UNIVERSITY OF CALIFORNIA

Los Angeles

Conversational Modeling with Human Values, Social Relations, Mental States, and Structure Learning

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

Liang Qiu

2022

ABSTRACT OF THE DISSERTATION

Conversational Modeling with Human Values, Social Relations, Mental States, and Structure Learning

by

Liang Qiu
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2022
Professor Achuta Kadambi, Co-Chair
Professor Song-Chun Zhu, Co-Chair

Teaching machines to speak and act like a human is one of the longest-running goals in Artificial Intelligence. This thesis tackles two important problems in building the next generation of dialogue systems: enhancing the emotional intelligence of social chatbots and learning semantic structures from dialogue corpus. On the one hand, taking into account emotional quotient in dialogue system design help machine to mimic human behavior and further improve the long-term user engagement. On the other hand, extracting structural information from dialogue data is critical for us to analyze user behavior and system performance. The technology could be applied to various areas in computational linguistics, such as dialogue management, discourse analysis, and dialogue summarization.

This thesis consists of two parts. In the first part, we aim to present our efforts at studying emotional intelligence in dialogues systems. We break down the problem into three subjects: 1) the modeling and incorporation of human values, *i.e.,* people tend to have common attitudes towards some statements or scenarios; 2) the inference of social relations between interlocutors

from dialogues. Chatbots with such inferring capability can understand human behavior better and act appropriately; and 3) the modeling and tracking of speakers' mental states. This is beyond understanding what users say to perceive what users imply, requiring agents to mentally simulate the evolution status of the environment.

In the second part of this thesis, we investigate how we can extract structural information from dialogue corpus. In particular, we pioneered two research directions: 1) we reconstruct the original dialogues with variational recurrent neural networks and structured attention, then we extract the structure by computing the transitions between latent states; and 2) we detect and track the status of potential slot token groups to approximate a representation of task-oriented dialogue structures. We explored the problem from both theoretical and practical perspectives.

The dissertation of Liang Qiu is approved.

Lin Yang

Ying Nian Wu

Song-Chun Zhu, Committee Co-Chair

Achuta Kadambi, Committee Co-Chair

University of California, Los Angeles

2022

To my parents and Yi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, my greatest thanks go to my advisor at the Computer Science and Statistics Department, Song-Chun Zhu. Prof. Zhu is a great thinker, an excellent explorer, and he shows me a high-level, visionary picture of AI. He is always supportive and encourages me to tackle the most challenging topic in the field. I am forever grateful to him, and I wish him all the best for starting a new journey in Beijing.

I would like to thank my advisor Prof. Achuta Kadambi. Dr. Kadambi is a young, talented professor at the Electrical and Computer Engineering department. I truly appreciate his support and guidance throughout my Ph.D. studies. His help got me out of the most difficult time of my Ph.D.

I thank Prof. Ying Nian Wu and Prof. Lin Yang for being on my committee. Dr. Wu is a great mathematician, but more importantly, an extremely kind and warm person. I enjoyed talking with him about research every late night, and I have been deeply touched by his passion for always chasing the truth. It is also my great honor to have Dr. Yang on my thesis committee. My research about dialogue systems is very relevant to his work in Reinforcement Learning. I look forward to working with him in the near future.

I have also worked closely with Dr. Zhou Yu and Weiyan Shi at the Columbia NLP group. I met Zhou and Weiyan when I tried to reproduce the results of one of their articles. They answered my questions in great detail, and since then, Dr. Yu has shown her unconditional support to my research and career, and she always provides me with insightful suggestions.

During my Ph.D., I have also done two wonderful internships at Microsoft Research and Salesforce Research. I would like to thank my mentors and managers Jianchao Li, Baolin Peng, Lars Liden, Jianfeng Gao, Chien-Sheng (Jason) Wu, Wenhao Liu, Caiming Xiong. My internship project at MSR is partially related to the VALUENET project and leads to a part of this dissertation. The work *Structure Extraction in Task-Oriented Dialogues with Slot Clustering* is done during my internship at Salesforce.

I also had a great time as a part-time engineer at DMAI. I thank them for being my colleagues,

friends, and teachers, including Linzuo Li, Siyuan (Don) Sun, Henry Chang, Wei Si, Wanyi Zhang, Xiaowen Feng, Mingqing Ye, Qiancheng Wu, Yutong Sun, Lei Gao, Ka-Man Leong, Jiajie Yan, Jiajie Huang, Divyansh Srivastava, Earle Aguilar, Ashwin Dharne, Nelson Solano, Nicole Liang, Nishant Shukla, Changsong Liu, Rui Fang, Mingtian Zhao, and Mark Nitzberg.

I thank all co-authors of my papers, including Yizhou Zhao, Pan Lu, Yuan Liang, Tao Yuan, Yaofang (Victor) Zhang, Ziheng Xu, Feng Shi, Yuanyi Ding, Lei He. Especially Yizhou and Pan, my dissertation would not be possible without your great contributions.

I thank the whole VCLA Lab, including Yixin Chen, Lifeng Fan, Shuwen Qiu, Ruiqi Gao, Hangxin Liu, Baoxiong Jia, Xiaodan Liang, Tian Han, Sirui Xie, Luyao Yuan, Zeyu Zhang, Bo Pang, Zilong Zheng, Arjun R. Akula, Qing Li, Keze Wang, Chi Zhang, Yixin Zhu, and so many others who helped me along the way.

Finally, I thank my parents for their unconditional love and the best support in every stage of my life. I thank my fiancee, Yi Huang, for her caring and love. Without her, I couldn't finish my Ph.D. through this pandemic. I thank her for everything she has done for me.

VITA

2017–2022 Graduate Research Assistant, VCLA@UCLA.

2021 Research Intern, Salesforce Research.

2021 Research Intern, Microsoft Research.

2017–2020 Software Engineer in Natural Language Processing, DMAI.

2012–2016 B.E. in Automation, Shanghai Jiao Tong University.

PUBLICATIONS

L. Qiu, Y. Zhao, J. Li, P. Lu, B. Peng, J. Gao, S.-C. Zhu. *ValueNet: A New Dataset for Human Value Driven Dialogue System.* Association for the Advancement of Artificial Intelligence (AAAI), 2022.

Y. Zhao, **L. Qiu**, P. Lu, F. Shi, T. Han, S.-C. Zhu. *Learning from the Tangram to Solve Mini Visual Tasks.* Association for the Advancement of Artificial Intelligence (AAAI), 2022.

P. Lu, **L. Qiu**, J. Chen, T. Xia, Y. Zhao, W. Zhang, Z. Yu, X. Liang, S.-C. Zhu. *IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning.* Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2021.

L. Qiu, Y. Liang, Y. Zhao, P. Lu, B. Peng, Z. Yu, Y.N. Wu, S.-C. Zhu. *SocAoG: Incremental Graph Parsing for Social Relation Inference in Dialogues.* Association for Computational Linguistics (ACL), 2021.

P. Lu, R. Gong, S. Jiang, **L. Qiu**, S. Huang, X. Liang, S.-C. Zhu. *Inter-GPS: Interpretable Geometry Problem Solving with Formal Language and Symbolic Reasoning.* Association for Computational Linguistics (ACL), 2021.

Y. Liang, W. Han, **L. Qiu**, C. Wu, Y. Shao, K. Wang, L. He. *Exploring Forensic Dental Identification with Deep Learning.* Advances in Neural Information Processing Systems (NeurIPS), 2021.

Y. Liang, **L. Qiu**, T. Lu, Z. Fang, D. Tu, J. Yang, T. Zhao, Y. Shao, K. Wang, X. Chen, L. He. *OralViewer: 3D Demonstration of Dental Surgeries for Patient Education with Oral Cavity Reconstruction from a 2D Panoramic X-ray.* Intelligent User Interfaces (IUI), 2021.

L. Qiu, Y. Zhao, W. Shi, Y. Liang, F. Shi, T. Yuan, Z. Yu, S.-C. Zhu. *Structured Attention for Unsupervised Dialogue Structure Induction.* Empirical Methods in Natural Language Processing (EMNLP), 2020.

Y. Liang, W. Song, J. Yang, **L. Qiu**, K. Wang, L. He. *Atlas-aware ConvNet for Accurate yet Robust Anatomical Segmentation.* Medical Image Computing & Computer Assisted Intervention (MICCAI), 2020.

L. Qiu. *Non-linguistic Vocalization Recognition Based on Convolutional, Long Short-Term Memory, Deep Neural Networks.* Diss. UCLA, 2018.

CHAPTER 1

Introduction

1.1 Motivation

Teaching machines to speak and act like a human is one of the longest-running goals in Artificial Intelligence. In this thesis, we try to address two critical problems towards building the next generation of dialogue systems.

1.1.1 Emotional Intelligent Dialogue Systems

Psychologist Nicholas Humphrey believes that it is social intelligence, rather than quantitative intelligence, that defines humans [GM15]. Enhancing the emotional intelligence of chatbots allows them to understand user behaviors better and act appropriately in various situations. We argue there are three critical aspects in this problem.

(*i*) **Values.** Value refers to desirable goals in human life. People tend to have common attitudes towards some statements or scenarios. For example, it gives a sense of achievement if you get a paper published. By considering values, we can estimate user behavior and cognitive patterns from their utterances and generate responses that conform to the robot's persona configuration. Figure 1.1 shows example dialogues of agents with value personalization and recognition. For instance, an agent who is set to value stimulation and self-direction will love skiing. After the user refuses the agent's invitation to drink some beers, the agent could recognize the user's value of security and steer its future recommendation to healthier options.

Figure 1.2 shows a sociological perspective of basic human values [Sch92]. It describes ten

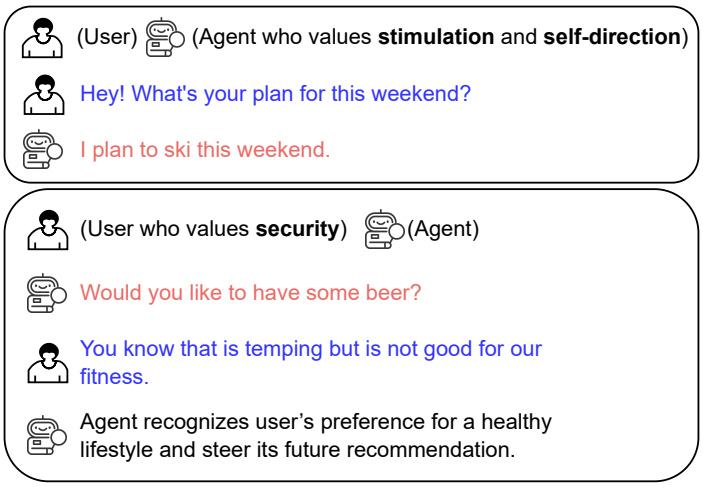

Figure 1.1: Socially intelligent agents with value personalization and recognition.

universal values that are recognized throughout all major cultures. The circular structure reflects the dynamic relations among these values, *i.e.*, the pursuit of some value may result in either accordance with another value or a conflict with another value.

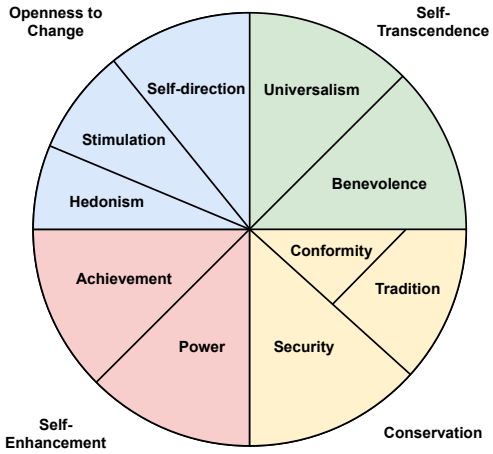

Figure 1.2: Theory of basic human values [Sch92].

(*ii*) **Social Relations.** Social relations form the basic structure of our society, defining not only

S1:	Well then we'll-we'll see you the day after tomorrow. Mom?! Dad?! What-what... what you guys doing here?!
S2:	Well you kids talk about this place so much, we thought we'd see what all the fuss is about.
S3:	I certainly see what the girls like coming here.
S1:	Why?!
S3:	The sexy blonde behind the counter.
S1:	Gunther?!
S2:	Your mother just added him to her list.
S1:	What? Your-your list?

	Argument Pair	**Trigger**	**Relation Type**
R1	(S2, S1)	dad	per:children
R2	(S3, Gunther)	sexy blonde	per:positive_impression
R3	(S3, S1)	mom	per:children
R4	(S1, S3)	mom	per:parents
R5	(S1, S2)	dad	per:parents

Table 1.1: A dialogue example from DialogRE [YSC20].

our self-images but also our relationships [Szt02]. To build an emotionally intelligent robot, it is vital to have the bot understand its contextual surroundings, including users' social relations. Table 1.1 shows an example in the dataset DialogRE [YSC20]. Given a dialogue as context and a set of entities, the task of Dialogue Relation Extraction predicts the relation types between the entities from a predefined relation set.

When we try to mimic humans' capability of inferring social relations, there are several aspects that we need to take into consideration:

- Human can reason an unknown relationship from the entities' relations with others. For example, given A is B's mother and C is B's father, we can easily infer A is C's wife. In another saying, each type of relation is governed by a special potential function — social norm.

- Human can incorporate personal attributes such as age, gender, and profession as cues to aid the relational inference.

- Social relations could evolve along with social interactions. For instance, strangers become friends over a good chat.

(*iii*) **Mental States.** The third is about constructing and maintaining the mental states reflected by an individual's *Theory of Mind* [PW78], *i.e.*, capability to understand others' thoughts and sense their emotions. This is beyond understanding what users say, but also to understand what users imply, requiring agents to mentally simulate and reason the evolution process of the social environment.

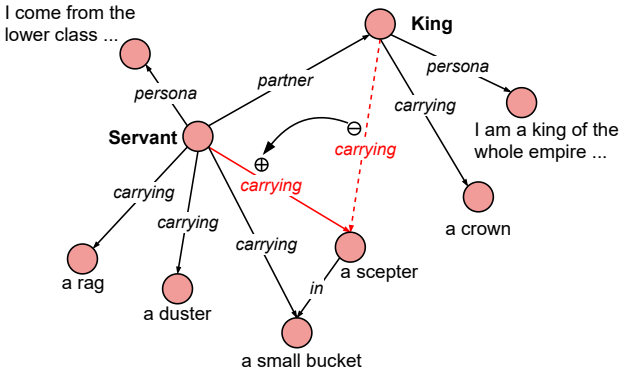

Figure 1.3: A graphical representation of the agent's mental state. Nodes are attributed with encoded descriptions of agents, objects and the environment. Agents' action trigger explicit topology changes of the graph.

For communication happening between agents A and B, the *Theory of Mind* describes the following recursive levels of their mental states:

- Level 0: Physical world;
- Level 1: A's belief and desires; B's belief and desires;
- Level 2: A's belief in A's mind, B's belief in B's mind (self-conscious); B's belief in A's mind, and A's belief in B's mind;
- ...

The reasoning capability on a deeper level of the mental states indicate a higher emotional quotient of a bot. It will further empower the machine to have the following cognitive capabilities:

- Attributing the causal effects and blames to actions;
- Predicting the possible actions and responses of other agents;
- Understanding the extended meaning, implicature, and irony of other agents.

1.1.2 Structure Learning in Dialogue Systems

The second central topic of this thesis is the structure learning of dialogues.

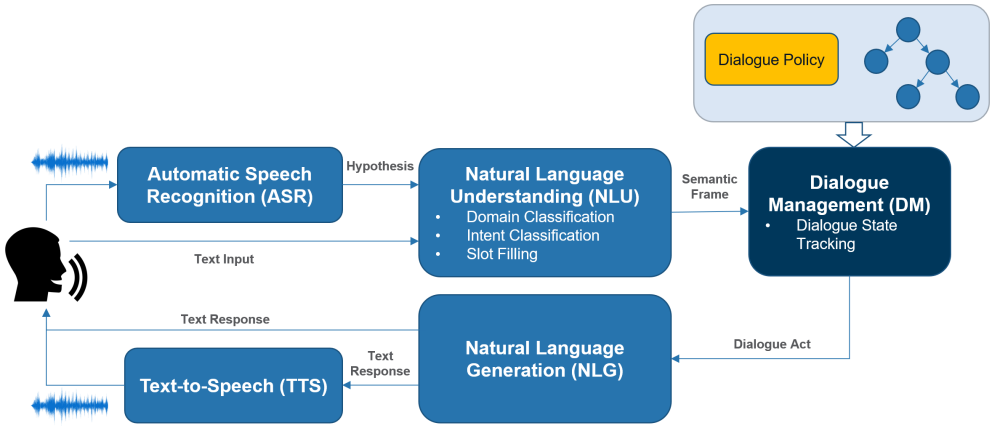

Figure 1.4: Common modularized dialogue systems that are widely used in industry.

Existing dialogue systems in the industry are mostly modularized systems that consist of components of Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), Dialogue Management (DM), Natural Language Generation (NLG), Text-to-Speech (TTS), as shown in Figure 1.4. While there has been remarkable progress in learning end-to-end ASR, NLU, NLG, and TTS, the design of DM still heavily relies on manually-crafted policies. The dialogue policy defines the system's action given a current dialogue state, which is based on recognized

user intent, slot values, and potentially other contexts from multi-modal sensors. In this way, the transition among the dialogue states composes a conversational graph or structure of the dialogue domain that the agent is handling.

As we can see in Figure 1.5, extracting a semantic structure from dialogue data provides us with a discourse skeleton of the domain, which is bus information request in this case. Another

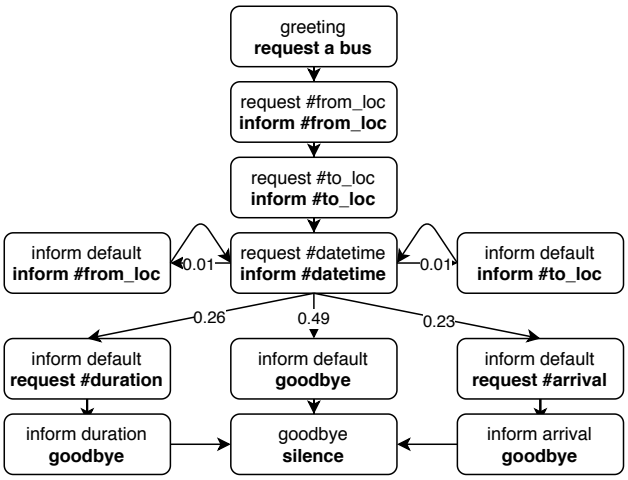

Figure 1.5: Original dialogue structure of the bus information request domain in SimDial [ZE18]. User intents are marked in **bold**.

interesting type of dialogue structure is the interactive structure in multi-party dialogues. Figure 1.6 illustrates the interactive structure learned from a dialogue sample in Ubuntu Chat Corpus [LPS15]. Each node represents an utterance from different speakers in the dialogue, with darker linkages representing stronger dependency relations between utterances. When speaker/addresses information is unavailable in the corpus, learning such a structure allows disentangling the conversation and estimating the speaker labels.

Therefore, extracting dialogue structures is an important topic for us to analyze user behavior and system performance, benefiting several downstream tasks such as dialogue system building, discourse analysis, and dialogue summarization.

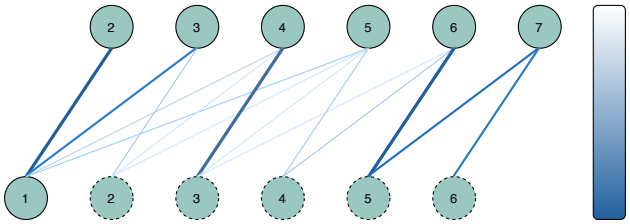

Figure 1.6: Learned interactive structure from a multi-party dialogue sample in Ubuntu Chat Corpus [UA13].

1.2 Thesis Outline

Following the two central topics that we just discussed, this thesis consists of two parts — PART I EMOTIONAL INTELLIGENT DIALOGUE SYSTEMS and PART II STRUCTURE LEARNING IN DIALOGUE SYSTEMS.

PART I focuses on modeling and incorporating emotional intelligence in dialogue systems. To be specific,

In Chapter 2, we talk about building human value driven dialogue systems. We begin with an introduction of the theory of basic human value and utility, existing social commonsense benchmarks, and emotionally intelligent dialogue datasets. Next, we present our curation procedure and the statistical specifications of a new dataset, VALUENET. We then formally define the problem formulation of value modeling and compare the performance between using Transformer [VSP17] variants. Finally, we demonstrate how to incorporate the value model into chatbots on existing dialogue datasets.

In Chapter 3, we discuss inferring social relations in dialogues. We begin by introducing the task of Dialogue Relation Extraction and a relevant dataset, DialogRE [YSC20]. We then talk about the difference between relation inference in documents and dialogues. The social relation is represented as an attributed And-Or graph [ZWM98, ZM07], and we describe our proposed α–β–γ approach to incrementally parse the graph. Finally, we present experimental results on the DialogRE and

another related dataset, MovieGraph [VTC18]. We conduct an in-depth analysis of the results and a case study of the model.

In Chapter 4, we explore how to combine mental state modeling and value modeling, and incorporate both into artificial agents. We argue that the problem needs to be studied in an embodied environment, such as the LIGHT [UFK19]. Then we introduce related works on text-based embodied AI, mental state modeling, and value modeling. We formally define the problem and present a holistic framework step by step as a proposal to address the problem. Finally, we compare the proposed framework with end-to-end neural-based models and conduct a detailed analysis of the results. We also discuss the future work in this area.

PART II explores two directions to learn and extract structures from dialogue data. Detailedly,

In Chapter 5, we begin with introducing the semantic structures in two-part dialogues and interactive structures in multi-party dialogues. We present related works about structured attention and previous works on structure learning. Then we formulate the learning of these two types of structures, and present our approach Variational Recurrent Neural Network (VRNN) [SZY19] with structured attention as a uniform solution. Finally, we compare the proposed algorithm with state-of-the-art models at that time and analyze the results.

In Chapter 6, we take a different path from the previous chapter. We begin with presenting a different representation of dialogue structures, where the nodes are simplified dialogue states in task-oriented dialogues. After discussing related works in this topic, we propose a two-stage approach, which is to detect and cluster potential slot tokens first, and then track the status of each cluster to build the graph. Besides evaluating the structure learning performance, we also show that data augmentation based on extracted structures improves the response generation quality.

We will finally conclude in Chapter 7.

Part I

Emotional Intelligent Dialogue Systems

CHAPTER 2

Human Value Driven Dialogue System

Building a socially intelligent agent involves many challenges, one of which is to teach the agent to speak guided by its value like a human. However, value-driven chatbots are still understudied in the area of dialogue systems. Most existing datasets focus on commonsense reasoning or social norm modeling. In this chapter, we present a new large-scale human value dataset called VALUENET, which contains human attitudes on 21,374 text scenarios. The dataset is organized in ten dimensions that conform to the basic human value theory in intercultural research. We further develop a Transformer-based value regression model on VALUENET to learn the utility distribution. Comprehensive empirical results show that the learned value model could benefit a wide range of dialogue tasks. For example, by teaching a generative agent with reinforcement learning and the rewards from the value model, our method attains state-of-the-art performance on the personalized dialog generation dataset: PERSONA-CHAT. With values as additional features, existing emotion recognition models enable capturing rich human emotions in the context, which further improves the empathetic response generation performance in the EMPATHETICDIALOGUES dataset. To the best of our knowledge, VALUENET is the first large-scale text dataset for human value modeling, and we are the first one trying to incorporate a value model into emotionally intelligent dialogue systems[1].

[1]The dataset is available at `https://liang-qiu.github.io/ValueNet/`.

2.1 Introduction

Value refers to desirable goals in human life. They guide the selection or evaluation of actions, policies, people, and events. A person's value priority or hierarchy profoundly affects his or her attitudes, beliefs, and traits, making it one core component of personality [Sch12]. In dialogue systems, modeling human values is a critical step towards building socially intelligent chatbots [QZL21]. By considering values, we can estimate user behavior and cognitive patterns from their utterances and generate responses that conform to the robot's persona configuration. For example, the robot is set to be aware of human values, and it invites Jerry to drink beers, but Jerry replies, "*You know that is tempting but is not good for our fitness*". The bot could read from the dialogue that Jerry prefers a healthy and self-disciplined lifestyle and steer its recommendation to healthier options in the future.

The development of socially intelligent chatbots has been one of the longest-running goals in artificial intelligence. Early dialogue systems such as Eliza [Wei66], Parry [CWH71], and more recent SimSimi[2], Panda Ichiro [OS18], Replika [FSR18], XiaoIce [ZGL20], were designed to mimic human behavior and incorporate emotional quotients (EQ) to some extent. There are also datasets and benchmarks for studying related problems, such as emotion recognition [MVC10, HCK18, PHM19, GMG20], personalized dialogue generation [ZDU18, LCC20], and empathetic dialogue generation [RSL19]. Even though value plays a fundamental and critical role in human EQ, there is a lack of explicit modeling of values in the dialogue domain, based on social domain theory. We have seen recent efforts about crowdsourcing social commonsense knowledge base or benchmarks [FHS20, SRC19, LBC21, HBB20, HBB21, GBS21]. However, it is not clearly shown how an agent can leverage this knowledge to estimate the users' value priorities or guide its own speaking and actions. In this work, we aim to alleviate this problem and investigate the usage of a learned value function.

We start the study by curating a knowledge base of human values called VALUENET. Samples with value-related scenarios were identified based on value-defined keyword searching. Next, we

[2] https://simsimi.com/

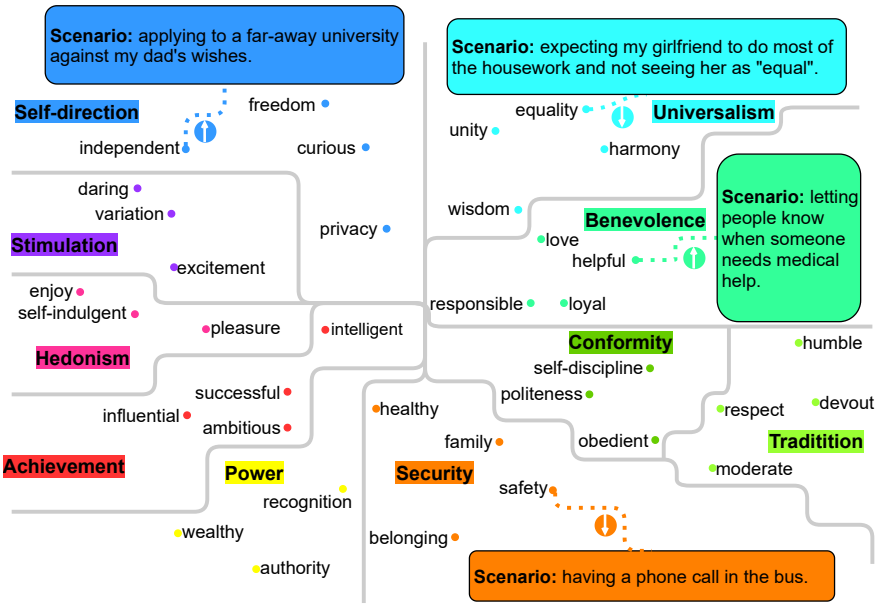

Figure 2.1: The presented VALUENET dataset with curated social scenarios organized by Schwartz values [Sch12].

asked Amazon Mechanical Turk workers about how the provided scenarios will affect one's value. This is based on the assumption that values underlie our attitudes; they are the guideline by which we evaluate things. Workers assess behaviors/events positively if they promote or protect the attainment of the goals we value. Behaviors/events are evaluated negatively if they hinder or threaten the attainment of these valued goals. The whole process gives us a large-scale (over 21k samples) multi-dimensional knowledge base of value. Figure 2.1 shows the overall structure of VALUENET. Each split represents a value dimension identified in the theory of basic human values [Sch12]. The figure also illustrates the value-related keywords and scenarios. The circular arrangement of the values represents a motivational continuum. By organizing data in such a structure, we anticipate the VALUENET to provide comprehensive coverage of different aspects of human values.

Next, we develop a Transformer-based value model to evaluate the utility score suggesting the positive or negative judgment given an utterance. We provide a detailed analysis of learning with multiple Transformer variants. Then we conduct a wide range of experiments to demonstrate that the value model could benefit EQ-related dialogue tasks: *(i)* By finetuning a generative agent

with reinforcement learning and the reward from our value model, the method achieves state-of-the-art performance on the personalized dialogue dataset: PERSONA-CHAT [ZDU18]; *(ii)* By incorporating values as additional features, in EMPATHETICDIALOGUES [RSL19], we improve the emotion classification accuracy of existing models, which further facilitates the empathetic response generation; *(iii)* Visualization of the value model shows that it provides a numerical way of user profile modeling from their utterances.

In all, our contributions are two-fold. First, we present a large-scale dataset VALUENET for the modeling of human values that are well-defined in intercultural research. Second, we initiate to develop the value model learned from VALUENET to several EQ-related tasks and demonstrate its usage for building a value-driven dialogue system. Our methodology can be generalized to a wide range of interactive situations in socially aware dialogue systems [ZRR18], and human-robot interactions [YL17, LHA21].

2.2 Related Work

An abundance of related work inspires our work. Our work aims to make contributions to dialogue systems by incorporating the theory of human value. The dataset we collect shares a similar nature with multiple social commonsense benchmarks and knowledge bases. Besides, we apply our VALUENET for various dialogue tasks related to EQ.

2.2.1 Theory of Human Value and Utility

In the field of intercultural research, [Sch12] developed the theory of basic human values. The theory identifies ten basic personal values that are recognized across cultures and explains where they come from, as shown in Figure 2.1. The closer any two values in either direction around the circle, the more similar their underlying motivations are; the more distant, the more antagonistic their motivations. Note that dividing the value item domain into ten distinct values is an arbitrary convenience. It is reasonable to partition the value items into more or less finetuned distinct values

according to the needs and objectives of one's analysis[3]. Similarly, in the economics field, the concept of utility [Fis70] is initially defined as a measure of pleasure or satisfaction in economics and ethics that drives human activities at all levels. Therefore, when we teach agents to speak and act in a socially intelligent way, an approach considering human value utilities should be adopted. In this work, we aim to learn a utility function for each dimension of value and steer the dialogue system response generation accordingly.

2.2.2 Social Commonsense Benchmarks

[HBB20] present the ETHICS dataset, a benchmark that assesses a language model's knowledge of basic concepts of morality. SCRUPLES [LBC21] is a large-scale dataset with ethical judgments over real-life anecdotes, motivated by descriptive ethics. SOCIAL-CHEM-101 presented by [FHS20] is a corpus that catalogs rules-of-thumb as basic concept units for studying people's everyday social norms and moral judgments. They also propose Neural Norm Transformer to reason about previously unseen situations, generating relevant social rules-of-thumb. SOCIAL IQA [SRC19] is a large-scale benchmark for commonsense reasoning about social situations. [HBE17] present a task and corpus for predicting the preferable options from two sentences describing the scenarios that may involve social and cultural situations. Instead, in this work, we release a new dataset VALUENET that provides annotation of human attitudes from different value aspects.

2.2.3 Emotionally Intelligent Dialogue Datasets

Several datasets are presented to study emotion dynamics in dialogues. DailyDialog [LSS17] is a multi-turn dialogue dataset, which reflects the way of daily communication and provides emotion labels for speakers. [HCK18] present EmotionLines with emotions labeling on all utterances in each dialogue based on their textual content. MELD [PHM19] is an extension of EmotionLines for

[3]A refinement of the theory [SCV12], partitions the same continuum into 19 more narrowly defined values that permit more precise explanation and prediction. We use the original 10-dimension version for simplicity in this work.

multi-modal multi-party emotion recognition. [MVC10] record a corpus SEMAINE of emotionally colored conversations. [GMG20] propose a framework COSMIC for emotion recognition in conversations by considering mental states, events, actions, and cause-effect relations. DialogRE [YSC20] is the first human-annotated dialogue-based dataset for social relation inference [QLZ21]. PERSONA-CHAT [ZDU18] (revised in ConvAI2 [DLM20]) provides natural language profiles of speakers. Based on PERSONA-CHAT, [LCC20] propose a transmitter-receiver-based framework with explicitly human understanding modeling to enhance the quality of personalized dialogue generation. EMPATHETICDIALOGUES [RSL19] is a dataset that provides 25k conversations grounded in emotional situations. Each dialogue is grounded in a specific situation where a speaker was feeling a given emotion.

2.3 The VALUENET Dataset

During decision-making, people tend to pick the choice that aligns more with their own values. This work aims to provide a transferable knowledge base for human value modeling in natural language. To collect the VALUENET dataset, we curated social scenarios with value-related keywords and further annotated them via Amazon Mechanical Turk. Each sample in VALUENET is a social scenario description labeled with the annotator's attitude through a specific value lens.

The entire dataset is organized in a circular structure as shown in Figure 2.1, aligning with the theory of basic human values [Sch12]. The theory identifies ten universal values that are recognized throughout major cultures. The circular structure reflects the dynamic relations among these values, *i.e.,* the pursuit of some value may result in either accordance with another value or a conflict with another value. The ten distinct values can be further organized into four higher-order groups.

- **Openness to change**: self-direction, stimulation
- **Self-enhancement**: hedonism, achievement, power
- **Conservation**: security, conformity, tradition

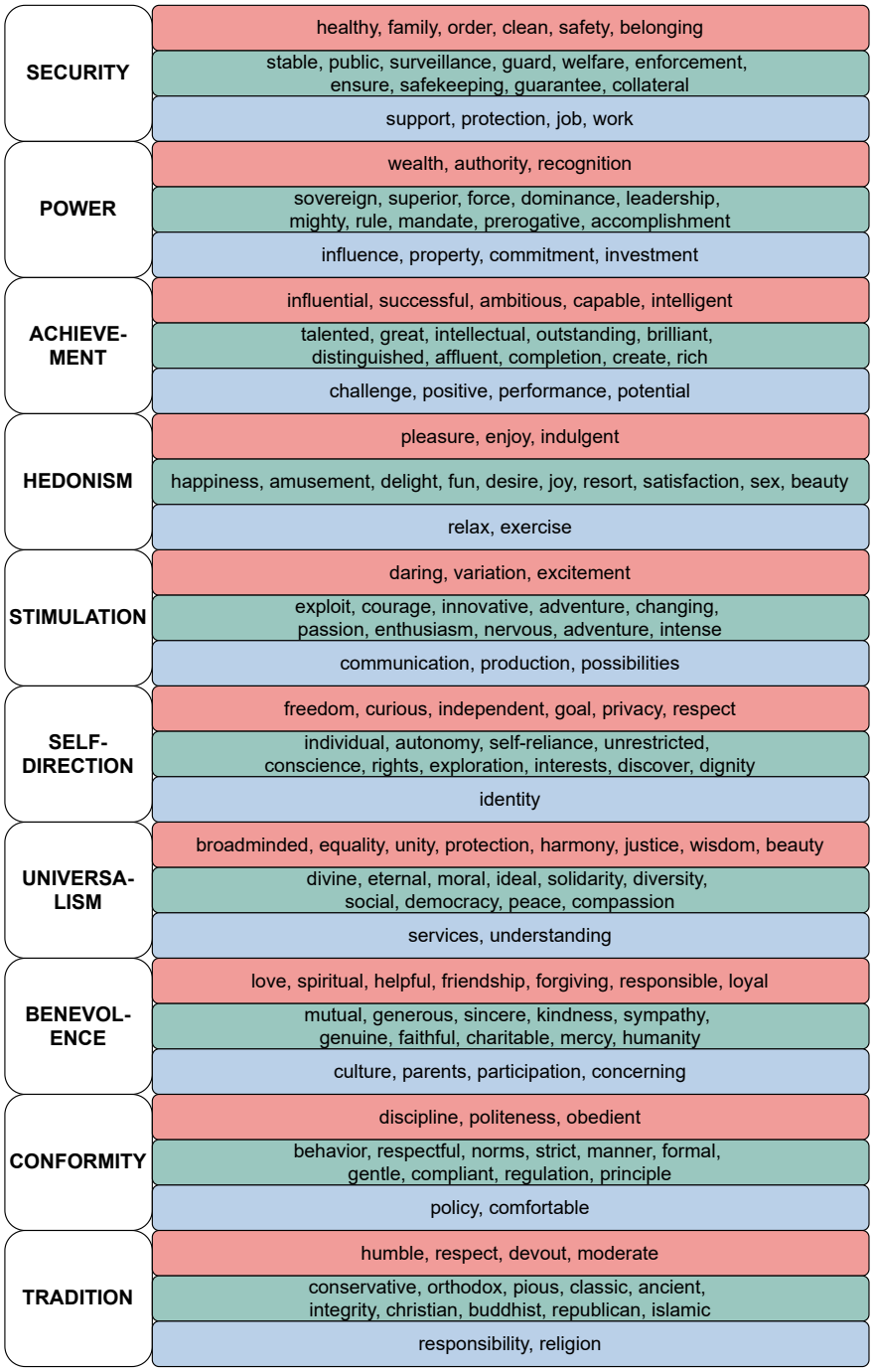

Figure 2.2: Ten universal human values and related keywords for social scenario curation. Red: keywords in the original value definition [Sch12]; Green: associated keywords found with datamuse; Blue: associated keywords found with GloVe embedding.

- **Self-transcendence**: benevolence, universalism

We describe the collection details of the VALUENET in the following sections.

2.3.1 Social Scenario Curation

We curated a set of 21,374 social scenarios from the large-scale social-related database SOCIAL-CHEM-101 [FHS20]. Value-related scenarios are retrieved with value keywords after lemmatization and stemming. There are three sets of keywords identified for each dimension of Schwartz value: (1) the keywords in the original definition of each value in Schwartz's paper [Sch12]; (2) words that share a similar meaning, words that are often used to describe the original keywords, and words that are triggered by (strongly associated with) the original keywords[4]; (3) words that are near the original keywords in the GloVe [PSM14] embedding space. The value keywords are verified and confirmed by humans as listed in Figure 2.2.

2.3.2 Value-Aspect Attitude Annotation

We crowdsourced people's attitudes to the curated scenarios on Amazon Mechanical Turk (AMT). Figure 2.3 shows an example.

We follow a strict procedure to select qualified workers and ensure the workers understand the concept of each value we ask. In Figure 2.3, the definition of `BENEVOLENCE` is shown to the workers throughout the entire annotation process. To further help the understanding, we include three examples in each assignment with correct answers being "yes", "no", and "unrelated", respectively. The worker is then required to answer a prerequisite question correctly to proceed to the formal survey. The formal survey is composed of ten questions, including two hidden qualification checking questions. Before publishing on the AMT, two Ph.D. students prepared the qualification questions by annotating a small subset of the curated scenarios. Their agreed samples

[4]We use datamuse (`https://www.datamuse.com/api/`) for this purpose.

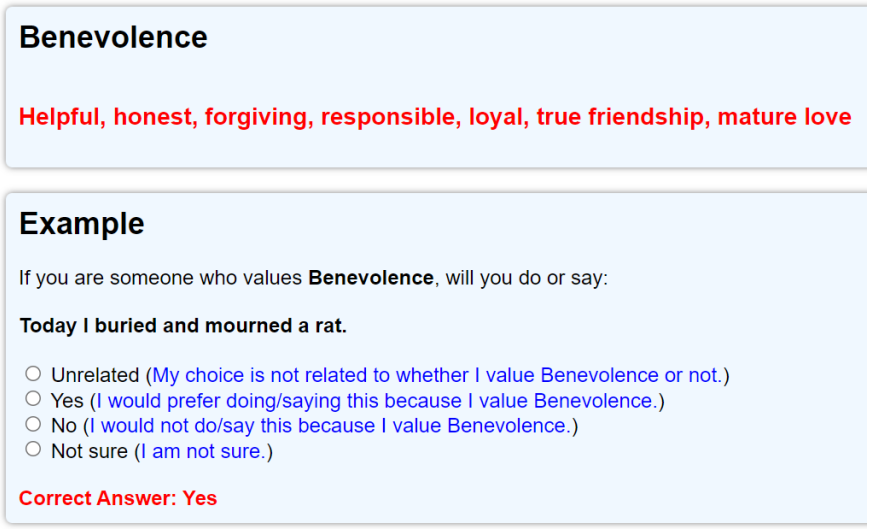

Figure 2.3: Value-aspect attitude annotation in AMT.

(100 in total) were randomly inserted into the survey for worker selection. The selection procedure was done in the value dimensions with more scenarios to get a large pool of qualified workers and a relatively balanced final dataset across different values. The complete Mechanical Turk interface is attached in the Appendix 2.8 for reference.

A total of 681 experienced AMT workers participated in our VALUENET annotation. 443 of them passed the qualification test. Each scenario is assigned to four different workers. The original inter-annotator agreement is 64.9%, and the Fleiss' kappa score [Fle71] among the workers is 0.48, which considers the possibility of the agreement by chance. Keeping the scope of VALUENET in commonly-agreed attitudes towards social scenarios, we only retain the samples with three or more agreements. Figure 2.4 shows the sample size of each value split and their label distribution.

The data is split into the train (75%), valid (15%), and test (10%). Similar to the polarity in sentiment analysis [KWM11], we quantify the annotated labels into numerical values: yes (positive): +1, no (negative): -1, unrelated (neutral): 0. We denote the numerical values as **utility** to describe the effect of a scenario on one's value. In other words, for people who appreciate a certain value, actions with a higher utility in this value dimension would be more desirable to them.

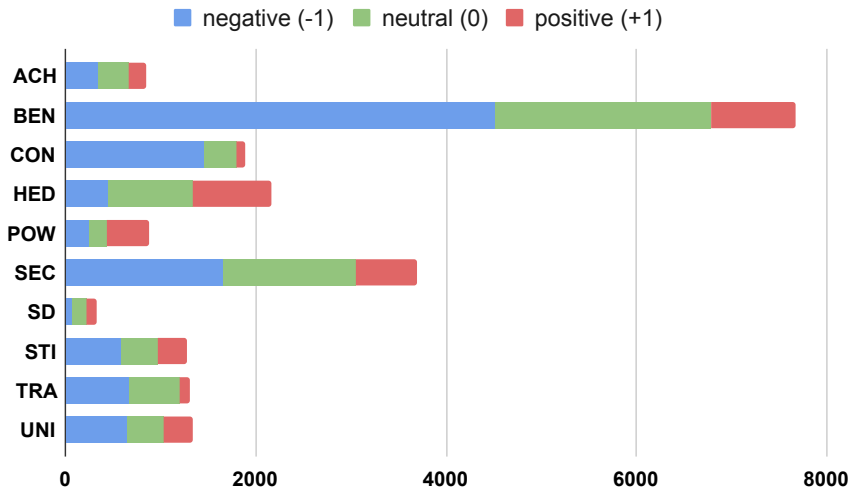

Figure 2.4: The sample number and label distribution of each value split in the VALUENET.

VALUENET	train	valid	test	total
# samples	16,030	3,206	2,138	21,374
average # tokens	12.05	12.09	12.26	12.07
unique # tokens	12,452	5,292	4,112	14,143

Table 2.1: Statistics of the VALUENET dataset.

Table 2.1 shows more statistical details about the VALUENET dataset. In total, we collected 21,374 samples covering a wide range of scenarios in daily social life.

2.4 Value Modeling

We experiment using Transformer-based pre-trained language models for modeling human values from the VALUENET dataset.

2.4.1 Task Formalization

Given a social scenario s, we wish to learn a value function that models the utility distribution of s from the ten Schwartz value dimensions: $\mathbf{V}(s) = [V_{\texttt{SEC}}(s), V_{\texttt{POW}}(s), V_{\texttt{ACH}}(s), V_{\texttt{HED}}(s), V_{\texttt{STI}}(s), V_{\texttt{SD}}(s), V_{\texttt{UNI}}(s), V_{\texttt{BEN}}(s), V_{\texttt{CON}}(s), V_{\texttt{TRA}}(s)]$, where $V_{\texttt{\$VALUE}}(\cdot) \in [-1, 1]$ and $V_{\texttt{\$VALUE}}(\cdot) \in \mathbb{R}$.

2.4.2 Model

Pre-trained language model variants: BERT [DCL18], RoBERTa [LOG19], DistilBERT [SDC19], BART [LLG19] are investigated for learning the value function. A custom input format constructed as $\langle$`[CLS] [$VALUE] s`$\rangle$ is fed into a Transformer encoder, *i.e.,*

$$V_{\texttt{\$VALUE}}(\texttt{s}) = \textbf{TRM}(\texttt{[CLS] [\$VALUE] s}), \tag{2.1}$$

where TRM denotes the Transformer encoder, `[CLS]` is the special token for regression or classification, and `[$VALUE]` are special tokens we define to prompt the language models the value dimension we are interested in [LL21, BMR20, LR21]. In order to get the ten-dimensional output $\mathbf{V}(s)$, a batch size of 10 is forwarded through the model. For the BERT, DistilBERT, and RoBERTa, a regression head is put on top of the models and they are trained with the Mean Squared Error (MSE) loss. We use the regression model with sigmoid activation to get a continuous estimation of the utility in the range of $[-1, 1]$. To evaluate the effect of different loss functions, we train the BART model with three output classes and the cross-entropy loss.

2.4.3 Result and Analysis

The learning performance of using fastText[5] [JGB17] and Transformer variants are reported in Table 2.2. All Transformers are trained for 40 epochs with a learning rate of $5\text{e}{-6}$. The prediction precision, recall, $\mathbf{F}1$ score, and accuracy for regression models are computed by the utility rounded to the nearest integer.

[5] `https://github.com/facebookresearch/fastText`

		F1(-1)	**F**1(0)	**F**1(1)	**P**(-1)	**P**(0)	**P**(1)	**R**(-1)	**R**(0)	**R**(1)	**Acc.**↑	**MSE**↓
VALUENET (original)	fastText	0.70	0.46	0.43	0.65	0.47	0.55	**0.76**	0.44	0.35	0.58	0.66
	BERT	**0.73**	0.50	0.51	0.72	0.46	0.71	0.74	0.55	0.39	0.61	0.39
	DistilBERT	0.71	0.52	0.47	0.74	0.45	0.69	0.68	0.62	0.36	0.60	**0.37**
	RoBERTa	0.65	0.51	0.34	0.74	0.40	0.71	0.58	0.69	0.22	0.55	0.41
	BART	0.00	**0.76**	0.54	0.00	0.70	0.60	0.00	**0.83**	**0.49**	**0.67**	0.52
VALUENET (balanced)	fastText	0.70	0.48	0.43	0.64	0.50	0.54	**0.76**	0.45	0.36	0.59	0.68
	BERT	0.67	0.48	0.51	0.73	0.42	0.61	0.62	0.58	0.43	0.57	0.40
	DistilBERT	0.66	0.49	0.50	0.74	0.41	0.61	0.60	0.60	0.43	0.57	0.40
	RoBERTa	0.65	0.51	0.34	0.74	0.40	0.71	0.58	0.69	0.22	0.55	0.41
	BART	0.00	0.75	0.51	0.00	0.72	0.57	0.00	0.77	0.47	0.65	0.55
VALUENET (augmented)	fastText	0.58	0.52	0.29	0.72	0.40	0.65	0.49	0.75	0.18	0.52	0.59
	BERT	0.67	0.55	0.41	0.78	0.43	**0.78**	0.58	0.76	0.28	0.58	0.38
	DistilBERT	0.68	0.57	0.41	**0.79**	0.44	**0.78**	0.59	0.78	0.28	0.60	0.38
	RoBERTa	0.70	0.56	0.41	0.78	0.45	0.75	0.64	0.74	0.28	0.61	0.40
	BART	0.00	0.74	**0.57**	0.00	**0.75**	0.49	0.00	0.73	0.66	0.64	0.46

Table 2.2: Value modeling performance in the VALUENET dataset. **Bold** items are the best in each metric column.

Acc.	ACH	BEN	CON	HED	POW	SEC	SD	STI	TRA	UNI
VALUENET (original)	**0.56**	**0.68**	0.82	**0.63**	0.35	**0.52**	0.45	**0.58**	0.60	**0.51**
VALUENET (balanced)	0.53	0.58	**0.83**	**0.63**	**0.41**	0.50	0.42	0.53	0.61	0.50
VALUENET (augmented)	0.48	0.66	0.82	0.58	0.33	0.47	**0.48**	0.49	**0.64**	0.42

Table 2.3: Accuracies of the BERT [DCL18] value model across different value dimensions in the VALUENET dataset.

In general, pre-trained language models perform better than the fastText baseline. However, there is not a noticeable difference between the Transformer variants. The prediction accuracy of BART is the highest among all models because it is explicitly trained for classification purposes. BERT and DistilBERT get the lowest MSE in terms of regression performance.

Observing the sample imbalance across different value splits and labels (Figure 2.4), we release another two versions of VALUENET: VALUENET (balanced) and VALUENET (augmented). The original dataset is balanced by subsampling the negative and neutral data of the largest value split (BENEVOLENCE). Moreover, we augment the neutral class of the original VALUENET by assigning AMT results with less worker agreement to "unrelated". Data distribution of the balanced and augmented versions of VALUENET are illustrated in the Appendix 2.8. By analyzing the prediction accuracy in different value splits (Table 2.3), we find that reducing the sample number of BENEVOLENCE hurts the model performance in that dimension. Looking at the $\mathbf{F}1$ score of each class in Table 2.2, we conclude that augmenting the neutral class improves the $\mathbf{F}1(0)$ but reduces $\mathbf{F}1(1)$ and $\mathbf{F}1(\text{-}1)$. We leave it a future work to further improve the value modeling performance.

In the next sections, we show how the learned value function could benefit EQ-related tasks and help build a value-driven dialogue system.

2.5 Application: PERSONA-CHAT

As values are closely related to one's personality, we first assess our value model on a personalized dialogue dataset: PERSONA-CHAT [ZDU18]. The PERSONA-CHAT dataset contains multi-turn dialogues conditioned on personas. Each persona is encoded by at least 5 sentences of textual description, termed a profile. Example profile sentences are "*I like to ski*", "*I enjoy walking for exercise*", "*I have four children*", *etc.* The dataset is composed of 8,939 dialogues for training, 1,000 for validation, and 968 for testing. It also provides *revised* personas by rephrasing, generalizing or specializing the *original* ones. The dataset we use for experiments is publicly available in ParlAI[6].

[6] https://parl.ai/projects/convai2/

Algorithm 1 Personalized Dialogue Value Matching

Input: $[\mathbf{V}(p_1), ..., \mathbf{V}(p_N)], [\mathbf{V}(\hat{x}_1^s), ..., \mathbf{V}(\hat{x}_T^s)]$

Output: reward R

for $t = 1, 2, ..., T$ **do**
  $r_t \leftarrow -1$
  $m_t \leftarrow -1$
  for $i = 1, 2, ..., N$ **do**
    if $\mathbf{V}(p_i) \cdot \mathbf{V}(\hat{x}_t^s) > r_t$ **then**
      $r_t \leftarrow \mathbf{V}(p_i) \cdot \mathbf{V}(\hat{x}_t^s)$
      $m_t \leftarrow i$
    end if
  end for
end for
$\gamma_i \leftarrow 1, i = 1, 2, ..., N$
for $t = 1, 2, ..., T$ **do**
  $\gamma_{m_t} \leftarrow \gamma_{m_t} + 1$
end for
$R \leftarrow 0$
for $t = 1, 2, ..., T$ **do**
  $R \leftarrow R + \text{sign}(r_t) \cdot |r_t|^{\text{sign}(r_t) \cdot \gamma_{m_t}}$
end for
return R/N

2.5.1 Task Formalization

Given the agent's self persona profile $\mathbf{p} = [p_1, p_2, ..., p_N]$ and the dialogue history up to the t-th turn $\mathbf{h}_t^s = (x_1^u, x_1^s, ..., x_t^u)$, x_i^u is the i-th utterance by Person 1 played by the user, x_i^s is the i-th utterance by Person 2 played by the system, we evaluate the model's performance on predicting the next

utterance x_t^s.

2.5.2 Model

A decoder-only Transformer-based model is used to estimate the generation distribution $p_\theta(x_t^s \mid \mathbf{h}_t^s, \mathbf{p})$, where θ is the model parameter. Following the practice proposed in [GLC18], the model is firstly trained with Maximum Likelihood Estimation (MLE) to ensure generating fluent responses. Then we took an interleaving of supervised training (MLE) and reinforcement learning. We use the REINFORCE policy gradient algorithm [Wil92] in our experiment, and the reward assignment is described as follows.

Denote $\mathbf{V}(p_i)$ and $\mathbf{V}(\hat{x}_i^s)$ to describe the estimation of the agent's value from its profile sentence p_i and generated response $\hat{x}_i^s$, respectively. We want the reward to promote the alignment of the agent's profile and utterances in the value space. For instance, if the agent has profile "*I like venture*" and "*I have a dog*", and it says "*I plan to ski this weekend*" and also "*Do you like skiing*". Both utterances should be aligned with the first persona. Here we propose a simple yet effective searching algorithm (Algorithm 1) to find a match between $[\mathbf{V}(p_1), \mathbf{V}(p_2), ..., \mathbf{V}(p_N)]$ and $[\mathbf{V}(\hat{x}_1^s), \mathbf{V}(\hat{x}_2^s), ..., \mathbf{V}(\hat{x}_T^s)]$ and return a reward R. N is the number of profile sentences and T is the length of the generated dialogue. $\mathbf{V}$ is normalized to ensure $|r_t| \leq 1$. Intuitively, the discount argument γ prevents the language model from repeating the same fact in the agent's profile.

2.5.3 Setup

We evaluate the same generative model in both generation and ranking settings. In the response ranking setup, the candidates are scored with their log-likelihood. For the GPT-2 [RWC19] and DIALOGPT [ZSG19] we have finetuned, we train them for 5k steps with a training batch size of 8. The learning rate is set to $2\mathrm{e}{-6}$. For an illustration of computational requirements, the training with MLE on 4 NVIDIA Tesla V100 takes $\sim$1 hours, and the reinforcement learning takes $\sim$30 minutes.

Model	Original			Revised		
	Hits@1(%) ↑	**Ppl.**↓	**F**1(%) ↑	**Hits**@1(%) ↑	**Ppl.**↓	**F**1(%) ↑
SEQ2SEQ-ATTN	12.5	35.07	16.82	9.8	39.54	15.52
$\mathcal{P}^2$BOT [LCC20]	–	15.12	19.77	–	18.89	19.08
GPT-2 (MLE) [RWC19]	$14.51_{[0.05]}$	$17.23_{[0.03]}$	$18.74_{[0.01]}$	$10.31_{[0.07]}$	$20.64_{[0.11]}$	$18.29_{[0.05]}$
GPT-2 + Value (Ours)	$16.44_{[0.10]}$	$16.83_{[0.06]}$	$18.76_{[0.02]}$	$12.19_{[0.03]}$	$19.98_{[0.06]}$	$17.88_{[0.05]}$
DIALOGPT (MLE) [ZSG19]	$20.20_{[0.04]}$	$14.38_{[0.05]}$	$20.16_{[0.04]}$	$15.80_{[0.03]}$	$17.35_{[0.05]}$	$19.08_{[0.08]}$
DIALOGPT + Value (Ours)	$\mathbf{20.97}_{[0.08]}$	$\mathbf{13.84}_{[0.03]}$	$\mathbf{20.22}_{[0.01]}$	$\mathbf{18.83}_{[0.03]}$	$\mathbf{17.01}_{[0.03]}$	$\mathbf{19.79}_{[0.10]}$

Table 2.4: Next utterance prediction performance on PERSONA-CHAT [ZDU18]. We report the standard deviation $[\sigma]$ (across 5 runs) of the models we trained.

2.5.4 Result and Analysis

Following [ZDU18] and [LCC20], we report the **Hits**@1, **Perplexity** and **F**1 to evaluate the methods in Table 2.4. By the submission of this dissertation, $\mathcal{P}^2$BOT [LCC20] is the state-of-the-art model reported in this task. We also include a generative baseline using SEQ2SEQ with attention mechanism [BCB14] for comparison. As observed, in terms of all the metrics we evaluated, finetuning GPT-2 or DIALOGPT models with our value function provides a significant performance boost compared to simply training them with MLE. Our DIALOGPT + Value model achieves new state-of-the-art performance on perplexity and **F**1.

2.6 Application: EMPATHETICDIALOGUES

EMPATHETICDIALOGUES [RSL19] provides 25k conversations grounded in emotional situations. It aims to test the dialogue system's capability to produce empathetic responses. Each dialogue is grounded in a specific situation where a speaker was feeling a given emotion, with a listener responding. In this section, we demonstrate how we could leverage VALUENET to improve the emotion classification accuracy and further improve the empathetic response generation.

2.6.1 Emotion Classification

An auxiliary task that is highly related to empathetic dialogue generation is emotion classification. In EMPATHETICDIALOGUES, each situation is written in association with a given emotion label. A total of 32 emotion labels were annotated to cover a broad range of positive and negative emotions.

2.6.1.1 Model

Given the situation context s, a pre-trained BERT model encodes s and gets the sentence representation from its pooling layer of the [CLS] token. The same context is parsed by our pre-trained value model to get a ten-dimensional vector, which serves as an additional feature for the classification:

$$\begin{aligned} h_s &= \text{BERT}(s), \\ v_s &= \mathbf{V}(s), \\ e &= \text{softmax}(\mathbf{W} \cdot ([h_s; v_s]) + \mathbf{b}), \end{aligned} \tag{2.2}$$

where $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters.

2.6.1.2 Result

We compare the performance between our implementation and the baseline that directly applies the BERT model for emotion classification. As shown in Table 2.5, the additional value information benefits emotion classification from both the DistilBERT and BERT models. Our method obtains a **relative** improvement of 5.2% on DistilBERT and 6.4% on BERT.

2.6.2 Empathetic Dialogue Generation

We further check whether our value model helps the empathetic dialogue generation. EMPATHETIC-DIALOGUES applies PREPEND-K, a strategy to add supervised information to data, when predicting the utterance given the dialogue history and the situation. We apply the strategy of prepending the top-k emotion labels for dialogue generation. The top predicted label from the classifiers of emotion

Model	**Accuracy** (σ)	
fastText	$42.27 \pm 0.3\%$	
DistilBERT	$41.81 \pm 0.2\%$	
DistilBERT + Value	$43.98 \pm 0.2\%$	$+\mathbf{2.17}\%$
BERT	$42.93 \pm 0.1\%$	
BERT + Value	$\mathbf{45.67} \pm 0.3\%$	$+\mathbf{2.74}\%$

Table 2.5: Emotion classification performance in EMPATHETICDIALOGUES [RSL19].

is prepended to the beginning of the token sequence as encoder input, as below:

- **Original**: "I finally got promoted!"
- **Prepend-1 emotion**: "*proud* I finally got promoted!"

2.6.2.1 Result

The results are shown in Table 2.6. As observed, prepending emotion tokens provides extra context and improves the generation performance of GPT-2 and DIALOGPT. Since incorporating value improves the emotion classification accuracy, it further improves the generation quality.

2.7 Application: Value Profiling

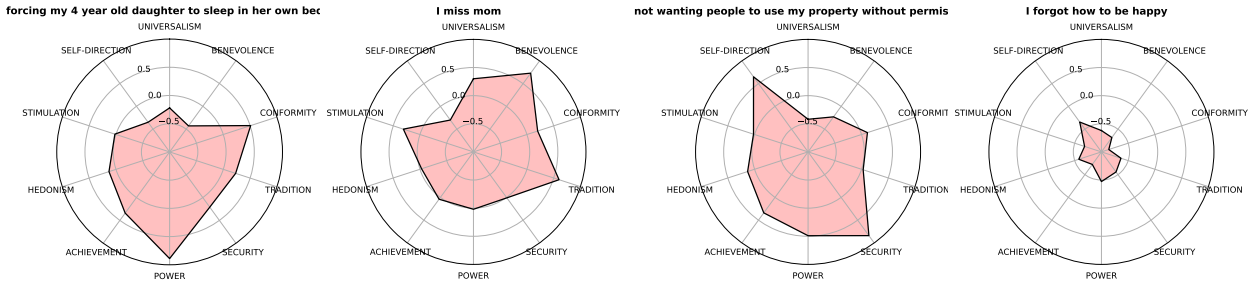

Figure 2.5: Value visualization of example utterances/scenarios.

Model	Ppl.↓
EmoPrepend-1 [RSL19]	24.30
GPT-2	14.74
GPT-2 + Emotion (w/o Value)	14.46
GPT-2 + Emotion (w/ Value)	14.01
DIALOGPT	13.48
DIALOGPT + Emotion (w/o Value)	12.32
DIALOGPT + Emotion (w/ Valued)	**12.12**

Table 2.6: Empathetic dialogue generation in EMPATHETICDIALOGUES [RSL19]. EmoPrepend-1: input prepending emotion from an external classifier.

For a more comprehensive understanding, we visualize the 10-dimensional value of four example scenarios in Figure 2.5. As shown, the value model provides a numerical speaker profile. For instance, saying ”forcing my daughter to sleep in her own bed” implies that the speaker values power and conformity; saying ”I miss mom” implies that the speaker values benevolence; saying ”not wanting people to use my property without permissions” implies the speaker is self-directed and values security. The last example ”I forgot how to be happy” results in a small radar graph. It suggests that even the model could predict the overall polarity pretty well, there is still space to improve its capability of distinguishing different values.

2.8 Appendix

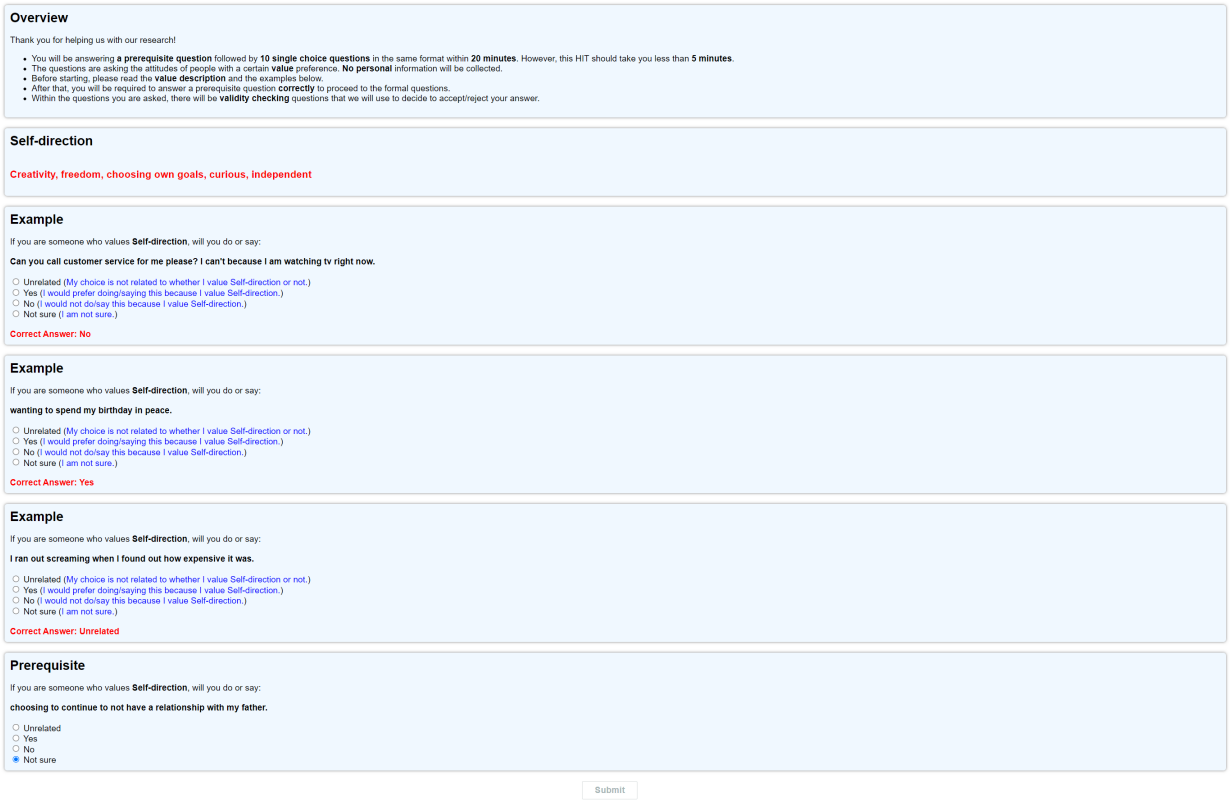

Figure 2.6: Amazon mechanical turk interface (prerequiste).

Figure 2.7: Amazon mechanical turk interface (formal).

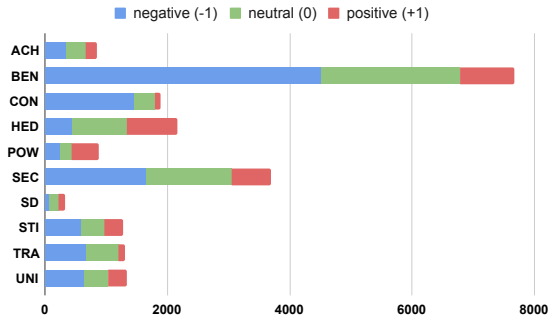

Figure 2.8: The sample number and label distribution of each value split in the VALUENET (original).

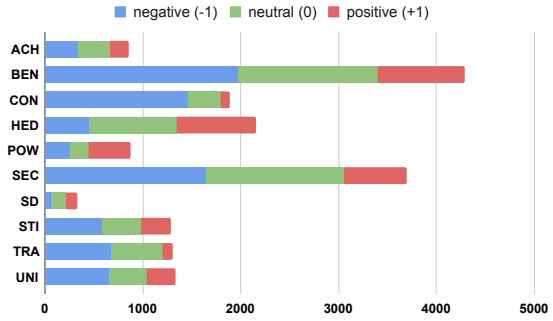

Figure 2.9: The sample number and label distribution of each value split in the VALUENET (balanced).

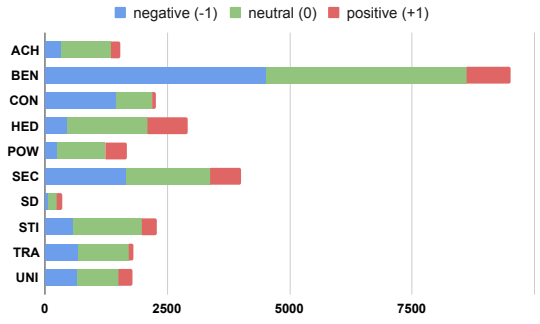

Figure 2.10: The sample number and label distribution of each value split in the VALUENET (augmented).

CHAPTER 3

Social Relation Inference in Dialogues

Inferring social relations from dialogues is vital for building emotionally intelligent robots to interpret human language better and act accordingly. We model the social network as an And-or Graph, named SocAoG, for the consistency of relations among a group and leveraging attributes as inference cues. Moreover, we formulate a sequential structure prediction task, and propose an α–β–γ strategy to incrementally parse SocAoG for the dynamic inference upon any incoming utterance: *(i)* an α process predicting attributes and relations conditioned on the semantics of dialogues, *(ii)* a β process updating the social relations based on related attributes, and *(iii)* a γ process updating individual's attributes based on interpersonal social relations. Empirical results on DialogRE and MovieGraph show that our model infers social relations more accurately than the state-of-the-art methods. Moreover, the ablation study shows the three processes complement each other, and the case study demonstrates the dynamic relational inference.

3.1 Introduction

Social relations form the basic structure of our society, defining not only our self-images but also our relationships [Szt02]. Robots with a higher emotional quotient (EQ) have the potential to understand users' social relations better and act appropriately. Given a dialogue as context and a set of entities, the task of Dialogue Relation Extraction (DRE) predicts the relation types between the entities from a predefined relation set. Table 3.1 shows such an example from the dataset DialogRE [YSC20].

Existing researches using BERT-based models [DCL18, YSC20, XSZ20a] or graph-based

S1:	Well then we'll-we'll see you the day after tomorrow. Mom?! Dad?! What-what... what you guys doing here?!
S2:	Well you kids talk about this place so much, we thought we'd see what all the fuss is about.
S3:	I certainly see what the girls like coming here.
S1:	Why?!
S3:	The sexy blonde behind the counter.
S1:	Gunther?!
S2:	Your mother just added him to her list.
S1:	What? Your-your list?

	Argument Pair	**Trigger**	**Relation Type**
R1	(S2, S1)	dad	per:children
R2	(S3, Gunther)	sexy blonde	per:positive_impression
R3	(S3, S1)	mom	per:children
R4	(S1, S3)	mom	per:parents
R5	(S1, S2)	dad	per:parents

Table 3.1: A dialogue example from DialogRE [YSC20]. Trigger word annotations are not used for training, but rather for illustrating purpose only.

models [XSZ20b, CHH20] focus on identifying entities' relations from the semantics of dialogues—they utilize either the attention mechanism or a refined token graph to locate informative words (*e.g.*, "dad" and "mom") that imply the argument pairs' relations. However, there are still three missing parts in current models for social relation inference, according to our observations. First, current models lack the explicit modeling of the relational consistency among a group of people—such consistency helps humans reason about the social relation of two targets by using their relations with a third person. For the example in Table 3.1, by knowing S2 and S3 are S1's parents and S3 is S1's mother, we can infer that S2 is S1's dad. Second, the personal attribute cues (*e.g.*, gender and profession) can also aid the relational inference but are not fully utilized. In the above example, besides inferring S3 is S1' mother according to S3's feminine attribute, we can also have a guess that Gunther is a waiter, which might be useful for the future social-relational inference. Third, since the BERT-based and token-graph-based models take dialogues as a whole for relation prediction, they cannot perform dynamic inference—updating the relational belief with an incoming dialogic

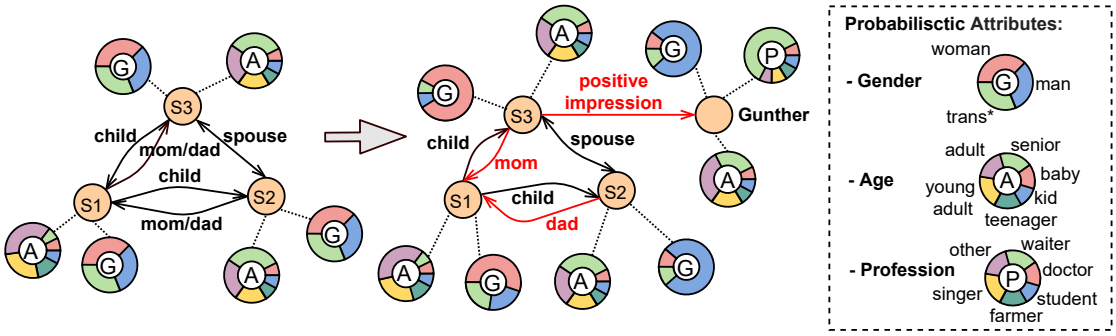

Figure 3.1: Our method iteratively updates the robot's belief of users' individual attributes and social relations, similar to human's reasoning process. The left and right graph show the established and updated belief, respectively.

utterance. This can limit their ability to track the evolving relations along with social interactions, *e.g.*, strangers become friends over a good chat [KTL20], unveiling intermediate reasoning results, or dealing with long dialogues.

Motivated by these observations, we propose to model social relation as an attributed And-Or graph (AoG) [ZWM98, ZM07, WZ11, SRZ16, QZH18], named SocAoG, and develop an incremental graph parsing algorithm to jointly infer human attributes and social relations from a dialogue. In specific, SocAoG describes social relations and personal attributes with contextual constraints of groups and hierarchical representations. To incrementally parse SocAoG and track social relations, we apply Markov Chain Monte Carlo (MCMC) to sample from the posterior probability calculated by three complementary processes (α–β–γ) [QGX20, Zay15]. Figure 3.1 schematically demonstrates a graph update of both relations (*i.e.*, disambiguating mom/dad and adding a new party) and attributes (*e.g.*, gender and profession) with the utterance "**S2**: *Your mother just added him to the list.*" from the example dialogue in Table 3.1.

We evaluate our method on two datasets of DialogRE [YSC20] and MovieGraph [VTC18] for relation inference, and the results show that our method outperforms the state-of-the-art (SOTA)

ones. Overall, we make the following contributions: (*i*) We propose to model and infer social relations and individuals' attributes jointly with SocAoG for the consistency of attributes and social relations among a group. To the best of our knowledge, it is the first time done in the dialogue domain; (*ii*) The MCMC sampling from α–β–γ posterior enables dynamic inference—incrementally parsing the social relation graph, which can be useful for tracking relational evolution, reflecting the reasoning process, and handling long dialogues; (*iii*) We perform an ablation study on each process of α–β–γ to investigate the information contribution, and perform case studies to show the effectiveness of our dynamic reasoning.

3.2 Related Work

We review the related works on the social relation inference from documents, which is a well-studied task, and those from dialogues, which is the emerging task that our work is focused on.

3.2.1 Relation Inference from Documents

Most of the existing literature focus on relation extraction from professional edited news reports or websites. They typically output a set of "subject-predicate-object" triples after reading the entire document [BB07, MBS09, Kum17]. While early works mostly utilize feature-based methods [Kam04, MS14, GYD15] and kernel-based methods [ZAR03, ZG05, MB06], more recent studies use deep learning methods such as recurrent neural networks or transformers [Kum17]. For example, [ZWX16] propose bidirectional LSTM model to capture the long-term dependency between entity pairs, [ZZC17] present PA-LSTM to encode global position information, and [AHH19, PRP19] fine-tune pre-trained transformer language models for relation extraction.

Two streams of work are closely related to our method. Regarding social network modeling, while most works treat pairs of entities isolated [YSC20, XSZ20b, CHH20], [SCM16] formulate the interpersonal relation inference as structured prediction [BM16, QZS20, ZQA20], inferring the collective assignment of relations among all entities from a document [LZW20, JYQ20]. Regard-

ing relation evolution, a few works are aimed to learn the dynamics in social networks, *i.e.,* the development of relations, from narratives by Hidden Markov Models [CID17], Recurrent Neural Networks [KK19], deep recurrent autoencoders [IGC16]. Our method differs from the aforementioned works by modeling the structured social relations and their changes concurrently, which can be useful for the task of tracking social network evolution [DS97] and unveiling the reasoning process of relations. We achieve this by parsing the graph incrementally per utterance with the proposed α–β–γ strategy.

3.2.2 Relation Inference from Dialogues

Recently, [YSC20] introduce the first human-annotated dialogue-based relation extraction dataset DialogRE, in which relations are annotated between arguments that appear in a dialogue session. Compared with traditional relation extraction tasks, DialogRE emphasizes the importance of tracking speaker-related information within the context across multiple sentences. SOTA methods can be categorized into token-graph models and pre-trained language models. For typical token-graph models, [CHH20] present a token graph attention network, and [XSZ20b] further generate a latent multi-view graph to capture relationships among tokens, which is then refined to select important words for relation extraction. For pre-trained models, [YSC20] evaluate a BERT-based baseline model [DCL18] and a modified version BERTs, which takes speaker arguments into consideration. [XSZ20a] propose a simple yet effective BERT-based model, SimpleRE, that takes a novel input format to capture the interrelations among all pairs of entities.

Both categories of SOTA models take a discriminative approach, whereas ignoring two key constraints on relations: *(i)* social relation consistency in a group and *(ii)* human attributes. Different from them, our method formulates the task as dialogue generation from an attributed relation graph, so that the posterior relation estimation models both two constraints. Moreover, SOTA models also assume the relations are static—they cannot learn the dynamics of the relations, while the incremental graph updating strategy naturally enables the dynamic relation inference.

3.3 Problem Formulation

Our goal is to construct a social network through utterances in dialogue. The network is a heterogeneous physical system [YBJ97] with particles representing entities and different types of edges representing social relations. Each entity is associated with multiple types of attributes, while each type of relation is governed by a potential function defined in human attribute and value space, acting as the social norm. The relations are often asymmetric, *e.g.*, A is B's father does not mean B is A's father. To model the network, we utilize an attributed And-Or Graph (A-AoG), a probabilistic grammar model with attributes on nodes. Such design takes advantage of the reconfigurability of its probabilistic context-free grammar to reflect the alternative attributes and relations, and the contextual relations defined on Markov Random Field to model the social norm constraints.

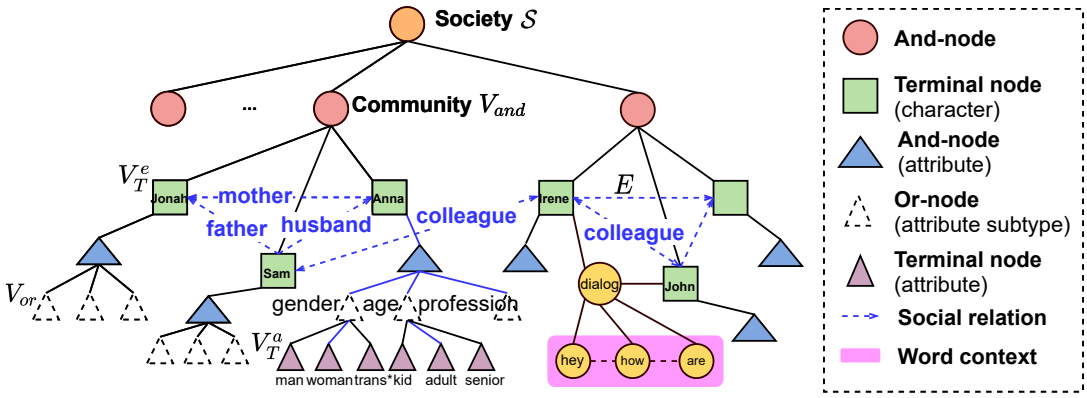

Figure 3.2: SocAoG: Attributed And-Or Graph representation of a social network. A parse graph determining each attribute and relation type is marked in blue lines. Dialogues are governed by the word context and associated human attributes and relations.

The social network graph, named SocAoG, is diagrammatically shown in Figure 3.2. Formally, SocAoG is defined as a 5-tuple:

$$\mathcal{G} =< S, V, E, X, P >, \tag{3.1}$$

where S is the root node for representing the interested society. $V = V_{and} \cup V_{or} \cup V_T^e \cup V_T^a$ denotes

all nodes' collection. Among them, And-nodes V_{and} represent the set of social communities, which can be decomposed to a set of entity terminal nodes, V_T^e, representing human members. Community detection is based on the social network analysis [BS16, DWP07], and can benefit the modeling of loosely connected social relations. Each human entity is associated with an And-node that breakdowns the attributes into subtypes such as gender, age, and profession. All the subtypes consist of an Or-node set, V_{or}, for representing branches to alternatives of attribute values. Meanwhile, all the attribute values are represented as a set of terminal nodes V_T^a. We denote E to be the edge set describing social relations, $X(v_i)$ to be the attributes associated with node v_i, and $X(\vec{e}_{ij})$ to be the social relation type of edge $\vec{e}_{ij} \in E$.

Given P to be the probability model defined on SocAoG, a parse graph pg is an instantiation of SocAoG with determined attribute selections for every Or-node and relation types for every edge. For a dialogue session with T turns $D_T = \{D^{(1)}, D^{(2)}, ..., D^{(T)}\}$, where $D^{(t)}$ is the utterance at turn t, our method infers the attributes and social relations incrementally over turns:

$$\mathcal{G}_T = \{pg^{(1)}, pg^{(2)}, ..., pg^{(T)}\}, \tag{3.2}$$

where $pg^{(t)}$ represents the belief of SocAoG at the dialogue turn t. We incrementally update the pg by maximizing the posterior probability:

$$pg^* = \arg\max_{pg} p(pg|D; \theta), \tag{3.3}$$

where pg^* is the optimum social relation belief, and θ is the set of model parameters.

3.4 Algorithm

3.4.1 α–β–γ for Graph Inference

For simplicity, we denote $X(v_i)$ as $\mathbf{v}_i$ and $X(\vec{e}_{ij})$ as $\mathbf{e}_{ij}$ in the rest of this chapter. We introduce three processes, *i.e.*, α, β, and γ process, to infer any SocAoG belief pg^*. We start by rewriting the

posterior probability as a Gibbs distribution:

$$\begin{aligned} p(pg|D;\theta) &\propto p(D|pg;\theta)p(pg;\theta) \\ &= \frac{1}{Z}exp\{-\mathcal{E}(D|pg;\theta) - \mathcal{E}(pg;\theta)\}, \end{aligned} \tag{3.4}$$

where Z is the partition function. $\mathcal{E}(D|pg;\theta)$ and $\mathcal{E}(pg;\theta)$ are dialogue- and social norm-based energy potentials respectively, measuring the cost of assigning a graph instantiation.

Denoting a dialogue as a sequence of words: $D = \{w_1, w_2, ..., w_\mathcal{T}\}$, the dialogue likelihood energy term $\mathcal{E}(D|pg;\theta)$ can be expressed with a language model conditioned on the parse graph:

$$\begin{aligned} \mathcal{E}(D|pg;\theta) &= \sum_{t=1}^{\mathcal{T}} \mathcal{E}(w_t|\mathbf{c}_t, pg) \\ &= \sum_{t=1}^{\mathcal{T}} -\log(p(w_t|\mathbf{c}_t, pg)), \end{aligned} \tag{3.5}$$

where $\mathbf{c}_t = [w_1, ..., w_{t-1}]$ is the context vector. Intuitively, the word selection depends on the word context, the entities' attributes and their interpersonal relations.

We approximate the likelihood by finetuning a BERT-based transformer with a customized input format $\langle$ `[CLS]D[SEP]` $v_{i_0}\mathbf{e}_{i_0j_0}v_{j_0}...v_{i_n}\mathbf{e}_{i_nj_n}v_{j_n}v_0\mathbf{v}_0...v_n\ \mathbf{v}_n$ `[SEP]` $\rangle$, which is a concatenation of the dialogue history `D` and a flattened parse graph string encoding the current belief. We call the estimation of pg from the dialogue likelihood $p(w_t|\mathbf{c}_t, pg)$ to be the α **process**. α process lacks the explicit constraints for social norms related to interpersonal relations and human attributes.

For the social norm-based potential, we design it to be composed of three potential terms:

$$\begin{aligned} \mathcal{E}(pg;\theta) = &- \beta \sum_{v_i, v_j \in V(pg)} \log(p(\mathbf{e}_{ij}|\mathbf{v}_i, \mathbf{v}_j)) \\ &- \gamma_l \sum_{\vec{e}_{ij} \in E(pg)} \log(p(\mathbf{v}_i|\mathbf{e}_{ij})) \\ &- \gamma_r \sum_{\vec{e}_{ij} \in E(pg)} \log(p(\mathbf{v}_j|\mathbf{e}_{ij})), \end{aligned} \tag{3.6}$$

where $V(pg)$ and $E(pg)$ are the set of terminal nodes and relations in the parse graph, respectively. We call the term $p(\mathbf{e}_{ij}|\mathbf{v}_i, \mathbf{v}_j)$ the β **process**, in which we bind the attributes of node v_i and v_j to

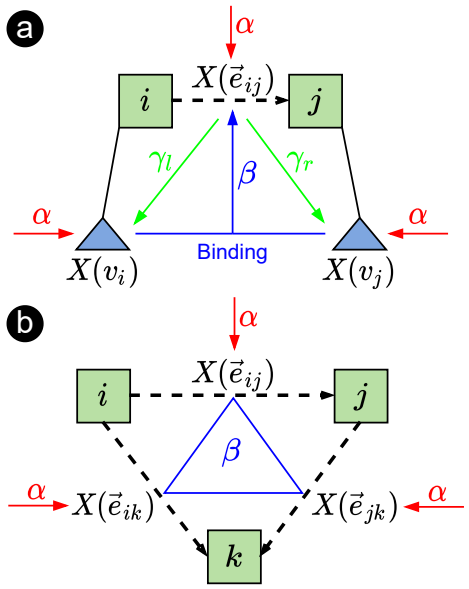

Figure 3.3: (a) α–β–γ process for SocAoG. (b) α–β process for reduced SocAoG without attributes. Note that this β is only modeling the interrelations among $X(\vec{e})$.

update their relation edge $\mathbf{e}_{ij}$, in order to model the constraint on relations from human attributes. Reversely, we call the terms $p(\mathbf{v}_i|\mathbf{e}_{ij})$ and $p(\mathbf{v}_j|\mathbf{e}_{ij})$ the γ **process**, in which we use the social relation edge $\mathbf{e}_{ij}$ to update the attributes of node v_i and v_j. This models the impact of relation to the attributes of related entities. β, γ_l, and γ_r are weight factors balancing α, β and γ processes. Figure 3.3ⓐ shows the graph inference schema with the three processes. Combining equation 3.4, 3.5, and 3.6, we get a posterior probability estimation $p(pg|D;\theta)$ of parse graph pg, with the guarantee of the attribute and social norm consistencies.

Here we also provide a reduced version of our model, $\text{SocAoG}_{\text{reduced}}$, which applies when characters' attributes annotation are not available for training[1]. With the same dialogue-based

[1]Both SocAoG and $\text{SocAoG}_{\text{reduced}}$ do not need attribute annotation during inference once trained.

Algorithm 2 Incremental SocAoG Parsing for Social Relation Inference

Input: dialogue $D_T = \{D^{(1)}, D^{(2)}, ..., D^{(T)}\}$, target argument pairs $\{a_1, a_2\}$.

Initialize $pg^{(0)}$. Initialize $\mathbf{v}_i$ and $\mathbf{e}_{ij}$.

for $t = 1, ..., T$ **do**

  for $s = 1, ..., S$ **do**

    Compute the posterior $p(pg|D^{(t)}; \theta)$.

    Make proposal moves with probabilities q_1, q_2 to get a new parse graph pg'.

    Compute the posterior $p(pg'|D^{(t)}; \theta)$.

    Compute acceptance rate $\alpha(pg'|pg, D^{(t)}; \theta)$.

    Accept/reject pg' according to the acceptance rate.

  end for

  return $\mathbf{e}_{a_1,a_2}$ from the average of accepted pg samples.

end for

energy potential, We define the parse graph prior energy over a set of relation triangles:

$$\mathcal{E}(pg; \theta) = -\beta \sum_{\vec{e}_{ij}, \vec{e}_{ik}, \vec{e}_{jk} \in E(pg)} \log(p(\mathbf{e}_{ij}|\mathbf{e}_{ik}, \mathbf{e}_{jk})). \tag{3.7}$$

The method directly models the constraint of two entities' relation from their relations to others, with the inference schema demonstrated in Figure 3.3ⓑ.

3.4.2 Incremental Graph Parsing

Incrementally parsing the SocAoG is accomplished by repeatedly sampling a new parse graph $pg^{(t)}$ from the posterior probability $p(pg^{(t)}|D^{(t)}; \theta)$. We utilize a Markov Chain Monte Carlo (MCMC) sampler to update our parse graph since the complexity of the problem caused by multiple energy terms.

At each dialogue turn t, we initialize the parse graph with the α classification process, by replacing all the Or-Node tokens with a special token `[CLS]`. We sample the parse graph for S steps and use the average value of obtained samples as an approximation of $pg^{(t)}$. We design two

types of Markov chain dynamics used at random probabilities $q_i, i = 1, 2$ to make proposal moves:

- Dynamics q_1: randomly pick a relation edge $\vec{e}_{ij}$ under the uniform distribution, flip its social relation type $\mathbf{e}_{ij}$ according to the prior distribution given by β process:

$$\prod_{v_i, v_j \in V(pg)} p(\mathbf{e}_{ij}|\mathbf{v}_i, \mathbf{v}_j). \tag{3.8}$$

- Dynamics q_2: randomly pick a terminal node v_i and its attribute subtype under the uniform distribution, and flip the one-hot value of attribute $\mathbf{v}_i$ according to the prior distribution given by γ process:

$$\prod_{\vec{e}_{ij} \in E(pg)} p(\mathbf{v}_i|\mathbf{e}_{ij}) \prod_{\vec{e}_{ji} \in E(pg)} p(\mathbf{v}_i|\mathbf{e}_{ji}). \tag{3.9}$$

Using the Metropolis-Hastings algorithm [CG95], the proposed new parse graph pg' is accepted according to the following acceptance probability:

$$\begin{aligned} \alpha(pg'|pg, D; \theta) &= \min(1, \frac{p(pg'|D;\theta)p(pg|pg')}{p(pg|D;\theta)p(pg'|pg)}) \\ &= \min(1, \frac{p(pg'|D;\theta)}{p(pg|D;\theta)}), \end{aligned} \tag{3.10}$$

where the proposal probability rate is canceled out since the proposal moves are symmetric in probability. We summarize the incremental SocAoG parsing in Algorithm 2. Dialogues give a continuously evolving energy landscape: at the beginning of iterations, $p(pg^{(0)}|D;\theta)$ is a "hot" distribution with a large energy value; by iterating the α–β–γ processes for pg updates through the dialogue, the pg converges to the pg^*, which is much cooler.

3.5 Experiments

3.5.1 Datasets

We use DialogRE (V2)[2] [YSC20] and MovieGraph[3] [VTC18] for evaluating our method. Detailed descriptions on the two datasets, *e.g.*, relation and attribute types, are provided in Appendix 3.6.

DialogRE contains 36 relation types (17 of them are interpersonal) that exist between pairs of arguments. For the joint parsing of relation and attribute, we further annotate the entity arguments with attributes from four subtypes (by following the practice of MovieGraph [VTC18]): gender, age, profession, and ethnicity, according to Friends Central in Fandom[4]. DialogRE is split into training (1073), validation (358), and test (357). Following previous works [YSC20, XSZ20b], we report macro $\mathbf{F}1$ scores in both the standard and conversational settings ($\mathbf{F}1_c$).

MovieGraph provides graph-based annotations of social situations from 51 movies. Each graph comprises nodes representing the characters, their emotional and physical attributes, relationships, and interactions. We use a subset (40) of MovieGraph with available full transcripts and split the dataset into training (26), validation (6), and test (8). For MovieGraph, we only evaluate with $\mathbf{F}1$ since the trigger word annotation for computing $\mathbf{F}1_c$ is not available.

3.5.2 Experiment Settings

We learn the SocAoG model with a contrastive loss [HCL06] comparing the posterior of a positive parse graph against a negative one. All parameters are learned by gradient descent using the Adam optimizer [KB14]. During the inference stage, for each utterance, we run the MCMC for $S = \min\{w \times (KM + K(K-1)N), S_{max}\}$ steps given K entities, M attributes, N relations, and a sweep number of w. The probability of flipping the relation q_1 is set to 0.7 to bias towards the

[2]https://github.com/nlpdata/dialogre

[3]http://moviegraphs.cs.toronto.edu/

[4]https://friends.fandom.com/wiki/Friends_Wiki

	DialogRE (V2)				MovieGraph	
	Dev		Test		Dev	Test
Methods	$\mathbf{F}1(\sigma)$	$\mathbf{F}1_c(\sigma)$	$\mathbf{F}1(\sigma)$	$\mathbf{F}1_c(\sigma)$	$\mathbf{F}1(\sigma)$	$\mathbf{F}1(\sigma)$
BERT [DCL18]	59.4 (0.7)	54.7 (0.8)	57.9 (1.0)	53.1 (0.7)	50.6 (1.2)	53.6 (0.3)
BERT_S [YSC20]	62.2 (1.3)	57.0 (1.0)	59.5 (2.1)	54.2 (1.4)	50.7 (1.1)	53.6 (0.4)
GDPNet [XSZ20b]	67.1 (1.0)	61.5 (0.8)	64.3 (1.1)	60.1 (0.9)	53.1 (1.1)	56.4 (0.8)
SimpleRE [XSZ20a]	68.2 (1.1)	63.4 (0.6)	66.7 (0.7)	63.3 (0.9)	55.2 (0.5)	58.1 (0.7)
$\text{SocAoG}_\text{reduced}$ (our method)	69.1 (0.4)	65.7 (0.5)	68.6 (0.9)	65.4 (1.1)	**60.7 (0.4)**	63.2 (0.3)
SocAoG (our method)	**69.5 (0.8)**	**66.1 (0.7)**	**69.1 (0.5)**	**66.5 (0.8)**	60.1 (0.6)	**64.1 (0.8)**

Table 3.2: Performance comparison between BERT, BERT_S, GDPNet, SimpleRE, $\text{SocAoG}_\text{reduced}$, and SocAoG. We report 5-run average results and the standard deviation (σ).

relation prediction at first.

3.5.3 Baseline Models

We compare our method with both transformer-based (**BERT**, $\mathbf{BERT_S}$, **SimpleRE**) and graph-based (**GDPNet**) models. Given dialogue history D and target argument pair (v_i, v_j), **BERT** [DCL18] takes input sequences formatted as $\langle \texttt{[CLS]}\, d\, \texttt{[SEP]}\, v_i\, \texttt{[SEP]}\, v_j\, \texttt{[SEP]} \rangle$. $\mathbf{BERT_S}$ [YSC20] is a speaker-aware modification of BERT, which also takes speaker information into consideration by converting it into a special token. **SimpleRE** [XSZ20a] models the relations between each pair of entities with a customized input format. **GDPNet** [XSZ20b] takes in token representations from BERT and constructs a multi-view graph with a Gaussian Graph Generator. The graph is then refined through graph convolution and DTWPool to identify indicative words.

3.5.4 Performance Comparison

Table 3.2 shows the performance comparison between different methods on the two datasets. It clearly shows that both of our models, SocAoG and $\text{SocAoG}_\text{reduced}$, outperform the existing methods

by all the metrics. In specific, without using any additional information of attributes, SocAoG$_{\text{reduced}}$ surpasses the state-of-the-art method (SimpleRE) by 1.9% ($\mathbf{F}1$)/2.1% ($\mathbf{F}1_c$) on DialogRE testing set, and by 5.1% ($\mathbf{F}1_c$) on MovieGraph testing set. Such improvement shows the importance of relational consistency for the modeling, and proves the effectiveness of our SocAoG formulation to introduce the social norm constraints.

Moreover, by comparing between SocAoG and SocAoG$_{\text{reduced}}$, we see that SocAoG further improves most of the metrics by leveraging the attribute information for relation reasoning, *e.g.*, 69.1% *vs.* 68.6% for DialogRE testing $\mathbf{F}1$ and 64.1% *vs.* 63.2% for MovieGraph testing $\mathbf{F}1$. The results demonstrate our method can effectively take advantage of the attributes as cues for social relation predictions. We compare our SocAoG model with the existing model of highest accuracy (SimpleRE) by relation types, and see consistent improvements for all types. A part of the results are shown in Figure 3.4. We also observe that there are larger accuracy boosts for relations between human entities than non-human entities (*e.g.*, human-place), by an average of +2.5% *vs.* +1.8% in $\mathbf{F}1$, which is also reflected from Figure 3.4 (left 10 bars *vs.* right 10 bars). This can be explained as relation/attribute constraints are more meaningful for interpersonal relations, *e.g.*, there are more constraints for the relation between three humans than the relation between two humans and a place.

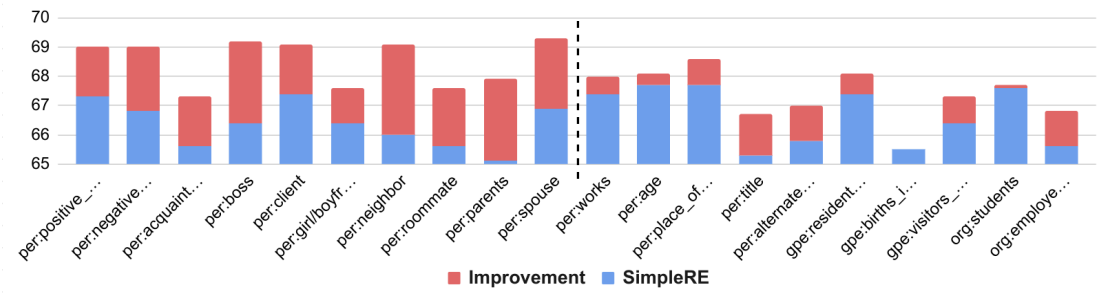

Figure 3.4: Performance boosts ($\mathbf{F}1$) of SocAoG compared to SimpleRE [XSZ20a] by relation type. The left bars to the dashed line are relations between humans, while the right ones are those between human and non-human entities.

Table 3.2 also sees more accuracy improvement on MovieGraph dataset than DialogRE (+3.2% *vs.* +6.0% in test $\mathbf{F}1_c$ using SimpleRE as baseline). This is possibly because the dynamic inference nature of our method makes it effective for dealing with dialogues with more turns: while existing methods either truncate dialogues or use sliding windows, our method continuously updates the relation graph given an incoming turn. We case study the dynamic inference in the next subsection.

3.5.5 Case Study on Dynamic Inference

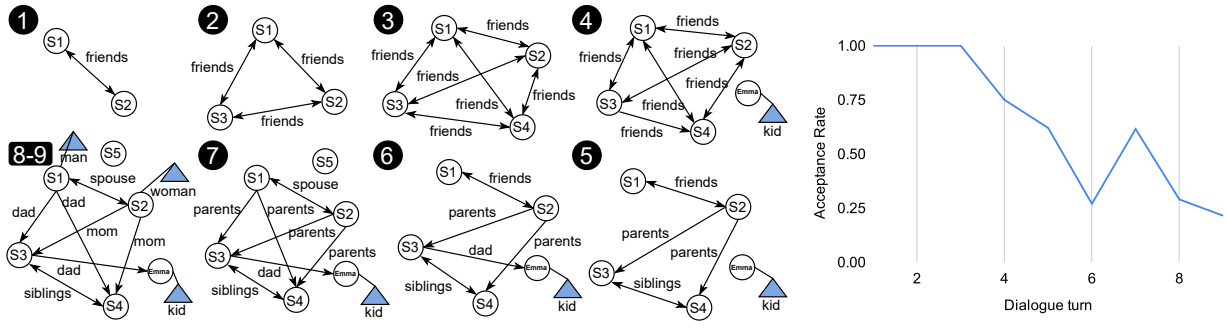

Figure 3.5: Left: inferred parse graph sequence from SocAoG based on the test dialogue in Table 3.3. Note that dad/mom are not distinguished in DialogRE. Right: model convergence measured by acceptance rate at each dialogue turn.

Our method incrementally updates the relation and attribute information for a group of entities upon per utterance input with the proposed α–β–γ strategy. Such dynamic inference can potentially help reflect the evolving relations, unveil the reasoning process, and deal with long dialogues. Figure 3.5 shows the parse graph sequence by SocAoG inferring from a DialogRE testing dialogue as shown in Table 3.3. We can see that the method continuously refines the relation/attributes from an initial guess with incoming contexts, *e.g.* S2-S3: friends→parents in turn 5. Besides, the case also shows that attributes can aid relation predictions, *e.g.*, the inferred age of Emma clarifies her relation with S3. Moreover, since our method models the relation consistency among a group, it can predict the relation between two humans that do not talk directly. For example, S1 and S2 are

①	**S1, S2:**	Hi!
②	**S3:**	Hey!
③	**S4:**	So glad you came!
④	**S1:**	I can't believe Emma is already one!
⑤	**S2:**	I remember your first birthday! Ross was jealous of all the attention we were giving you. He pulled on his testicles so hard! We had to take him to the emergency room!
⑥	**S3:**	There's something you didn't know about your dad!
⑦	**S5:**	Hey Mr. and Mrs. Geller! Let me help you with that.
⑧	**S1:**	Thank you!
⑨	**S5:**	Oh man, this is great, uh? The three of us together again! You know what would be fun? If we gave this present to Emma from all of us!

Table 3.3: Dialogue example from the testing set of DialogRE [YSC20].

inferred to be a couple by their dialogues with S5 in turn 7.

Figure 3.5 also plots the average MCMC acceptance rate for the case, as defined in Formula 3.10, indicting the convergence of the inference. We see that the algorithm only needs to update the current graph belief slightly with a new perceived utterance. A peak in the curve can indicate that a key piece of information is detected that contradicts the existing belief: *e.g.,* there is a peak of convergence curve in turn 7, which corresponds to *"S5: Hey Mr. and Mrs. Geller!"*, indicating that S1 and S2 are a couple rather than friends. As such, we can see the algorithm get several relations updated accordingly. We also show the convergence plots for 50 random testing cases from DialogRE in Figure 3.6, and the mean/standard deviation convergence rate as the black line/blue shade. We prove that our updating algorithm is robust for the converged results.

3.5.6 Ablation Study on α–β–γ

The α–β–γ strategy is designed to update relations and attributes jointly, having the input information flowing through the parse graph for the consistency of predictions. To validate the design,

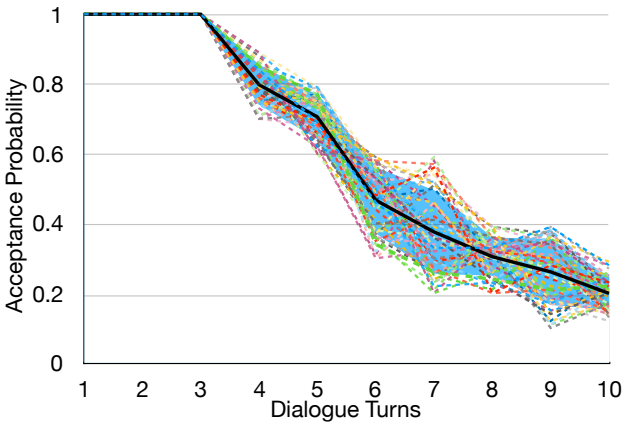

Figure 3.6: MCMC acceptance rate of the incremental parsing process. Dotted lines, black line, and blue shade are for samples, mean, and standard deviation, respectively.

Processes			$\mathbf{F}1(\sigma)$	$\mathbf{F}1_c(\sigma)$
α	β	γ		
✓			67.1 (0.5)	64.2 (1.1)
✓	✓		68.4 (0.8)	65.3 (0.6)
✓		✓	68.3 (0.4)	65.2 (0.7)
✓	✓	✓	**69.1 (0.5)**	**66.5 (0.8)**

Table 3.4: An ablation study on our parsing algorithm.

we ablate the processes on DialogRE to evaluate their impact on performance. Table 3.4 shows that α process, which is the discriminative model, makes the fundamental contribution, whereas β and γ processes alone cannot recognize social relations since they cannot perceive information from dialogues. Significantly, removing either one of the two processes will decrease the overall performance since the inference efficiency is reduced.

3.6 Appendix

ID	Subject	Relation Type	Object	Inverse Relation
1	PER	per:positive_impression	NAME	
2	PER	per:negative_impression	NAME	
3	PER	per:acquaintance	NAME	per:acquaintance
4	PER	per:alumni	NAME	per:alumni
5	PER	per:boss	NAME	per:subordinate
6	PER	per:subordinate	NAME	per:boss
7	PER	per:client	NAME	
8	PER	per:dates	NAME	per:dates
9	PER	per:friends	NAME	per:friends
10	PER	per:girl/boyfriend	NAME	per:girl/boyfriend
11	PER	per:neighbor	NAME	per:neighbor
12	PER	per:roommate	NAME	per:roommate
13	PER	per:children	NAME	per:parents
14	PER	per:other_family	NAME	per:other_family
15	PER	per:parents	NAME	per:children
16	PER	per:siblings	NAME	per:siblings
17	PER	per:spouse	NAME	per:spouse
18	PER	per:place_of_residence	NAME	gpe:residents_of_place
19	PER	per:place_of_birth	NAME	gpe:births_in_place
20	PER	per:visited_place	NAME	gpe:visitors_of_place
21	PER	per:origin	NAME	
22	PER	per:employee_or_member_of	NAME	org:employees_or_members
23	PER	per:schools_attended	NAME	org:students
24	PER	per:works	NAME	
25	PER	per:age	VALUE	
26	PER	per:date_of_birth	VALUE	
27	PER	per:major	STRING	
28	PER	per:place_of_work	STRING	
29	PER	per:title	STRING	
30	PER	per:alternate_names	NAME/STRING	
31	PER	per:pet	NAME/STRING	
32	GPE	gpe:residents_of_place	NAME	per:place_of_residence
33	GPE	gpe:births_in_place	NAME	per:place_of_birth
34	GPE	gpe:visitors_of_place	NAME	per:visited_place
35	ORG	org:employees_or_members	NAME	per:employee_or_member_of
36	ORG	org:students	NAME	per:schools_attended
37	NAME	unanswerable	NAME/STRING/VALUE	

Table 3.5: Relation types in DialogRE [YSC20].

attributes	gender	male, female
	age	adult, kid, young adult, teenager, senior, baby
	ethnicity	caucasian, asian, arab, south-asian, hispanic, african, native american, other, aboriginal, african-american
	profession	photographer, cab driver, priest, writer, receptionist, delivery man, yoga instructor, chef, bartender, waitress, tailor, parking attendant, student, professional, lawyer, teacher, businessman, secretary, model, prince, banker, court reporter, intern, police officer, child psychologist, doctor, salesman/woman, hustler, bull rider, worker, doctors, businessman/woman, nurse, barman, janitor, policeman, inspector, FDA agent, counselor, waiter, judge, magician, prostitute, doorman, elevator operator, hotel manager, maid, bellhop, saleswoman, salesman, politician, driver, usher, actress, actor, florist, pilot, flight attendant, film/tv producer, building manager, paramedic, federal agent, postal worker, comic book artist, singer, executive, hockey player, referee, waiter/waitress, ex-soldier, receptionist, mafia boss, mafia member, musician, drug lord, fruit vendor, barber, masseuse, mental patient, mental patient, bus driver, night guard, housewife, editor, gardener, publisher, builder, elf, security guard, security chief, pedicurist, professor of defense against the dark arts, wandmaker, wizard, caretaker, ghost, villain, Philadelphia Eagles fan, cowboys America fan, bookmaker, unemployed, high school principal, jobless, racists, nuclear physicist, surgeon, soldier, colonel, professor, engineer, military officer, technician, game show host, police, robber, waiter/waitress, hitman, actor/actress, criminal, boxer, drug dealer, restaurant host, impersonator, military, trainer, manager, housekeeper, veterinarian, sportsperson, sports coach, sports agent, accountant, personal assistant, nanny, reporter, tv host, cameraman, tv presenter, cashier, artist, chauffeur, video artist, private investigator, administrator, tennis instructor, professional tennis player, detective, ticket collector, director, medical workers, hospital orderly, pharmacist, security officer, dental assistant, dentist, drug addict, registered sex offender, fetish worker, customer support, policemen, CEO, babysitter, assistant, principal, guidance counselor, farmer, entertaining, domestic worker, fisherman, author, psychologist, security person, tv personality, zeppelin crewman, king/queen, knight, journalist, assistant, weatherman, show host, make-up artist, seller, agent, tv show host, makeup artist, treasure hunter, naval officer, steward, ship captain, ship designer, sailor, designer, carpenter, valet, bail bondsman, court bailiff, court clerk, blackjack dealer, movie star, casino owner, casino manager, art director, executive recruiter, sports editor, cowboy, cowboy employer, hacker, investment counselor, hairdresser, sports commentator, chemist, government rep, vicar, robot, hotline agent, cook, surrogate date, philosopher, architect, record store owner, movie reviewer, call operator, bride, dog sitter, newspaper employer, vet, insurance broker, union leader, tv reporter, senator, rancher, locksmith, district attorney, store owner, smuggler, insurance agent, video editor, bouncer, trainee, real estate agent, prison guard, tour guide, mobster
relations		sibling, parent, cousin, customer, friend, stranger, spouse, colleague, boss, would like to know, lover, mentor, engaged, knows by reputation, acquaintance, roommate, best friend, antagonist, employed by, business partner, student, classmate, patient, teacher, child, heard about, enemy, employer of, psychiatrist, doctor, collaborator, ex-lover, landlord, superior, supervisor, grandchild, divorced, sponsor, ex-boyfriend, neighbor, fan, close friend, sister/brother-in-law, uncle, host, employer, step-mother, foster-son, family friend, godfather, godson, brother-in-law, nanny, grandparent, aunt, aide, students, family, customers, classmates, alleged lover, trainer, slave, hostage, robber, owner, instructor, competitor, fiancee, aunt/uncle, mother-in-law, girlfriend, killer, babysitter, one-night stand, boyfriend, tenant, distant cousin, father-in-law, mistress, agent, replacement, argue about the relationship, lawyer, ex-spouse, ex-girlfriend/ex-boyfriend, niece/nephew, parent-in-law, guardian, operative system, couple, goddaughter, customer, ex-neighbor, worker, vet, apprentice, public official, nurse, supporter, interviewee, interviewer, supporters, ex-fiance, fiance

Table 3.6: Attribute and relation types in MovieGraph [VTC18].

CHAPTER 4

Mental State Transition and Human Value

Building a socially intelligent agent involves many challenges. One of which is to track the agent's mental state transition and teach the agent to make decisions guided by its value like a human. Towards this end, we propose to incorporate mental state simulation and value modeling into dialogue agents. First, we build a hybrid mental state parser that extracts information from both the dialogue and event observations and maintains a graphical representation of the agent's mind; Meanwhile, the transformer-based value model learns human preferences from a curated human value dataset, VALUENET. Empirical results show that the proposed model attains state-of-the-art performance on the dialogue/action/emotion prediction task in the fantasy text-adventure game dataset, LIGHT. We also show example cases to demonstrate: (*i*) how the proposed mental state parser can assist the agent's decision by grounding on the context like locations and objects, and (*ii*) how the value model can help the agent make decisions based on its personal priorities.

4.1 Introduction

Recently, there has been remarkable progress in language modeling with large-scale pre-trained models [VSP17, DCL19, RWC19]. Such models are used to build either general chatbots [ZSG20] or task-oriented dialogue systems [PLL20, AAA21, QZS20]. While most of these systems have been able to generate fluent sentences, there are two major challenges towards building socially intelligent agents. First, considering dialogues as a "meeting of minds" [Gar14] or achieving some alignment of the interlocutors' mental models [RSM86, SVT16], few existing works are explicitly tracking the mental state transition of agents [AYC20]. Endowing current dialogue systems with

Figure 4.1: Socially intelligent agents with mental state simulation and human values.

such capability would allow the agent to condition its utterance on the context, simulate the effect of its actions, and further help understand the extended meaning, implicature, and irony expressed by the user [Gri81, Gri89]. Second, it remains under-explored that teaching agents to make a rational decision guided by its value. From a social and cultural perspective, humans tend to have a common preference described by the utility function related to individual values, common sense, and social awareness. For the example in Figure 4.1, someone who values personal security prefers staying at home rather than going outside at night.

Our work aims to alleviate the aforementioned problems, based on Embodied Cognitive Linguistics (ECL) [LJ80, Gar14] and established value theories in sociology [Sch12]. The ECL states that natural language is inherently executable, driven by mental simulation and metaphoric inference [LJ80], and learned through embodied interaction [FN04, TSH20]. Following its tenents, we present a hybrid mental state parser that converts dialogue and event observations into a graphical representation of the agents' minds. Initialized with the location and object description, the interpretable representation is updated through the interaction history to track the evolving process of an agent's belief about surroundings and other agents.

In the field of intercultural research, [Sch92, SCV12] identify basic individual values that are recognized across cultures. Inspired by the theory, we propose to incorporate a value model that learns social common preferences on a curated knowledge base, VALUENET. We perform experiments on a large-scale text-based embodied environment LIGHT [UFK19]. Empirical results show that the model with our mental state emulator and value function achieves the highest performance that aligns with human annotation among existing transformer-based models. Moreover, case studies further demonstrate that the mental state provides extra context information, while the value model helps agents make value-driven decisions.

Our contributions are two-fold. First, we propose to rethink the design of current dialogue systems and suggest a new paradigm from the perspective of cognitive science and contemporary sociology. Second, we present a new framework for building socially intelligent agents by incorporating mental state simulation and human value modeling into dialogue generation and decision making. Our methodology can be generalized to a wide range of interactive social situations in dialogue systems [Zha19], virtual reality [LSY19], and human-robot interactions [YL17].

4.2 Related Work

4.2.1 Text-based Embodied AI

Most recent works in dialogues only study the statistical regularities of language data, without an explicit understanding of the underlying world. The virtual embodiment [KP19] was proposed as a strategy for language research by several previous works [Bro91, KBV16, GM16, MJB16, LUT17]. It implies that the best way to acquire human knowledge is to have the agent learn through experience in a situated environment. [UFK19] introduce LIGHT as a research platform for studying grounded dialogue [Gri81, Gri89, Sta02], where agents can perceive, emote, and act when conducting dialogues with other agents. [AUL20] extend LIGHT with a dataset of ”quests”, aiming to create agents that both act and communicate with other agents in pursuit of a goal. Instead of guiding the agent to complete an in-game goal, our work aims to teach agents to speak/act in a

socially intelligent way. Besides LIGHT, there are also other text-adventure game frameworks, such as [NKB15] and TextWorld [CKY18], but no human dialogues are incorporated in them. Based on TextWorld, there are recent works [YCS18, YM19, AH19, AYC20] on building agents trained with reinforcement learning.

4.2.2 Mental State Transition

An important hypothesis in the ECL [LJ80, FN04] is that humans understand the meaning of language by mentally simulating its content. Great efforts have been made to model human mental states. For example, [DRS19] design a memory network capable of storing knowledge and generating natural responses conditioning on retrieved entries. [AYC20] propose a graph-aided transformer agent (GATA) that infers and updates latent belief graphs during planning to enable effective action selection. However, GATA is designed for capturing game dynamics not dialogues, and our method is more flexible to encode both explicit environmental changes caused by agents' actions and implicit mental state updates triggered by agents' utterances. Such hybrid approaches mixing fixed symbolic states with deep continuous states are studied in recent neural-symbolic research [Sun94, GLG08, BGB17, YWG18]. The result interpretable graphs have two benefits: *(i)* the mental state parsing could be viewed as a form of executable semantic parses [Lia16], so it is easy to write programs to simulate the mind transition. A real-world application leveraging similar approaches is seen in [ABB20]. *(ii)* the unified graphical representation can be extended to model higher-order mental states, *i.e., Theory of Mind* (ToM) [PW78]. ToM is defined as the ability to impute mental states to oneself and others. It enables humans to make inferences about what other people believe in a given situation and predict what they will do [App10, GH17, ALS19]. ToM is thus impossible without the capacity to form "second-order representations" [Den78, Pyl78, GM15].

4.2.3 Human Value

When teaching agents to speak and act in a socially intelligent way, an approach considering values should be adopted. The theory of basic human values, developed by [Sch92, Sch12], tries to measure universal values that are recognized throughout major cultures. A set of 10 basic values[1] are identified and serve as the guiding principles in the life of a person or group [CD12], as shown in Figure 4.2. Similarly, in economics and ethics, the concept of utility was developed as a measure

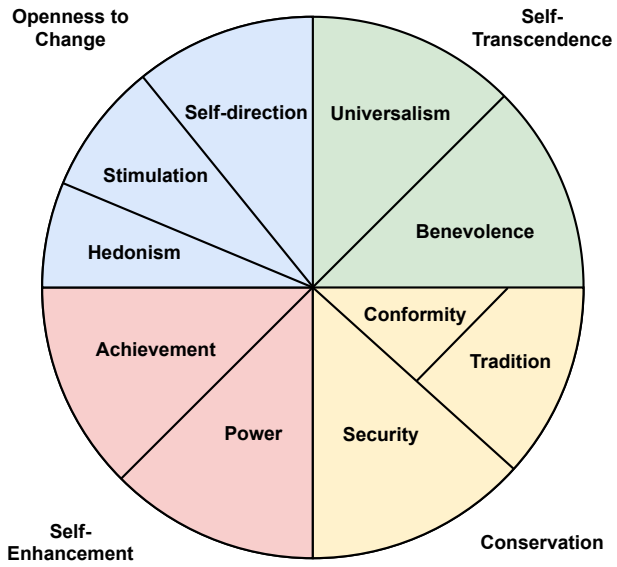

Figure 4.2: Theory of basic human values [Sch92].

of pleasure or satisfaction that drives human activities at all levels. Derived from the rational choice theory [Abe09], the utilitarianism states that human decision-making could be viewed as a two-step procedure. First, we select a feasible region based on the financial, legal, physical, or emotional restrictions we are facing. Then we make a choice based on the preference order [All02, Jon12]. In this work, we learn a transformer-based utility function of human values from a self-curated knowledge base VALUENET. Inspired by descriptive ethics, VALUENET provides social scenarios

[1] A refinement of the theory [SCV12], partitions the same continuum into 19 more narrowly defined values that permit more precise explanation and prediction.

and annotated human preference to teach the agent human attitudes to various ethical situations. The dataset is curated from the widely used social commonsense dataset SOCIAL-CHEM-101 [FHS20] and labeled with Amazon Mechanical Turk.

4.3 Problem Formulation

We will first briefly introduce the text-adventure environment LIGHT, followed by the mental state modeling and value utility formulation.

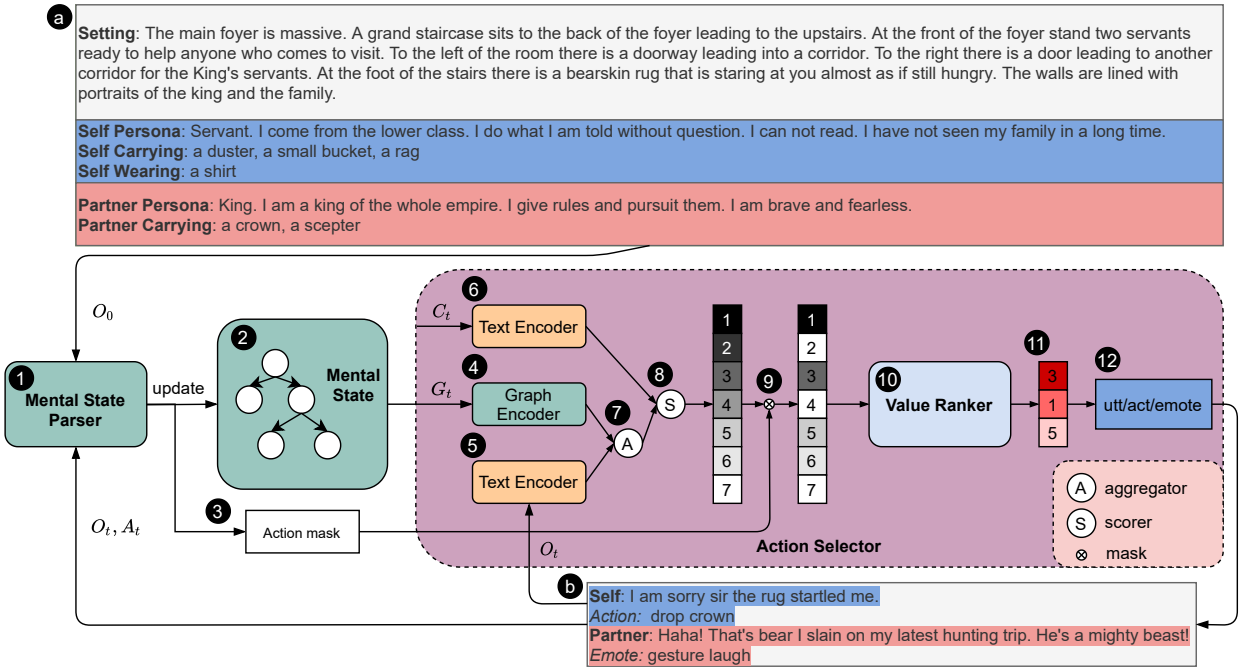

Figure 4.3: Socially intelligent agent architecture with mental state parser and value model.

LIGHT [UFK19] is a large-scale crowd-sourced fantasy text-adventure platform for studying grounded dialogues. Figure 4.3ⓐ shows a typical local environment setting, including location description, objects (and their affordances), characters, and their personas. Agents can talk to other agents in free-form text, take actions defined by templates, or express certain emotions (Figure

4.3ⓑ). Given the environmental setting and observation history, our task is to predict the agent's utterance/action/emotion for the next turn. To achieve this goal in a socially intelligent manner, we model the agent's mental state transition and incorporate human values. The mind model is proposed to depict the agent's belief about the underlying states of the text world. Meanwhile, a utility function of human values is designed to describe human preferences in common social situations. We experiment on the text-adventure game for simplicity, but the proposed architecture supports richer environments.

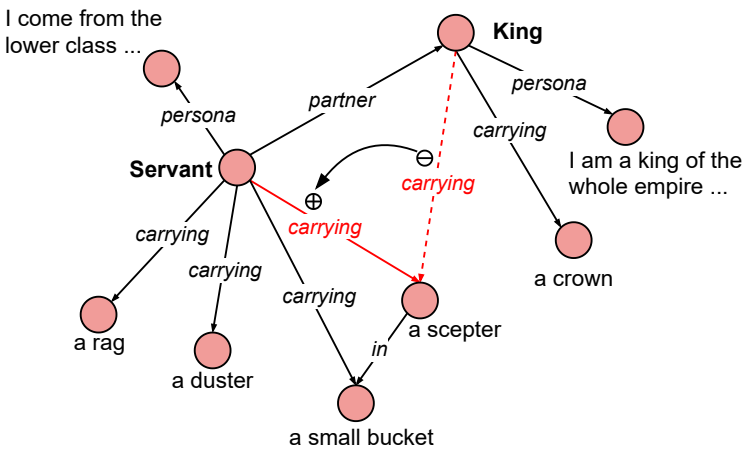

Figure 4.4: A graphical representation of the agent's mental state. Nodes are attributed with encoded natural language description of agents, objects and the environment. Agents' action trigger explicit topology changes of the graph.

4.3.1 Mental State Modeling

Our goal is to parse, construct and maintain the mental states in dialogues. With the mental state grounding on the details of the local environment, the agent could simulate and reason the evolutionary status of the world and condition its speaking and actions. A graphical representation of the mental state is proposed, as illustrated in Figure 4.4. Nodes in the graph represent the involved

agents, persona descriptions, objects, objects' descriptions, and setting descriptions, which will change as the game setting switches. The relational edges between these nodes describe the state of mind. The mental state is updated with the observed dialogue history or actions, *e.g.*, *King gives the scepter to the servant* will result in the scepter being moved from the King to the servant.

4.3.2 Human Value Modeling

We assume that the agent in the fantasy world would make near-optimal choices to maximize the utility of its preferred values. We denote the available alternatives to be a set of n exhaustive and exclusive utterances or actions $A = \{a_1, ..., a_i, ..., a_n\}$. The function $f_v(\cdot)$ describes the utility score of the alternative from the value dimension $v, v \in V = \{$*achievement, power, security, conformity, tradition, benevolence, universalism, self-direction, stimulation, hedonism*$\}$. For example, if a_i is more preferred than a_j in terms of *security*, then $f_{security}(a_i) > f_{security}(a_j)$. Usually, we cannot find an analytical form of the utility function. However, what matters for preference ordering is which of the two options gives the higher expected utility, not the numerical values of those expected utilities.

In LIGHT, the agent's value priority is reflected by its persona description. For the example in Figure 4.3ⓐ, the servant is a person who values conformity and tradition and has a lower priority on *self-direction* and *stimulation*. Using the same value function to approximate a value priority parser: $f_v(p)$, where p is the persona description, the utility or the desirability of candidate a_i to person p is the Euclidean distance between its value priority and the candidate's utility score:

$$u(a_i) = \sqrt{\sum_{v \in V} (f_v(p) - f_v(a_i))^2}. \tag{4.1}$$

Since some actions could be impossible physically (*e.g., one cannot drop an object if the agent is not carrying the object*), the decision making process becomes a problem of maximizing the utility function that is subject to some constraints from the mental state, *i.e.*, $u(a|c)$, where c represents the context or constraints.

4.4 Algorithms

The overall architecture of our proposed framework is illustrated in Figure 4.3. For each scenario, a setting description (Figure 4.3ⓐ) is provided by the LIGHT environment, which can include a description of the location, object affordances, agents' personas, and the objects that agents are carrying, wearing, or wielding. The free-form conversations, actions, and emotions are logged during the communication as the observation history (Figure 4.3ⓑ). To begin with, a mental state parser will parse the setting descriptions into graph representation and initialize the agent's mental state (steps ① and ②). Besides the mental state updating, the parser also outputs an action mask that is aimed to rule out actions that are physically or causally impossible to take (step ③). A graph encoder (step ④) and a text encoder (step ⑤) will convert the mental state graph G_t and the dialogue observation O_t into vector representations, respectively. The same text encoder will be used to encode the candidates C_t (step ⑥). In step ⑦, the context vectors are combined by a bi-directional attention aggregator [YDL18, SKF16], and each candidate is assigned a score with a Multi-Layer Perceptron (MLP) (step ⑧). The action mask is then applied to get the feasible candidates under current mental state constraints (step ⑨). In steps ⑩ and ⑪, the top three candidates from the last step will be fed into the value model and re-ranked. Finally, the selected utterance/action/emotion is executed by the agent (step ⑫) and fed back to the environment. Upon receiving the response from other agents in the environment, the new observation will be again parsed and used to update the agent's state of mind, and the cycle repeats. In the following, we will describe each component in more detail.

4.4.1 Mental State Modeling (steps ①-②)

Figure 4.5 describes the architecture of the mental state parser. We define the mental state graph $G \in [-1, 1]^{R \times N \times N}$, where R is the maximum number of relation types and N is the maximum number of entities. The initial mental state graph G_0 is constructed by a ruled-based parser from the setting description O_0. The graph is encoded by function f_e to a hidden state h_0 that is later used for

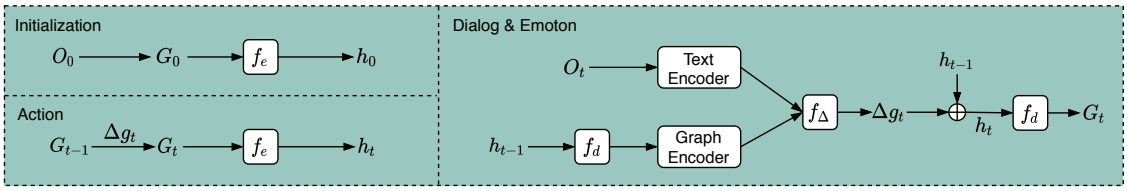

Figure 4.5: Overall architecture of the hybrid mental state parser.

graph update. At game step t, the mental state parser parses relevant information from observation O_t and update the agent's mental state from G_{t-1} to G_t. Considering that observation O_t typically conveys incremental information from step $t-1$ to t, we generate the graph update Δg_t instead of the whole graph at each step

$$G_t = G_{t-1} \oplus \Delta g_t, \tag{4.2}$$

where $\oplus$ is the graph update operation. The graph update can be either discrete or continuous, and there have been studies on the pros and cons of each updating method [AYC20]. The discrete approach may suffer from an accumulation of errors but benefit from its interpretability. The continuous graph model needs to be trained from data, but it is more robust to possible errors. In this work, we propose a hybrid (discrete-continuous) method for updating the agent's state of mind by considering there exists a mixture of discrete events and continuous information in typical human-machine interactive environments. In the specific example of our tested LIGHT, the actions or events are template-based, it is more appropriate to adopt a discrete method for parsing; meanwhile, since utterances are challenging to be encoded into discrete representations, we apply a continuous update method instead.

4.4.1.1 Discrete Graph Definition & Update

To update the graph, we define Δg_t as a sequence of update operations of the following two atomic types:

- ADD(`src`, `dst`, `relation`): add a directed edge, named `relation`, from node `src` to

node `dst`.

- DEL(`src`, `dst`, `relation`): delete a directed edge, named `relation`, from node `src` to node `dst`.

LIGHT defines various actions including *get*, *drop*, *put*, *give*, *steal*, *wear*, *remove*, *eat*, *drink*, *hug* and *hit*, and each taking either one or two arguments, *e.g.*, *give scepter to servant*. Every action could be parsed as one or a sequence of update operators that act on G_{t-1}. For example, actor performing "*give object to agent*" can be parsed into DEL(*actor*, *object*, *carrying*) and ADD(*agent*, *object*, *carrying*). The rule-based parsing of the setting description and the discrete events could also be replaced by a seq2seq decoding process. Since both strings are well-structured in LIGHT, we omit training such a decoder for simplicity. Note that actions in LIGHT could only be executed when constraints are met, so we also generate an action mask according to the current mental state. By checking the adjacency matrix, we rule out action candidates conducted on objects that are inaccessible.

4.4.1.2 Continuous Graph Definition & Update

Besides the actions taken by the agents, their utterances could also have an implicit impact on the agents' mental states. To handle the continuous dialogue observation, we use a recurrent neural network as the graph update operation $\oplus$.

$$\begin{aligned} \Delta g_t &= f_\Delta(h_{G_{t-1}}, h_{O_t}), \\ h_t &= \text{RNN}(\Delta g_t, h_{t-1}), \\ G_t &= \text{MLP}(h_t). \end{aligned} \tag{4.3}$$

The function f_Δ aggregates the information from the previous mental state G_{t-1} and observation O_t to generate the graph update Δg_t. $h_{G_{t-1}}$ denotes the representation of G_{t-1} from the graph encoder. h_{O_t} is the output of the text encoder. h_t is a hidden state acting as the memory, from which we decode the new mental state G_t using a MLP. For the recurrent operator, we could either use

LSTM [HS97] or GRU [CVB14]. More details on the graph encoder and text encoder we applied are presented in the section 4.4.2.

4.4.2 Action Selector (steps ④-⑪)

Conditioned on the agent's mental state, the action selector chooses the optimal candidate based on the prediction task (*i.e.*, utterance, action or emotion). The selector consists of five components: a graph encoder (Fig. 4.3④) to convert the state-of-mind graph to a hidden state vector; a text encoder (Fig. 4.3(⑤, ⑥)) to encode the dialogue history and text candidates; an aggregator (Fig. 4.3⑦) to fuse the two context representations; a general scorer (Fig. 4.3⑧) to assign a score to each candidate; and a value model (Fig. 4.3⑩) to re-rank the candidates based on the assigned persona.

1. Graph Encoder. We use relational graph convolutional networks (R-GCNs) [SKB18] to encode the graph representation of mental states. The R-GCN is adapted from Graph Convolutional Networks (GCNs) so that it could embed the edge attributes (relational text embedding) in the mental state graph.

2. Text Encoder. A BERT-based [DCL19] encoder converts the text-based dialogue history into a vector representation, using the last hidden state corresponding to the `[CLS]` token; We also use the same encoder to encode the text response candidates.

3. Aggregator. A bi-directional attention layer [YDL18, SKF16] is adopted to fuse the information from the mental state and the contextualized text hidden state. The co-attention allows the agent to focus on the memory part that has been mentioned in the dialogue.

4. Scorer. The full context representation vector is concatenated with each candidate, and an MLP layer with softmax activation generates a score for each of them.

5. Value Ranker. After all the candidates are ranked, we select the top three candidates and then re-rank them according to the proposed value model. The value model is a BERT-based utility scorer trained on a self-curated knowledge base VALUENET. A custom input format constructed as

$\langle$`[CLS][$VALUE]s`$\rangle$ is fed into the BERT, *i.e.,*

$$f_v(\texttt{s}) = \text{BERT}(\texttt{[CLS][\$VALUE]s}), \tag{4.4}$$

where `[CLS]` is the special token for regression, `s` is the scenario, and `[$VALUE]` are special tokens we define to prompt [LL21, BMR20] the transformer the interested value dimension v. A regression head is put on top of the model to get a continuous estimation of the utility in the range of $[-1, 1]$.

The VALUENET is organized in 10 dimensions of Schwartz values. It consists of social scenarios curated from SOCIAL-CHEM-101 [FHS20]. And the samples are annotated by Amazon Mechanical Turk workers, who are asked about their attitudes towards provided scenarios. For example, if you are someone who values *benevolence*, will you do or say: "today I buried and mourned a rat"? Their choices (yes, no, unrelated) are then quantified to numerical utilities: +1, -1, 0, respectively.

4.5 Experiments

We conduct experiments on the LIGHT dataset and compare our model with state-of-the-art methods based on two variants of BERT models. An ablation study is carried out to justify our model design, and a case study is performed to demonstrate how the proposed framework could help the agent ground upon the environment details and make value-driven decisions.

4.5.1 Experimental Setup and Implementation

The dialogues in LIGHT are split into *train* (8539), *valid* (500), *seen test* (1000), and *unseen test* (739) as the dataset is released. The *unseen test* set consists of dialogues collected on a set of scenarios that have not appeared in the training data. We use the history of dialogues, actions, and emotions to predict the agent's next turn. Note that the original paper manually filters out actions with no affordance leveraging the object annotation, while we provide all candidates to demonstrate our model's capability of reasoning feasible actions automatically from the agent's mental state.

Here we describe the implementation details of the proposed framework. The mental state graph is initialized with a structured setting string including all involved elements in the scenario (an example is attached in Appendix 4.6). The setting parser is based on general parsing tools like regular expression and spaCy [HM17, CM16, HJ15], resulting in the initial mental state graph as shown in Figure 4.6. For the functions f_e and f_d, we use two-layer MLPs with

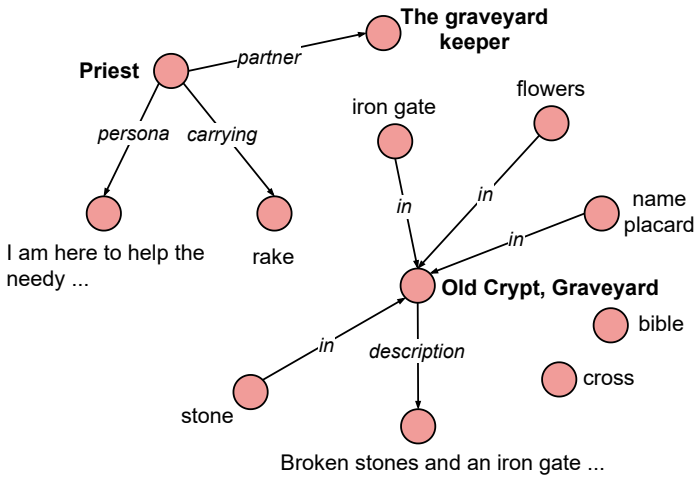

Figure 4.6: Initial mental state graph parsed from the example setting string in Appendix 4.6. The nodes of objects' descriptions are omitted to save space.

tanh [KO11] and ReLU [Aga18] activations. The **Text Encoder** is a pre-trained BERT (base-uncased) model [WDS20]. The **Graph Encoder** is an R-GCN with six layers and a hidden size of 64. We also adopt the highway connections between consecutive layers for faster convergence and 3-basis decomposition to reduce the parameters and prevent overfitting.

4.5.2 Baseline Models

Two BERT-based models [UFK19] are used as strong baselines, which have kept the state-of-the-art performance on this task. **BERT Bi-Ranker** produces a vector representation for the context and each candidate. Each candidate is assigned a score by the dot product between the context

embedding and the candidate embedding. **BERT Cross-Ranker** concatenates the context string with each candidate and feeds the string to the BERT model instead. Compared with the bi-ranker, The cross-ranker allows the model to attend to the context when encoding each candidate.

4.5.3 Results and Analysis

	Seen Test			*Unseen Test*		
	Dialogue	Action	Emotion	Dialogue	Action	Emotion
Method	R@1/20	Acc	Acc	R@1/20	Acc	Acc
BERT-based Bi-Ranker	76.5	42.5	25.0	70.5	38.8	25.7
BERT-based Cross-Ranker	74.9	50.7	25.8	69.7	51.8	28.6
discrete mental state	75.8	52.1	25.1	69.9	53.4	25.5
continuous mental state	77.3	49.3	**26.2**	72.1	45.2	29.1
hybrid mental state	78.4	53.5	26.1	72.3	54.3	29.5
hybrid+mask	78.5	54.5	26.1	72.3	55.4	29.4
hybrid+mask+value	**78.8**	**56.4**	26.1	**72.6**	**57.5**	**30.1**
Human Performance*	87.5	62.0	27.0	91.8	71.9	34.4

Table 4.1: Model performance on the LIGHT *Seen Test* and *Unseen Test*. For dialogue prediction, Recall@1/20 is reported for ranking the ground truth among 19 other randomly chosen candidates. Percentage accuracy is calculated for action and emotion prediction. (*) Human performance is reported by the original paper [UFK19] on a subset of data.

Table 4.1 shows the results, where our model outperforms the state-of-the-art models by a large margin. To understand the results, we first compare mental state graph designs using discrete, continuous, and the proposed hybrid parser.

The discrete mental state parser uses actions to explicitly update the graph to augment the context representation. In the action prediction task, the discrete parser outperforms the purely continuous method (+2.8% (seen), +8.2% (unseen)), the BERT Bi-Ranker (+9.6% (seen), +14.6%

(unseen)), and the BERT Cross-Ranker (+1.4% (seen), +1.6% (unseen)). While the continuous mental state parser misses the hard constraints introduced by less frequent actions, it updates the graph implicitly with the dialogues and shows a better result than the discrete one on dialogue prediction (+1.5% (seen), +2.2% (unseen)) and emotion prediction (+1.1% (seen), +3.6% (unseen)).

The hybrid mental state parser performs the best among the three according to almost all metrics, mainly because it aggregates the soft update from the dense dialogue and the hard constraints from the sparse actions. We also notice that the emotion prediction in LIGHT is a hard task because it is not strictly constrained by the context. Even humans can only achieve 27.0% (seen) and 34.4% (unseen) accuracy. Nevertheless, our model provides a relatively 1.2% (seen) and 3.1% (unseen) performance boost compared to the best BERT baseline.

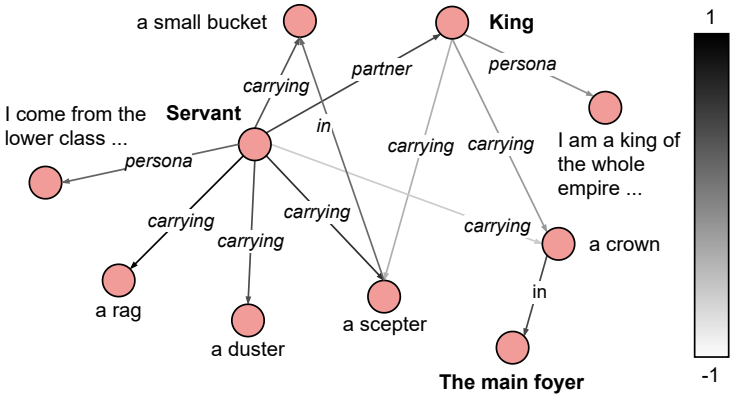

Figure 4.7: Intermediate mental state for the agent **Servant** in the dialogue example of Figure 4.3. The adjacency matrix of the mental state graph is visualized and the darkness of the edges represent the relation strength. Only critical relation types between nodes are shown for illustration purpose.

Then, with the ablation study of our proposed action mask (hybrid mental state *vs.* hybrid+mask), we prove the effectiveness of it for improving action accuracy by $\sim$1% in action prediction. Figure 4.7 demonstrates how the mental state could help agent ground on the context. We can see a very weak relation of the type "*carrying*" between the agent servant and the object crown. Thus the servant should not be able to give the crown to others at this time step. Though our model does not

rely on annotated action affordances during action predicting, an action mask can be reasoned from such a mental state, which helps filter out physical or causally impossible actions.

Lastly, we analyze the results after introducing the value model. We first compute the value priority of the agent by applying the value function to its persona description. For example, given the servant's persona description in Figure 4.3, it shows *conformity, tradition*, and *security* have higher utility scores to the agent than other dimensions. Then we calculate utility scores of the top three candidates based on Equation 4.1. This teaches the agent to make decisions that align with the assigned role and further improves the overall performance, (+0.3% (seen), +0.3% (unseen)) for dialogue prediction, (+1.9% (seen), +2.1% (unseen)) for action prediction, and +0.7% (unseen) for emotion prediction.

4.6 Appendix

An example setting string for the utterance prediction is:

"**_task_speech**

_setting_name Old Crypt, Graveyard

_setting_desc Broken stones and a iron gate closing the entrance with a name placard that the name is worn off.

_partner_name the graveyard keeper who lives across the yard **_self_name** priest

_self_persona I am here to help the needy. I am well respected in the town. I can not accept lying.

_object_desc a gate : The gate is made out of rusty metal. It squeaks as it swings on its hinges.

_object_desc a flowers : you can see them up close but not afar. when noticed, you realize that they are old.

_object_desc a name placard : The placard is made of wood witha clear name on it.

_object_desc a stone : The stone is chipped from being used as target practice from soldier trainees

_object_desc a placard : A sign used to display names of buildings or notices.

_object_desc an iron gate : The gate is ornate, with complicated iron scrollwork patterns.

_object_desc a Rake : This rake is made of carefully split wood with a sturdy looking handle. Seems useful for keeping the leaves under control.

_object_desc a Cross : The cross is broken and with a few dents in the sides.

_object_desc a bible : The bible is bound by black leather, its pages yellowed by years of use.

_object_in_room a gate

_object_in_room a flowers

_object_in_room a name placard

_object_in_room a stone

_object_in_room a placard

_object_in_room an iron gate

_object_carrying a Rake".

The result mental state graph parsed from this setting is illustrated in Figure 4.6.

Part II

Structure Learning in Dialogue Systems

CHAPTER 5

Structured Attention for Unsupervised Dialogue Structure Induction

Inducing a meaningful structural representation from one or a set of dialogues is a crucial but challenging task in computational linguistics. Advancement made in this area is critical for dialogue system design and discourse analysis. It can also be extended to solve grammatical inference. In this work, we propose to incorporate structured attention layers into a Variational Recurrent Neural Network (VRNN) model with discrete latent states to learn dialogue structure in an unsupervised fashion. Compared to a vanilla VRNN, structured attention enables a model to focus on different parts of the source sentence embeddings while enforcing a structural inductive bias. Experiments show that on two-party dialogue datasets, VRNN with structured attention learns semantic structures that are similar to templates used to generate this dialogue corpus. While on multi-party dialogue datasets, our model learns an interactive structure demonstrating its capability of distinguishing speakers or addresses, automatically disentangling dialogues without explicit human annotation[1].

5.1 Introduction

Grammatical induction for capturing a structural representation of knowledge has been studied for some time [Hig10]. Given the achievement in related areas like learning *Hidden Markov* acoustic models in speech recognition [BBD86] and sentence dependency parsing in language understanding [Cov01], our work aims to explore a more sophisticated topic: learning structures in dialogues.

[1]The code is released at `https://github.com/Liang-Qiu/SVRNN-dialogues`.

Figure 5.1 shows the underlying semantic structure of conversations about bus information request from SimDial dataset [ZE18], with one example dialogue as shown in Table 5.1. Another interesting

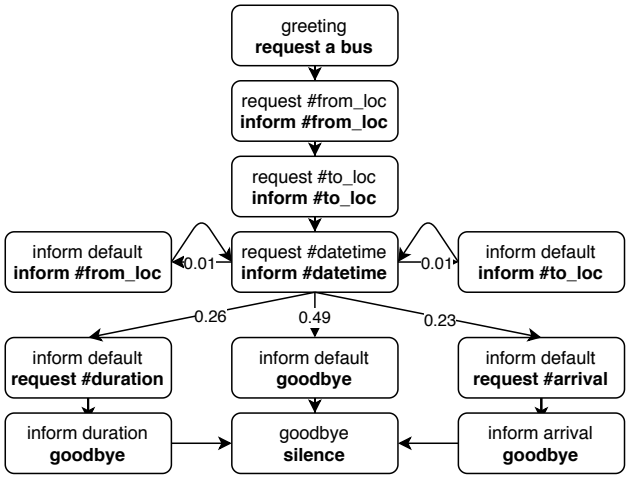

Figure 5.1: Original dialogue structure of the bus information request domain in SimDial [ZE18]. User intents are marked in **bold**.

type of dialogue structure is the interactive structure in multi-party dialogues. Figure 5.2 illustrates the interactive structure we learned from a dialogue sample in Ubuntu Chat Corpus [LPS15]. Each node represents an utterance from different speakers in the dialogue, with darker linkages representing stronger dependency relations between utterances. When speaker/addressee information is unavailable in the corpus, learning such a structure allows disentangling the conversation [SP15] and estimating the speaker labels. Discovering dialogue structures is crucial for various areas in computational linguistics, such as dialogue system building [You06], discourse analysis [GS86], and dialogue summarization [MRC05, LSZ10]. By looking into this topic, we can further improve the capability of machines to learn more generalized, interpretable knowledge representation from data.

However, capturing structure from the conversation is still much under-explored. The complexity of dialogues could range from several-round task-oriented dialogues to tens-round multi-party chitchat. It is unclear that for these different categories of dialogues, what types of inductive biases

From	Utterance
SYS:	Ask me about bus information. How can I help?
USR:	Hi. I need a bus.
SYS:	Where do you want to take off?
USR:	Going to Lawrance.
SYS:	What time do you need the bus?
USR:	Departure time is 9.
SYS:	Bus 137 can take you there. What else can I do?
USR:	Not done yet. How long will it take?
SYS:	The ride is 45 minutes long. What else can I do?
USR:	No more questions. Thank you.
SYS:	Goodbye.

Table 5.1: An example two-party bus information request dialogue in SimDial [ZE18].

or constraints we could add to reduce the search space. It also remains an unsolved question for formally evaluating the performance of dialogue structure induction algorithms. In this work, we propose to use a combination of structured attention and unsupervised generative model to infer the latent structure in a dialogue.

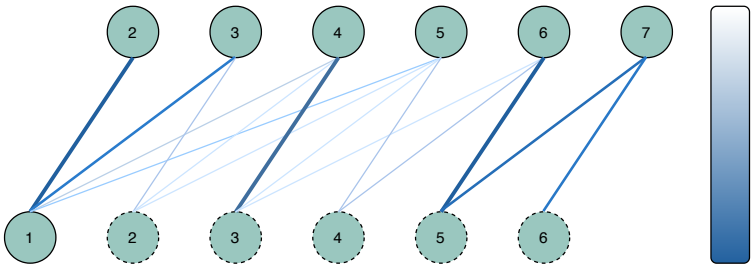

Figure 5.2: Learned interactive structure from a multi-party dialogue sample in Ubuntu Chat Corpus [UA13].

Specifically, instead of simply applying a softmax function on potentials between a decoder query and encoder hidden states, dynamic programming algorithms like *Forward-Backward* [Dev85]

and *Inside-Outside* [LY90] could be used to efficiently calculate marginal probabilities from pairwise potentials with a structural constraint. Through embedding such structured attention layers in a *Variational Recurrent Neural Network* (VRNN) model, we can learn latent structures in dialogues by jointly re-generating training dialogues. Such a process requires no human annotation and is useful for dialogue analysis. In addition, by selecting appropriate structural biases or constraints, we can learn not only semantic structures but also interactive structures. A linear *Conditional Random Field* (CRF) attention layer is used in two-party dialogues to discover semantic structures. A non-projective dependency tree attention layer is embedded to learn an interactive structure that could help identify speaker/addressee information in multi-party dialogues that have tangled conversation threads, such as forum discussions.

This work makes the following contributions. We propose to incorporate a structured attention layer in VRNN to learn latent structures in dialogues. To our knowledge, no work connecting structured attention with unsupervised dialogue structure learning has been done. We prove our proposed VRNN-LinearCRF learns better structures than the baseline VRNN on the SimDial dataset for semantic structure learning in two-party dialogues. For interactive structure learning in multi-party dialogues, we combine VRNN with a non-projective dependency tree attention layer. It achieves similar generation performance as the baseline GSN model [HCL19] on Ubuntu Chat Corpus [UA13, LPS15], while our model can identify the speaker/addressee information without trained on explicit labels.

5.2 Related Work

Attention mechanism [VSP17] has been widely adopted as a way for embedding categorical inference in neural networks for performance gain and interpretability [JW19, WP19]. However, for many tasks, we want to model richer structural dependencies without abandoning end-to-end training. *Structured Attention Networks* [KDH17] can extend attention beyond the standard soft-selection approach by attending to partial segments or subtrees. People have proven its effectiveness on a

variety of synthetic and real tasks: tree transduction, neural machine translation, question answering, and natural language inference [Rus20]. In this work, we propose to utilize structured attention to explore dialogue structures. Specifically, we work on two types of dialogue structures, semantic structures (dialogue intent transitions), and interactive structures (addressee/speaker changes).

Semantic structures have been studied extensively. Some previous works, such as [Jur97], learned semantic structures relying on human annotations, while such annotations are costly and can vary in quality. Other unsupervised studies used *Hidden Markov Model* (HMM) [Cho08, RCD10, ZW14]. Recently, *Variational Autoencoders* (VAEs) [KW13] and their recurrent version, *Variational Recurrent Neural Networks* (VRNNs) [CKD15], connects neural networks and traditional Bayes methods. Because VRNNs apply a point-wise non-linearity to the output at every timestamp, they are also more suitable to model highly non-linear dynamics over the simpler dynamic Bayesian network models. [SSL17] proposed the VHRED model by combining the idea of VRNNs and *Hierarchical Recurrent Encoder-Decoder* (HRED) [SBV15] for dialogue generation. Similarly, [ZLE18] proposed to use VAEs to learn discrete sentence representations. [SZY19] used two variants of VRNNs to learn the dialogue semantic structures and discussed how to use learned structure to improve reinforcement learning-based dialogue systems. But none of the previous work has tried to incorporate structured attention in VRNNs to learn dialogue structure.

Compared to semantic structures, the interactive structure of dialogues is not clearly defined. [EC08] initiated some work about dialogue disentanglement, which is defined as dividing a transcript into a set of distinct conversations. [SP15] tested standard RNN and its conditional variant for turn taking and speaker identification. Both of the tasks are highly related to understanding the interactive structure but not identical. Our task, different from both of them, aims to construct an utterance dependency tree to represent a multi-party dialogue's turn taking. The tree can not only be used to disentangle the conversations but also to label each utterance's speakers and addressees. We compare our model with *Graph Structured Network* (GSN), recently proposed by [HCL19]. GSN builds a conversation graph utilizing explicit speaker/addressee information in Ubuntu Chat Corpus [UA13] to improve the dialogue generation performance. Our model shows similar generation

performance as them while demonstrating its capability of learning the utterance dependency tree.

5.3 Problem Formulations

We discuss the semantic and interactive dialogue structure learning separately. In task-oriented two-party dialogues (between system and user), we want to discover a probabilistic semantic grammar shared by dialogues in the same domain. While for multi-party dialogues, *e.g.*, conversations in a chatroom, which may have multiple conversations occur simultaneously, we are more interested in finding an interactive structure that could help disentangle the conversation and identify the speakers/addressees. Our method of structure learning is flexible to handle both problems with the formulations as shown below.

For semantic dialogue structure learning, we formulate the problem as labeling the dialogue with a sequence of latent states. Each conversational exchange x_i (a pair of system and user utterances at time step i) belongs to a latent state z_i, which has an effect on the future latent states and the words the interlocutors produce. The latent dialogue state is defined to be discrete, *i.e.*, $z_i \in \{1, 2, ..., N\}$, where N is the number of states predefined from experience. Our goal is to generate the current sentence pair x_i that maximizes the conditional likelihood of x_i given the dialogue history while jointly learning a latent state sequence $\mathbf{z} = [z_1, z_2, ..., z_n]$:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \sum_{i=1}^{|\mathbf{x}|} \log(p(\mathbf{z}_{<i}|\mathbf{x}_{<i})p(x_i|\mathbf{z}_{<i})). \tag{5.1}$$

Then, we can induce a probabilistic dialogue grammar by estimating the state transition probabilities through maximizing the likelihood of the parsed latent state sequences.

A multi-party dialogue session can be formulated as an utterance-level dependency tree $\mathbf{T}(V, E)$, where V is the set of nodes encoding the utterances, $E = \{e_{i,j}\}_{i<j}^{m} \in \{0, 1\}$ indicates whether

utterance i is the parent of utterance j, and m is the maximum number of possible edges.

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\max_{\mathbf{x}} \sum_{i=1}^{|\mathbf{x}|} \log(p(\mathbf{T}|\mathbf{x}_{<i})p(x_i|\mathbf{T})) \\ &= \arg\max_{\mathbf{x}} \sum_{i=1}^{|\mathbf{x}|} \log(\prod_{j<k}^{i-1} p(e_{j,k}=1|\mathbf{x}_{<i}) \cdot p(x_i|\mathbf{T})) \\ &= \arg\max_{\mathbf{x}} \Big[\sum_{i=1}^{|\mathbf{x}|} \sum_{j<k}^{i-1} \log(p(e_{j,k}=1|\mathbf{x}_{<i}) + \sum_{i=1}^{|\mathbf{x}|} \log(p(x_i|\mathbf{T}) \Big]\end{aligned} \tag{5.2}$$

Each path of the dependency tree represents a thread in the multi-party conversation in chronological order. Our goal is to generate the response $\hat{\mathbf{x}}$ that maximizes the conditional likelihood of the response given the dialogue history while jointly learning a latent utterance dependency tree as shown in Equation 5.2. The conditional likelihood is factorized into two parts, representing the encoding and decoding processes, respectively. We can further reason about the speaker/addressee labels or disentangle the conversation by clustering the utterances from the learned tree.

5.4 Variational Recurrent Neural Network with Structured Attention

The overall architecture of Structured-Attention Variational Recurrent Neural Network (SVRNN) is illustrated in Figure 5.3. The LSTM [HS01] word-level encoder marked in pink encodes each utterance into a sentence embedding. Then an utterance-level encoder VRNN with different structured attention layers encodes the dialogue history into a latent state z. A decoder marked in blue will decode the next utterances from the latent state. We describe more details about the key components of our model in the following subsections.

5.4.1 Variational Recurrent Neural Network

The pursuit of using an autoencoder like *Variational Recurrent Neural Network* (VRNN) is to compress the essential information of the dialogue history into a lower-dimensional latent code. The latent code z is a random vector sampled from a prior $p(z)$ and the data generation model is

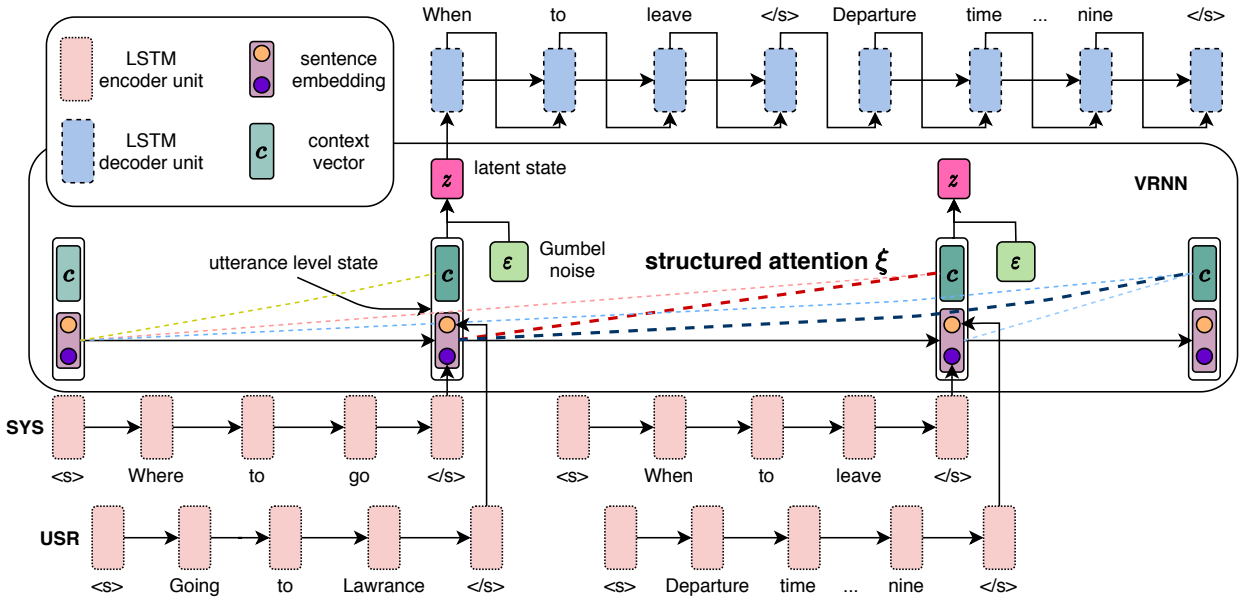

Figure 5.3: Structured-Attention Variational Recurrent Neural Network (SVRNN).

described by $p(x|z)$. The VRNN contains a *Variational Autoencoder* (VAE) at each time step. The VAE consists of an encoder $q_\lambda(z|x)$ for approximating the posterior $p(z|x)$, and a decoder $p_\theta(x|z)$ for representing the distribution $p(x|z)$. The variational inference attains its maximum likelihood by maximizing *evidence lower bound* (ELBO):

$$\mathbb{E}\left[\log p_\theta(x|z)\right] - \mathrm{KL}\left(q_\lambda(z|x)\|p(z)\right) \leq \log p(x). \tag{5.3}$$

For sequential data, the parameterization of the generative model is factorized by the posterior $p\left(z_t|x_{<t}, z_{<t}\right)$ and the generative model $p\left(x_t|z_{\leq t}, x_{<t}\right)$, *i.e.*,

$$p(x \leq T, z \leq T) = \prod_{t=1}^{T} \left[p\left(x_t|z_{\leq t}, x_{<t}\right) \cdot p\left(z_t|x_{<t}, z_{<t}\right)\right]. \tag{5.4}$$

The learning objective function becomes maximizing the ELBO for all time steps

$$\mathbb{E}\Big[\sum_{t=1}^{T}(\log p(x_t|z_{\leq t}, x_{<t}) - \mathrm{KL}\left(q(z_t|x_{\leq t}, z_{<t})\|p(z_t|x_{<t}, z_{<t})\right))\Big]. \tag{5.5}$$

In addition, to mitigate the *vanishing latent variable problem* in VAE, we incorporate Bag-of-Words (BOW) loss and Batch Prior Regularization (BPR) [ZZE17] with a tunable weight λ. By adjusting the λ, the VRNN based models can achieve a balance between clustering the utterance surface formats and attention on the context.

5.4.2 Linear CRF Attention

As we formulate the semantic structure learning in two-party dialogues as a state tagging problem, we find it suitable to use a linear-chain *Conditional Random Field* (CRF) attention layer with VRNN. Define $\boldsymbol{\xi}$ to be a random vector $\boldsymbol{\xi} = [\xi_1, ..., \xi_n]$ with $\xi_i \in \{0, 1\}$. n is the number of utterances in a dialogue. The context vector $\mathbf{c}_j$ given the current sentence hidden state $\mathbf{h}_j$ and hidden state history $\mathbf{h}$ can thus be written as:

$$\mathbf{c}_j = \sum_{i=1}^{j-1} p(\xi_i = 1|\mathbf{h}, \mathbf{h}_j)\mathbf{h}_i. \tag{5.6}$$

We model the distribution over the latent variable $\boldsymbol{\xi}$ with a linear-chain CRF with pairwise edges,

$$p(\xi_1, ..., \xi_n|\mathbf{h}, \mathbf{h}_j) = \text{softmax}(\sum_{i=1}^{j-2} \theta_{i,i+1}(\xi_i, \xi_{i+1})), \tag{5.7}$$

where $\theta_{i,i+1}(k, l)$ is the pairwise potential for $\xi_i = k$ and $\xi_{i+1} = l$. The attention layer is a two-state CRF where the unary potentials at the i-th dialogue turn are:

$$\theta_i(k) = \begin{cases} \mathbf{h}_i \mathbf{W}_1 \mathbf{h}_j, k = 0 \\ \mathbf{h}_i \mathbf{W}_2 \mathbf{h}_j, k = 1 \end{cases}, \tag{5.8}$$

where $[\mathbf{h}_1, ..., \mathbf{h}_n]$ are utterance level hidden states and $\mathbf{W}_1, \mathbf{W}_2$ are parameters. The pairwise potentials can be parameterized as

$$\theta_{i,i+1}(\xi_i, \xi_{i+1}) = \theta_i(\xi_i) + \theta_{i+1}(\xi_{i+1}) + \mathbf{h}_i^\top \mathbf{h}_{i+1}. \tag{5.9}$$

The marginal distribution $p(\xi_i = 1|x)$ can be calculated efficiently in linear-time for all i using message-passing, *i.e.*, the *forward-backward* shown in Algorithm 3.

Algorithm 3 Forward-Backward for LinearCRF Attention

Input: potential θ

$\alpha[0, \langle t \rangle] \leftarrow 0$

$\beta[n+1, \langle t \rangle] \leftarrow 0$

for $i = 1, ..., n; c \in \mathcal{C}$ **do**

$\alpha[i, c] \leftarrow \bigoplus_y \alpha[i-1, y] \otimes \theta_{i-1,i}[y, c]$

end for

for $i = n, ..., 1; c \in \mathcal{C}$ **do**

$\beta[i, c] \leftarrow \bigoplus_y \beta[i+1, y] \otimes \theta_{i,i+1}[c, y]$

end for

$A \leftarrow \alpha[n+1, \langle t \rangle]$

for $i = 1, ..., n; c \in \mathcal{C}$ **do**

$p(\xi_i = c|x) \leftarrow \exp(\alpha[i, c] \otimes \beta[i, c] \otimes -A)$

end for

return p

$\mathcal{C}$ denotes the state space and $\langle t \rangle$ is the special start/stop state. Typically the forward-backward with marginals is performed in the log-space semifield $\mathbb{R} \cup \{\pm\infty\}$ with binary operations $\oplus =$ logadd and $\otimes = +$ for numerical precision. These marginals allow us to calculate the context vector. Crucially, the process from vector softmax to *forward-backward* algorithm is a series of differentiable steps, and we can compute the gradient of the marginals with respect to the potentials [KDH17]. This allows the linear CRF attention layer to be trained end-to-end as a part of the VRNN.

5.4.3 Non-projective Dependency Tree Attention

For interactive structure learning in multi-party dialogues, we want to learn an utterance dependency tree from each dialogue. Therefore, we propose to use a non-projective dependency tree attention layer with VRNN for this purpose. The potentials $\theta_{i,j}$, which reflect the score of selecting the i-th sentence being the parent of the j-th sentence (*i.e.*, $x_i \rightarrow x_j$), can be calculated by

$$\theta_{i,j} = \tanh(\mathbf{s}^\top \tanh\left(\mathbf{W}_1 \mathbf{h}_i + \mathbf{W}_2 \mathbf{h}_j + \mathbf{b}\right)), \tag{5.10}$$

where $\mathbf{s}, \mathbf{b}, \mathbf{W}_1, \mathbf{W}_2$ are parameters, $\mathbf{h}_i, \mathbf{h}_j$ are sentence hidden states.

The probability of a parse tree ξ given the dialogue $x = [x_1, ..., x_n]$ is,

$$p(\xi|x) = \text{softmax}(\mathbb{1}\{\xi \text{ is valid}\} \cdot \sum_{i \neq j} \mathbb{1}\{\xi_{i,j} = 1\}\theta_{i,j}), \tag{5.11}$$

where the latent variable $\xi_{i,j} \in \{0, 1\}$ for all $i \neq j$ indicates that the i-th sentence is the parent of the j-th sentence; and $\mathbb{1}\{\xi \text{ is valid}\}$ is a special global constraint that rules out configurations of $\xi_{i,j}$'s that violate parsing constraints. In our case, we specify each sentence has one parent and that must precede the child sentence, i.e,

$$\sum_{i=1}^{n} \xi_{i,j} = 1, \xi_{i,j} = 0(i \geq j). \tag{5.12}$$

It is possible to calculate the marginal probability of each edge $p(\xi_{i,j} = 1|x)$ for all i, j in $O(n^3)$ time using the *inside-outside* algorithm with details explained in Appendix 5.6, which is a generalization of the *forward-backward* algorithm.

Then the soft-parent or the context vector of the j-th sentence is calculated using parsing marginals, *i.e.*,

$$\mathbf{c}_j = \sum_{i=1}^{n} p(\xi_{i,j} = 1|\mathbf{h}, \mathbf{h}_j)\mathbf{h}_i. \tag{5.13}$$

The original embedding is concatenated with its context vector to form the new representation

$$\hat{\mathbf{h}}_j = [\mathbf{h}_j; \mathbf{c}_j]. \tag{5.14}$$

5.4.4 Decoder

In order to generate a response to an utterance i, the decoder calculates a distribution over the vocabulary then sequentially predicts word w_k using a softmax function:

$$\begin{aligned}
p(\mathbf{w}|\hat{\mathbf{h}}) &= \prod_{k=1}^{|\mathbf{w}|} p(w_k|\hat{\mathbf{h}}, \mathbf{w}_{<k}) = \prod_{k=1}^{|\mathbf{w}|} \text{softmax}(\text{MLP}(\mathbf{h}_k^{dec}, \mathbf{c}_k^{dec})) \\
\mathbf{h}_0^{dec} &= \hat{\mathbf{h}}_i \\
\mathbf{h}_k^{dec} &= \text{LSTM}(\mathbf{h}_{k-1}^{dec}, \text{MLP}(\mathbf{e}_{w_{k-1}}; \mathbf{c}_{k-1}^{dec})) \\
\mathbf{c}_k^{dec} &= \sum_{j=1}^{i} \text{softmax}(\mathbf{h}_k^{dec}\mathbf{W}_a\hat{\mathbf{h}}_j)\hat{\mathbf{h}}_j,
\end{aligned} \tag{5.15}$$

where $\hat{\mathbf{h}}_i$ is the hidden state for utterance i with structured attention, $\mathbf{h}_k^{dec}$ is the hidden state of the decoder LSTM, $\mathbf{e}_{w_{k-1}}$ is the embedding of the predicted word at decoding time stamp $(k-1)$, and $\mathbf{c}_k^{dec}$ is the attention-based context vector at decoding time stamp k. Note that the context vector here is calculated with the simple attention different from the structured attention we described before. $\mathbf{W}_a$ is a matrix to learn the match degree of $\mathbf{h}_k^{dec}$ and $\hat{\mathbf{h}}_j$.

5.5 Experiments

We incorporate structured attention in VRNNs to explore two types of dialogue structure, semantic structure, and interactive structure.

5.5.1 Semantic Structure Learning in Two-party Dialogues

5.5.1.1 Datasets

We test the VRNN with Linear CRF Attention on the SimDial dataset [ZE18] of simulated conversations. Dialogues are generated for information requests in four domains: bus, restaurant, weather, and movie. Table 5.1 shows an example dialogue in bus schedule request domain. Although significant variations exist between dialogues of the same domain, we aim to explore a shared semantic structure among each dialogue domain. We validate our algorithm on this simulated dataset because these dialogues are generated using pre-defined templates that make recovering ground truth structures much easier. One recovered ground truth structure with transition probabilities is shown in Figure 5.1. We have 800 dialogue samples for training, 100 for validation, and 100 for testing in each dialog domain. The length of the dialogues ranges from 6 to 13 utterances. The maximum length of an utterance is 33 words.

5.5.1.2 Evaluation Metrics

Since the number of states is unknown during unsupervised training, we set the state number empirically to 10. Then the learned structure is essentially a state transition matrix of size 10×10. However, the original structure could be another state transition matrix of any size depending on the domain complexity. This makes the model evaluation on the ground truth a problem because it requires us to measure the difference between two state transition matrices of different sizes. To alleviate this problem, we define two metrics: Structure Euclidean Distance (SED) and Structure Cross-Entropy (SCE). We first estimate a probabilistic mapping P_{s_i,s'_i} between the learned states $\{s'_i, i = 1, 2, ..., M\}$ and the true states $\{s_i, i = 1, 2, ..., N\}$, through dividing the number of utterances that have the ground truth state s_i and learned state s'_i by number of utterances with the ground truth state s_i. And we let the reversed mapping probability $P_{s'_i,s_i}$ be the normalized

transpose of P_{s_i,s'_i}. Then SED and SCE are defined as:

$$
\begin{aligned}
T'_{s_a,s_b} &= \sum_{i,j\in\{1,2,...,M\}} P_{s_a,s'_i} \cdot T_{s'_i,s'_j} \cdot P_{s'_j,s_b} \\
\text{SED} &= \frac{1}{N}\sqrt{\sum_{a,b\in\{1,2,...,N\}} (T'_{s_a,s_b} - T_{s_a s_b})^2} \\
\text{SCE} &= \frac{1}{N} \sum_{a,b\in\{1,2,...,N\}} -\log(T'_{s_a,s_b}) T_{s_a s_b},
\end{aligned}
\tag{5.16}
$$

where T'_{s_a,s_b} is the learned transition probability from state s_a to state s_b and T_{s_a,s_b} is the true transition probability.

5.5.1.3 Results and Analysis

We compare the proposed VRNN-LinearCRF against other unsupervised methods: K-means clustering, Hidden Markov Model, D-VRNN [SZY19], and VRNN with vanilla attention. D-VRNN is similar to our work but without structured attention. We use a bidirectional LSTM with 300 hidden units as the sentence encoder and a forward LSTM for decoding. 300-dimensional word embeddings are initialized with GloVe word embedding [PSM14]. A dropout rate of 0.5 is adopted during training. We set the BOW-loss weight λ to be 0.5. The whole network is trained with the Adam [LH17] optimizer with a learning rate of 0.001 on GTX Titan X GPUs for 60 epochs. The training takes on average 11.2 hours to finish.

To evaluate the learned structure, we compare VRNN-LinearCRF's output in Figure 5.4 with the ground truth dialogue structure in Figure 5.1. A dialogue structure learned by VRNN without structured attention is also shown in the Appendix 5.6. We find our method generates a similar structure compared to ground truth in the bus domain. Figure 5.5 shows all models' quantitative results. Having a lower value in SED and SCE indicates the learned structure is closer to the ground truth and better. Our method with BERT, VRNN-LinearCRF-BERT performs the best. K-means clustering performs worse than VRNN-based models because it only considers utterances' surface format and ignores the context information. Hidden Markov Model is similar to VRNN but lacks a

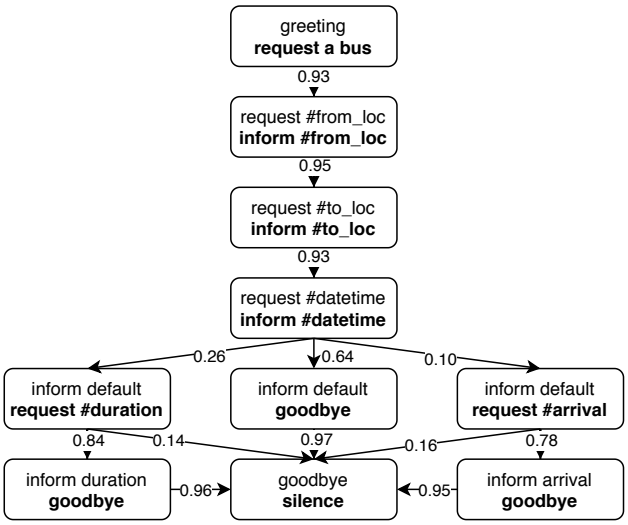

Figure 5.4: Learned semantic structure of SimDial bus domain [ZE18]. User intents are marked in **bold**. Transitions with $P < 0.1$ are omitted.

continuous propagating hidden state layer. VRNN-LinearCRF observes the entire history of latent states but ignores the redundant transitions due to the structure attention. The model's performance further improves when replacing the vanilla LSTM encoder with a large-scale pre-trained encoder like BERT [DCL19], as BERT provides better representations.

5.5.2 Interactive Structure Learning in Multi-party Dialogues

We extend our method to learn interactive structure in multi-party dialogues. Specifically, we detect each utterance's speaker and addressee by constructing an utterance dependency tree.

5.5.2.1 Datasets

We use Ubuntu Chat Corpus [UA13] as the dataset to study interactive structure since it provides the ground truth of speaker/addressee information for evaluation. Though every record of Ubuntu Chat Corpus contains a clear speaker ID, only part of the data has an implicit addressee ID, coming as

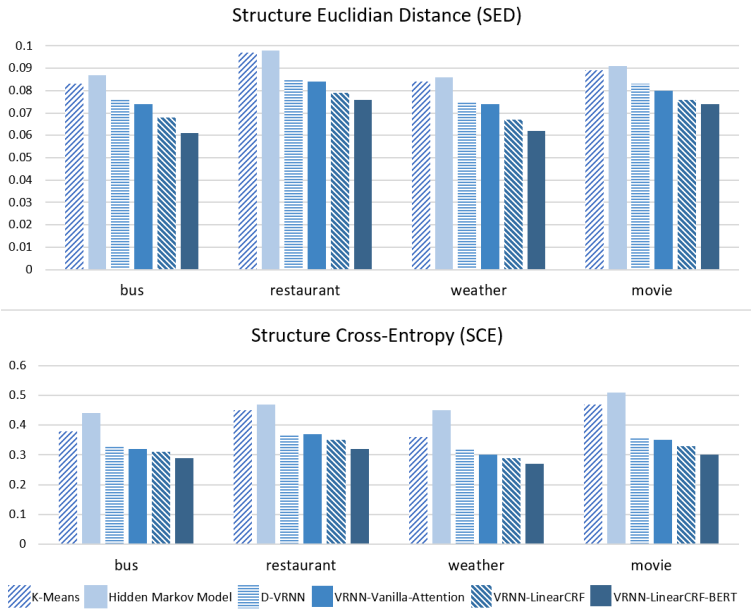

Figure 5.5: All models' performance in (a) Structure Euclidean Distance (SED) and (b) Structure Cross-Entropy (SCE) in four dialogue domains.

the first word in the utterance. We select addressee ID that appeared in a limited context and extract dialogue sessions with all utterances having verified speaker ID and addressee ID. We extract 20k dialogues with lengths ranging from 7 to 8 turns. Table 5.2 shows an example dialogue.

5.5.2.2 Results and Analysis

Considering Ubuntu Chat Corpus have a large number of technical terminologies, we use a relatively larger vocabulary size of 30k. We use LSTMs and BERT as the sentence embedding encoder and two GRU [CGC14] layers with 300 hidden units each as the decoder. The model converges after 100 epochs on GTX Titan X GPUs. The training procedure takes about 54 hours.

To evaluate the learned utterance dependency tree, we compare it with the annotated speaker-addressee relation and find 68.5% utterances are assigned the correct parents. This is a reasonable number because the dependency relationship does not fully rely on the speaker/addressee informa-

From	To	Utterance
p_1	p_2	I know upgrading always got hardon settings to new system..
p_3	–	And the description of the settings is even wrong
p_1	p_2	So these days i always clean install
p_2	p_1	Yeah, i think i will end up doing it
p_2	p_1	Do you happen to know if 12.10 install will let me install grub2 to partition instead of mbr without any extra tweaks?
p_1	p_2	I think default clean install will install grub2 on first section of your hd
p_4	p_2	No

Table 5.2: Multi-party dialogue example in Ubuntu Chat Corpus [UA13].

tion in a chatroom. A different interlocutor could answer others' questions even when the questions were not addressed to him/her. Figure 5.2 visualizes the learned interactive structure from the example in Table 5.2. Specifically, utterance 4 largely depends on utterance 3, while utterances 6 and 7 answer the question from utterance 5.

Model	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L
HRED	10.54	4.63	2.67	1.53	4.22	10.14
GSN No-speaker (1-iter)	9.23	3.32	1.89	1.24	3.57	8.12
GSN No-speaker (2-iter)	11.32	4.89	2.94	1.54	4.12	10.15
GSN No-speaker (3-iter)	11.42	4.81	3.11	1.87	4.51	10.29
GSN W-speaker (1-iter)	10.11	3.75	1.93	1.31	3.56	9.89
GSN W-speaker (2-iter)	11.43	4.90	2.99	1.63	4.32	10.34
GSN W-speaker (3-iter)	**11.52**	**4.93**	3.23	1.91	**4.77**	**11.21**
VRNN-Dependency-Tree	11.23	4.92	**3.24**	**1.92**	4.69	10.88

Table 5.3: Different methods' experiment results on Ubuntu dataset.

We also compare the model's generation performance with *Hierarchical Recurrent Encoder-Decoder* (HRED) and *Graph-Structured Network* (GSN) [HCL19]. The GSN model uses the annotated speaker/addressee information to construct a dialogue graph for utterance encoding

iteration. However, this is not required by our VRNN-Dependency-Tree since we generate the original dialogues while learning a dependency structure. For consistent comparison with previous work, we evaluate all models with BLEU 1 to 4, METEOR, and ROUGE_L with the package in [CFL15]. All results are shown in Table 5.3. We observe that the proposed VRNN-Dependency-Tree model without using any speaker annotation achieves similar generation performance compared to the state-of-the-art method, GSN with speaker annotation.

5.6 Appendix

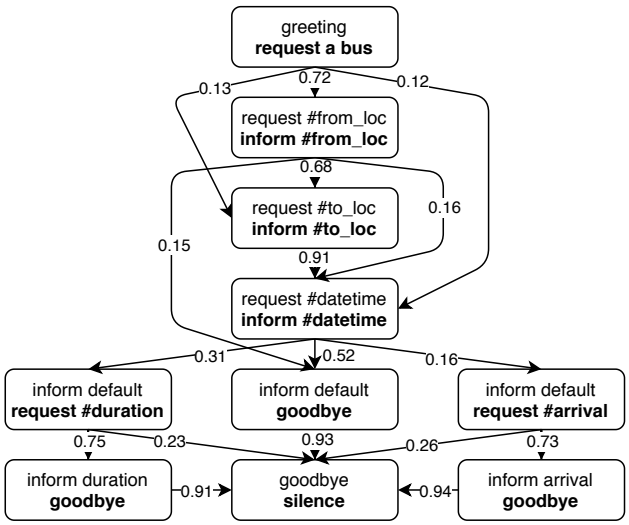

Figure 5.6: Learned dialogue structure from VRNN without structured attention in SimDial bus domain.

Algorithm 4 Inside-Outside for Non-projective Dependency Tree Attention

Input: potential θ_{ij}

$\alpha, \beta \leftarrow -\infty$

for $i = 1, \ldots, n$ **do**

  $\alpha[i, i, L, 1] \leftarrow 0$

  $\alpha[i, i, R, 1] \leftarrow 0$

end for

$\beta[1, n, R, 1] \leftarrow 0$

for $k = 1, \ldots, n$ **do**

  for $s = 1, \ldots, n - k$ **do**

    $t \leftarrow s + k$

    $\alpha[s, t, R, 0] \leftarrow \bigoplus_{u \in [s, t-1]} \alpha[s, u, R, 1] \otimes \alpha[u + 1, t, L, 1] \otimes \theta_{st}$

    $\alpha[s, t, L, 0] \leftarrow \bigoplus_{u \in [s, t-1]} \alpha[s, u, R, 1] \otimes \alpha[u + 1, t, L, 1] \otimes \theta_{ts}$

    $\alpha[s, t, R, 1] \leftarrow \bigoplus_{u \in [s+1, t]} \alpha[s, u, R, 0] \otimes \alpha[u, t, R, 1]$

    $\alpha[s, t, L, 1] \leftarrow \bigoplus_{u \in [s, t-1]} \alpha[s, u, L, 1] \otimes \alpha[u, t, L, 0]$

  end for

end for

for $k = n, \ldots, 1$ **do**

  for $s = 1, \ldots, n - k$ **do**

    $t \leftarrow s + k$

    for $u = s + 1, \ldots, t$ **do**

      $\beta[s, u, R, 0] \leftarrow_{\oplus} \beta[s, t, R, 1] \otimes \alpha[u, t, R, 1]$

      $\beta[u, t, R, 1] \leftarrow_{\oplus} \beta[s, t, R, 1] \otimes \alpha[s, u, R, 0]$

    end for

    if $s > 1$ **then**

      for $u = s, \ldots, t - 1$ **do**

        $\beta[s, u, L, 1] \leftarrow_{\oplus} \beta[s, t, L, 1] \otimes \alpha[u, t, L, 0]$

        $\beta[u, t, L, 0] \leftarrow_{\oplus} \beta[s, t, L, 1] \otimes \alpha[s, u, L, 1]$

      end for

    end if

    for $u = s, \ldots, t - 1$ **do**

      $\beta[s, u, R, 1] \leftarrow_{\oplus} \beta[s, t, R, 0] \otimes \alpha[u + 1, t, L, 1] \otimes \theta_{st}$

      $\beta[u + 1, t, L, 1] \leftarrow_{\oplus} \beta[s, t, R, 0] \otimes \alpha[s, u, R, 1] \otimes \theta_{st}$

    end for

    if $s > 1$ **then**

      for $u = s, \ldots, t - 1$ **do**

        $\beta[s, u, R, 1] \leftarrow_{\oplus} \beta[s, t, L, 0] \otimes \alpha[u + 1, t, L, 1] \otimes \theta_{ts}$

        $\beta[u + 1, t, L, 1] \leftarrow_{\oplus} \beta[s, t, L, 0] \otimes \alpha[s, u, R, 1] \otimes \theta_{ts}$

      end for

    end if

  end for

end for

$A \leftarrow \alpha[1, n, R, 1]$

for $s = 1, \ldots, n - 1$ **do**

  for $t = s + 1, \ldots, n$ **do**

    $p[s, t] \leftarrow exp(\alpha[s, t, R, 0] \otimes \beta[s, t, R, 0] \otimes -A)$

    if $s > 1$ **then**

      $p[t, s] \leftarrow exp(\alpha[s, t, L, 0] \otimes \beta[s, t, L, 0] \otimes -A)$

    end if

  end for

end for

CHAPTER 6

Structure Extraction in Task-Oriented Dialogues with Slot Clustering

Extracting structure information from dialogue data can help us better understand user and system behaviors. In task-oriented dialogues, dialogue structure has often been considered as transition graphs among dialogue states. However, annotating dialogue states manually is expensive and time-consuming. In this chapter, we propose a simple yet effective approach for structure extraction in task-oriented dialogues. We first detect and cluster possible slot tokens with a pre-trained model to approximate dialogue ontology for a target domain. Then we track the status of each identified token group and derive a state transition structure. Empirical results show that our approach outperforms unsupervised baseline models by far in dialogue structure extraction. In addition, we show that data augmentation based on extracted structures enriches the surface formats of training data and can achieve a significant performance boost in dialogue response generation.

6.1 Introduction

There is a long trend of studying the semantic state transition in dialogue systems. For example, modeling dialogue states in a deep and continuous space has been shown to be beneficial in response generation task [SSL17, WVM17, WHS20]. While in a modular system [BFP17] that is more preferred in industry, dialogue states are explicitly defined as the status of a set of slots. The domain-specific slots are often manually designed, and their values are updated through the interaction with users, as shown in Table 6.1. Extracting structure information from dialogue data is an important

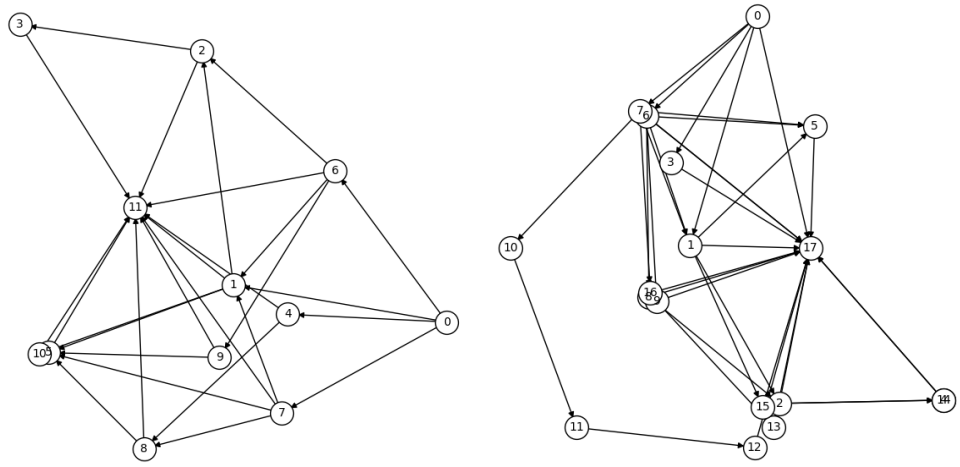

Figure 6.1: Dialogue structure in the *attraction* domain of the MultiWOZ [BWT18]. The structure on the left is from annotated dialogue states, while the right one is extracted by our approach. Structures for other domains are attached in Appendix 6.6.

topic for us to analyze user behavior and system performance. It also provides us with a discourse skeleton for data augmentation. Figure 6.1 shows an example of dialogue structure in the *attraction* domain of the MultiWOZ dataset [BWT18]. Each node represents a distinct dialogue state in Table 6.1, where the three dialogue turns correspond to node ⓪, ① and ② respectively. And the edges indicate transitions between pairs of states.

However, high-quality dialogue data with complete dialogue state annotation is limited. Existing works put more emphasis on unsupervised learning of dialogue structures. Representative ones include training language models based on Hidden Markov Models (HMMs) [Cho08] or Variational Recurrent Neural Networks (VRNNs) [SZY19, QZS20] to reconstruct the original dialogues. The structure built upon the latent states is then evaluated in downstream tasks like dialogue policy learning. Since the latent states are implicitly defined, there is a gap between the learned structure and the canonical dialogue states in task-oriented dialogues, making the structure hard to interpret and analyze. What's more, it remains unclear how we should choose the number of states during extraction. The state number directly decides the structure granularity, but it is not available in practice.

Dialogue	Dialogue State Slot Value
[usr] Can you please help me find a place to go? [sys] I've found 79 places for you to go. Do you have any specific ideas in your mind?	[0, 0, 0] → ⓪ [‘’, ‘’, ‘’]
[usr] I'd like a **sports** place in the **centre** please. [sys] There are no results matching your query. Can I try a different area or type?	[0, 1, 1] → ① [‘’, ‘sports’, ‘centre’]
[usr] Okay, are there any **cinemas** in the centre? [sys] We have vue cinema.	[0, 2, 1] → ② [‘’, ‘cinemas’, ‘centre’]

Table 6.1: Example dialogue in the *attraction* domain of the MultiWOZ [BWT18]. **Bold** tokens are detected by our algorithm as potential slots and used to update the dialogue state. The dialogue state vectors record how many times each slot is updated.

To alleviate these problems, we propose a simple yet effective approach for structure extraction in task-oriented dialogues. First, we define a task called Slot Boundary Detection (SBD). Utterances from training domains are tagged with the conventional BIO schema but without the slot names. A Transformer-based classifier is trained to detect the boundary of potential slot tokens in the test domain. Second, while the state number is usually unknown, it is more reasonable for us to assume the slot number can be estimated by checking just a few chat transcripts. We therefore cluster the detected tokens into groups with the same number of slots. Finally, the dialogue state is represented with a vector recording the modification times of every slot. We track the slot values through each dialogue session in the corpus and label utterances with their dialogue states accordingly. The semantic structure is portrayed by computing the transition frequencies among the unique states.

We evaluate our approach against baseline models that directly encode utterances or use rule-based slot detectors, besides the afore-mentioned latent variable model VRNN. Empirical results in the MultiWOZ dataset [BWT18] show that the proposed method outperforms the baselines by a large margin in all clustering metrics. By creating a state-utterance dictionary, we further

demonstrate how we could augment original data by following the extracted structure. The extra training data is coherent logically but creates more variety in surface formats, thus provides a significant performance boost for end-to-end response generation. The proposed Multi-Response Data Augmentation (MRDA) beats recent work [GLI21] using Most Frequent Sampling in a single-turn setting without annotated states.

6.2 Related Works

Extensive works have been done on studying the structures of dialogues. [Jur97] learned semantic structures based on human annotations. While such annotations are expensive and vary in quality, recent research shifted their focus to unsupervised approaches. By reconstructing the original dialogues with discrete latent variable models, we can extract a structure representing the transition among the variables. In this direction, people have tried Hidden Markov Models [Cho08, RCD10, ZW14], Variational Auto-Encoders (VAEs) [KW13], and its recurrent version Variational Recurrent Neural Networks (VRNNs) [CKD15, SZY19]. Based on VRNNs, [QZS20] tried to incorporate structured attention in VRNNs to inject structural inductive bias. More recently, [SST21] proposed an Edge-Enhanced Graph Auto-Encoder (EGAE) architecture to model local-contextual and global structural information. Meanwhile, [XLW21] integrates Graph Neural Networks into a Discrete Variational Auto-Encoder to discover structures in open-domain dialogues. However, since the latent variables are defined implicitly, it is hard to interpret or evaluate the extracted structure directly.

Benefiting from the pre-training technique [MBX17, HR18, PNI18, DCL19], the Transformer architecture [VSP17] can be trained on generic corpora and adapted to specific downstream tasks. In dialogue systems, [WHS20] pre-trained the BERT model [DCL19] on task-oriented dialogues for intent recognition, dialogue state tracking, dialogue act prediction, and response selection. [PLL21, HMW20] parameterize classical modular task-oriented dialogue system with an autoregressive language model GPT-2 [RWC19]. DIALOGPT [ZSG20] extends the GPT-2 to conversational

response generation in single-turn dialogue settings. In this work, we demonstrate how we can adapt a pre-trained Transformer for structure extraction in task-oriented dialogues. Our approach of detecting slot boundaries first is also related to the work of [HDY21] but uses a different model.

The extracted structures are proved useful in multiple downstream tasks. [XLW21] use the structure to guide coherent dialogue generation in open domains. [ABB20] synthesize a dataflow as task-oriented dialogue going on to improve representability and predictability. [SZY19] and [ZXE19] use the learned structures for dialogue policy learning. [GLI21] augment training data with the proposed Most Frequent Sampling (MFS) to improve the success rate of task-oriented dialog systems. [ZOY20] propose a Multi-Action Data Augmentation (MADA) framework guiding the dialog policy to learn a balanced action distribution. Nevertheless, both MFS and MADA are based on annotated dialogue states. Our work shows that the extracted structure can also be leveraged for data augmentation and alternative sampling strategies could be used.

6.3 Methodology

6.3.1 Problem Formulation

We aim to discover a probabilistic semantic structure shared by dialogues from the same domain. We formulate the problem as labeling each dialogue with a sequence of dialogue states. A structure is then extracted by calculating the transition frequencies between pairs of states. Each conversational exchange x_i, a pair of system and user utterances at time step i, corresponds to a dialogue state $\mathbf{z}_i$, which tracks the status of the task and guide the upcoming dialogue.

Commonly, dialogue states in task-oriented dialogue systems are defined as a set of slot-value pairs, which results in a huge amount of distinct states in total. To make the problem tractable, we count how many times each slot is modified without considering the actual slot values. Specifically,

$$\mathbf{z}_i = [M(S_1), M(S_2), ..., M(S_N)], \tag{6.1}$$

where S_j is a domain-specific slot, $M(S_j)$ is the number of changes of the slot S_j from the beginning

of the dialogue session, and N is the number of slots in the given domain. Although it is hard to determine the number of states because the value of each slot could be updated for infinite times, it is reasonable to assume that the number of slots is available during inference. For example, by checking a few transcripts, a bot builder for the MultiWOZ *attraction* domain can easily identify there are three slot types (name, type, area) that they need to fill in.

6.3.2 Slot Boundary Detection and Clustering

In a task-oriented dialogue system, slots are predefined in a domain ontology, and the system needs to identify their values to accomplish users' intents. For example, in order to book a taxi service, we need to fill the values of four slots: *"leave-at": 4 p.m., "arrive-by": 6 p.m., "departure": Palo Alto*, and *"destination": San Jose*. However, such a slot ontology is usually not available in a real scenario. We thus define a preliminary sub-task of slot boundary detection (SBD) and clustering for dialogue structure extraction. Given a target domain G, a set of dialogues D, and the number of slots N, the sub-task is to find all token spans that are possible slots in the domain G, and assign them into N separate slot groups $\{S'_1, S'_2, ..., S'_N\}$. As mentioned, we assume we do not have the slot ontology but N is available.

For the SBD task, we retain the slot annotation in the conventional BIO scheme but drop the slot name labels. Table 6.2 shows examples in three task-oriented dialogue datasets. We hypothesize that the capability to identify slot tokens is transferable across domains. We train a BERT-based slot detector on some domains and apply it to an unseen domain. A `[CLS]` classification embedding is inserted as the first token and a `[SEP]` token is added as the final token. Given an input token sequence $\mathbf{x} = (x_1, ..., x_T)$, the final hidden states of BERT ($\mathbf{h}_t$) is fed into a softmax layer to classify over three labels: "B", "I", and "O".

$$y_t = \text{softmax}(\mathbf{W}\mathbf{h}_t + \mathbf{b}). \tag{6.2}$$

To make the process compatible with the BERT WordPiece tokenizer [WSC16], we assign the

MultiWOZ	
Utt	[usr] a train to London King Cross that departs after 08:15
Slots	O O O O B I I O O O B
ATIS	
Utt	i want to fly from baltimore to dallas round trip
Slots	O O O O O B O B B I
Snips	
Utt	book a restaurant for eight people in six years
Slots	O O B O B O B I I

Table 6.2: Slot boundary annotation in the BIO scheme. Examples are from the MultiWOZ [BWT18], ATIS [THH10], and Snips [CSB18] datasets.

original label of a word to all its sub-tokens. This model is trained end-to-end to minimize with cross-entropy loss. For each token span $\mathbf{T}_i = [T_{i1}, ..., T_{ik}]$, if their slot labels predicted are $[\mathrm{B}, \mathrm{I}, ..., \mathrm{I}](k > 1)$ or $\mathrm{B}(k = 1)$, and the label of token T_{ik+1} is predicted as B or O, then $\mathbf{T}_i$ is considered as a slot token span.

The pre-trained BERT model provides a powerful contextualized token representation. Therefore, we reuse the final hidden states for slot clustering. Mathematically, the token span $\mathbf{T}_i$ is encoded as

$$\overline{\mathbf{h}}_i = \frac{1}{k} \sum_{j=1}^{k} \mathbf{h}_{ij}, \tag{6.3}$$

where $\mathbf{h}_{i1}, ..., \mathbf{h}_{ik}$ are the final hidden states of $\mathbf{T}_i = [T_{i1}, ..., T_{ik}]$. The BERT representations are contextualized, so the same token spans appearing in different contexts have different encodings. By doing so, one token span can be assigned to multiple slot clusters simultaneously. For example, "Palo Alto" can be both a departure city and an arrival city, depending on its context. By clustering the token span encodings, we can assign each of them into one of the N groups and derive a fake slot ontology.

$$S'_j = \text{clustering}(\overline{\mathbf{h}}_i), j \in \{1, ..., N\}, \tag{6.4}$$

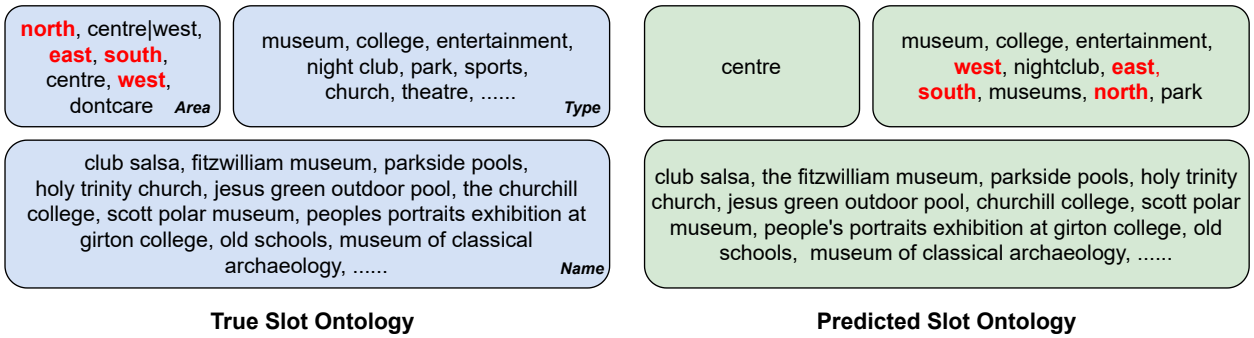

Figure 6.2: True slot ontology *v.s.* predicted slot ontology of the *attraction* domain in the MultiWOZ. Mis-clustered tokens are marked in **bold** and red. Slot names are unknown but it will not affect the structure extraction procedure.

where S'_j is the j-th predicted slot group. Three clustering algorithms including KMeans [AV06], Birch [ZRL96], and Agglomerative Clustering [Mul11] are evaluated. Note that there is no guarantee that S'_j can be mapped to any predefined slot type S_j. A clustered example is shown in Figure 6.2. More details about clustering are explained in section 6.4.

6.3.3 Deterministic Dialogue State Labeling

The slot boundary detection and clustering are followed by a deterministic procedure to construct the dialogue structure. To begin with, we initialize the dialogue state as $\mathbf{z}_0 = [0, 0, ..., 0]$. Then in dialogue turn k, for each slot token span $\mathbf{T}_i$ detected, if the clustering algorithm determines $\mathbf{T}_i \in S'_j$, we increment $M(S'_j)$ by one. Table 6.1 demonstrates this procedure. In this way, we label each dialogue session with its extracted dialogue states without any state annotation. The dialogue structure is then depicted by representing distinct dialogue states as nodes. Due to the variety of $M(S'_j)$, the number of dialogue states is always larger than the number of slots, as shown in Table 6.3. We connect an edge between a pair of nodes if there is such a transition in the data, and the edge is labeled as the normalized transition probability from the parent node.

Domain	Taxi	Restaurant	Hotel	Attraction	Train
#samples	435	1,311	635	135	345
#slots	4	7	10	3	6
#states	29	206	734	11	85

Table 6.3: Statistics of the MultiWOZ [BWT18] dataset. **#states** are number of annotated distinct dialogue states.

6.4 Experiment

6.4.1 Datasets

MultiWOZ [BWT18] is a common benchmark for task-oriented dialogues. It has 8,420/1,000/1,000 dialogues for train, validation, and test, respectively. We use its revised version MultiWOZ 2.1 [EGP20], which has the same dialogue transcripts but with cleaner state label annotation. Table 6.3 shows some statistics of the MultiWOZ dataset. The MultiWOZ has five domains of dialogues: *taxi, restaurant, hotel, attraction*, and *train*. We hold out each of the domain for testing and use the remaining four domains for SBD training. Note that some of the target slots are not presented in the training slots, *e.g.,* "*stay*", "*stars*", and "*internet*" only appear in the *hotel* domain.

To evaluate the transferability of the approach, we also tried to train the slot boundary detector on another two public datasets, ATIS [THH10, GGH18] and Snips [CSB18]. The ATIS dataset includes recordings of people making flight reservations and contains 4,478 utterances in its training set. The Snips dataset is collected from the Snips personal voice assistant and contains 13,084 training utterances. We train the SBD model on their training split and test on the selected domain of MultiWOZ. Examples are shown in Table 6.2.

6.4.2 Setup

We conduct extensive experiments to compare our approach with different baseline models. The ground truth construction follows the same deterministic procedure by counting the modification times of annotated slot values, instead of the spans predicted by our algorithm. We describe details of the baseline models as follows.

- **Random** Every conversational turn is randomly assigned a state by selecting a number from 1 to the ground truth #states in Table 6.3.

- **VRNN** Dialogues are reconstructed with Variational Recurrent Neural Networks [SZY19, QZS20], which is a recurrent version of Variational Auto-Encoder (VAE). The extracted structure represents the transition among discrete latent variables.

- **BERT-KMeans/Birch/Agg** Each conversational turn is encoded by BERT with the final hidden state of `[CLS]` token. The utterance encodings are then clustered with Kmeans, Birch, and Aggolomerative clustering methods.

$$\mathbf{z} = \text{clustering}(\mathbf{h}_{\text{CLS}}) \tag{6.5}$$

  Number of clusters are directly set to the #states.

- **TOD-BERT-mlm/jnt** This is similar to the previous baseline but encoding utterances with TOD-BERT. TOD-BERT [WHS20] is based on BERT architecture and trained on nine task-oriented datasets using two loss functions: masked language modeling (MLM) loss and response contrastive loss (RCL). TOD-BERT-mlm only uses the MLM loss, while TOD-BERT-jnt is jointly trained with both loss functions. The utterance encodings are clustered with KMeans.

- **(TOD-)BERT-spaCy** Instead of training a slot boundary detector based on BERT, we implement a heuristic-based detector with spaCy[1]. Words are labeled as slot spans if they are

[1] `https://spacy.io/`

nouns. Suppose it detects n slot words $\{w_1..., w_n\}$ in the u_i utterance, the j-th word has $|w_j|$ sub-tokens, the BERT/TOD-BERT encoding of the k-th sub-token of this word is $\mathbf{h}_{jk}$. Then we represent this turn as:

$$\mathbf{u_i} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{|w_j|}\sum_{k=1}^{|w_i|}\mathbf{h}_{jk}. \tag{6.6}$$

In this method, we do not cluster slot representations, but we use average slot embedding to represent the whole utterance. Then $\mathbf{u_i}$ are clustered to #states clusters with KMeans:

$$\mathbf{z_i} = \text{clustering}(\mathbf{u_i}). \tag{6.7}$$

- **TOD-BERT-SBD**$_{\text{MWOZ}}$ This is similar to the previous approach. But instead of using a heuristic-based detector, the TOD-BERT is trained for SBD in training domains of MultiWOZ and detect slot tokens in the test domain, and then we use those detected slot embeddings to represent each utterance.

- **TOD-BERT-DET**$_{\text{ATIS/SNIPS/MWOZ}}$ The TOD-BERT is trained for SBD in the ATIS, Snips, or the MultiWOZ training domains. Then in the test domain of MultiWOZ, we follow the deterministic dialogue state labeling process described in section 6.3.3, instead of clustering utterance embeddings, to extract a structure.

We use English uncased BERT-Base model, which has 12 layers, 12 heads, and 768 hidden states. We train BERT (or TOD-BERT) on the Slot Boundary Detection (SBD) task with AdamW [LH17] optimizer using a dropout rate of 0.1. The model is trained with an initial learning rate of $5\text{e}{-5}$ for 5 epochs on two NVIDIA Tesla V100 GPUs.

6.4.3 Results and Analysis

Table 6.4 shows the empirical results of Slot Boundary Detection. We report the $\mathbf{F}1$ score in both slot level ($\mathbf{F}1_{slot}$) and token level ($\mathbf{F}1_{token}$). In the slot level, a slot prediction is considered correct only when an exact match is found, which doesn't reward token overlap (partial match). In general,

	$\mathbf{F1}_{slot}$					$\mathbf{F1}_{token}$				
Method	Taxi	Restaurant	Hotel	Attraction	Train	Taxi	Restaurant	Hotel	Attraction	Train
spaCy	0.43	0.48	0.47	0.33	0.39	0.28	0.21	0.21	0.16	0.23
TOD-BERT$_{\text{ATIS}}$	0.57	0.56	0.52	0.45	0.62	0.57	0.54	0.44	0.43	0.60
TOD-BERT$_{\text{SNIPS}}$	0.50	0.53	0.48	0.41	0.52	0.55	0.49	0.41	0.37	0.51
TOD-BERT$_{\text{MWOZ}}$	**0.90**	**0.89**	**0.84**	**0.91**	**0.84**	**0.90**	**0.89**	**0.82**	**0.91**	**0.84**

Table 6.4: Slot boundary detection results tested in the MultiWOZ.

BERT-based slot boundary detectors perform better than the heuristic-based detector. Because utterances in MultiWOZ share similar interaction behaviors and utterance lengths, it makes the model easier to transfer from one domain to another within MultiWOZ than from the ATIS and Snips to the MultiWOZ.

	ARI					AMI					SC				
	Taxi	Rest.	Hotel	Attr.	Train	Taxi	Rest.	Hotel	Attr.	Train	Taxi	Rest.	Hotel	Attr.	Train
Random	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-	-	-	-
VRNN	0.05	0.00	0.00	0.00	0.00	0.05	0.02	0.00	0.01	0.06	-	-	-	-	-
BERT-KMeans	0.02	0.01	0.01	0.01	0.01	0.11	0.09	0.02	0.03	0.06	0.11	0.08	0.06	0.13	0.09
TOD-BERT-mlm	0.02	0.01	0.01	0.03	0.02	0.13	0.11	0.03	0.06	0.10	0.12	0.08	0.06	0.17	0.09
TOD-BERT-jnt	0.03	0.02	0.02	0.03	0.03	0.16	0.13	0.06	0.08	0.14	0.09	0.08	0.06	0.13	0.07
BERT-spaCy	0.01	0.06	0.04	0.01	0.01	0.09	0.18	0.12	0.06	0.08	-	-	-	-	-
TOD-BERT-spaCy	0.01	0.03	0.05	0.02	0.01	0.09	0.15	0.12	0.05	0.05	-	-	-	-	-
TOD-BERT-SBD$_{\text{MWOZ}}$	**0.15**	0.00	0.00	0.00	0.05	0.17	0.13	0.04	0.06	0.16	**0.39**	**0.34**	**0.27**	**0.44**	**0.34**
TOD-BERT-DET$_{\text{ATIS}}$	0.08	0.05	0.09	0.03	0.06	0.26	0.22	0.25	0.15	0.26	-	-	-	-	-
TOD-BERT-DET$_{\text{SNIPS}}$	0.06	0.05	0.11	0.03	0.04	0.25	0.23	0.22	0.09	0.22	-	-	-	-	-
TOD-BERT-DET$_{\text{MWOZ}}$	**0.15**	**0.22**	**0.24**	**0.33**	**0.24**	**0.39**	**0.48**	**0.44**	**0.44**	**0.44**	-	-	-	-	-

Table 6.5: Structure extraction results using clustering metrics in the MultiWOZ dataset. SC is omitted for methods that do not encode utterances directly. Results using BERT-Birch and BERT-Agg are reported in Appendix 6.6.

We further analyze the performance of structure extraction, as shown in Table 6.5. We evaluate the model performance with clustering metrics, testing whether utterances assigned to the same state are more similar than utterances of different states. Given the knowledge of the ground truth

dialogue state assignments and the model assignments of the same utterances, the Rand Index (RI) is a function that measures the similarity of the two assignments. Mathematically,

$$\begin{aligned} \textbf{RI} &= \frac{a+b}{C_2^{n_{\text{samples}}}}, \\ \textbf{ARI} &= \frac{\textbf{RI} - E[\textbf{RI}]}{max(\textbf{RI}) - E[\textbf{RI}]}, \end{aligned} \tag{6.8}$$

where a is the number of pairs of elements that are assigned to the same set by both the ground truth and the model, b is the number of pairs of elements that are assigned to different sets by both, $C_2^{n_{\text{samples}}}$ is the total number of pairs in the dataset. The Adjusted Rand Index (ARI) corrects for chance and guarantees that random assignments have an ARI close to 0. For a comprehensive analysis, we also report Adjusted Mutual Information (AMI) and Silhouette Coefficient (SC). While both ARI and AMI require the knowledge of the ground truth classes, the Silhouette Coefficient (SC) evaluates the model itself but needs utterance representations to compute the distance. Thus, we do not report SC for methods such as TOD-BERT-DET.

$$\textbf{SC} = \frac{b-a}{max(a,b)}, \tag{6.9}$$

where a is the mean distance between the sample and all other points in the same class, b is the mean distance between a sample and all other points in the next nearest cluster.

We observe a negligible effect of using different clustering algorithms on the structure extraction performance. As we can see in Table 6.5, the VRNN baseline performs not so well, because their dialogue states are defined in a latent space while the ground truth we compare with is based on the accumulative status of slots. Switching the encoder from the original BERT to TOD-BERT provides a slight improvement. Using a spaCy-based detector can have inaccurate slot detection, so the performance of (TOD-)BERT-spaCy are worse than TOD-BERT-SBD$_{\text{MWOZ}}$. Simply averaging the detected slot token encodings for utterance clustering will also lose the information of individual slot changes. Compared with these baselines, our approach TOD-BERT-DET$_{\text{MWOZ/ATIS/SNIPS}}$ based on slot boundary detection and deterministic dialogue state labeling outperforms others by a large margin. In Figure 6.3 and Figure 6.4, we show the robustness of the proposed TOD-BERT-DET$_{\text{MWOZ}}$

to an inaccurate estimation of #slots. In Appendix 6.6, we show example utterances that are predicted as the same state in different domains.

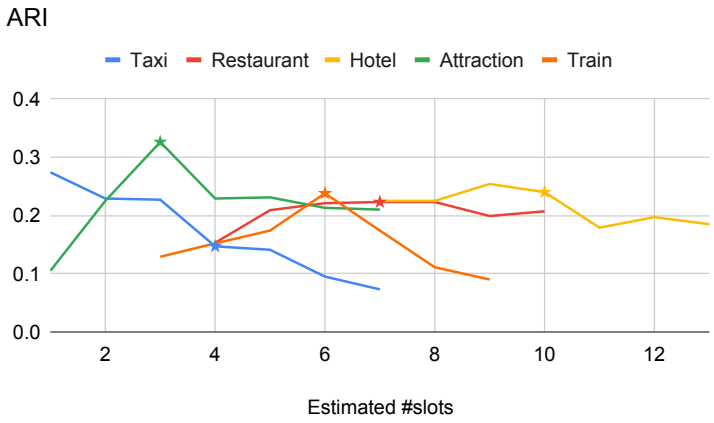

Figure 6.3: Evaluation of the proposed TOD-BERT-DET$_{\text{MWOZ}}$'s robustness to estimated #slots. Stars are the ground truth.

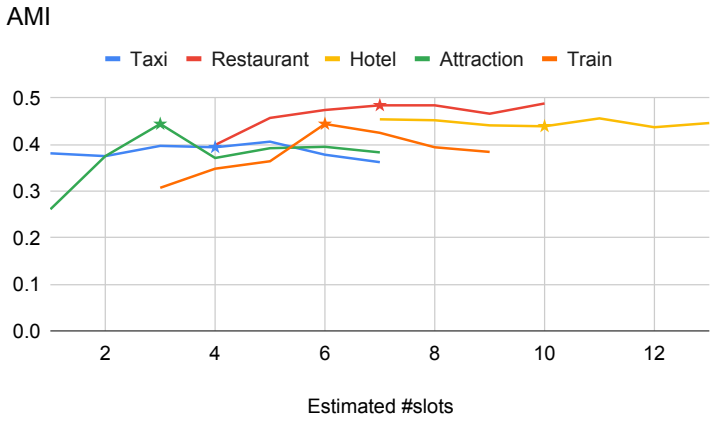

Figure 6.4: Evaluation of the proposed TOD-BERT-DET$_{\text{MWOZ}}$'s robustness to estimated #slots. Stars are the ground truth.

6.5 Data Augmentation

Conversations have an intrinsic one-to-many property, meaning that multiple responses can be appropriate for the same dialog context [ZOY20]. Leveraging this property, we augmented training data to improve end-to-end dialogue response generation based on the extracted structure. Specifically, we build a dictionary mapping from the dialogue state to its different valid utterances. Then we enable this dictionary to create additional data during training, which allows a language model to learn a balanced distribution. In the following sections, we will briefly introduce the task of single-turn dialogue response generation, the baseline augmentation approach called Most Frequent Sampling [GLI21], and the proposed Multi-Response Data Augmentation.

6.5.1 Single-Turn Dialogue Generation

Training a single-turn dialogue response generative model is to learn an autoregressive (AR) model that maximize the log-likelihood $\mathcal{L}$ of ground truth response $R = x_{n+1}, ..., x_T$ conditioned on dialogue history $C = x_1, ..., x_n$, which is encoded by dialogue state $\mathbf{z}$:

$$\begin{aligned} \mathcal{L} &= \sum_{i \in D} \log P(R_i | C_i) \\ &= \sum_{i \in D} \log \prod_{t=n+1}^{T} p(x_t | x_1, ..., x_{t-1}), \end{aligned} \tag{6.10}$$

where i is each turn in dialogue corpus D. For a number of dialogue history C_i belonging to the same state $\mathbf{z}$, there exits K different system responses $R^{(1)}, ..., R^{(K)}$ that are valid, *i.e.,* for $j = 1, ..., K, \exists i \in D$ *s.t.* $(\mathbf{z}_i, R_i) = (\mathbf{z}, R^{(j)})$. We denote the valid system response set for dialogue state $\mathbf{z}$ as $\mathcal{V}(\mathbf{z})$.

6.5.2 Most Frequent Sampling

[GLI21] proposed Most Frequent Sampling (MFS) as a data augmentation strategy based on the annotated conversational graph. MFS generates novel training instances so that the most frequent

agent actions are preceded by new histories, which is one or more original paths leading to common actions. The authors observed the best performance when they combined extra data augmented from MFS with the baseline training data.

6.5.3 Multi-Response Data Augmentation

However, augmented with the Most Frequent Sampling, it may exaggerate the frequency imbalance among valid responses, resulting in a lower response diversity. The original MFS also depends on annotated dialogue states from the MultiWOZ. To alleviate the problems, we propose Multi-Response Data Augmentation (MRDA) to balance the valid response distribution of each state $\mathbf{z}$ based on our extracted dialogue structure. Concretely, for each dialog turn i with state-response pair $(\mathbf{z}_i, R_i)$, we incorporate other valid system responses under the same state, *i.e.,* $R_{i'}, i' \neq i$ with $\mathbf{z}_{i'} = \mathbf{z}_i$, as additional training data for turn i. The new objective function becomes:

$$\mathcal{L}_{\text{aug}} = \sum_{i \in D} \sum_{R_{i'} \in \mathcal{V}^*(\mathbf{z}_i)} \log P(R_{i'}|C_i), \tag{6.11}$$

where $\mathcal{V}^*(\mathbf{z}_i) \subseteq \mathcal{V}(\mathbf{z}_i)$ is a subset of the valid response set $\mathcal{V}(\mathbf{z}_i)$ of dialogue state $\mathbf{z}_i$, $\mathbf{z}_i$ is the predicted dialogue state of history C_i. The idea is similar to the Multi-Action Data Augmentation (MADA) proposed by [ZOY20], but our method doesn't need to train an action decoder to generate responses and no annotated dialogue states are required.

6.5.4 Setup

We compare our MRDA approach with the MFS baseline in the MultiWOZ dataset. We used the ground truth dialogue states for MFS as in its original paper. For MRDA, we hold out each of the domains for testing and use the remaining four domains for SBD training and dialogue state prediction. The data of each held-out domain is split into train (60%), valid (20%), and test (20%), which are used for language model training and testing. To evaluate both methods in a realistic setting where training data is limited and augmentation is required, we adjust the ratio between actually used training data and total training data, denoted by r_{train}. Moreover, to explore the impact

Perplexity↓	Taxi	Rest.	Hotel	Attr.	Train
Original train	4.88	4.46	6.16	7.75	6.58
+ MFS	5.34	6.54	7.17	8.39	6.91
+ MRDA	**4.64**	**3.69**	**5.57**	**3.91**	**5.83**
BLEU↑	Taxi	Rest.	Hotel	Attr.	Train
Original train	9.88	12.54	8.68	8.46	8.93
+ MFS	9.73	12.56	7.54	9.21	7.64
+ MRDA	**22.13**	**18.77**	**10.66**	**47.79**	**9.20**

Table 6.6: Response generation in the MultiWOZ with data augmentation ($r_{\text{train}} = 1.0, r_{\text{aug}} = 1.0$).

of augmented data size, we define r_{aug} as the ratio between the size of augmented samples and used training samples. The DIALOGPT [ZSG20] model is trained with the data for 5 epochs with a learning rate of $3\text{e}{-5}$ to generate single-turn responses.

6.5.5 Results and Analysis

The generation perplexity and BLEU scores in the five domains of the MultiWOZ are reported in Table 6.6. Both augmentation methods first double the original training samples, *i.e.,* $r_{\text{train}} = 1.0, r_{\text{aug}} = 1.0$. By augmenting the data, we reduce the perplexity by an average of 1.24 and improve the BLEU score by an average of 12.01. The results also demonstrate our approach outperforms the MFS baseline by an average of 2.14 in perplexity and 12.37 in BLEU, because the MRDA balances the valid response distribution. Our approach also doesn't require any annotation of the test domain.

To explore the impact of available training data size and augmented data size, we try different combinations of the r_{train} and r_{aug}, and illustrate the results in Figure 6.5 and Figure 6.6 (numbers attached in Appendix 6.6). The figures show that: *(i)* Our MRDA approach constantly improves the generation performance in both metrics, and it outperforms the MFS baseline regardless

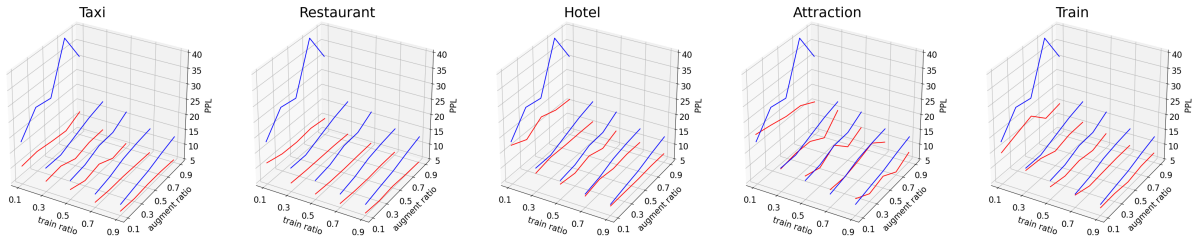

Figure 6.5: Data Augmentation (perplexity↓) in the MultiWOZ. Blue: MFS. Red: MRDA (ours).

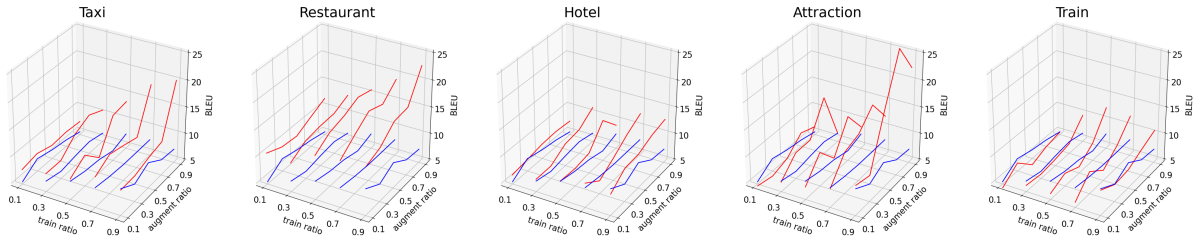

Figure 6.6: Data Augmentation (BLEU↑) in the MultiWOZ. Blue: MFS. Red: MRDA (ours).

of the original data size. *(ii)* Data augmentation based on a larger training set provides more performance boost because the language model is trained with more data and different valid responses are balanced. These observations suggest that our extracted dialogue structure can successfully augment meaningful dialogue for response generation, with the potential to improve other dialogue downstream tasks such as policy learning and summarization. We also include example augmented dialogues in the Appendix 6.6.

Table 6.7 reports how many states are overlapped in the MultiWOZ, using the slot value annotation and our dialogue state definition. It shows that our test set has no distinct dialogue state that never appears in the train or valid sets, while this may not be the case in practice. The MRDA method creates new instances that follow existing dialogue flows but with different surface formats, while it remains a compelling direction to create completely new state sequences by discovering causal dependencies in the extracted structures.

6.6 Appendix

State Overlap	Taxi	Rest.	Hotel	Attr.	Train
Train Only	9	97	387	0	25
Valid Only	0	12	51	0	7
Test Only	**0**	**0**	**0**	**0**	**0**
Train & Valid	18	68	148	10	38
Train & Test	16	55	99	9	32
Test & Valid	16	62	141	9	37
Train & Valid & Test	16	55	99	9	32

Table 6.7: Annotated dialogue state overlap across train, valid, and test splits in the MultiWOZ dataset.

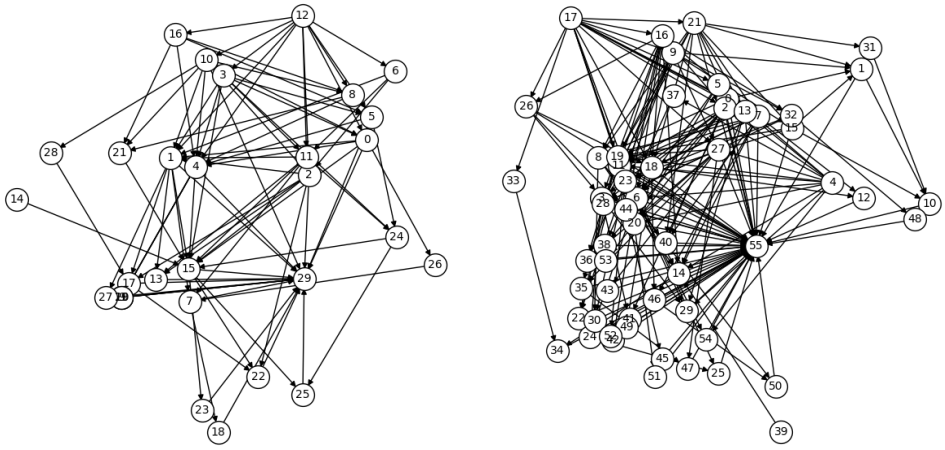

Figure 6.7: Dialogue structure in the *taxi* domain of the MultiWOZ. The structure on the left is from annotated dialogue states, while the right one is extracted by our approach.

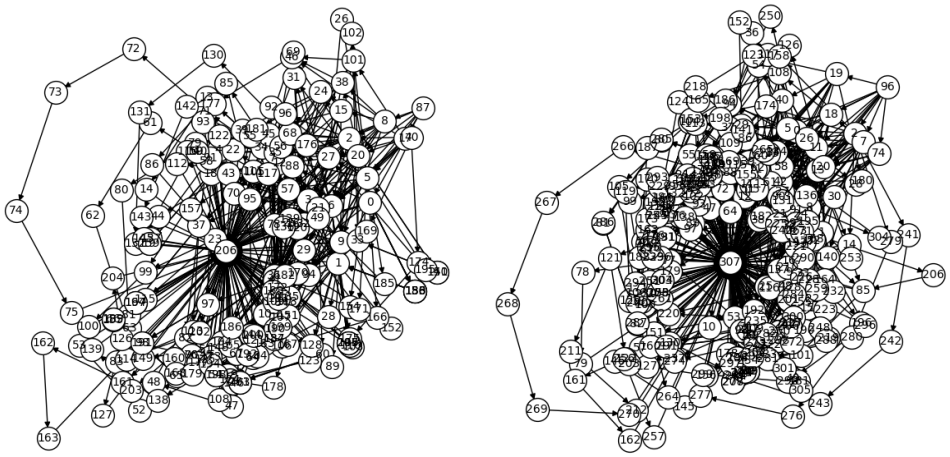

Figure 6.8: Dialogue structure in the *restaurant* domain of the MultiWOZ. The structure on the left is from annotated dialogue states, while the right one is extracted by our approach.

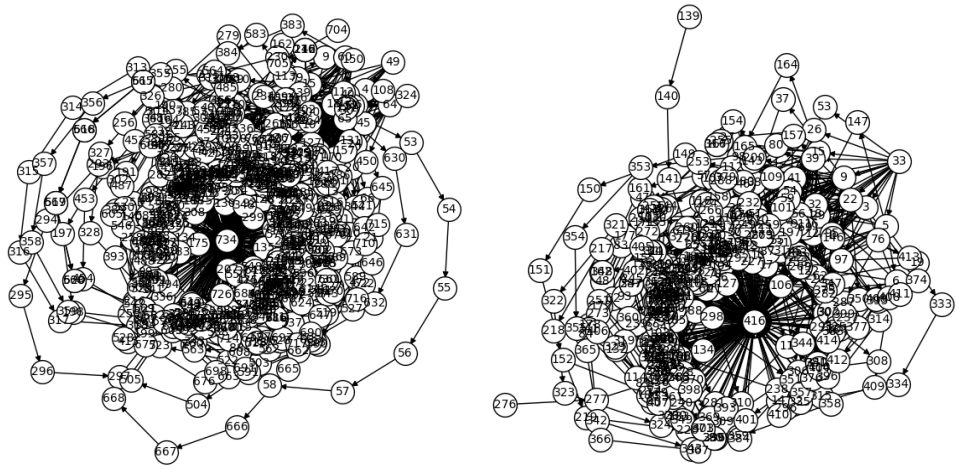

Figure 6.9: Dialogue structure in the *hotel* domain of the MultiWOZ. The structure on the left is from annotated dialogue states, while the right one is extracted by our approach.

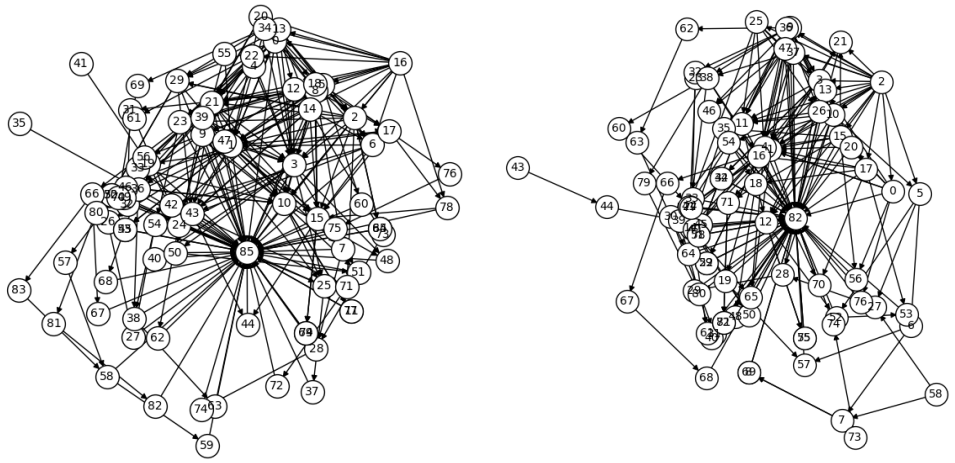

Figure 6.10: Dialogue structure in the *train* domain of the MultiWOZ. The structure on the left is from annotated dialogue states, while the right one is extracted by our approach.

$r_{\text{train}} \vert r_{\text{aug}}$	0.1	0.3	0.5	0.7	0.9
0.1	10.28 (11.96)	10.90 (14.07)	10.17 (13.68)	9.67 (15.39)	12.25 (14.42)
0.3	7.77 (8.48)	8.44 (9.17)	6.66 (9.16)	8.05 (9.89)	8.27 (9.45)
0.5	7.60 (7.87)	6.09 (8.34)	7.28 (8.05)	5.00 (9.01)	5.85 (8.56)
0.7	6.01 (6.65)	5.58 (6.61)	5.65 (6.83)	5.40 (7.96)	5.13 (7.03)
0.9	5.75 (6.17)	5.28 (6.79)	5.62 (6.51)	5.52 (7.18)	5.08 (7.14)

Table 6.8: Data Augmentation with MRDA (perplexity↓) in the *Taxi* domain of the MultiWOZ. Numbers in the parenthesis are using MFS.

$r_{\text{train}} \vert r_{\text{aug}}$	0.1	0.3	0.5	0.7	0.9
0.1	11.36 (10.39)	10.09 (12.18)	9.91 (14.97)	10.57 (14.53)	9.90 (14.81)
0.3	7.17 (7.65)	6.96 (8.64)	6.96 (9.34)	6.99 (8.26)	6.78 (8.04)
0.5	6.10 (6.47)	5.70 (6.78)	5.66 (7.27)	5.85 (7.07)	5.96 (7.93)
0.7	5.61 (5.84)	5.13 (5.86)	5.00 (6.64)	5.29 (6.19)	5.27 (6.46)
0.9	5.14 (5.32)	4.58 (5.64)	4.96 (5.78)	4.50 (5.75)	4.28 (6.18)

Table 6.9: Data Augmentation with MRDA (perplexity↓) in the *Restaurant* domain of the MultiWOZ. Numbers in the parenthesis are using MFS.

$r_{\text{train}} \vert r_{\text{aug}}$	0.1	0.3	0.5	0.7	0.9
0.1	17.17 (18.31)	14.90 (25.47)	18.39 (24.60)	16.75 (40.37)	16.52 (30.99)
0.3	10.45 (12.12)	11.45 (12.15)	11.22 (14.33)	11.21 (16.43)	11.08 (17.94)
0.5	9.38 (10.57)	8.60 (11.35)	10.06 (13.52)	8.40 (13.81)	10.22 (16.71)
0.7	8.05 (8.71)	9.12 (9.80)	8.28 (11.69)	8.85 (12.66)	8.62 (13.31)
0.9	7.45 (7.98)	7.36 (9.08)	7.22 (10.04)	7.39 (11.65)	7.20 (12.98)

Table 6.10: Data Augmentation with MRDA (perplexity↓) in the *Hotel* domain of the MultiWOZ. Numbers in the parenthesis are using MFS.

$r_{\text{train}} \vert r_{\text{aug}}$	0.1	0.3	0.5	0.7	0.9
0.1	20.70 (21.61)	19.74 (21.30)	18.57 (30.25)	18.26 (23.63)	15.67 (30.17)
0.3	11.88 (14.29)	12.69 (15.65)	12.52 (16.07)	9.61 (17.93)	14.98 (18.71)
0.5	12.06 (12.18)	10.41 (14.43)	13.66 (13.49)	8.95 (14.35)	11.37 (15.27)
0.7	10.17 (10.66)	9.81 (11.70)	10.97 (11.74)	12.59 (11.27)	8.78 (13.61)
0.9	9.84 (10.00)	6.89 (10.46)	8.26 (12.42)	5.28 (11.52)	7.68 (11.20)

Table 6.11: Data Augmentation with MRDA (perplexity↓) in the *Attraction* domain of the MultiWOZ. Numbers in the parenthesis are using MFS.

$r_{\text{train}} \vert r_{\text{aug}}$	0.1	0.3	0.5	0.7	0.9
0.1	14.71 (18.41)	16.74 (18.31)	18.67 (22.44)	13.96 (21.89)	15.05 (28.61)
0.3	11.36 (11.29)	10.98 (14.00)	9.25 (14.07)	10.68 (15.73)	10.92 (13.93)
0.5	8.51 (9.90)	7.62 (9.26)	8.47 (10.11)	7.70 (11.13)	9.30 (12.30)
0.7	8.38 (8.49)	6.88 (9.46)	7.24 (10.28)	7.58 (10.78)	7.61 (10.18)
0.9	6.85 (8.30)	6.93 (8.80)	6.33 (8.70)	7.31 (8.42)	7.43 (9.16)

Table 6.12: Data Augmentation with MRDA (perplexity↓) in the *Train* domain of the MultiWOZ. Numbers in the parenthesis are using MFS.

$r_{\text{train}} \vert r_{\text{aug}}$	0.1	0.3	0.5	0.7	0.9
0.1	7.34 (6.69)	7.97 (5.73)	7.15 (6.18)	7.41 (5.57)	7.19 (5.67)
0.3	8.02 (7.79)	8.02 (7.75)	10.67 (7.05)	12.04 (7.51)	10.77 (7.11)
0.5	8.49 (7.04)	10.47 (8.77)	7.55 (6.96)	13.27 (5.86)	13.70 (4.63)
0.7	10.34 (8.34)	12.03 (7.77)	11.29 (5.60)	10.44 (4.97)	18.15 (6.39)
0.9	9.35 (7.79)	9.46 (8.93)	10.36 (6.58)	10.98 (7.20)	20.16 (8.31)

Table 6.13: Data Augmentation with MRDA (BLEU↑) in the *Taxi* domain of the MultiWOZ. Numbers in the parenthesis are using MFS.

$r_{\text{train}} \vert r_{\text{aug}}$	0.1	0.3	0.5	0.7	0.9
0.1	10.47 (9.45)	9.24 (9.32)	8.96 (9.30)	10.39 (6.20)	11.76 (9.44)
0.3	10.04 (9.98)	11.74 (10.33)	12.20 (10.57)	12.35 (11.29)	13.63 (11.32)
0.5	12.84 (12.14)	14.70 (11.89)	15.46 (13.45)	16.71 (12.78)	15.96 (13.58)
0.7	13.33 (12.33)	15.61 (13.29)	17.67 (12.49)	16.68 (12.80)	19.13 (12.13)
0.9	13.64 (13.84)	14.71 (11.99)	17.13 (12.93)	17.39 (13.03)	22.87 (13.66)

Table 6.14: Data Augmentation with MRDA (BLEU↑) in the *Restaurant* domain of the MultiWOZ. Numbers in the parenthesis are using MFS.

$r_{\text{train}} \vert r_{\text{aug}}$	0.1	0.3	0.5	0.7	0.9
0.1	6.32 (5.00)	6.80 (6.97)	7.35 (6.27)	7.54 (5.84)	7.28 (5.14)
0.3	6.94 (6.60)	6.98 (5.96)	8.79 (6.48)	9.22 (6.86)	11.28 (6.21)
0.5	8.07 (8.05)	7.91 (7.43)	8.43 (7.17)	12.32 (6.45)	9.12 (7.39)
0.7	9.16 (8.32)	7.05 (7.88)	8.92 (7.64)	11.00 (7.62)	12.67 (7.62)
0.9	8.61 (9.60)	10.89 (7.90)	11.06 (9.43)	12.24 (7.52)	12.89 (7.15)

Table 6.15: Data Augmentation with MRDA (BLEU↑) in the *Hotel* domain of the MultiWOZ. Numbers in the parenthesis are using MFS.

$r_{\text{train}} \vert r_{\text{aug}}$	0.1	0.3	0.5	0.7	0.9
0.1	4.20 (4.36)	3.61 (2.59)	4.60 (4.60)	7.24 (3.15)	6.16 (5.52)
0.3	6.40 (5.26)	9.50 (9.40)	9.55 (6.60)	15.35 (7.39)	6.21 (4.17)
0.5	6.96 (5.90)	10.96 (5.77)	7.46 (7.36)	13.06 (7.25)	8.81 (7.45)
0.7	8.14 (4.93)	12.40 (7.65)	13.41 (7.76)	16.54 (2.58)	12.19 (7.14)
0.9	10.22 (6.94)	9.52 (8.06)	19.13 (6.51)	27.91 (8.71)	22.50 (10.01)

Table 6.16: Data Augmentation with MRDA (BLEU↑) in the *Attraction* domain of the MultiWOZ. Numbers in the parenthesis are using MFS.

$r_{\text{train}} \vert r_{\text{aug}}$	0.1	0.3	0.5	0.7	0.9
0.1	3.83 (3.77)	6.13 (3.72)	3.35 (4.14)	4.51 (2.94)	5.09 (3.31)
0.3	5.78 (5.66)	5.42 (5.33)	5.84 (3.55)	6.48 (4.17)	9.73 (4.88)
0.5	5.70 (6.24)	7.02 (6.69)	5.79 (4.67)	8.02 (6.66)	12.52 (5.21)
0.7	5.58 (3.81)	8.79 (6.83)	6.56 (5.44)	8.85 (6.34)	12.21 (7.43)
0.9	9.34 (6.45)	7.82 (6.61)	8.08 (4.25)	9.51 (6.84)	10.79 (6.70)

Table 6.17: Data Augmentation with MRDA (BLEU↑) in the *Train* domain of the MultiWOZ. Numbers in the parenthesis are using MFS.

	ARI					AMI					SC				
	Taxi	Rest.	Hotel	Attr.	Train	Taxi	Rest.	Hotel	Attr.	Train	Taxi	Rest.	Hotel	Attr.	Train
Random	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-	-	-	-
VRNN	0.05	0.00	0.00	0.00	0.00	0.05	0.02	0.00	0.01	0.06	-	-	-	-	-
BERT-KMeans	0.02	0.01	0.01	0.01	0.01	0.11	0.09	0.02	0.03	0.06	0.11	0.08	0.06	0.13	0.09
BERT-Birch	0.02	0.01	0.01	0.01	0.01	0.12	0.09	0.02	0.03	0.06	0.07	0.06	0.08	0.09	0.07
BERT-Agg	0.02	0.01	0.01	0.01	0.01	0.11	0.09	0.02	0.05	0.07	0.08	0.05	0.08	0.11	0.07
TOD-BERT-mlm	0.02	0.01	0.01	0.03	0.02	0.13	0.11	0.03	0.06	0.10	0.12	0.08	0.06	0.17	0.09
TOD-BERT-jnt	0.03	0.02	0.02	0.03	0.03	0.16	0.13	0.06	0.08	0.14	0.09	0.08	0.06	0.13	0.07
BERT-spaCy	0.01	0.06	0.04	0.01	0.01	0.09	0.18	0.12	0.06	0.08	-	-	-	-	-
TOD-BERT-spaCy	0.01	0.03	0.05	0.02	0.01	0.09	0.15	0.12	0.05	0.05	-	-	-	-	-
TOD-BERT-SBD_{MWOZ}	**0.15**	0.00	0.00	0.00	0.05	0.17	0.13	0.04	0.06	0.16	**0.39**	**0.34**	**0.27**	**0.44**	**0.34**
TOD-BERT-DET_{ATIS}	0.08	0.05	0.09	0.03	0.06	0.26	0.22	0.25	0.15	0.26	-	-	-	-	-
TOD-BERT-DET_{SNIPS}	0.06	0.05	0.11	0.03	0.04	0.25	0.23	0.22	0.09	0.22	-	-	-	-	-
TOD-BERT-DET_{MWOZ}	**0.15**	**0.22**	**0.24**	**0.33**	**0.24**	**0.39**	**0.48**	**0.44**	**0.44**	**0.44**	-	-	-	-	-

Table 6.18: Complete structure extraction results using clustering metrics in the MultiWOZ dataset. SC is omitted for methods that do not encode utterances directly. BERT with different clustering methods are included.

Domain	Original	Augmented
Taxi	[USR] I would like a taxi from Saint John's college to Pizza Hut Fen Ditton. [SYS] What time do you want to leave and what time do you want to arrive by? [USR] I want to leave after 17:15. [USR] Booking completed! Your taxi will be blue honda Contact number is 07218068540 [USR] Thank you for all the help! I appreciate it. [SYS] You are welcome. Is there anything else I can help you with today? [USR] No, I am all set. Have a nice day. Bye. [SYS] you too! thank you	[USR] From museum of classical archaeology to london kings cross train station, please [SYS] I have ordered one for you their pick up time is 1:00pm. Is this time okay or do you need a different time? [USR] Please pick me up at the scott polar museum [SYS] Okay, I've booked a taxi for you. [USR] Could you give me the car type and contact number? [SYS] Certainly, the car is a white Honda, and the contact number is 07135860962. [USR] Got it. Thanks for helping me. [SYS] Your welcome. Have a great time.
Restaurant	[USR] Can you help me find Restaurant One Seven? [SYS] I have Restaurant One Seven located in Centre, phone number is 01223337766. Would you like for me to make a reservation for you? [USR] I would like to book for two on Wednesday, at 17:15. [SYS] I have you booked Wednesday at 17:15 a table for 2, this table will hold for 15 mins and your reference number is GXO64LG8. [USR] Great, thanks! That's all I needed! [SYS] You're welcome! Have a great day. Goodbye.	[USR] I am looking for a particular restaurant. It is called pizza hut city centre. [SYS] It's a cheap italian place in the centre at Regent Street City Centre cb21ab. The phone number is 01223323737. Would you like more info? [USR] That sounds like just the place! Can you book me a table for 2 at 19:45 on Thursday? [SYS] You're all set. I've got a table for 2 at 19:45 this Thursday. Your reference number is F3K2PQZZ. [USR] Ok sounds good. Thank you! [SYS] You are welcome. Is there anything else I can assist you with today? [USR] That is all. Thank you for your help! [SYS] You're welcome. Have a great day.
Hotel	[USR] I need a place to stay in the north please. [SYS] There are 13 places to stay in the north. Do you have any other criteria? [USR] I don't want to have to pay for parking. [SYS] I have 5 different gueshouses listed. For what day and for how many please? [USR] For 3 people starting on Wednesday and staying 2 nights. [SYS] There are 11 hotels available actually. Would you like to narrow it to a specific price range or star rating? [USR] Just a guesthouse in the north with free parking. I'll need a reference number then too please. [SYS] I was able to get you in at Archway House. Your reference number is WUMUZ0JX. Can I help you with anything else? [USR] That's all I need today. Thanks! Bye! [SYS] Thank you for using our system!	[USR] north area of town [SYS] There are two hotels in that area with free parking: Ashley Hotel and the Lovell lodge. Would you like to book either of those? [USR] Actually, I'd like a guest house. [SYS] how about hamilton lodge? It sounds lovely. [USR] Can you see if either of them are available for 2 nights for 5 people starting on Thursday, I have no preference for which one. [SYS] Both of those hotels are completely booked up when trying to fit that criteria. Perhaps we can try different specifications?
Attraction	[USR] Someone told me about a place called Old Schools. It's supposed to be some kind of attraction. Can you tell me anything about it? [SYS] Old schools is an architecture museum in the centre area. The admission is free. Would you like tickets? [USR] Yes, please. That would be great. [SYS] I just checked and it is free. The address is trinity lane and the phone number is 01223332320. Can I help you with anything else? [USR] No, that is all for today. Thanks so much! [SYS] You're welcome! Have a great day!	[USR] Yes, hello. I'm looking for a particular attraction called Downing College. [SYS] t is located on \tregent street, in the centre. [USR] Thank you so much! [SYS] You are welcome! Did you need anything else today? [USR] No, that was all. Good bye. [SYS] Thanks and enjoy!
Train	[USR] am looking for a train, it should go to cambridge and should depart from norwich [SYS] What time and day are you looking to travel? [USR] Yes, I would like travel on Monday and I would need to arrive by 08:30. [SYS] TrainID TR9020 leaves norwich on monday and arrives in cambridge at 06:35. [USR] What is the price for the train? [SYS] The cost will be 17.60 pounds. Would you like more information about that? [USR] What time does the train leave? [SYS] it leaves at 05:16", [USR] Thanks, that's all. Good bye.\n [SYS] Goodbye and thanks for using the help desk.	[USR] I am looking for a train out of kings lynn please. [SYS] Where will you be going? [USR] I want to go to Cambridge. [SYS] When will you be leaving? [USR] yes. i should leave after 13:45 and should leave on thursday [SYS] I have a train leaving Thursday at 13:59, would you like me to book it for you? [USR] Could you tell me when the train arrives in London Liverpool Street? [SYS] That train arrives at 15:27. Would you like me to book it for you? [USR] yes pliz.may i also get the arrival time [SYS] The arrival time is by 15:27, do you want to book a seat? [USR] Not yet. I just needed to get the details. Thanks for helping me. Goodbye. [SYS] Thank you for using our services. Do you need any further assistance?

Table 6.19: Examples of generated dialogues by the Multi-Response Data Augmentation in the MultiWOZ.

State/Utts	State 0	State 1	...
Taxi			
Utt. 0	[usr] I would like a taxi from Saint John's college to Pizza Hut Fen Ditton. [sys] What time do you want to leave and what time do you want to arrive by?	[usr] I would like to be picked up at the cambridge belfy and go to the cambridge shop. [sys] I'm sorry, are you going to the Cambridge shop house, or did you mean the cambridge shop?	...
Utt. 1	[usr] I need to book a taxi to come to Express by Holiday Inn Cambridge to take me to the Oak Bistro. [sys] Okay, waht time do you want to leave by?	[usr] I want to be picked up at frankie and bennys please. [sys] I've booked a black Honda, the contact number is 07796011098. [usr] Great, thank you for your help. [sys] No problem. Are you finished?	...
Utt. 2	[usr] I want to depart from sidney sussex college. [sys] Great. Now I'll just need a time from you please.	[usr] I am departing from la tasca. [sys] I have booked a grey BMW, the contact number os 07618837066. [usr] Thanks so much for your help. [sys] Thank you goodbye.	...
Utt. 3	...	...	...
Restaurant			
Utt. 0	[usr] I need help finding a place to eat called curry garden. [sys] Curry garden is an expensive restaurant that serves Indian food. Do you want me to book it for you?	[usr] Yes, please.I need one for friday at 13:45. I'll be dining alone. [sys] I apologize but I was unable to book you for the restaurant. Would you like for me to find you another restaurant?	...
Utt. 1	[usr] I'm trying to find a restaurant called the Slug and Lettuce. Do you know where that is? [sys] Yes it is in the centre area.	[usr] 5 people on a Wednesday at 19:45. [sys] You got it. Here is your reference number: P9D58C0O. [usr] Thank you for help. That's everything I needed. [sys] Have a great day!	...
Utt. 2	[usr] I need to get to a restaurant known as the Lucky star. [sys] It is located at Cambridge Leisure Park Clifton Way Cherry Hinton.	[usr] Yes I would, for 3 people Wednesday at 18:15. [sys] Unfortunately, the restaurant is full at this time. Is there another time or day I can reserve for you?	...
Utt. 3	...	...	...
Hotel			
Utt. 0	[usr] I would really like something expensive. [sys] Unfortunately, I can't find any that are expensive. Let's try a different price range.	[usr] Yes, please. 6 people 3 nights starting on tuesday. [sys] I am sorry but I wasn't able to book that for you for Tuesday. Is there another day you would like to stay or perhaps a shorter stay?	...
Utt. 1	[usr] I need a guesthouse with a moderate price. [sys] Do you have a preferred area of the city you'd like to stay in? [usr] No, I don't have a preference. I'd like the guesthouse to have free parking though! [sys] I found acorn guest house. It is moderately priced and has four stars! Would you like me to book that for you?	[usr] Yes. I have heard great things about that guest house. [sys] Your booking is complete. Your reference number is 33ZFXQ8P. Is there anything else I can help you with today? [usr] No, that's all I needed today. Thank you! [sys] Thank you for using our service today."	...
Utt. 2	[usr] I'm looking for something that would be in the middle price point. [sys] I have located 15 guesthouse types in the moderate price range. Do you have a preference for a particular area of town?	[usr] That sounds perfect. Can I book it for 5 nights starting on Monday for 3 people? [sys] The Avalon doesn't have room for 5 nights starting Monday. How about a different day or a shorter stay?	...
Utt. 3	...	...	...
Attraction			
Utt. 0	[usr] Can you please help me find a place to go? [sys] I've found 79 places for you to go. Do you have any specific ideas in mind?	[usr] I'd like a sports place in the centre please. [sys] There are no results matching your query. Can I try a different area or type?	...
Utt. 1	[usr] I'm a tourist from out of town. But, I was trying to find something fun to do near my hotel. Could you recommend a place? [sys] I would be more than happy to recommend an attraction, first could you tell me in what part of town your hotel is located?	[usr] I'm looking for a place in the centre of town that is a nightclub. [sys] There are 5 nightclubs in the centre area, the one with the lowest entrance fee is soul tree nightclub which is 4 pounds to get in.	...
Utt. 2	[usr] I am looking for a place to go in town. [sys] There are many places, do you have a particular destination type in mind?	[usr] Can you found me a swimming pool in the south part of town? [sys] Unfortunately I was unable to find a pool at the south part of town, however we have some north east and at the centre.	...
Utt. 3	...	...	...
Train			
Utt. 0	[usr] I am looking for a train departing from london liverpool please. [sys] I'll be glad to help. You would like to from london liverpool street to what destination, please?	[usr] I want to leave on Friday. [sys] Unfortunately, the only train I have matching your criteria is one leaving 23:59 and arriving at 01:27 in the morning. Did you want to book that?	...
Utt. 1	[usr] I am looking for a train to go to London Kings Cross. [sys] There are several available option to travel to Kings cross today. What time would you like to travel?	[usr] I want to leave on Monday. [sys] Great, TR5720 from london kings cross to cambridge leaves monday at 11:17. Can I book this for you?	...
Utt. 2	[usr] I am looking for a train from Cambridge to Birmingham New Street. [sys] The next train leaving from Cambridge for Birmingham New Street departs Friday at 5:01, and will arrive by 7:44.	[usr] I would like to leave after 13:45. [sys] There is a train that leaves at 15:00 would you like me to book that train for you?	...
Utt. 3	...	...	...

Table 6.20: Predicted dialogue states for dialogues in the five domains of the MultiWOZ dataset.

CHAPTER 7

Conclusion

In this dissertation, we discussed two important topics towards building the next generation of dialogue systems: PART I: Emotional Intelligent Dialogue Systems and PART II: Structure Learning in Dialogue Systems.

In Chapter 2, we introduce a new dataset for human value modeling, VALUENET, which contains 21,374 scenarios in ten distinct human values. We also apply the learned value model from VALUENET to several EQ-related dialogue tasks. Our experiments show our approach and dataset provide a new way to control the dialogue system speaking style and numerically estimate one's value preference.

In Chapter 3, We propose a SocAoG model with α–β–γ processes for the consistent inference of social relations in dialogues. The model can also leverage attribute information to assist the inference. MCMC is proposed to parse the relation graph incrementally, enabling the dynamic inference upon any incoming utterance. Experiments show that our model outperforms state-of-the-art methods; case studies and ablation studies are provided for analysis.

In Chapter 4, we propose to build a socially intelligent agent by incorporating mind simulation and human values. We explore using a hybrid parser to track agents' mental state transition. The value model pre-trained on VALUENET brings social preference to help the agent make decisions. The model is proved to have a better performance than the state-of-the-art models on LIGHT.

In Chapter 5, we propose to inject structured attention into variational recurrent neural network models for unsupervised dialogue structure learning. We explore two different structure inductive biases: linear CRF for utterance-level semantic structure induction in two-party dialogues; and non-

projective dependency tree for interactive structure learning in multi-party dialogues. Both models are proved to have a better structure learning performance over the state-of-the-art algorithms.

In Chapter 6, we propose a simple yet effective approach for structure extraction in task-oriented dialogues. We define a task of Slot Boundary Detection and clustering to approximate the dialogue ontology. We extract a semantic structure that explicitly depicts the state transitions in task-oriented dialogues, without using state annotation during inference. Extensive experiments demonstrate that our approach is superior to the baseline models in all the domains of the MultiWOZ dataset. In addition, we demonstrate how to augment dialogue data based on our extracted structures to improve end-to-end response generation remarkably.

Altogether, we are excited about the progress that has been made in the field of dialogue systems and have been glad to be able to contribute to this area. At the same time, we believe there is still a long way towards building human-like chatbots. A lot of open questions need to be addressed in the future:

- **Values.** One challenge is to capture and encode the complete context for making value-driven decisions. We are also interested in improving the modeling performance and extending current formalism to non-English speaking cultures.

- **Social relations.** We will further explore how different initialization of the parse graph could help warm start the inference under various situations and how multi-modal cues could be leveraged.

- **Mental states.** Modeling and maintaining the mental states is another key challenge and our approach is just one step towards solving the problem. Right now, we are not clear how to model the deeper levels in the *Theory of Mind* and how to avoid the error propagation throughout the interaction.

- **Structure learning.** We will further explore how to explicitly incorporate linguistics information, such as named entities into the latent states. We also hope to encourage more researchers

to work on comprehensive analysis and downstream application study of these extracted dialogue structures.

REFERENCES

[AAA21] Anish Acharya, Suranjit Adhikari, Sanchit Agarwal, Vincent Auvray, Nehal Belgamwar, Arijit Biswas, Shubhra Chandra, Tagyoung Chung, Maryam Fazel-Zarandi, Raefer Gabriel, et al. "Alexa Conversations: An Extensible Data-driven Approach for Building Task-oriented Dialogue Systems." *arXiv preprint arXiv:2104.09088*, 2021.

[ABB20] 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. "Task-Oriented Dialogue as Dataflow Synthesis." *Transactions of the Association for Computational Linguistics*, **8**:556–571, 2020.

[Abe09] Alex Abella. *Soldiers of reason: The RAND corporation and the rise of the American empire*. Houghton Mifflin Harcourt, 2009.

[Aga18] Abien Fred Agarap. "Deep learning using rectified linear units (relu)." *arXiv preprint arXiv:1803.08375*, 2018.

[AH19] Leonard Adolphs and Thomas Hofmann. "Ledeepchef: Deep reinforcement learning agent for families of text-based games." *arXiv preprint arXiv:1909.01646*, 2019.

[AHH19] Christoph Alt, Marc Hübner, and Leonhard Hennig. "Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1388–1398, Florence, Italy, July 2019. Association for Computational Linguistics.

[All02] Michael Allingham. *Choice theory: A very short introduction*. OUP Oxford, 2002.

[ALS19] Arjun R Akula, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, Sinisa Todorovic, Joyce Y Chai, and Song-Chun Zhu. "X-tom: Explaining with theory-of-mind for gaining justified human trust." *arXiv preprint arXiv:1909.06907*, 2019.

[App10] Ian Apperly. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press, 2010.

[AUL20] Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li, Arthur Szlam, Tim Rocktäschel, and Jason Weston. "How to Motivate Your Dragon: Teaching Goal-Driven Agents to Speak and Act in Fantasy Worlds." *arXiv preprint arXiv:2010.00685*, 2020.

[AV06] David Arthur and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding." Technical report, Stanford, 2006.

[AYC20] Ashutosh Adhikari, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, and Will Hamilton. "Learning dynamic belief graphs to generalize on text-based games." *Advances in Neural Information Processing Systems*, **33**, 2020.

[BB07] Nguyen Bach and Sameer Badaskar. "A review of relation extraction." *Literature review for Language and Statistics II*, **2**:1–15, 2007.

[BBD86] Lalit Bahl, Peter Brown, Peter De Souza, and Robert Mercer. "Maximum mutual information estimation of hidden Markov model parameters for speech recognition." In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pp. 49–52. IEEE, 1986.

[BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473*, 2014.

[BFP17] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. "Rasa: Open source language understanding and dialogue management." *arXiv preprint arXiv:1712.05181*, 2017.

[BGB17] Tarek R Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Daniel Lowd, Priscila Machado Vieira Lima, et al. "Neural-symbolic learning and reasoning: A survey and interpretation." *arXiv preprint arXiv:1711.03902*, 2017.

[BM16] David Belanger and Andrew McCallum. "Structured prediction energy networks." In *International Conference on Machine Learning*, pp. 983–992. PMLR, 2016.

[BMR20] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165*, 2020.

[Bro91] Rodney A Brooks. "Intelligence without representation." *Artificial intelligence*, **47**(1-3):139–159, 1991.

[BS16] Punam Bedi and Chhavi Sharma. "Community detection in social networks." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **6**(3):115–135, 2016.

[BWT18] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. "MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[CD12] Jan Cieciuch and Eldad Davidov. "A comparison of the invariance properties of the PVQ-40 and the PVQ-21 to measure human values across German and Polish samples." In *Survey Research Methods*, volume 6, pp. 37–48, 2012.

[CFL15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco captions: Data collection and evaluation server." *arXiv preprint arXiv:1504.00325*, 2015.

[CG95] Siddhartha Chib and Edward Greenberg. "Understanding the metropolis-hastings algorithm." *The american statistician*, **49**(4):327–335, 1995.

[CGC14] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint arXiv:1412.3555*, 2014.

[CHH20] Hui Chen, Pengfei Hong, Wei Han, Navonil Majumder, and Soujanya Poria. "Dialogue Relation Extraction with Document-level Heterogeneous Graph Attention Networks." *arXiv preprint arXiv:2009.05092*, 2020.

[Cho08] Ananlada Chotimongkol. *Learning the structure of task-oriented conversations from the corpus of in-domain dialogs*. PhD thesis, Carnegie Mellon University, Language Technologies Institute, School of . . . , 2008.

[CID17] Snigdha Chaturvedi, Mohit Iyyer, and Hal Daume III. "Unsupervised learning of evolving relationships between literary characters." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[CKD15] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. "A recurrent latent variable model for sequential data." In *Advances in neural information processing systems*, pp. 2980–2988, 2015.

[CKY18] Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. "Textworld: A learning environment for text-based games." In *Workshop on Computer Games*, pp. 41–75. Springer, 2018.

[CM16] Kevin Clark and Christopher D. Manning. "Deep Reinforcement Learning for Mention-Ranking Coreference Models." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2256–2262, Austin, Texas, November 2016. Association for Computational Linguistics.

[Cov01] Michael A Covington. "A fundamental algorithm for dependency parsing." In *Proceedings of the 39th annual ACM southeast conference*, pp. 95–102. Citeseer, 2001.

[CSB18] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces." *arXiv preprint arXiv:1805.10190*, 2018.

[CVB14] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. "On the properties of neural machine translation: Encoder-decoder approaches." *arXiv preprint arXiv:1409.1259*, 2014.

[CWH71] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. "Artificial paranoia." *Artificial Intelligence*, **2**(1):1–25, 1971.

[DCL18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*, 2018.

[DCL19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Den78] Daniel C Dennett. "Beliefs about beliefs [P&W, SR&B]." *Behavioral and Brain sciences*, **1**(4):568–570, 1978.

[Dev85] Pierre A Devijver. "Baum's forward-backward algorithm revisited." *Pattern Recognition Letters*, **3**(6):369–373, 1985.

[DLM20] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. "The second conversational intelligence challenge (convai2)." In *The NeurIPS'18 Competition*, pp. 187–208. Springer, 2020.

[DRS19] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. "Wizard of Wikipedia: Knowledge-Powered Conversational Agents." In *International Conference on Learning Representations*, 2019.

[DS97] Patrick Doreian and Frans N Stokman. "The dynamics and evolution of social networks." *Evolution of social networks*, **1**(1), 1997.

[DWP07] Nan Du, Bin Wu, Xin Pei, Bai Wang, and Liutong Xu. "Community detection in large-scale social networks." In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 16–25, 2007.

[EC08] Micha Elsner and Eugene Charniak. "You Talking to Me? A Corpus and Algorithm for Conversation Disentanglement." In *Proceedings of ACL-08: HLT*, pp. 834–842, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[EGP20] Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. "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*, pp. 422–428, Marseille, France, May 2020. European Language Resources Association.

[FHS20] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. "Social Chemistry 101: Learning to Reason about Social and Moral Norms." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 653–670, Online, November 2020. Association for Computational Linguistics.

[Fis70] Peter C Fishburn. "Utility theory for decision making." Technical report, Research analysis corp McLean VA, 1970.

[Fle71] Joseph L Fleiss. "Measuring nominal scale agreement among many raters." *Psychological bulletin*, **76**(5):378, 1971.

[FN04] Jerome Feldman and Srinivas Narayanan. "Embodied meaning in a neural theory of language." *Brain and language*, **89**(2):385–392, 2004.

[FSR18] Denis Fedorenko, Nikita Smetanin, and Artem Rodichev. "Avoiding echo-responses in a retrieval-based conversation system." In *Conference on Artificial Intelligence and Natural Language*, pp. 91–97. Springer, 2018.

[Gar14] Peter Gardenfors. *The geometry of meaning: Semantics based on conceptual spaces*. MIT press, 2014.

[GBS21] Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. "Paragraph-level commonsense transformers with recurrent memory." In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[GGH18] Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. "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)*, pp. 753–757, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[GH17] Andrew S Gordon and Jerry R Hobbs. *A formal theory of commonsense psychology: How people think people think*. Cambridge University Press, 2017.

[GLC18] Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. "Long text generation via adversarial training with leaked information." In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[GLG08] Artur SD'Avila Garcez, Luis C Lamb, and Dov M Gabbay. *Neural-symbolic cognitive reasoning*. Springer Science & Business Media, 2008.

[GLI21] Milan Gritta, Gerasimos Lampouras, and Ignacio Iacobacci. "Conversation Graph: Data Augmentation, Training, and Evaluation for Non-Deterministic Dialogue Management." *Transactions of the Association for Computational Linguistics*, **9**:36–52, 2021.

[GM15] MY Ganaie and Hafiz Mudasir. "A study of social intelligence & academic achievement of college students of district Srinagar, J&K, India." *Journal of American Science*, **11**(3):23–27, 2015.

[GM16] Jon Gauthier and Igor Mordatch. "A paradigm for situated and goal-driven language learning." *arXiv preprint arXiv:1610.03585*, 2016.

[GMG20] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. "COSMIC: COmmonSense knowledge for eMotion Identification in Conversations." In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2470–2481, Online, November 2020. Association for Computational Linguistics.

[Gri81] H Paul Grice. "Presupposition and conversational implicature." *Radical pragmatics*, **183**, 1981.

[Gri89] H Paul Grice. "Indicative conditionals." *Studies in the Way of Words*, pp. 58–85, 1989.

[GS86] Barbara J. Grosz and Candace L. Sidner. "Attention, Intentions, and the Structure of Discourse." *Computational Linguistics*, **12**(3):175–204, 1986.

[GYD15] Matthew R. Gormley, Mo Yu, and Mark Dredze. "Improved Relation Extraction with Feature-Rich Compositional Embedding Models." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1774–1784, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[HBB20] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. "Aligning ai with shared human values." *arXiv preprint arXiv:2008.02275*, 2020.

[HBB21] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. "Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs." In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[HBE17] He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. "Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1766–1776, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[HCK18] Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. "EmotionLines: An Emotion Corpus of Multi-Party Conversations." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[HCL06] Raia Hadsell, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping." In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

[HCL19] Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. "Gsn: A graph-structured network for multi-party dialogues." *arXiv preprint arXiv:1905.13637*, 2019.

[HDY21] Vojtěch Hudeček, Ondřej Dušek, and Zhou Yu. "Discovering Dialogue Slots with Weak Supervision." 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)*, pp. 2430–2442, Online, August 2021. Association for Computational Linguistics.

[Hig10] Colin De la Higuera. *Grammatical inference: learning automata and grammars*. Cambridge University Press, 2010.

[HJ15] Matthew Honnibal and Mark Johnson. "An Improved Non-monotonic Transition System for Dependency Parsing." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[HM17] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." To appear, 2017.

[HMW20] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. "A simple language model for task-oriented dialogue." *arXiv preprint arXiv:2005.00796*, 2020.

[HR18] Jeremy Howard and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory." *Neural computation*, **9**(8):1735–1780, 1997.

[HS01] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory." 2001.

[IGC16] Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. "Feuding Families and Former Friends: Unsupervised Learning for Dynamic Fictional Relationships." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1534–1544, San Diego, California, June 2016. Association for Computational Linguistics.

[JGB17] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. "Bag of Tricks for Efficient Text Classification." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431. Association for Computational Linguistics, April 2017.

[Jon12] Jan de Jonge. "Rational and Moral Action." In *Rethinking Rational Choice Theory*, pp. 199–206. Springer, 2012.

[Jur97] Dan Jurafsky. "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual." *Institute of Cognitive Science Technical Report*, 1997.

[JW19] Sarthak Jain and Byron C. Wallace. "Attention is not Explanation." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[JYQ20] Zhijing Jin, Yongyi Yang, Xipeng Qiu, and Zheng Zhang. "Relation of the Relations: A New Paradigm of the Relation Extraction Problem." *arXiv preprint arXiv:2006.03719*, 2020.

[Kam04] Nanda Kambhatla. "Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Information Extraction." In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 178–181, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*, 2014.

[KBV16] Douwe Kiela, Luana Bulat, Anita L Vero, and Stephen Clark. "Virtual embodiment: A scalable long-term strategy for artificial intelligence research." *arXiv preprint arXiv:1610.07432*, 2016.

[KDH17] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. "Structured attention networks." *arXiv preprint arXiv:1702.00887*, 2017.

[KK19] Evgeny Kim and Roman Klinger. "Frowning Frodo, Wincing Leia, and a Seriously Great Friendship: Learning to Classify Emotional Relationships of Fictional Characters." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 647–653, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[KO11] Bekir Karlik and A Vehbi Olgac. "Performance analysis of various activation functions in generalized MLP architectures of neural networks." *International Journal of Artificial Intelligence and Expert Systems*, **1**(4):111–122, 2011.

[KP19] Nikhil Krishnaswamy and James Pustejovsky. "Multimodal continuation-style architectures for human-robot interaction." *arXiv preprint arXiv:1909.08161*, 2019.

[KTL20] Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. "Learning Interactions and Relationships between Movie Characters." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9849–9858, 2020.

[Kum17] Shantanu Kumar. "A survey of deep learning methods for relation extraction." *arXiv preprint arXiv:1705.03645*, 2017.

[KW13] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114*, 2013.

[KWM11] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. "Twitter sentiment analysis: The good the bad and the omg!" In *Fifth International AAAI conference on weblogs and social media*, 2011.

[LBC21] Nicholas Lourie, Ronan Le Bras, and Yejin Choi. "Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-Life Anecdotes." In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[LCC20] Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. "You Impress Me: Dialogue Generation via Mutual Persona Perception." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1417–1427, Online, July 2020. Association for Computational Linguistics.

[LH17] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization." *arXiv preprint arXiv:1711.05101*, 2017.

[LHA21] Yuan Liang, Lei He, and Xiang Anthony'Chen. "Human-Centered AI for Medical Imaging." *Artificial Intelligence for Human Computer Interaction: A Modern Approach*, pp. 539–570, 2021.

[Lia16] Percy Liang. "Learning executable semantic parsers for natural language understanding." *Communications of the ACM*, **59**(9):68–76, 2016.

[LJ80] George Lakoff and Mark Johnson. "The metaphorical structure of the human conceptual system." *Cognitive science*, **4**(2):195–208, 1980.

[LL21] Xiang Lisa Li and Percy Liang. "Prefix-tuning: Optimizing continuous prompts for generation." *arXiv preprint arXiv:2101.00190*, 2021.

[LLG19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461*, 2019.

[LOG19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692*, 2019.

[LPS15] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems." In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 285–294, Prague, Czech Republic, September 2015. Association for Computational Linguistics.

[LR21] Teven Le Scao and Alexander M Rush. "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*, pp. 2627–2636, 2021.

[LSS17] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset." In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[LSY19] Hsu-Chao Lai, Hong-Han Shuai, De-Nian Yang, Jiun-Long Huang, Wang-Chien Lee, and Philip S Yu. "Social-aware VR configuration recommendation via multi-feedback coupled tensor factorization." In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1773–1782, 2019.

[LSZ10] Jingjing Liu, Stephanie Seneff, and Victor Zue. "Dialogue-Oriented Review Summary Generation for Spoken Dialogue Recommendation Systems." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 64–72, Los Angeles, California, June 2010. Association for Computational Linguistics.

[LUT17] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. "Building machines that learn and think like people." *Behavioral and brain sciences*, **40**, 2017.

[LY90] Karim Lari and Steve J Young. "The estimation of stochastic context-free grammars using the inside-outside algorithm." *Computer speech & language*, **4**(1):35–56, 1990.

[LZW20] Zuchao Li, Hai Zhao, Rui Wang, and Kevin Parnow. "High-order Semantic Role Labeling." In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1134–1151, Online, November 2020. Association for Computational Linguistics.

[MB06] Raymond J Mooney and Razvan C Bunescu. "Subsequence kernels for relation extraction." In *Advances in neural information processing systems*, pp. 171–178, 2006.

[MBS09] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. "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*, pp. 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[MBX17] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. "Learned in translation: Contextualized word vectors." *arXiv preprint arXiv:1708.00107*, 2017.

[MJB16] Tomas Mikolov, Armand Joulin, and Marco Baroni. "A roadmap towards machine intelligence." In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 29–61. Springer, 2016.

[MRC05] Gabriel Murray, Steve Renals, and Jean Carletta. "Extractive summarization of meeting recordings." 2005.

[MS14] Makoto Miwa and Yutaka Sasaki. "Modeling Joint Entity and Relation Extraction with Table Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1858–1869, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Mul11] Daniel Müllner. "Modern hierarchical, agglomerative clustering algorithms." *arXiv preprint arXiv:1109.2378*, 2011.

[MVC10] Gary McKeown, Michel F Valstar, Roderick Cowie, and Maja Pantic. "The SEMAINE corpus of emotionally coloured character interactions." In *2010 IEEE International Conference on Multimedia and Expo*, pp. 1079–1084. IEEE, 2010.

[NKB15] Karthik Narasimhan, Tejas Kulkarni, and Regina Barzilay. "Language understanding for text-based games using deep reinforcement learning." *arXiv preprint arXiv:1506.08941*, 2015.

[OS18] Takuma Okuda and Sanae Shoda. "AI-based chatbot service for financial industry." *Fujitsu Scientific and Technical Journal*, **54**(2):4–8, 2018.

[PHM19] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics.

[PLL20] Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. "SOLOIST: Few-shot Task-Oriented Dialog with A Single Pre-trained Auto-regressive Model." *arXiv preprint arXiv:2005.05298*, 2020.

[PLL21] Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. "Soloist: BuildingTask Bots at Scale with Transfer Learning and Machine Teaching." *Transactions of the Association for Computational Linguistics*, **9**:807–824, 2021.

[PNI18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep Contextualized Word Representations." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[PRP19] Yannis Papanikolaou, Ian Roberts, and Andrea Pierleoni. "Deep Bidirectional Transformers for Relation Extraction without Supervision." In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 67–75, Hong Kong, China, November 2019. Association for Computational Linguistics.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[PW78] David Premack and Guy Woodruff. "Does the chimpanzee have a theory of mind?" *Behavioral and brain sciences*, **1**(4):515–526, 1978.

[Pyl78] Zenon W Pylyshyn. "When is attribution of beliefs justified?[P&W]." *Behavioral and brain sciences*, **1**(4):592–593, 1978.

[QGX20] Meng Qu, Tianyu Gao, Louis-Pascal Xhonneux, and Jian Tang. "Few-shot relation extraction via bayesian meta-learning on relation graphs." In *International Conference on Machine Learning*, pp. 7867–7876. PMLR, 2020.

[QLZ21] Liang Qiu, Yuan Liang, Yizhou Zhao, Pan Lu, Baolin Peng, Zhou Yu, Ying Nian Wu, and Song-Chun Zhu. "SocAoG: Incremental Graph Parsing for Social Relation Inference in Dialogues." 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)*, pp. 658–670, Online, August 2021. Association for Computational Linguistics.

[QZH18] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu. "Human-centric indoor scene synthesis using stochastic grammar." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5899–5908, 2018.

[QZL21] Liang Qiu, Yizhou Zhao, Yuan Liang, Pan Lu, Weiyan Shi, Zhou Yu, and Song-Chun Zhu. "Towards Socially Intelligent Agents with Mental State Transition and Human Utility." *arXiv preprint arXiv:2103.07011*, 2021.

[QZS20] Liang Qiu, Yizhou Zhao, Weiyan Shi, Yuan Liang, Feng Shi, Tao Yuan, Zhou Yu, and Song-Chun Zhu. "Structured Attention for Unsupervised Dialogue Structure Induction." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1889–1899, Online, November 2020. Association for Computational Linguistics.

[RCD10] Alan Ritter, Colin Cherry, and Bill Dolan. "Unsupervised Modeling of Twitter Conversations." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 172–180, Los Angeles, California, June 2010. Association for Computational Linguistics.

[RSL19] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics.

[RSM86] David E Rumelhart, Paul Smolensky, James L McClelland, and G Hinton. "Sequential thought processes in PDP models." *Parallel distributed processing: explorations in the microstructures of cognition*, **2**:3–57, 1986.

[Rus20] Alexander M. Rush. "Torch-Struct: Deep Structured Prediction Library.", 2020.

[RWC19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. "Language models are unsupervised multitask learners." *OpenAI blog*, **1**(8):9, 2019.

[SBV15] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion." In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 553–562, 2015.

[Sch92] Shalom H Schwartz. "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries." In *Advances in experimental social psychology*, volume 25, pp. 1–65. Elsevier, 1992.

[Sch12] Shalom H Schwartz. "An overview of the Schwartz theory of basic values." *Online readings in Psychology and Culture*, **2**(1):2307–0919, 2012.

[SCM16] Shashank Srivastava, Snigdha Chaturvedi, and Tom Mitchell. "Inferring interpersonal relations in narrative summaries." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[SCV12] Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. "Refining the theory of basic individual values." *Journal of personality and social psychology*, **103**(4):663, 2012.

[SDC19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108*, 2019.

[SKB18] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. "Modeling relational data with graph convolutional networks." In *European Semantic Web Conference*, pp. 593–607. Springer, 2018.

[SKF16] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. "Bidirectional attention flow for machine comprehension." *arXiv preprint arXiv:1611.01603*, 2016.

[SP15] Iulian V Serban and Joelle Pineau. "Text-based speaker identification for multi-participant open-domain dialogue systems." In *NIPS Workshop on Machine Learning for Spoken Language Understanding. Montreal, Canada*, 2015.

[SRC19] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. "Social IQa: Commonsense Reasoning about Social Interactions." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[SRZ16] Tianmin Shu, Michael S Ryoo, and Song-Chun Zhu. "Learning social affordance for human-robot interaction." *arXiv preprint arXiv:1604.03692*, 2016.

[SSL17] Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. "A hierarchical latent variable encoder-decoder model for generating dialogues." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[SST21] Yajing Sun, Yong Shan, Chengguang Tang, Yue Hu, Yinpei Dai, Jing Yu, Jian Sun, Fei Huang, and Luo Si. "Unsupervised Learning of Deterministic Dialogue Structure with Edge-Enhanced Graph Auto-Encoder." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13869–13877, 2021.

[Sta02] Robert Stalnaker. "Common ground." *Linguistics and philosophy*, **25**(5/6):701–721, 2002.

[Sun94] Ron Sun. *Integrating rules and connectionism for robust commonsense reasoning*. John Wiley & Sons, Inc., 1994.

[SVT16] Arjen Stolk, Lennart Verhagen, and Ivan Toni. "Conceptual alignment: How brains achieve mutual understanding." *Trends in cognitive sciences*, **20**(3):180–191, 2016.

[Szt02] Piotr Sztompka. "Socjologia." *Analiza społeczeństwa, Znak, Kraków*, p. 324, 2002.

[SZY19] Weiyan Shi, Tiancheng Zhao, and Zhou Yu. "Unsupervised Dialog Structure Learning." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1797–1807, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[THH10] Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. "What is left to be understood in ATIS?" In *2010 IEEE Spoken Language Technology Workshop*, pp. 19–24. IEEE, 2010.

[TSH20] Ronen Tamari, Chen Shani, Tom Hope, Miriam R L Petruck, Omri Abend, and Dafna Shahaf. "Language (Re)modelling: Towards Embodied Language Understanding." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6268–6281, Online, July 2020. Association for Computational Linguistics.

[UA13] David C Uthus and David W Aha. "The ubuntu chat corpus for multiparticipant chat analysis." In *2013 AAAI Spring Symposium Series*, 2013.

[UFK19] Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. "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)*, pp. 673–683, Hong Kong, China, November 2019. Association for Computational Linguistics.

[VSP17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[VTC18] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. "Moviegraphs: Towards understanding human-centric situations from videos." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8581–8590, 2018.

[WDS20] 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. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

[Wei66] Joseph Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine." *Communications of the ACM*, **9**(1):36–45, 1966.

[WHS20] Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. "TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 917–929, Online, November 2020. Association for Computational Linguistics.

[Wil92] Ronald J Williams. "Simple statistical gradient-following algorithms for connectionist reinforcement learning." *Machine learning*, **8**(3):229–256, 1992.

[WP19] Sarah Wiegreffe and Yuval Pinter. "Attention is not not Explanation." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics.

[WSC16] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint arXiv:1609.08144*, 2016.

[WVM17] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. "A Network-based End-to-End Trainable Task-oriented Dialogue System." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 438–449, Valencia, Spain, April 2017. Association for Computational Linguistics.

[WZ11] Tianfu Wu and Song-Chun Zhu. "A numerical study of the bottom-up and top-down inference processes in and-or graphs." *International journal of computer vision*, **93**(2):226–252, 2011.

[XLW21] Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. "Discovering Dialog Structure Graph for Coherent 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 (Volume 1: Long Papers)*, pp. 1726–1739, Online, August 2021. Association for Computational Linguistics.

[XSZ20a] Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. "An Embarrassingly Simple Model for Dialogue Relation Extraction." *arXiv preprint arXiv:2012.13873*, 2020.

[XSZ20b] Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. "GDPNet: Refining Latent Multi-View Graph for Relation Extraction." *arXiv preprint arXiv:2012.06780*, 2020.

[YBJ97] Xue Yongqiang, Gao Baojiao, and Gao Jianfeng. "The theory of thermodynamics for chemical reactions in dispersed heterogeneous systems." *Journal of colloid and interface science*, **191**(1):81–85, 1997.

[YCS18] Xingdi Yuan, Marc-Alexandre Côté, Alessandro Sordoni, Romain Laroche, Remi Tachet des Combes, Matthew Hausknecht, and Adam Trischler. "Counting to explore and generalize in text-based games." *arXiv preprint arXiv:1806.11525*, 2018.

[YDL18] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. "Qanet: Combining local convolution with global self-attention for reading comprehension." *arXiv preprint arXiv:1804.09541*, 2018.

[YL17] Wang Yuan and Zhijun Li. "Development of a human-friendly robot for socially aware human-robot interaction." In *2017 2nd International Conference on Advanced Robotics and Mechatronics (ICARM)*, pp. 76–81. IEEE, 2017.

[YM19] Xusen Yin and Jonathan May. "Comprehensible context-driven text game playing." In *2019 IEEE Conference on Games (CoG)*, pp. 1–8. IEEE, 2019.

[You06] Steve Young. "Using POMDPs for dialog management." In *2006 IEEE Spoken Language Technology Workshop*, pp. 8–13. IEEE, 2006.

[YSC20] Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. "Dialogue-Based Relation Extraction." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4927–4940, Online, July 2020. Association for Computational Linguistics.

[YWG18] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. "Neural-symbolic vqa: Disentangling reasoning from vision and language understanding." *arXiv preprint arXiv:1810.02338*, 2018.

[ZAR03] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. "Kernel methods for relation extraction." *Journal of machine learning research*, **3**(Feb):1083–1106, 2003.

[Zay15] Godandapani Zayaraz et al. "Concept relation extraction using Naïve Bayes classifier for ontology-based question answering systems." *Journal of King Saud University-Computer and Information Sciences*, **27**(1):13–24, 2015.

[ZDU18] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. "Personalizing Dialogue Agents: I have a dog, do you have pets too?" In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[ZE18] Tiancheng Zhao and Maxine Eskenazi. "Zero-Shot Dialog Generation with Cross-Domain Latent Actions." In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 1–10, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[ZG05] Shubin Zhao and Ralph Grishman. "Extracting Relations with Integrated Information Using Kernel Methods." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 419–426, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[ZGL20] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot." *Computational Linguistics*, **46**(1):53–93, March 2020.

[Zha19] Ran Zhao. *Socially-Aware Dialogue System*. PhD thesis, Carnegie Mellon University, 2019.

[ZLE18] Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. "Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1098–1107, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[ZM07] Song-Chun Zhu and David Mumford. *A stochastic grammar of images*. Now Publishers Inc, 2007.

[ZOY20] Yichi Zhang, Zhijian Ou, and Zhou Yu. "Task-oriented dialog systems that consider multiple appropriate responses under the same context." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9604–9611, 2020.

[ZQA20] Yizhou Zhao, Liang Qiu, Wensi Ai, Feng Shi, and Song-Chun Zhu. "Vertical-Horizontal Structured Attention for Generating Music with Chords." *arXiv preprint arXiv:2011.09078*, 2020.

[ZRL96] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. "BIRCH: an efficient data clustering method for very large databases." *ACM sigmod record*, **25**(2):103–114, 1996.

[ZRR18] Ran Zhao, Oscar J Romero, and Alex Rudnicky. "SOGO: a social intelligent negotiation dialogue system." In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 239–246, 2018.

[ZSG19] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. "Dialogpt: Large-scale generative pre-training for conversational response generation." *arXiv preprint arXiv:1911.00536*, 2019.

[ZSG20] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. "DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 270–278, Online, July 2020. Association for Computational Linguistics.

[ZW14] Ke Zhai and Jason D. Williams. "Discovering Latent Structure in Task-Oriented Dialogues." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36–46, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[ZWM98] Song Chun Zhu, Yingnian Wu, and David Mumford. "Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling." *International Journal of Computer Vision*, **27**(2):107–126, 1998.

[ZWX16] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. "Attention-based LSTM Network for Cross-Lingual Sentiment Classification." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 247–256, Austin, Texas, November 2016. Association for Computational Linguistics.

[ZXE19] Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. "Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1208–1218, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[ZZC17] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. "Position-aware Attention and Supervised Data Improve Slot Filling." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 35–45, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[ZZE17] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. "Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders." In *Proceedings*

of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 654–664, Vancouver, Canada, July 2017. Association for Computational Linguistics.