational Linguistics

2018), EmoContext (Chatterjee et al., 2019), EmpatheticDialogues (Rashkin et al., 2018), and ESConv (Liu et al., 2021) to enable training dialogue systems that can generate emotion-aware and empathetic responses. However, the above datasets include only a limited amount of dialogues dealing with distress. The dialogues in the first three datasets are more open-domain and span across topics less related to distress. The ESConv dataset that is more focussed on distress contains only 1.3K dialogues covering only 13 distress-related topics. Recent research has curated and conducted analysis on real counseling conversations (Althoff et al., 2016; Zhang and Danescu-Niculescu-Mizil, 2020). But these datasets are not publicly accessible due to privacy and ethical reasons.

To address the above limitations, we curated a large-scale dialogue dataset, named RED (Reddit Emotional Distress), containing ≈1.3M dialogues spanning across more than 4K distress-related topics. The dialogues are scraped from the popular peer support forum, Reddit. Peers are seen to actively engage in such forums to support others undergoing distress and thus they contain distress-related dialogues in abundance spanning a wide range of topics. Prior work has also found that responses from peers contain higher empathic concern for posts seeking help as many peers share similar distressful experiences (Hodges et al., 2010). But as these conversations are available as long threads, the turn-taking structure has to be explicitly extracted and the conversations have to undergo a rigorous pre-processing pipeline including the removal of profanity before they are used to train chatbots. Even then, the dataset can still possess less ideal responses to distress since peers are not trained in delivering distress support as professionals. We take steps to address this by making use of existing empathetic frameworks based on psychology that can be used to identify highly empathetic responses in such dialogues and enabling chatbot models to favor such responses over others.

Empathy is a complex multi-dimensional construct with two broad aspects related to emotion and cognition. The emotion aspect refers to the ability to share the feelings of another person and the cognition aspect refers to the ability to understand and acknowledge how a person feels. In mental health therapy, both emotional and cognitive empathy are equally important (Selman, 1981). Thus, for the development of distress support chatbots, it is vital to understand these types of empathy and the techniques by which these different types of empathy can be elicited. We apply such empathy recognition frameworks on RED to develop several distress support chatbots models. Figure 1 shows an example. In the first instance, identification of the context emotion enables the chatbot to produce a suitable generic response. In the second instance, the chatbot's response is conditioned on a specific empathetic response intent that helps to generate a diversified response. In the third instance, training the model to favour more cognitive and emotional empathy helps in generating lengthier responses containing specific cognitive and emotional empathy communication strategies.

Our contributions are three folds. 1) We curate a large-scale dialogue dataset containing ≈1.3M distress support dialogues spanning across more than 4K distress topics, from a set of carefully selected subreddits. 2) We describe the empathetic dialogue characteristics between the speakers and the listeners in this dataset using statistical and visual means. 3) Using this dataset as a benchmark, we develop four baseline chatbot models. The first two baseline models adapt existing empathetic response generation architectures. On top of them, we develop two new baselines by incorporating a framework that can identify levels of emotional and cognitive empathy in responses contained in RED. Automatic and human evaluation of the models' responses validate the utility of the RED dataset in facilitating empathetic response generation and show that identifying different levels of emotional and cognitive empathy enables generating responses that are lengthier, richer in empathy, and closer to the ground-truth. The code and the datasets are available at `https://github.com/yehchunhung/EPIMEED`

2 Related Work

Many dialogue datasets such as IEMOCAP (Busso et al., 2008), SEMAINE (McKeown et al., 2011), and MELD (Poria et al., 2019) are developed to make chatbots understand users' emotions and respond appropriately. These datasets contain visual, acoustic, and textual signals. More recent work such as EmotionLines (Hsu et al., 2018), OpenSubtitles (Lison et al., 2019), and EDOS (Welivita et al., 2021) are conversation datasets containing TV and movie transcripts translated from voice to text. Though these works intend to build dialogue

datasets by improving the sentence quality, they are still unable to fully model interactions occurring only via text. And most of the dialogues contained in these datasets represent generic day-to-day situations and not psychological distress in particular.

Rashkin et al. (2018) developed the EmpatheticDialogues dataset, inclusive of 25K dialogues grounded on 32 positive and negative emotions. Liu et al. (2021) developed the ESConv dataset, containing ≈1.3K dialogues discussing emotional distress and whose responses are grounded on the Helping Skills Theory (Hill, 2009). But the crowdsourced artificial setting used to curate them makes the dialogue prompts less authentic and the responses less genuine. Because of the cost of crowdsourcing, it also limits the size of these datasets as well as their topic coverage. Thus, a large-scale topically diverse dataset focused on textual conversations between speakers who are emotionally distressed and listeners who actively offer emotional support is lacking in the literature. This type of conversation could be available as recorded therapy sessions between psychologically distressed patients and therapists. However, such counseling datasets used to conduct recent research (Althoff et al., 2016; Zhang and Danescu-Niculescu-Mizil, 2020) are not directly accessible to the public due to ethical reasons. To address these limitations, we curate a large dataset containing peer support dialogues related to a variety of distress-related topics and validate that combined with existing empathy-identifying frameworks, it can potentially be used to develop chatbots that can offer empathetic support to distressful user prompts.

3 Reddit Emotional Distress Dataset

3.1 Data Curation and Preprocessing

Online peer support forums encourage open discussion of often stigmatized psychological concerns and personal distress (De Choudhury and De, 2014; Sharma et al., 2017). They provide alternative means for connection and support when other means of care are less accessible. The anonymity in such platforms facilitates self-disclosure and such discussions help people to feel more supported and less stressed in times of crisis (De Choudhury and De, 2014; Smith-Merry et al., 2019). Reddit is one such platform, which ranks among the most visited websites in the world (Sharma et al., 2017). Reddit users can create community forums called "subreddits" to discuss and support each other on a breadth of topics. Reddit policies also allow researchers to scrape its data and use them for research. Since many people interact in Reddit in a day-to-day basis, the distress-related topics it covers are abundant and have a wide variety. Because of these reasons we chose Reddit to curate conversations that provide support for people in distress.

For this purpose, we choose 8 subreddits: *depression*; *depressed*; *Off My Chest*; *SuicideWatch*; *Depression Help*; *sad*; *Anxiety Help*; and *Mental Health Support*, where such conversations were abundantly present. We used the Pushshift API (Baumgartner et al., 2020) to scrape English textual conversations from the above subreddits. We extracted one dyadic dialogue per conversation thread, selected randomly, thereby diversifying the conversation topics in the dataset. To preserve anonymity, we replaced the usernames with *speaker* and *listener*. The *speaker* here is the user who posted the Reddit post and the *listener* here is the person who commented on it. Dyadic conversations were extracted by selecting comment threads in which only the poster (speaker) and one other commenter (listener) were engaged. For simplicity, we call the original post by the speaker or the first turn in the conversation as the *distress prompt*. Next, we removed HTML tags and URLs from the data, and replaced numerals with a special tag <NUM>. But punctuation marks, emoticons, and emojis were preserved as they can be useful indicators to identify users' emotions.

3.2 Removal of Profanity

To remove profanity from the dataset, we applied `profanity-check` (Zhou et al., 2020), a fast and robust library to detect offensive language. Instead of using hard-coded lists of profane words, it makes use of a linear Support Vector Machine (Cortes and Vapnik, 1995) trained on 200k human-labeled samples of clean and profane text. It is simple but surprisingly effective generalized approach towards profanity checking. When it is applied to a text message, it returns the probability of predicting profanity. Thus, we could set up a threshold to classify the message as profane or not. In our case, we manually set the threshold to be as high as 0.95 because the users sometimes express their feeling aggressively but with no mean intention. This threshold was determined after a thorough inspection of the profane text returned at different thresholds. We removed profane lis-

teners' utterances above this threshold, however, retained profane speakers' utterances as they contain cues about the speakers' state of mind. All the dialogue turns following a removed utterance were also removed to maintain consistency.

3.3 Descriptive Statistics

The resultant RED dataset contains ≈1.3 million dyadic conversations. Table 1 displays the summary of descriptive statistics of conversations present in the dataset as well as in individual subreddits. We used Agglomerative clustering (Murtagh and Legendre, 2014) to cluster distress prompts and recognize clearly identifiable topic clusters. At an optimal clustering threshold of 0.85, the prompts were separated into 4,363 topic clusters. By applying TF-IDF based topic modeling on these clusters, we uncovered some clearly distinguishable distress-related topics. Some of the most common topics identified were *Suicidal ideation*, *Anxiety attacks*, *Weight gain*, *Loneliness*, *Failing college*, and *Covid19*. The topics and their associated keywords are included in the appendices.

3.4 Emotion and Intent Analysis

To analyse the emotions and intents expressed in the RED dataset, we used a BERT transformer-based classifier proposed by Welivita and Pu (2020) and classified the utterances in RED into one of 32 fine-grained emotions and 8 empathetic response intents. This classifier was trained on the EmpatheticDialogues dataset and has a classification accuracy of 65.88% on the EmpatheticDialogues test set, which is comparable with the state-of-the-art emotion classifiers. Manual validation of the labels proposed by the classifier on a random subset of 100 utterances from the RED datasets yielded an accuracy of 64%, which allows us to have reasonable judgments about the RED dataset using the predicted labels. In Figure 2, we visualize the emotion and intent distributions in speaker and listener turns in the RED dataset. It could be seen that the speakers' emotions are mostly centered around negative emotions. The most frequent speakers' emotions that can be observed are *ashamed* (9.98%), *lonely* (8.41%), *sad* (7.52%), and *apprehensive* (5.32%).

A significant proportion of the listener turns contain empathetic response intents. The listeners' intents are mostly centered around *questioning* (10.26%), *agreeing* (7.98%), *suggesting* (5.49%), and *sympathizing* (4.56%). Though empathetic response intents take prominence in the listener turns, they also contain emotional statements that mostly reflect the *sad* emotion (4.98%). This can possibly be explained by the study of affective asymmetry by Vaish et al. (2008) that states negative emotional experiences have more power in triggering negative emotions in the listener as humans are more sensitive to negative emotions.

Figure 3 shows the conversational dynamics in terms of emotion-intent flow patterns that could be observed in the first four dialogue turns. The first and the third turns represent the speaker turns, while the second and the fourth turns represent the listener turns. According to statistics, 93.71% dialogues in the dataset start with a negative emotion. Then in the next turn, the listeners tend to show empathy by means of intents such as *questioning* (35%), *agreeing* (12.43%), *suggesting* (8.11%), and *sympathizing* (7.23%). As the dialogues proceed, we can observe a 278.59% increase of positive emotions expressed in the third turn compared to the first. The speakers mostly express emotions such as *grateful* (7.50%), *trusting* (7.26%), and *hopeful* (6.56%) as a result of the support offered by the listeners. Such conversational dynamics further validate the use of RED in applications concerning empathetic chatbots that can lift up the emotions of people suffering from distress.

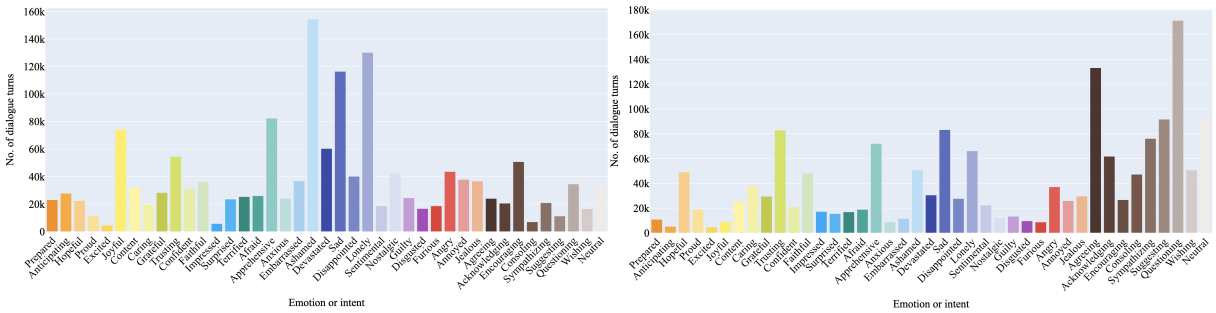
4 Conversational Baselines

Using the RED dataset as a benchmark, we trained four baseline dialogue models. The first two baselines adapted the architecture of EmoPrepend (Rashkin et al., 2018) and MEED (Xie and Pu, 2021), which are state-of-the-art empathetic chatbot models. We also examined different ways existing models can be combined to produce more empathetic responses for distress prompts. For this purpose, we developed another two experimental baselines, EPIMEED and EPIMEED+, by combining MEED with EPITOME (Sharma et al., 2020), which is a theoretically-grounded framework that can identify levels of cognitive and emotional empathy in text-based conversations and extract rationales underlying its predictions. All the models were trained on 80% of RED conversations, leaving 10% of the conversations each for validation and testing. Figure 4 show the architecture of the different models we used for evaluation.

EmoPrepend: This model proposed by Rashkin et al. (2018) is a transformer based encoder-decoder model. During training and inference, the

Subreddit	# Dialogues	# Turns	# Tokens	Avg. # turns per dialog	Avg. # tokens per dialogue	Avg. # tokens per turn
r/depression	510,035	1,396,044	106,967,833	2.74	209.73	76.62
r/depressed	10,892	23,804	1,940,000	2.19	178.11	81.50
r/offmychest	437,737	1,064,467	109,459,738	2.43	250.06	102.83
r/sad	18,827	42,293	3,088,562	2.25	164.05	73.03
r/SuicideWatch	262,469	791,737	59,267,000	3.02	225.81	74.86
r/depression_help	23,678	51,849	5,412,390	2.19	228.58	104.39
r/Anxietyhelp	8,297	18,351	1,428,287	2.21	172.14	77.83
r/MentalHealth Support	3,551	7,931	772,952	2.23	217.67	97.46
All	**1,275,486**	**3,396,476**	**88,336,762**	**2.66**	**226.06**	**84.89**

Table 1: Descriptive statistics of the conversations in the RED dataset.



(a) Speaker turns

(b) Listener turns

Figure 2: Emotion and intent distributions in speaker and listener turns in the RED dataset. The last 9 bars depict empathetic intents and the rest depict emotional statements.

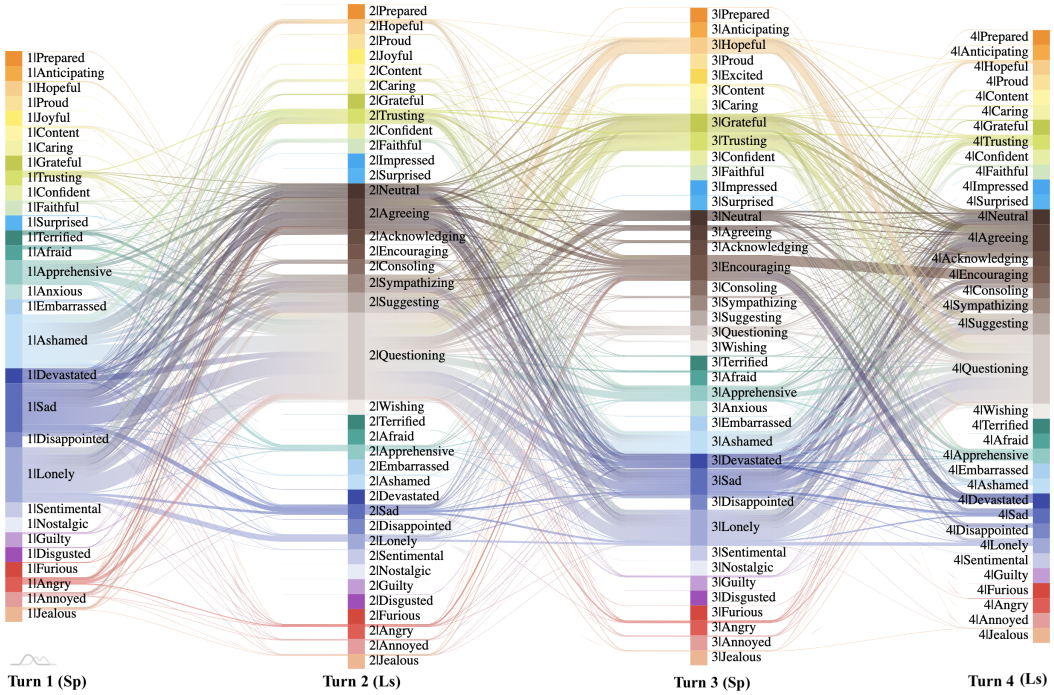


Figure 3: Frequent emotion-intent flow patterns in the RED dataset. For simplicity, only the first four dialogue turns are visualized.

top-k predicted emotion labels from a supervised classifier for the corresponding dialogue context is prepended to the beginning of the token sequence as encoder input. We initialized the encoder of this model with weights from the pre-trained language model RoBERTa (Liu et al., 2019) and trained it on RED, prepending the top-1 emotion or intent predicted by the BERT transformer-based classifier

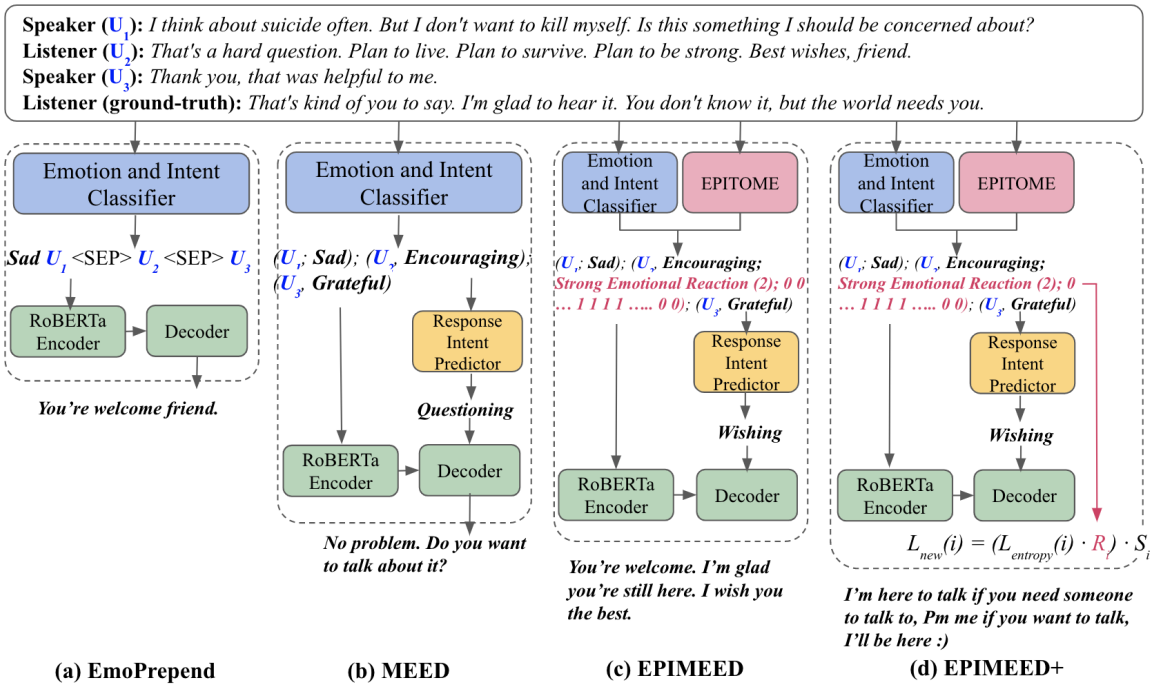


Figure 4: The four models EmoPrepend, MEED, EPIMEED, and EPIMEED+ used for evaluation.

proposed by Welivita and Pu (2020).

MEED: This model proposed by Xie and Pu (2021) consists of two modules: 1) a response emotion/intent prediction module; and 2) a response generation module. The response generation module is an encoder-decoder model that uses the transformer architecture, in which the encoder is initialized with weights from RoBERTa. The response emotion/intent prediction module takes the dialogue context as input and predicts what the emotion or intent of the response should be. This prediction is used to condition the response generated by the decoder in the first module.

EPIMEED: In therapy, interacting empathetically with clients is fundamental to success (Bohart et al., 2002; Elliott et al., 2018). Even though empathy can be interpreted as reacting with emotions of warmth and compassion (Buechel et al., 2018), a separate but key aspect of empathy is also to communicate a cognitive understanding of others, referred to as cognitive empathy. For mental health support, both emotional and cognitive empathy are equally important (Selman, 1981). Thus, it is important to identify such emotional and cognitive empathetic responses amongst other responses that appear in RED and train models in such a way that they favor such responses that reflect cognitive and emotional empathy over others. To support this, we experimented with a new model EPIMEED, by combining MEED with an existing text-based cognitive and emotional empathy identifying framework named EPITOME (Sharma et al., 2020). EPITOME recognizes three empathetic communication mechanisms 1) Emotional reactions (emotional empathy); 2) Interpretations (cognitive empathy); and 3) Explorations (cognitive empathy). For each of these mechanisms, it predicts a numerical value, 0, 1, or 2 — 0: peers not expressing them at all (no communication); 1: peers expressing them to some weak degree (weak communication); 2: peers expressing them strongly (strong communication). Table 2 shows some examples of these communication levels identified in peer support communications.

Empathy type	Communication mechanism	Examples
Emotional	Emotional reactions	- *Everything'll be fine.* (weak) - *I really hope things would improve.* (strong)
Cognitive	Interpretations	- *I realize how you feel.* (weak) - *If that happened to me, I would feel really isolated.* (strong)
Cognitive	Explorations	- *What happened?* (weak) - *I wonder if this makes you feel isolated.* (strong)

Table 2: Examples of emotional and cognitive empathy communication mechanisms identified by EPITOME.

We use this framework to assign a numerical value to each token contained in the listener re-

sponses of the RED dataset. This numerical value is the total of the values predicted by the EPITOME framework for emotional reactions, interpretations, and explorations. This is termed the rationale mask. Next, we feed this information as an additional embedding (in addition to the token embeddings, segment embeddings, position embeddings and emotion embeddings) to the encoder of the response emotion/intent prediction module and response generation module in MEED. We call this additional embedding the *communication embedding*. The rationale behind incorporating this communication embedding is to recognize and give more weight to the parts of the conversation history that expresses empathy. The accuracy, precision, and recall of the response emotion/intent predictor of MEED were increased by 22.88%, 62.65%, and 22.89%, respectively after incorporating this additional information.

EPIMEED+: To enable the model to favour responses containing stronger emotional reactions, interpretations, and explorations while decoding, we further tweaked the loss function associated with MEED such that it incorporates levels of emotional and cognitive empathy predicted by EPITOME. We modified the loss function to be the dot product between the cross entropy loss and the rationale mask predicted by EPITOME. The rationale mask predicted by EPITOME may assign 1 to each token in a text subsequence that may be considered more empathetic than the rest of the text. It acts as an amplifier to the loss so that the model will predict better the tokens with larger empathetic values as predicted by EPITOME. Compared to the original loss $L_{old\,(i)}$, the new loss $L_{new\,(i)}$ given an input sequence i can be written as:

$$L_{old\,(i)} = L_{entropy\,(i)} \cdot S_i$$
$$L_{new\,(i)} = (L_{entropy\,(i)} \cdot R_i) \cdot S_i$$

where $L_{entropy\,(i)}$, R_i, and S_i represent the cross entropy between the predicted and the ground-truth responses, the rationale mask, and the segment mask (the segment mask recognizes the speaker's tokens as 0 and the listener's tokens as 1) of the input i, respectively. By doing so, it facilitates the model to have a higher tendency to generate tokens with stronger levels of emotional and cognitive empathy as recognized by EPITOME.

5 Automatic Evaluation

Automatic evaluation of the models was conducted using a variety of automatic metrics used in evaluating chatbots. They are grouped into diversity-based, word-overlap-based, and embedding-based metrics (details in appendices). Table 3 shows results on the RED test dataset. Accordingly, MEED ranks the top in terms of distinct-unigram and distinct-bigram scores that measures the diversity of the responses. EPIMEED+ ranks the top in majority of word-overlap based metrics and also in embedding average cosine similarity, indicating that responses generated by EPIMEED+ are most likely to contain words from the ground-truth. We also computed the average no. of tokens contained in the responses and EPIMEED+ ranked at the top generating lengthier responses closer to the average length of the ground-truth.

The levels of emotional reactions, interpretations, and explorations computed by EPITOME in the responses generated by the four models are denoted in Table 4. Accordingly, EPIMEED+ generates responses that contain stronger levels of cognitive empathy (as means of interpretations and explorations) than the rest.

6 Human Evaluation

A human evaluation experiment was designed to evaluate the empathetic appropriateness of the responses generated by the four models, by recruiting workers from Amazon Mechanical Turk. We randomly selected 200 dialogue prompts from the RED test dataset and the responses generated by the four models for these prompts to be evaluated by the crowdworkers. The workers were asked to drag and drop the responses generated by the models into areas *Good*, *Okay*, and *Bad*, depending on how empathetically appropriate those responses were to the given prompt. This new way of rating makes it easy to compare many models at once instead of traditional A/B testing, which only allows the comparison of a pair of models at a time. Three workers rated the same response and the final results were computed based on the majority vote.

The human evaluation scores for each of the models is denoted in Table 5. Accordingly, it could be observed that ≈83% of the responses generated by MEED trained on the RED dataset and ≈74% of the responses generated by EPIMEED are rated *Good* with above 90% majority agreement between the workers. None of the responses

Model	Diversity metrics		Word-overlap metrics				Embedding-based metrics		Avg. length (# tokens)
	D1	D2	B1	B2	ROUGE-L	METEOR	Skip Thought	Embedding Average	
EmoPrepend	0.0317	0.1178	0.0513	0.0157	0.0662	0.0434	0.4842	0.7346	16.55
MEED	**0.0618**	**0.2889**	0.0283	0.0123	0.0690	0.0331	0.4874	0.7408	9.68
EPIMEED	0.0487	0.1912	0.0271	0.011	**0.0746**	0.0365	**0.4911**	0.7285	10.30
EPIMEED+	0.0039	0.0181	**0.0543**	**0.0191**	0.0559	**0.0637**	0.4268	**0.7650**	**40.82**

Table 3: Automatic evaluation metrics computed on the RED test dataset. D1 and D2 stands for Distinct-1 and Distinct-2 metrics and B1 and B2 stands for BLEU-1 and BLEU-2 metrics.

Model	Emotional Reactions	Interpre -tations	Explora -tions	Total
EmoPrepend	**1.148**	0.216	0.364	1.728
MEED	0.464	0.412	0.520	1.396
EPIMEED	0.480	0.168	0.732	1.380
EPIMEED+	1.036	**0.728**	**0.948**	**2.712**

Table 4: Average levels of empathy computed for the responses generated by each model, as computed by the EPITOME framework. The maximum value for each of *emotional reactions*, *interpretations*, and *explorations* is 2, making the maximum value for *total* 6.

Model	Good	Okay	Bad	2/3 agree -ment
EmoPrepend	51.82%	21.95%	26.22%	82.00%
MEED	**83.42%**	16.58%	**0.00%**	**99.50%**
EPIMEED	74.07%	**25.93%**	**0.00%**	94.50%
EPIMEED+	42.14%	25.16%	30.70%	79.50%

Table 5: Human evaluation results.

generated by the above models were rated *Bad*. This further validates the potential of the RED dataset in generating more emotionally supportive and empathetic responses for distress-related prompts. Compared to the above models, there is more disagreement between the workers on the ratings of the responses produced by EPIMEED+ (79.5% compared to 99.50% majority agreement in MEED), leading to EPIMEED+ scoring low on the percentage of *Good* ratings. We elaborate on this under the case study.

6.1 Case Study

Table 6 shows some example responses produced by the four models for two different distress prompts. It could be observed that in both cases, the four models generate responses that are emotionally supportive and empathetic to the given prompt. In the first case, the response generated by EPIMEED+ stands out from the rest since it is lengthier and contains stronger emotional reactions and explorations than the responses generated by the other three models. The same observation could be made for EPIMEED+ when responding to the second dialogue prompt, however, it could be noticed that the response is composed of repetitive statements, which was observed in a number of other responses as well. We believe that this is why workers might have more disagreement when assigning a rating to the responses produced by EPIMEED+. Future work should address such repetitive statements in the responses produced by EPIMEED+, while preserving the empathy-richness brought by emotional reactions, interpretations, and explorations.

Prompt 1: *My cat died today, and I can't get it out of my mind. I keep think of him running on a sunny hill but that only makes me cry more.*	
EmoPrepend	*I'm so sorry you're going through this.*
MEED	*I'm sorry for your loss.* **(Sympathizing)**
EPIMEED	*I'm so sorry for your loss.* **(Sympathizing)**
EPIMEED+	*I'm so sorry for what you're going through. I hope you'll be able to reach out to someone who's going through something similar. Have you looked into getting help?* **(Sympathizing)**
Prompt 2: *School is almost over. Things still haven't fixed with my best friend. I want to talk to her about it but I fear it'll just make my situation worse than it already is.*	
EmoPrepend	*I'm so sorry you're going through this. I hope things get better for you soon.*
MEED	*How long have you been together?* **(Questioning)**
EPIMEED	*Do you want to talk about it?* **(Questioning)**
EPIMEED+	*Wanna talk about it? I'm here if you want to talk about it. I'm here to listen if you need someone to talk to. You're not alone.* **(Questioning)**

Table 6: Examples of responses generated by the models for dialogue prompts from the RED test dataset. The sentences highlighted in red are repetitive statements.

7 Discussion and Conclusion

In this paper, we addressed the problem of responding empathetically to distress by curating dialogues from the peer support forum, Reddit, and develop-

ing models by adapting and hybridizing existing empathetic response generation architectures and empathy identifying frameworks. The RED dataset can be used as benchmark to develop similar and better performing chatbot models that can respond to distress. The results of the emotion and intent analysis as well as the automatic and human evaluation results of the experiments conducted on the four baseline chatbot models validate the utility of this dataset in generating emotionally supportive and empathetic responses for distress-related dialogue prompts.

But there are some limitations to this work. Since users responding to distress-related posts in Reddit are not professionals, caution must be taken if these conversations are directly used for training automatic systems that can offer emotional support. Removal of profanity is one step that we have taken towards making such systems reliable and fail-safe. The shift in the emotion of the speaker towards more positive emotions such as gratefulness is also another indicator that the responses do help the speaker lift his/her mood. But deeper analysis such as measuring the level of speaker satisfaction in subsequent dialogue turns and identifying the specific communication techniques that lead to positive outcomes are required when developing an emotionally supportive chatbot based on these conversations. We showed that incorporating existing empathetic frameworks such as EPITOME (Sharma et al., 2020) and conditioning the response on specific empathetic response intents such as in MEED (Xie and Pu, 2021) are good advances in addressing such limitations.

8 Ethics Statement

Data curation: In social sciences, analysis of posts of a website like Reddit is likely considered "fair play" as individuals are anonymous, and users can understand their responses remain archived on the site unless taken action to delete them. The Reddit privacy policy states it allows third parties to access public Reddit content through the Reddit API and other similar technologies and users should take that into consideration when posting.[*] And Reddit data is already widely available in larger dumps such as Pushshift (Baumgartner et al., 2020). We collected only publicly available data in Reddit and it did not involve any interaction with Reddit users. But a study on user perceptions on social media research ethics (Fiesler and Proferes, 2018) highlights some potential harms that can be caused due to social computing research as internet users rarely read or could fully understand website terms and conditions and are unaware that the data they share publicly could be used for research. In particular, this dataset contains sensitive information. So, as suggested by Benton et al. (2017)'s guidelines for working with social media data in health research, in this paper, we share only anonymized and paraphrased excerpts from the dataset. The shared dataset will also contain anonymized usernames and post identifiers. References to usernames and URLs are removed from dialogue content for de-identification. The dataset as well as the models are intended for research purposes only.

Distress support agents: The idea of supportive chatbots for distress is not a new concept. Chatbots such as SimSensei (DeVault et al., 2014), Dipsy (Xie, 2017), Emma (Ghandeharioun et al., 2019), Woebot (`woebothealth.com`), and Wysa (`www.wysa.io`) are some examples. As Czerwinski et al. (2021) state, *About 1 billion people globally are affected by mental disorders; a scalable solution such as an AI therapist could be a huge boon*. Thus, even though empathetic and distress support chatbots may encompass certain ethical implications as pointed out by several researchers (Lanteigne, 2019; Montemayor et al., 2021; Tatman, 2022), based on previous studies we already can acknowledge that the use of chatbots has the potential to improve mental health services notably in relation to accessibility and anonymity. It should be noted that we only address the empathetic component of such distress support agents in this paper. Delivery of therapeutic interventions for distress support should be addressed separately and does not fall under the scope of this paper. And with the significant performance achieved by recent pre-trained language models, going for a deep learning-based solution is one of the choices that can be taken when developing such an agent. But it should not be undermined that because of the unpredictability associated with generative models, they always carry a risk when delivering emotional support to those undergoing distress. Thus, caution should be taken to avoid the delivery of inappropriate responses.

[*] `www.redditinc.com/policies/privacy-policy-october-15-2020`

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102.

Arthur C Bohart, Robert Elliott, Leslie S Greenberg, and Jeanne C Watson. 2002. Empathy.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and Joao Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *arXiv preprint arXiv:1808.10399*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Mary Czerwinski, Javier Hernandez, and Daniel McDuff. 2021. Building an ai that feels: Ai systems with emotional intelligence could learn faster and be more helpful. *IEEE Spectrum*, 58(5):32–38.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399.

Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1).

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.

Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. Emma: An emotion-aware wellbeing chatbot. In *International Conference on Affective Computing and Intelligent Interaction*.

Clara E Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.

Sara D Hodges, Kristi J Kiel, Adam DI Kramer, Darya Veach, and B Renee Villanueva. 2010. Giving birth to empathy: The effects of similar experience on empathic accuracy, empathic concern, and perceived empathy. *Personality and Social Psychology Bulletin*, 36(3):398–409.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Becky Inkster, Shubhankar Sarda, Vinod Subramanian, et al. 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.

Camylle Lanteigne. 2019. Social robots and empathy: The harmful effects of always getting what we want.

Pierre Lison, Jörg Tiedemann, Milen Kouylekov, et al. 2019. Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.

Carlos Montemayor, Jodi Halpern, and Abrol Fairweather. 2021. In principle obstacles for empathic ai: why we can't replace human empathy in healthcare. *AI & society*, pages 1–7.

Seyed Mahed Mousavi, Alessandra Cervone, Morena Danieli, and Giuseppe Riccardi. 2021. Would you like to tell me more? generating a corpus of psychotherapy dialogues. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9.

Fionn Murtagh and Pierre Legendre. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, 31(3):274–295.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Robert L Selman. 1981. The development of interpersonal competence: The role of understanding in conduct. *Developmental review*, 1(4):401–422.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Ratika Sharma, Britta Wigginton, Carla Meurk, Pauline Ford, and Coral E Gartner. 2017. Motivations and limitations associated with vaping among people with mental illness: A qualitative analysis of reddit discussions. *International journal of environmental research and public health*, 14(1):7.

Jennifer Smith-Merry, Gerard Goggin, Andrew Campbell, Kirsty McKenzie, Brad Ridout, Cherry Baylosis, et al. 2019. Social connection and online engagement: insights from interviews with users of a mental health online forum. *JMIR mental health*, 6(3):e11084.

Zachary Steel, Claire Marnane, Changiz Iranpour, Tien Chey, John W Jackson, Vikram Patel, and Derrick Silove. 2014. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International journal of epidemiology*, 43(2):476–493.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Rachael Tatman. 2022. [link].

Richard Thwaites and James Bennett-Levy. 2007. Conceptualizing empathy in cognitive behaviour therapy: Making the implicit explicit. *Behavioural and Cognitive Psychotherapy*, 35(5):591–612.

Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

Amrisha Vaish, Tobias Grossmann, and Amanda Woodward. 2008. Not all emotions are created equal: the negativity bias in social-emotional development. *Psychological bulletin*, 134(3):383.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Xing Xie. 2017. Dipsy: A digital psychologist.

Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 133–147.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. *arXiv preprint arXiv:2005.04245*.

Victor Zhou, Domitrios Mistriotis, and Vadim Shestopalov. 2020. profanity-check.

A Topic Coverage

We used automatic clustering to identify clearly distinguishable topics present in the Reddit distress dialogues. For this purpose, we used "Agglomerative Clustering" tuned for large datasets (Murtagh and Legendre, 2014). It recursively merges pairs of clusters that minimally increase a given linkage distance. The linkage distance was computed using the cosine similarity between pairs of embeddings generated by Sentence-BERT (Reimers and Gurevych, 2019) since the resulting embeddings have shown to be of high quality and working substantially well for document-level embeddings.

We experimented with 8 similarity thresholds from 0.6 to 0.95 with 0.05 increments to cluster distress prompts. At an optimal threshold of 0.85 identified by manual inspection of a randomly selected subset of 10 clusters resulted in 4.93% of the distress prompts (47, 109 prompts in total) getting clustered into 4, 363 clearly identifiable clusters. After applying TF-IDF-based topic modeling to these clusters, clearly distinguishable topics were uncovered. Table 7 shows some distress-related topics and their corresponding keywords.

Distress topic	Keywords
Suicidal	*commit, killing, death, painless, option*
Anxiety attacks	*anxiety, anxious, attacks, social, attack*
Weight gain	*eating, weight, eat, lose, fat*
Loneliness	*lonely, surround, connect, isolated, social*
Failing college	*study, college, class, semester, failing*
Alchoholic	*drinking, drink, alcohol, drunk, sober*
US election	*trump, president, donald, election, war*
Covid19	*covid, 19, pandemic, shambolic, brought*

Table 7: Some distress-related topics identified in the RED dataset along with corresponding keywords.

B Human Evaluation Experiment

In the human evaluation experiment, randomly selected 200 dialogues were bundled into 20 HITs (Human Intelligent Tasks) with each HIT containing 10 such dialogues. Three workers were assigned per HIT. To evaluate the workers' attentiveness to the task, we randomly inserted 3 checkpoints among the 10 dialogues by including the ground-truth response to be rated among the other chatbot-generated responses. Ideally, the ground-truth response should be rated either as *Good* or *Okay* by the workers. If a worker was able to pass at least 2 out of the 3 checkpoints, he was offered a bonus of $0.1 in addition to the compensation of $0.4 offered for completing an assignment. The bonus compensation was used as a mean to motivate workers to pay attention to the task and deliver quality work. In addition, workers who accepted more than 10 HITs were blocked automatically to avoid one worker monopolising the task. We also monitored the total time a worker takes to complete an assignment and rejected all the assignments that were completed in less than 2 minutes and added them back to the pool of available assignments. This enabled us to avoid ratings generated by inattentive workers and spammers. Figure 5 shows the task interface. An overall inter-annotator agreement score (Fleiss' kappa) of 0.2294 was observed, indicating fair agreement between the workers.

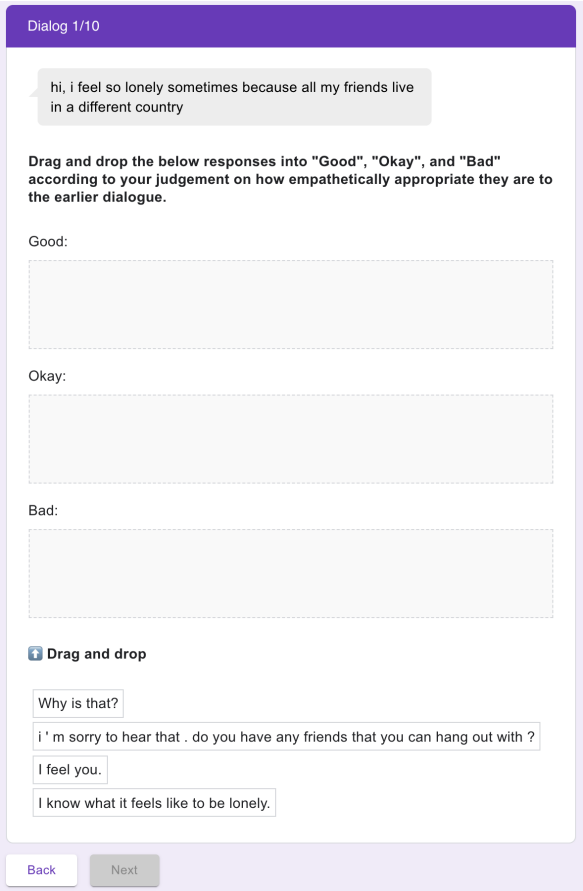


Figure 5: Dialogue response rating interface. A drag and drop mechanism was implemented for ease of rating dialogue responses by the crowdworkers.

C Example Dialogue Responses

Table 8 shows some additional example dialogue responses generated by the four models EmoPrepend, MEED, EPIMEED, and EPIMEED+ for dialogue prompts from the RED test dataset.

Speaker: *Back then I used to be all cheery, and went through my day. Now I just sit at my chair, looking down on myself, and sleeping days away. I reflect on myself no matter what, especially how I am around other people. I just lost all drive to do anything. I just want to lay down, and wish death.*	
EmoPrepend	*I'm so sorry you're going through this. You talked to anyone in your life about any of the stuff you're going through?*
MEED	*What kind of stuff is making you feel this way?* **(Questioning)**
EPIMEED	*Do you want to talk about it?* **(Questioning)**
EPIMEED+	*I've been in a similar situation. You're not alone. Did you tell anyone what you're going through or looked into getting help?* **(Questioning)**
Speaker: *I want out. I want to shoot someone, I want to be shot. I don't know what the fuck I want, I know exactly what I want. I want someone to shoot me, I don't want a new prescription doc, I want to be f**ing shot.*	
EmoPrepend	*What country do you live in?*
MEED	*Why do you want to do this?* **(Questioning)**
EPIMEED	*Why would you want to do this?* **(Questioning)**
EPIMEED+	*What's going on in your life? I'm here to talk if you want to talk. PM me if you need someone to talk to.* **(Questioning)**
Speaker: *I don't have one friend, when my family come into this house all they do is look threw me. No one talks to me unless it's something perverted online. I feel terrible.*	
EmoPrepend	*I'm sorry you're going through this.*
MEED	*How old are you?* **(Questioning)**
EPIMEED	*I'm here if you want to talk. What's going on?* **(Questioning)**
EPIMEED+	*What's making you feel this way? I'm here if you need someone to talk to.* **(Questioning)**
Speaker: *I'm dealing with severe anxiety. I feel like I'm failing my fiance. Sorry for rambling but I just feel like giving up. I just feel like I can't win.* **Listener:** *I'm going through a similar situation. I also feel like I'm failing at life.* **Speaker:** *Well, I guess we can at least go down swinging. At least I'm not alone.*	
EmoPrepend	*I'm sorry you're going through this.*
MEED	*You're not alone.* **(Agreeing)**
EPIMEED	*That's a good point.* **(Acknowledging)**
EPIMEED+	*You're not alone. I'm in the same boat. I can relate to this. I wish you'd be happy to talk to me if you want to talk about it.* **(Acknowledging)**
Speaker: *I just want to be dead, I don't want to live in this shitty world, I want to be released. I'm planning to kill myself later in life. Nobody will notice about my success.*	
EmoPrepend	*I'm so sorry you're going through this. I hope things get better for you soon.*
MEED	*Why do you want to die?* **(Questioning)**
EPIMEED	*Please don't. Why do you feel this way?* **(Questioning)**
EPIMEED+	*Why do you want to kill yourself? Why do you feel like you want to do this? Have you told anyone what you're going through or looked into getting help? If you want to talk I'm here.* **(Questioning)**
Speaker: *They say one person eventually gets attached, Funny how it's always me..* **Listener:** *How did it become that?* **Speaker:** *Cos I'm a very emotional person who gets attached very easily. I eventually become some side pet in any relationship.*	
EmoPrepend	*I'm so sorry you're going through this.*
MEED	*Are you in an open relationship?* **(Questioning)**
EPIMEED	*That's why I don't want any relationship.* **(Neutral)**
EPIMEED+	*I'm sure you're not the only one. I'm here to talk if you need to talk.* **(Neutral)**

Table 8: Examples of responses generated by the models for dialogue prompts from the RED test dataset.

Reasoning before Responding: Integrating Commonsense-based Causality Explanation for Empathetic Response Generation

Yahui Fu, Koji Inoue, Chenhui Chu, and Tatsuya Kawahara
Graduate School of Informatics, Kyoto University, Japan
[fu, inoue, kawahara]@sap.ist.i.kyoto-u.ac.jp
chu@i.kyoto-u.ac.jp

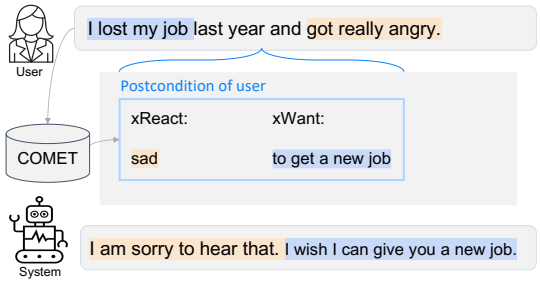
Abstract

Recent approaches to empathetic response generation try to incorporate commonsense knowledge or reasoning about the causes of emotions to better understand the user's experiences and feelings. However, these approaches mainly focus on understanding the causalities of context from the user's perspective, ignoring the system's perspective. In this paper, we propose a commonsense-based causality explanation approach for diverse empathetic response generation that considers both the user's perspective (user's desires and reactions) and the system's perspective (system's intentions and reactions). We enhance ChatGPT's ability to reason for the system's perspective by integrating in-context learning with commonsense knowledge. Then, we integrate the commonsense-based causality explanation with both ChatGPT and a T5-based model. Experimental evaluations demonstrate that our method outperforms other comparable methods on both automatic and human evaluations.

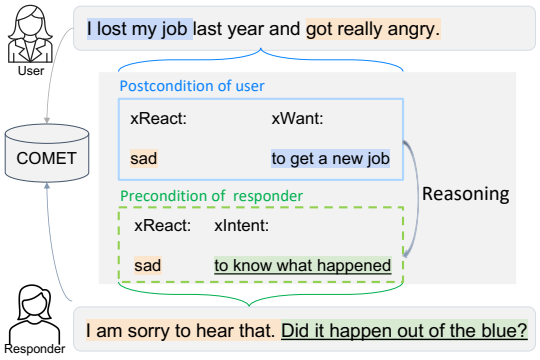
1 Introduction

Empathy is a desirable capacity of humans to place themselves in another's position to show understanding of his/her experience and feelings and respond appropriately. Empathy involves both cognitive and affective aspects (Davis, 1983), including the ability to perceive the user's situation and express appropriate emotions.

Previous work on empathetic response generation has primarily focused on the affective aspect of emotional expression (Lin et al., 2019; Majumder et al., 2020; Li et al., 2020) by emotion detection, without sufficient consideration of context understanding. Recently, there has been a growing interest in exploring context understanding by leveraging external commonsense knowledge for reasoning emotion causes-effects or the user's desires, such as Sabour et al. (2022) and Wang et al. (2022b,a).



(a) Example of using commonsense from COMET to generate a response from the user's perspective.



(b) Example of a response from the actual responder's perspective, based on reasoning reaction and intent to mimic humans.

Figure 1: Two examples to produce a response from different perspectives. The blue solid box contains "xReact" and "xWant" representing the user's emotional reaction and desires. The green dotted box comprises "xReact" and "xIntent," representing the emotional reaction and intention of the actual responder.

However, these approaches focus on understanding the causalities from the user's perspective.

Exploring the causality within the user's context and reasoning his/her desires can be helpful so that the system's intention is aligned with the user's desires, and the response is generated from the user's perspective (Figure 1(a)). However, in real human communication, the responder's intention is not always confined to the user's desires, as shown in Figure 1(b). Relying solely on the user's desire to generate a response may not fully

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 645–656
September 11–15, 2023. ©2023 Association for Computational Linguistics

understand the user's experience, and leads to weak empathy, as shown in Figure 1(a). Therefore, it is necessary to incorporate both the user's perspective (exploring his/her desire and reaction) and the system's perspective (reasoning its intention and reaction to mimic humans) for empathetic response generation.

Through the utilization of COMET (Bosselut et al., 2019), which is a pre-trained GPT-2 model (Radford et al. 2018) fine-tuned on the if-then reasoning graph from ATOMIC (Sap et al., 2019), the system's possible intentions can be predicted to align with the user's desires. However, the system's intention may not be constrained by the user's desire. Therefore, we do not adopt COMET for the system's intention reasoning.

ChatGPT[1] has shown its efficacy in several tasks (Zhao et al., 2023). Bang et al. (2023) introduced ChatGPT's potential in causal reasoning on human-annotated explainable CAusal REasoning dataset (E-CARE) (Du et al., 2022). However, it is based on whether the model can make a judgment on correct causes or effects instead of generating causality explanations. In this paper, we propose to enhance it by incorporating in-context learning with commonsense reasoning for causality explanation. Our main contributions are as follows:

- We propose to integrate a commonsense-based causality reasoning for empathetic response generation, which takes the system's intention and reaction, along with the user's desire and reaction.
- We propose to enhance ChatGPT's capability for causality explanation through the integration of in-context learning with commonsense knowledge (desire, reaction, and intention).
- We present experimental results to demonstrate both ChatGPT and a T5-based model, integrated with the proposed commonsense-based causality explanation, outperform other competitive methods based on both automatic and human evaluations.

2 Related Work

2.1 Commonsense and Causality Reasoning for Empathetic Response Generation

Kim et al. (2021) extracted emotion causes from the dialogue context by utilizing a rational speech act framework. Sabour et al. (2022); Wang et al. (2022b) utilized ATOMIC-2020 (Hwang et al., 2021), which is a collection of commonsense reasoning inferences about everyday if-then events, to enrich context understanding with information on the user's reactions, intentions, effects, needs, and desires. However, these approaches only focus on understanding the causalities within the context from the user's perspective for empathetic response generation, ignoring the system's perspective.

2.2 Large Language Models for Empathetic Response Generation

With the development of large language models such as GPT-3 (Brown et al., 2020) and ChatGPT, many studies have shown their ability on various NLP tasks with either a few-shot or zero-shot setting (Madotto et al., 2021; Lee et al., 2022; Zhao et al., 2023). Lee et al. (2022) introduced two selection methods that choose in-context examples based on emotion and situation information to generate empathetic responses by GPT-3. Zhao et al. (2023) showed ChatGPT's ability on empathetic response generation. In this study, we enhance ChatGPT with a commonsense-based causality explanation prompt for empathetic response generation.

3 Preliminaries

3.1 Knowledge Acquisition

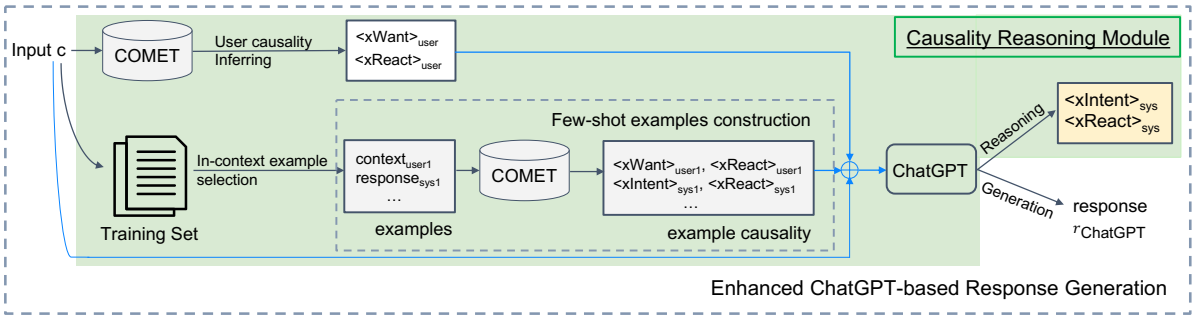
In order to generate commonsense inferences for given events, we adopt a modified BART-based (Lewis et al., 2019) variation of COMET, which was trained on the ATOMIC-2020 dataset (Hwang et al., 2021). This model is suitable for inferring knowledge regarding unseen events (Hwang et al., 2021), like events in the EmpatheticDialogue dataset (Rashkin et al., 2018).

In the training process, we leverage this model to infer the relations of *xWant* and *xReact* for each user's utterance in the training set and the relations of *xIntent* and *xReact* for the system's utterance, which are inferred from the ground-truth response in training. In the testing, we only infer the relations of *xWant* and *xReact* for the user's utterance. The system's *xIntent* and *xReact* will be inferred by the proposed causality reasoning module.

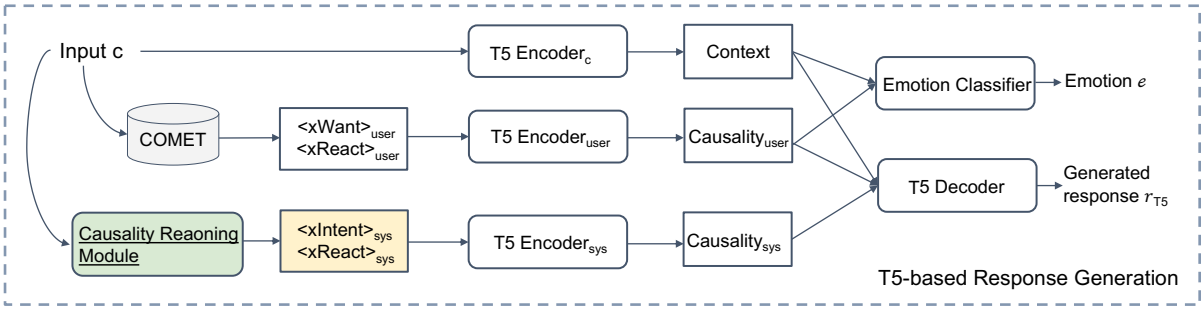
3.2 In-Context Example Selection

We enhance ChatGPT's causality explanation based on the few-shot setting. Given the sensitivity

[1] https://chat.openai.com/



(a) Proposed causality reasoning module and enhanced ChatGPT-based empathetic response generation method.



(b) Integrating the causality reasoning module into a T5-based encoder-decoder for empathetic response generation.

Figure 2: Overview of our proposed model. The input c ends with the user's utterance. The generated response r_{T5} and $r_{ChatGPT}$ are in the role of the system (sys).

of large language models such as ChatGPT to in-context examples (Liu et al., 2021; Lee et al., 2022), we adopt a method similar to Lee et al. (2022) to select top-k examples from the training set based on the similarity between the test conversation and the training conversations. Specifically, we adopt Sentence BERT introduced by Reimers and Gurevych (2019) to encode the sentence semantics of the conversation. In this study, we compute the cosine similarity between the situation utterance of the training set and the test sample, which is annotated in the dataset. Top-k samples are chosen from the training set for each test sample as in-context few shot examples for ChatGPT.

4 Proposed Method

Figure 2 shows an overview of our proposed method. It consists of three components: (1) Causality reasoning module, which aims to enhance the ChatGPT or T5 decoder with a causality explanation for empathetic response generation. (2) Enhanced ChatGPT-based response generation. (3) T5-based response generation, which is based on a trained T5 encoder-decoder to be compared with other approaches that have developed their own model using the EmpatheticDialogue dataset (Lin et al., 2019; Majumder et al., 2020; Li et al., 2020; Sabour et al., 2022; Majumder et al., 2022).

4.1 Causality Reasoning Module based on ChatGPT

As outlined in Algorithm 1, this module consists of four steps. Initially, for a test input c, we employ the method outlined in Section 3.2 to select the top-k relevant training samples, denoted as $\mathcal{S}$, for in-context learning, such as (context1, response1) and (context2, response2) as exemplified in Table 13 in Appendix B.

In the second step, for each selected sample $(c_n, r_n) \in \mathcal{S}$, we leverage the COMET model to infer the *xWant* (c_{nWant}) and *xReact* (c_{nReact}) knowledge corresponding to the user's utterance c_n. Additionally, we extract the *xIntent* ($r_{nIntent}$) and *xReact* (r_{nReact}) knowledge pertaining to the ground truth system response r_n. This information is then concatenated as few-shot examples (Table 13 in Appendix B), denoted as $\mathcal{M}_{prompt}$.

Thirdly, for the test input c, we obtain the *xWant* (c_{Want}) and *xReact* (c_{React}) knowledge using COMET. Finally, they are appended to $\mathcal{M}_{prompt}$ as the prompt to ChatGPT, which reasons *Intent* (r_{Intent}) and *React* (r_{React}) from the system's perspective based on the few-shot learning.

4.2 Enhanced ChatGPT-based Response Generation

The prompt provided to ChatGPT encompasses two components: causality explanation from the user's perspective, predicted by COMET, and causality explanation from the system's perspective, derived through the causality reasoning module described in Section 4.1. These components, along with the few-shot examples, are integrated into ChatGPT to generate empathetic responses.

Algorithm 1 Commonsense-based causality explanation prompt

Require: A training set $\mathcal{D}$=$\{(c_n,r_n)\}_{n=1}^{N}$, N is the number of training samples; a test input (c); c, r represents context, ground truth response, respectively; COMET model $f_\theta(\cdot)$

/*Step 1: In-context examples selection*/

$\mathcal{M}_{sim} \leftarrow$ empty list

for each d=$(c_n,r_n) \in \mathcal{D}$ **do**

Get similarity score: sim_n

$\mathcal{M}_{sim}$.append(sim_n)

end for

$\mathcal{S}$=$\{(c_n,r_n)\}_{n=1}^{k}$=max($\mathcal{M}_{sim}$,k), k is the number of in-context examples

/*Step 2: Get the commonsense knowledge for the selected examples */

$\mathcal{M}_{prompt} \leftarrow$ empty list

for each $s \in \mathcal{S}$ **do**

Get causality information (desire and reaction of user, intent, and reaction of sys) for the sample in $\mathcal{S}$ inferred by COMET

c_{nWant}= $f_\theta(c_n + [xWant])$

c_{nReact}=$f_\theta(c_n + [xReact])$

r_{nItent}=$f_\theta(r_n + [xIntet])$

r_{nReact}=$f_\theta(r_n + [xReact])$

k_n=c_{nWant}+c_{nReact}+$r_{nIntent}$+r_{nReact}

$\mathcal{M}_{prompt}$.append(c_n,k_n,r_n)

end for

/*Step 3: Get the commonsense knowledge for the test sample */

Get causality information (desire and reaction of user) for the test sample c

c_{Want}= $f_\theta(c + [xWant])$

c_{React}=$f_\theta(c + [xReact])$

/*Step 4: prompting ChatGPT, and output the reasoned *Intent*, *React* for generating a empathetic response*/

Input: $\mathcal{M}^{+}_{prompt}$=$\mathcal{M}_{prompt}$+c+c_{Want}+c_{React}

Output: $r_{Itent}, r_{React}, r_{ChatGPT}$

4.3 T5-Based Response Generation

Context and Causality Encoding For a test input c, we use the COMET model to infer the user's causality information, which are desire and reaction of the user (k_{user}: c_{Want} and c_{React}), and use the causality reasoning module based on ChatGPT to infer the system's causality information, which are intention and reaction of the system (k_{sys}: r_{Itent}, r_{React}). We utilize three T5 encoders for encoding input context, the user's causality information, and the system's causality information.

$$\begin{aligned} z_c &= T5^{c}_{enc}(c) \\ z_{user} &= T5^{user}_{enc}(k_{user}) \\ z_{sys} &= T5^{sys}_{enc}(k_{sys}) \end{aligned} \tag{1}$$

Emotion Classification In order to detect the user's affective state, we concatenate the context representations and the user's causality information, and then pass them through a linear layer followed by a softmax operation to produce the emotion category distribution:

$$p_e = softmax(W_e(z_c \oplus z_{user})) \tag{2}$$

where W_e is the weight vector of the linear layer. Given the ground-truth emotion label e^* for each conversation, the cross-entropy loss is computed to optimize the process of emotion classification:

$$\mathcal{L}e = -\log(p_e(e^*)) \tag{3}$$

Response Generation We fuse and feed the information of the user's context and the corresponding causality explanation of the user and the system to a fully-connected (FC) layer.

$$z_{fused} = FC([z_c \oplus z_{user} \oplus z_{sys}]) \tag{4}$$

Subsequently, the target response r_{T5} = [y_1,...,y_T] with length T, is generated by the T5 decoder token by token:

$$p(y_t|c, y_{<t}) = T5^{c}_{dec}(E_{y<t}, z_{fused}) \tag{5}$$

where $E_{y<t}$ denotes the embeddings of the tokens that have been generated. The negative log-likelihood for generation is defined as:

$$\mathcal{L}_{gen} = -\sum_{t=1}^{T} \log p(y_t|c, y_{<t}) \tag{6}$$

The combined loss is defined as:

$$\mathcal{L} = \mathcal{L}_e + \mathcal{L}_{gen} \tag{7}$$

Table 1: Evaluations of reaction and intention reasoned by ChatGPT+Causality$_{user,sys}$, and we set the corresponding knowledge of ground-truth response inferred by COMET as the reference. PBert, RBERT, and FBert represent Bertscore in terms of precision, recall, and F1, respectively.

	Reaction							Intention						
k	F1	BLEU-2	BLEU-3	BLEU-4	PBert	RBert	FBert	F1	BLEU-2	BLEU-3	BLEU-4	PBert	RBert	FBert
2	19.32	6.81	3.16	1.56	91.92	92.60	92.25	13.29	14.65	6.39	3.49	88.90	89.17	89.02
3	21.83	7.12	3.25	1.34	92.28	92.74	92.50	14.49	17.39	8.91	5.37	89.13	89.40	89.26
4	25.83	8.74	3.72	1.48	92.55	92.92	92.73	15.14	19.05	10.07	6.14	89.30	89.54	89.41
5	27.87	8.52	3.55	**1.69**	92.76	92.95	92.85	15.00	19.74	10.69	6.51	89.29	89.46	89.37
6	**29.53**	**9.43**	**4.14**	0.00	**93.15**	**93.22**	**93.18**	**15.71**	**20.72**	**11.55**	**7.25**	**89.62**	**89.76**	**89.68**

5 Evaluation of Causality Explanation based on ChatGPT

We first evaluate how the output of the causality reasoning module is matched with the reaction and intention of the actual (ground-truth) response.

5.1 Dataset

The EmpatheticDialogues dataset of 25k empathetic conversations is used. The ratio for training/validation/test is 8:1:1.

5.2 Setting

For the experiments based on ChatGPT, we used the "gpt-3.5-turbo" engine version with a temperature of 0. We used the 10% of the EmpatheticDialogue test set for this evaluation (250 samples for single-turn and multi-turn settings, respectively).

5.3 Automatic Metrics

(Macro-averaged) F1 score (Rajpurkar et al., 2016), precision, and recall are computed by matching the portion of words in the generation and ground truth that overlap after removing stopwords.
BLEU (Papineni et al., 2002) evaluates the matching between n-grams of the generated response to the ground truth. We utilize BLEU-2, BLEU-3, and BLEU-4 scores.
BERTScore (Zhang et al., 2019) is a BERT-based evaluation measure for text generation, which focuses on lexical semantic similarity between the generated response and the ground truth. We adopt its precision, recall, and F1 score (PBERT, RBERT, FBERT). We used the RoBERTa-Large (Liu et al., 2019) version.

5.4 Results

We evaluate the performance of the system's intention/reaction reasoning under a different number of in-context examples. Experimental results in Table 1 show that increasing the value of k allows for ChatGPT to generate reactions and intentions that are more closely aligned with those inferred by COMET from the ground truth response.

6 Evaluations on ChatGPT-Based Response Generation

Then, we evaluate the responses generated by ChatGPT.

6.1 Evaluation Models

ChatGPT: The prompt given to ChatGPT includes only the chosen in-context raw examples $\mathcal{S}$ from the training set, along with the test sample.
ChatGPT+Causality$_{user,sys}$: The commonsense-based causality explanation prompt $\mathcal{M}^{+}_{prompt}$ is utilized to generate a response by ChatGPT, as illustrated in Algorithm 1.

6.2 Evaluation Metrics

6.2.1 Automatic Metrics

EMOACC: Following Welivita and Pu (2020); Lee et al. (2022), we utilize the EMOACC [2] to measure the emotion accuracy of the generated responses, which is a fine-tuned BERT-base (Devlin et al., 2018) model on the EmpatheticDialogue dataset.
EMPTOME (Sharma et al., 2020): It consists of three empathy metrics: **Interpretations (IP)**, which represent expressions of acknowledgments or understanding of the interlocutor's emotion or situation. For example, a response like *"I also worked hard for the math exam, which made me anxious,"* is considered a stronger interpretation than *"I understand how you feel."* **Explorations (EX)**, which represent expressions of active interest in the interlocutor's situation. For instance, a statement like *"Are you feeling terrified right now?"* exhibits stronger exploration compared to *"What happened?"* **Emotional Reactions (ER)**, which

[2] https://github.com//passing2961/EmpGPT-3

represent expressions of explicit emotions. They are computed by pre-trained empathy identification models.[3] Specifically, RoBERTa (Liu et al., 2019) models are separately fine-tuned for each metric by evaluating the generated response to the number of 0, 1, or 2, a higher value means stronger empathy.

Coherence: We leverage the BERTScore (Zhang et al., 2019) to quantify coherence by computing the semantic similarity between the generated response and the input context.

6.2.2 Human A/B Test

We also conducted A/B test to compare the performance of *ChatGPT+Causality*$_{user,sys}$ and *ChatGPT*. For each comparison, three crowd-workers are asked to choose the better one or select "Tie" based on three aspects: Empathy, Coherence, and Informativeness (Sabour et al., 2022). (1) **Empathy (Emp.)** measures whether the generated response understands the user's feelings and experiences. (2) **Coherence (Coh.)** measures whether the response is coherent/relevant in context. (3) **Informativeness (Inf.)** evaluates whether the generated response conveys more information corresponding to the context.

6.3 Results and Analysis

6.3.1 Number of In-context Examples

We investigate the effect of the number of in-context examples using our proposed commonsense-based causality explanation prompt. Table 2 shows that setting k to 4 results in the highest emotion accuracy, and setting k to 2 yields better exploration and emotional reactions. Therefore, we select k values of 2 and 4 for the experiments.

Table 2: Ablation study on the number of in-context examples k in the prompt.

	EMOACC	IP	EX	ER
k=2	0.24	0.08	**0.57**	**1.10**
k=3	0.25	0.09	0.48	1.05
k=4	**0.27**	0.09	0.40	1.04
k=5	0.25	**0.10**	0.33	1.00
k=6	0.25	0.08	0.32	1.01

[3] https://github.com/behavioral-data/Empathy-Mental-Health

6.3.2 Experimental Results

Table 3 and Table 4 present the results of *ChatGPT* and *ChatGPT+Causality*$_{user,sys}$ with k set to 2 and 4, under the single-turn and multi-turn settings, respectively. In the single-turn setting, a test sample consists of one utterance, while in the multi-turn setting, a test sample contains multiple turns. From the four comparisons, we observe that *ChatGPT+Causality*$_{user,sys}$ outperforms *ChatGPT* in at least 5 out of 7 evaluation metrics. Notably, *ChatGPT+Causality*$_{user,sys}$ significantly outperforms *ChatGPT* on *EMOACC* and *ER*, indicating that *ChatGPT+Causality*$_{user,sys}$ can generate responses with appropriate emotions. This can be attributed to the inclusion of inferred user emotions and reasoned system emotions, which provide appropriate affective information for generating empathetic responses. This improvement addresses the limitation of *ChatGPT* on emotion recognition, as highlighted in Zhao et al. (2023).

ChatGPT+Causality$_{user,sys}$ performs better when k is set to 2 under the single-turn setting. Overall, the performance of *ChatGPT+Causality*$_{user,sys}$ is superior in the single-turn setting compared to the multi-turn setting. This discrepancy can be attributed to COMET, which is trained based on events, not context, making it less effective in predicting causality for long context. To solve the limitation of COMET will be placed on our future work.

The results of the human A/B test in Table 5 show that *ChatGPT+Causality*$_{user,sys}$ is better than *ChatGPT* on the aspects of *Empathy* and *Informativeness* because of the enriched knowledge by the commonsense-based causality explanations.

7 Experiments on T5-Based Response Generation

Finally, we evaluate the responses generated by the T5-based model.

7.1 Evaluation Metrics

(1) Perplexity (PPL) (Vinyals and Le, 2015) which measures the confidence of the generated response. (2)BLEU. (3) D1/D2 (Distinct-1/ Distinct-2) (Li et al., 2016) which evaluates the diversity aspect. (4)BERTscore. (5) Human A/B Test.

7.2 Evaluation Models

Affection-based Methods: MoEL (Lin et al., 2019); MIME (Majumder et al., 2020); EmpDG

Table 3: Evaluations on the effectiveness of causality$_{user,sys}$ when k set to 2 and 4 with the single-turn setting for our ChatGPT-based methods.

	Method	Empathy				Coherence		
		EMOACC	IP	EX	ER	PBERT	RBERT	FBERT
k=2	ChatGPT	0.060	0.073	0.341	0.923	0.877	0.872	0.875
	ChatGPT+Causality$_{user,sys}$	**0.280**	**0.104**	**0.768**	**1.116**	**0.886**	**0.878**	**0.882**
k=4	ChatGPT	0.036	0.081	0.323	0.867	0.882	**0.875**	0.879
	ChatGPT+Causality$_{user,sys}$	**0.280**	**0.120**	**0.528**	**1.076**	**0.888**	0.874	**0.881**

Table 4: Evaluations on the effectiveness of causality$_{user,sys}$ when k set to 2 and 4 with the multi-turn setting for our ChatGPT-based methods.

	Method	Empathy				Coherence		
		EMOACC	IP	EX	ER	PBERT	RBERT	FBERT
k=2	ChatGPT	0.083	**0.065**	0.318	0.917	0.891	0.902	0.894
	ChatGPT+Causality$_{user,sys}$	**0.199**	0.058	**0.397**	**1.094**	**0.899**	**0.907**	**0.901**
k=4	ChatGPT	0.062	**0.072**	**0.297**	0.866	0.896	0.904	0.898
	ChatGPT+Causality$_{user,sys}$	**0.256**	0.065	0.282	**1.007**	**0.902**	**0.904**	**0.901**

Table 5: Human A/B test when k set to 2 and 4 with the single-turn setting for our ChatGPT-based methods.

Comparisons	Aspects	Win	Loss	Tie
ChatGPT+Causality$_{user,sys}$ vs. ChatGPT (k=2)	Emp.	**50.7**	36.0	13.3
	Coh.	**42.7**	42.0	15.3
	Inf.	**51.3**	37.3	11.3
ChatGPT+Causality$_{user,sys}$ vs. ChatGTP (k=4)	Emp.	**49.3**	32.7	18.0
	Coh.	20.0	24.0	**56.0**
	Inf.	**43.3**	40.7	16.0

(Li et al., 2020).

COMET-based Method: CEM (Sabour et al., 2022), which employs commonsense knowledge, such as the user's reactions, intentions, desires, needs, and effects, to enhance its understanding of the interlocutor's situations and emotions.

T5-based Method: LEMPEx (Majumder et al., 2022), which adopts T5 as the encoder-decoder and utilizes a combination of exemplar-based retrieval, a response generator, and an empathy control module to generate empathetic responses.

T5 (Raffel et al., 2020): We utilize the T5 model as our base encoder-decoder architecture, integrating with the emotion classifier. We train it from scratch on the EmpatheticDialogue dataset.

T5+Causality$_{user}$: The T5 model is extended with an additional T5 encoder for user's desires/reactions.

T5+Causality$_{user,sys}$: The T5 model is extended with two T5 encoders for the user's causality attributes (desires/reactions) and the system's causality attributes (intentions/reactions), respectively.

7.3 Settings

We trained T5-small (Raffel et al., 2020) from scratch on the EmpatheticDialogues dataset. The learning rate is set to 0.00001, the batch size is set to 8, we utilize the top-k search decoding strategy with k set to 20, and sampling with the temperature set to 0.2, the max generation length set to 40.

7.4 Results and Analysis

Previous studies (Sabour et al., 2022; Majumder et al., 2022) have shown that CEM and LEMPEx outperformed MoEL, MIME, and EmpDG. Therefore, we compared our method with CEM and LEMPEx in the human A/B test. Automatic evaluation results shown in Table 6 and human A/B test results shown in Table 7 demonstrate the effectiveness of the proposed commonsense-based causality explanation (Causality$_{user,sys}$). The performance comparison presented in Table 8 demonstrates the superiority of our method over the baselines in terms of emotion accuracy (EMOACC), interpretation (IP), and emotion reaction (EX) when compared to the ground truth.

7.5 Comparison between T5-based and ChatGPT-based Response Generation

We conducted a performance comparison between the T5-based and ChatGPT-based response generation, as presented in Table 9. In terms of "Em-

Table 6: Automatic evaluation results of baselines and our T5-based method. Bold denotes the best score.

	Methods	PPL ↓	BLEU-2	BLEU-3	BLEU-4	D1	D2	PBERT	RBERT	FBERT
Baselines	MOEL	37.63	8.63	4.25	2.43	0.38	1.74	86.19	85.67	85.91
	MIME	36.84	8.37	4.31	2.51	0.28	0.95	86.27	85.59	85.92
	EmpDG	38.08	7.74	4.09	2.49	0.46	1.90	86.09	85.49	85.78
	CEM	36.36	6.35	3.55	2.26	0.54	2.38	86.61	85.39	85.98
	LEMPEx	30.42	2.1	0.8	0.35	1.02	**10.81**	83.60	83.09	83.34
Ours	T5	46.13	3.59	1.94	1.15	0.49	2.82	86.69	84.07	85.35
	T5+Causality$_{user}$	15.26	4.84	1.97	0.89	**1.08**	10.75	90.16	89.48	89.80
	T5+Causality$_{user,sys}$	**13.07**	**10.53**	**6.34**	**4.06**	0.75	5.52	**92.24**	**90.76**	**91.48**

Table 7: Results of human A/B test for our T5-based model.

Comparisons	Aspects	Win	Loss	Tie
T5+Causality$_{user,sys}$ vs. CEM	Emp.	**42.0**	40.0	18.0
	Coh.	**38.7**	33.3	28.0
	Inf.	38.3	**44.3**	17.3
T5+Causality$_{user,sys}$ vs. LEMPEx	Emp.	**53.0**	35.0	12.0
	Coh.	**39.0**	33.3	27.7
	Inf.	**50.0**	38.0	12.0

Table 8: Evaluation results of the responses generated by our T5-based method and baselines. The closest to the ground truth is marked as bold.

Methods	EMOACC	IP	EX	ER
MoEL	0.103	0.184	0.209	1.166
MIME	0.076	0.099	0.207	1.256
EmpDG	0.091	0.150	0.169	1.270
CEM	0.091	0.091	0.569	0.950
LEMPEx	0.090	0.135	0.861	**0.575**
T5	0.049	0.110	0.408	1.299
T5+Causality$_{user}$	0.093	0.172	**0.685**	0.784
T5+Causality$_{user,sys}$	**0.125**	**0.271**	0.498	0.751
Ground Truth	0.190	0.279	0.688	0.501

pathy," *ChatGPT+Causality*$_{user,sys}$ outperforms *T5+Causality*$_{user,sys}$ for EMOACC, EX, and ER, but performs worse for IP. Stronger interpretation (IP), which involves understanding and empathizing through shared experiences (Sharma et al., 2020), is more frequently observed in the T5-based model, which was trained from the ground truth. In contrast, ChatGPT-based generation is not constrained by the ground truth and tends to respond from the perspective of a machine.

In terms of "Diversity" and "BLEU," it is evident that *ChatGPT+Causality*$_{user,sys}$ exhibits a larger diversity but results in a higher degree of mismatch with the ground truth (lower BLEU scores), indicating a potential need of balancing the response diversity and the accuracy in generating empathetic responses.

Table 9: Automatic evaluation results of T5+Causality$_{user,sys}$ and ChatGPT+Causality$_{user,sys}$ (k=2, with whole test set and both single and multi-turn settings).

Evaluations		T5+ Causality$_{user,sys}$	ChatGPT+ Causality$_{user,sys}$
Empathy	EMOACC	0.125	**0.235**
	IP	**0.271**	0.046
	EX	0.498	**0.668**
	ER	0.751	**1.109**
Diversity	D1	0.75	**2.91**
	D2	5.52	**16.44**
BLEU	BLEU-2	**10.53**	3.95
	BLEU-3	**6.34**	2.17
	BLEU-4	**4.06**	1.32

Comparative case studies between T5-based and ChatGPT-based models with corresponding baselines can be seen in Appendix C.

8 Conclusions and Future Work

We have proposed a commonsense-based causality explanation approach for diverse empathetic response generation that considers the system's intentions and reactions as well as the user's desires and reactions. Specifically, we enhance ChatGPT's ability to reason the system's intentions and reactions by integrating in-context learning with commonsense knowledge (desire, reaction, and intention). We have integrated the commonsense-based causality explanation with both ChatGPT and a trained T5 model. The experimental results demonstrate that our method outperforms other competitive methods on both automatic and human evaluations.

In the future, we will explore fine-grained approaches for causality explanation from the perspective of both the user and the system.

Acknowledgements

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011. This work was also supported by JST, the establishment of university fellowships towards the creation of science and technology innovation, Grant Number JPMJFS2123.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. *arXiv preprint arXiv:2205.05849*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, volume 35, pages 6384–6392.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. *arXiv preprint arXiv:2109.08828*.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *29th COLING*, pages 669–683.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, pages 110–119.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multi-resolution interactive empathetic dialogue generation. In *28th COLING*, pages 4454–4466.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *EMNLP-IJCNLP*, pages 121–132.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2022. Exemplars-guided empathetic response generation controlled by the elements of human communication. *IEEE Access*, 10:77176–77190.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. In *EMNLP*, pages 8968–8979. Association for Computational Linguistics (ACL).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *AAAI*, volume 36, pages 11229–11237.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*, volume 33, pages 3027–3035.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Jiashuo Wang, Yi Cheng, and Wenjie Li. 2022a. Care: Causality reasoning for empathetic responses by conditional graph generation. *EMNLP findings*.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022b. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. *arXiv preprint arXiv:2210.11715*.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

A Case Analysis on the COMET

We evaluate the effectiveness of COMET in inferring intents and reactions since ChatGPT's ability to reason them is sensitive to the given in-context examples. We assess 60 samples from the EmpatheticDialogue dataset based on two evaluation metrics: (1) Whether the inferred intents or reactions capture the context; (2) whether there are any conflicts among the generated intents or reactions. We find that 51 out of 60 intent predictions and 46 out of 60 reaction predictions are acceptable. Table 10 and 11 show the example of reasoned intentions and reactions, respectively.

Table 10: Example intents inferred from COMET

An accepted example: sys: Did you suffer any injuries? sys's intents: to make sure they are ok; to know if you are ok.
An unaccepted example that does not satisfy metric (1) sys: I understand that one, they are my favorite place to eat. sys's intents: to eat food; to eat good.
An unaccepted example that does not satisfy metric (2) sys: Jeez! It's so unfortunate... very sad really. sys's intents: to be sad; to be happy.

Table 11: Example reactions referred by COMET

An accepted example sys: That's not good. Do you own a gun? sys's reactions: scared; worried; nervous; fearful; angry
An unaccepted example that does not satisfy metric (2) sys: oh man. I'm all about discipline! I don't like spoiled bratty kids. sys's reactions: angry; good; happy; controlling; bad

B Introduction in the prompt for ChatGPT

The introduction in the prompt for ChatGPT is shown in Table 12, and the few-shot examples construction is in Table 13.

C Case Studies and Error Analysis

Table 14 shows a case about the comparison between *ChatGPT* and *ChatGPT+Causality*$_{user,sys}$, and illustrates the impact of our proposed commonsense-based causality explanation. We can see that both the responses by *ChatGPT* and *ChatGPT+Causality*$_{user,sys}$ show emotion reactions to the user's context. However, *ChatGPT+Causality*$_{user,sys}$ outperforms *ChatGPT* by providing detailed suggestions that align with the user's desires based on reasoned intentions. As discussed in Section A, COMET is not always reliable in its predictions. This sensitivity is evident in Table 15, where the user's inferred desires mislead the reasoned intentions of the system.

Table 16 further shows comparative case studies between T5-based and ChatGPT-based models with corresponding baselines.

Table 12: Introduction template to ChatGPT for causality reasoning and empathetic response generation.

Introduction: Assuming that you are sys, who is a friend of the user. You are empathetic sometimes. In this task, you are given the user's input and the information of "user wants to:" and "user reacts to:": "user wants to:", which means what the user wants to do after the input; "user reacts to:", which means how the user react to the input. After that, please reason about the following two parts: "sys's intent:": which means what the sys wants to do after the input, or what's the intent of sys to respond to the input; "sys reacts to:", which means how the sys reacts to the input. Then you respond (should be concise, no more than 30 words) to the input based on the information of user's input, "user wants to:", "user reacts to:", "sys's intent:", "sys reacts to:". "sys:": which means the response of sys. Please generate the following three parts in the format below: sys's intent: sys reacts to: sys:

Table 13: Few-shot examples (top-2 examples).

Test input	user: I'm so excited because I'm finally going to visit my parents next month! I didn't see them for 3 years.	
Few-shot1	context1	user1: Someone is visiting me soon and I can't wait! sys1: Who is it? user1: My mom, she is amazing.
	example causality	<xWant>$_{user1}$: to have a good time. to talk to their mom. to have fun with Mom. <xReact>$_{user1}$: excited. happy. satisfied. good. loved. <xIntent>$_{sys}$: to be with her. to be loved. to be nice. happy. <xReact>$_{sys}$: happy. excited. proud. good. loving.
	response1	sys1: I bet she is! I am so glad you get to see her. Mom's are awesome!
Few-shot2	context2	user2: My family is coming to visit! sys2: Awesome. When are they coming and for how long? user2: They are coming next year from Africa!
	example causality	<xWant>$_{user2}$: to have a good time. to go to the airport. to have fun with the family. <xReact>$_{user2}$: happy. excited. happy. excited. loved. <xIntent>$_{sys2}$: to see the sights. to be with family. to be with them. to have fun. <xReact>$_{sys2}$: happy. excited. satisfied. tired. relieved.
	response2	sys2: That's a long trip. I hope they have a good time.

Table 14: Case studies of the proposed method and baseline. Blue texts represent cognition, and orange texts show affection. Desires$_{user}$ and Reactions$_{user}$ are inferred by COMET, while Intentions$_{sys}$ and Reactions$_{sys}$ are reasoned by ChatGPT+Causality$_{user,sys}$. "sys" is the abbreviation of system.

Context	user: I'm so scared of bugs! i found one in my hair yesterday and almost died.
ChatGPT	sys: That's so scary! Did you manage to get it out?
Desires$_{user}$	user wants: to get rid of bugs; to go to the doctor; to run away.
Reactions$_{user}$	user reacts to: scared; frightened; terrified; afraid; nervous.
Intentions$_{sys}$	sys's intent: to comfort; to help; to offer advice; to be supportive; to avoid bugs.
Reactions$_{sys}$	sys reacts to: empathetic. caring. helpful. understanding. supportive.
ChatGPT+Causality$_{user,sys}$	sys: I know how you feel, bugs are the worst. Have you tried using bug spray or wearing a hat outside?

Table 15: Error analysis of COMET knowledge's negative impact on the causality explanation.

Context	user: I'm so grateful for my family due to an incident with my friend.
ChatGPT	sys: What happened with your friend?
Desires$_{user}$	user wants: to make amends to their friend; to make amends for the incident.
Intentions$_{sys}$	sys's intent: to help; to support; to be a good friend; to be there for user.
ChatGPT+Causality$_{user,sys}$	sys: I'm here for you if you need any help making amends with your friend.

Table 16: Comparative case studies between T5-based and ChatGPT-based models with corresponding baselines. Meanwhile, Causality$_{user,sys}$ includes both Causality$_{user}$ and Causality$_{sys}$.

	Context		user: I'm so sad because I've read an article about a newborn girl who died because her parents didn't believe in medication and doctors.
Baselines	MOEL		sys: I am so sorry to hear that. I hope you are able to find her.
	MIME		sys: That is sad. I am sorry to hear that.
	EmpDG		sys: I am sorry to hear that. I hope you find out of it.
	CEM		sys: Oh no, I am so sorry to hear that.
	LEMPEx		sys: I'm sorry to hear that. I'm so sorry to hear that.
Ours	T5		sys: Oh no! I'm so sorry to hear that. What happened?
	T5+Causality$_{user}$	Causality$_{user}$	user wants: to tell their friends; to cry; to find out what happened. user reacts to: sad; sorrowful; depressed; upset.
		Response	sys: Oh, that's terrible. I am so sorry for your loss.
	T5+Causality$_{user,sys}$	Causality$_{sys}$	sys's intent: to comfort spk1; to show empathy; to be supportive. sys reacts to: sad; upset; empathetic; sympathetic; concerned.
		Response	sys: I know what you mean, I would have been so mad!
Baseline	ChatGPT		sys: That's heartbreaking. It's important to trust medical professionals and seek help when needed.
Ours	ChatGPT+Causality$_{user}$		sys: That's really heartbreaking. I'm here for you if you need to talk or just need someone to listen.
	Context		user: I think that the girl of my dreams likes somebody else. I feel very sad about it.
Baselines	MOEL		I am sorry to hear that. I hope you can find a new one.
	MIME		I am sorry to hear that. I hope you can find a good relationship with her.
	EmpDG		I am sorry to hear that.
	CEM		I am sorry to hear that.
	LEMPEx		I understand that. it's really stressful.
Ours	T5		I'm sorry to hear that. I hope you can find a new job.
	T5+Causality$_{user}$	Causality$_{user}$	user wants: to find a new girlfriend; to ask her out. user reacts to: sad; upset; unhappy; disappointed; angry.
		Response	sys: what did he buy you?
	T5+Causality$_{user,sys}$	Causality$_{sys}$	sys's intent: to comfort; to encourage; to help find a new love. sys reacts to: sympathetic; caring; supportive; encouraging.
		Response	sys: I know how you feel. I have been there before and it is hard to find a new one.
Baseline	ChatGPT		sys: I'm sorry to hear that. It's tough when we have feelings for someone who doesn't feel the same way.
Ours	ChatGPT+Causality$_{user,sys}$		sys: I'm sorry to hear that. Maybe it's time to move on and find someone who appreciates you.

Author Index