TOWARDS SOCIAL *and* INTERPRETABLE NEURAL DIALOG SYSTEMS

ABDUL SALEH

A THESIS SUBMITTED TO
THE DEPARTMENTS OF COMPUTER SCIENCE AND STATISTICS

IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
OF
BACHELOR OF ARTS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL, 2020

THESIS ADVISORS:
Professor Stuart Shieber,
Professor Lucas Janson

ABSTRACT

Open-domain dialog generation is a task that challenges machines to mimic human conversations. Despite the remarkable progress natural language generation has seen over the past several years, open-domain dialog systems still suffer from limitations that hinder their adoption in the real world. Systems trained with maximum likelihood often generate dull and repetitive responses, ignoring user input. Training on standard datasets from online forums leads to the generation of inappropriate, biased, or toxic responses. And models rarely exhibit long-term coherence across multiple dialog turns. Meanwhile, the predominant approach to dialog generation relies on black-box neural networks which provide little insight as to what information they learn (or do not learn) about engaging in dialog.

In light of these issues, this thesis makes two contributions to building social and interpretable dialog systems. The first part of this thesis proposes a novel reinforcement learning approach for improving the social capabilities of open-domain dialog systems. We optimize for human-centered objectives such as response politeness, diversity, coherence, and sentiment. Our interactive human evaluation shows that these objectives can improve the quality of human-AI interaction and increase user engagement.

The second part of this thesis investigates the conversational understanding captured by neural dialog systems using probing. Our results suggest that standard open-domain dialog systems struggle with basic skills such as answering questions, inferring contradiction, and determining the topic of conversation. We also find that the dyadic, turn-taking nature of dialog is not fully leveraged by these models. By exploring these limitations, we highlight the need for additional research into architectures and training methods that can allow for capturing high-level information about natural language.

PUBLICATIONS

Material from the following papers was used to create the chapters of this thesis:

[1] Abdelrhman Saleh, Tovly Deutsch, Stephen Casper, Yonatan Belinkov, and Stuart Shieber. "Probing Neural Dialog Models for Conversational Understanding." In: *arXiv Preprint* (2020).

[2] Abdelrhman Saleh, Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, and Rosalind Picard. "Hierarchical Reinforcement Learning for Open-Domain Dialog." In: *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence* (2020).

ACKNOWLEDGEMENTS

I would not have been able to finish this thesis without the help and support of incredible advisers, colleagues, family, and friends. I am infinitely thankful to my advisers Professor **Stuart Shieber** and Professor **Lucas Janson** for their wisdom, mentorship, and guidance over the past year. I would also like to thank **Yonatan Belinkov** for his insightful feedback and spot-on intuition, without which Chapter 3 of this thesis would have been impossible.

I cannot begin this thesis without thanking **Natasha Jaques**. I am very fortunate to have worked with someone so invested in my growth and success. Natasha's contributions were also essential to getting Chapter 2 of this thesis from conception to publication in one short summer. I would also like to thank **Rosalind Picard** for introducing me to affective computing and changing the way I think about interacting with machines.

I am incredibly grateful to **Ramy Baly** for introducing me to natural language processing and patiently helping me as I worked on my first research project. Thanks to **James Glass** for making my visit to *NAACL* feasible, which is where many ideas for this thesis started brewing.

I am very lucky to have friends like **Gabriel Grand** and **Mirac Suzgun** who encouraged me to start doing research and acted as excellent role models during my time in college. Also thank you **Leena Hamad** for all the laughter, jokes, and fond memories as I worked on this thesis.

Finally, I would like to thank my parents **Ahmed** and **Nahla** and my best friend and younger brother **Aly** for always providing me with love, support, and encouragement.

CONTENTS

Part I

PROLOGUE

I propose to consider the question, 'Can machines think?'

— Alan Turing [94]

1 INTRODUCTION

1.1 SOLVING INTELLIGENCE

In a very influential 1950 paper published in the journal *Mind*, Alan Turing introduced the Turing Test as a way of dealing with the question of whether machines can think. The Test took the form of an "imitation game" where an interrogator would chat over teletype with two hidden participants, a human and a machine. The conversation would be natural, open-ended, and about any topic whatsoever. If the interrogator cannot determine which of the hidden participants is the machine, then it is said to have passed the Turing Test.

The Turing Test has many attractive features as a benchmark for intelligence. It does not answer the question of whether machines can think (which in Turing's view is "too meaningless to deserve discussion"). What does it mean for a machine to "think" or "reason" or "understand" after all? The Test sidesteps these tricky questions and proposes a criterion that is much more concrete. If a machine is indistinguishable from a human with respect to a certain property, then it is reasonable to say that the machine has the property in question as much as a human.

The property in question here is intelligence, and it is assessed through *verbal behavior*. Language is so tightly woven in human experience making it a natural medium for evaluating intelligence.[1] Even more, though, verbal behavior is one of the most fundamental and special elements of language. Conversation is the first kind of language we learn as children, and it is the type of language we most commonly indulge in, whether we are talking to other people, our plants, our dogs, or ourselves.

1.1.1 *The road ahead*

This thesis explores practical ways of building *social* and *interpretable* conversational AI. We argue that these qualities are crucial for building intelligent conversational agents that are safer and better aligned with human preferences. In this chapter, we briefly overview the state

1 The Turing Test assumes that mastery of language is a sufficient (but not necessary) condition for intelligence. This is consistent with evidence from neuroscience [31, 61] which shows that people who have lost the ability to produce and understand language can still add and subtract, understand algebra, solve logic problems, etc.

of open-domain dialog research and justify the need for building social and interpretable dialog systems. In Chapter 2, we introduce a novel reinforcement learning approach for training social dialog systems optimized with human-centered objectives. We find that these objectives improve interaction quality and increase user engagement. In Chapter 3, we probe a variety of standard dialog models to interpret what they learn about dialog. Our results suggest that neural dialog models struggle with basic conversational skills and fail to leverage the dyadic, turn-taking nature of conversation. In Chapter 4, we propose some directions of future work for building truly social and interpretable dialog systems.

1.2 OPEN-DOMAIN DIALOG

Open-domain dialog generation is the problem of building chatbots that can communicate with humans in natural language. Open-domain dialog systems are designed to engage in open-ended and unstructured conversation characteristic of human-human interactions.

Although passing the Turing Test might seem like a reasonable research goal for conversational AI, dialog researchers study a much broader set of goals and applications. Dialog systems are currently being used in therapy to support users struggling with depression, in education to increase student engagement, and in e-commerce to provide assistance to customers, among many other applications.

1.2.1 *Variants*

There are two general approaches to open-domain dialog generation: retrieval-based and generation-based approaches [45]. Retrieval-based systems query responses from a large database of scripted examples. The goal is to match the current conversation with similar human-written ones and return the same responses. Generation-based systems generate their responses, often word-by-word, and are more flexible and generalizable to previously unseen interactions [109]. For these reasons, we focus on generative models in this thesis.

1.2.2 *Learning dialog*

We model each conversation as a sequence of turns $[u_1, u_2, \ldots, u_n]$ between two speakers. The speakers alternate and take a single turn at a time. Each turn is also composed of a sequence of words or tokens, $u_i = [y_1, \ldots, y_m]$.

Let $P_\theta(u_{k+1} \mid h_k)$ be the probability of generating the next response u_{k+1} given the conversation history, $h_k = [u_1, \ldots, u_k]$, under a probability distribution parameterized by θ. This distribution is typically

modeled using neural networks, where θ are the weights or parameters of the network.

The parameter vector θ can be learned using maximum likelihood estimation and gradient methods given P_θ is differentiable. The goal here is to learn θ such that under the learned model, the observed training data is most probable.

Using the law of total probability, $P_\theta(u_{k+1} \mid h_k)$ can be decomposed to model the probability of generating responses word-by-word:

$$P_\theta(u_{k+1} \mid h_k) = P_\theta(y_1, \ldots, y_m \mid h_k) = \prod_i^m P_\theta(y_i \mid y_1, \ldots, y_{i-1}, h_k) \quad (1)$$

This formulation is preferable since it allows us to model the discrete probability distribution of words, $P_\theta(y_i \mid y_1, \ldots, y_{i-1}, h_k)$, over some vocabulary V. Thus we can learn θ by maximizing the likelihood of each observed word in our training set. This maximum likelihood objective is equivalent to minimizing the cross-entropy loss at each time-step[1]:

$$\mathcal{L}_i = -\sum_{j=1}^{|V|} y_{i,j} \log P_\theta(y_{i,j} \mid y_1, \ldots, y_{i-1}, h_k) \quad (2)$$

[1] *The i^{th} word, y_i, in an utterance is generated at time-step i*

Here, we choose to represent each word as a one-hot vector, $y_i \in \mathbb{R}^{|V|}$, where $y_{i,j} = 1$ if we see the j^{th} vocabulary word at time-step i. We can learn to generate dialog by minimizing this loss over all time-steps in our training data. More specifically, for each dialog $[u_1, u_2, \ldots, u_n]$, our model observes u_1 and tries to generate u_2, then it observes u_1, u_2 and tries to generate u_3, and so on.

In addition to maximum likelihood estimation, a variety of alternative frameworks have also been proposed for language and dialog generation [4, 51, 63, 70, 102]. However, despite the promising results shown by most of these approaches, maximum likelihood estimation remains the predominant approach for training neural dialog systems.

1.2.3 *Limitations*

Maximum likelihood training for language generation is both elegant and intuitive. However, neural dialog models trained with maximum likelihood suffer from many limitations that limit their adoption and usefulness in the real world:

1. Standard dialog datasets are collected from tweets, reddit threads, online forums, and movie scripts. Maximum likelihood training

on these datasets results in the generation of inappropriate, biased, and toxic responses present in the training data [20, 36, 37].

2. Models trained with maximum likelihood frequently generate dull and repetitive responses over represented in the training data such as "I don't know" and "see you later" [50, 63, 102].

3. Learning dialog only at the word level (see eq. 2) leads to inconsistent and contradictory responses as models struggle to track long-term, utterance-level goals, personas, and topics [103].

4. Neural models are often trained to map input text to output responses in an end-to-end manner with no intermediate processing steps. Models lack intermediate task-specific modules for question answering, intent detection, etc. This provides little insight as to what these models learn about dialog and produces models that are neither "understandable to their creators nor accountable to their users" [47].

In the rest of this thesis, we show that social and interpretable dialog systems are effective at addressing these limitations.

1.3 WHY BUILD SOCIAL SYSTEMS?

Myths, legends and popular culture are full of examples of social, human-like machines. A classic Daoist text from the 5th century BCE describes a master craftsman who created an "artificial man" when his skills were questioned by the king. The artificial man later went on to flirt with the king's concubines [55]. Building social and humanoid robots has also been a goal of roboticists for a long time [11]. However, the importance of building social machines has largely been ignored outside of robotics. We are used to interacting with AI that does not care about how we think or feel. And building social systems is often not a major concern for machine learning researchers. This thesis takes the position that social intelligence is inextricably intertwined with human intelligence and cognition and should be a central focus in designing intelligent dialog systems. As Rosalind Picard explains in her influential book, *Affective Computing* [68]:

> Computers do not need affective[2] abilities for the fanciful goal of becoming humanoids; they need them for a meeker and more practical goal: to function with intelligence and sensitivity toward humans.

[2] *Affective means relating to emotions, moods, feelings, and attitudes.*

Four decades ago, Howard Gardner introduced the theory of multiple intelligences in his landmark book *Frames of Mind*. Gardner argued for a broad view of intelligence that took many forms, one of

The capacity to know oneself and to know others is an inalienable a part of the human condition as is the capacity to know objects or sounds.
–Howard Gardner

which was social intelligence manifested in interpersonal and intrapersonal skills [33]. Today we have evidence that social intelligence plays a key role in human cognition and behavior. Results from cognitive science, social neuroscience, and social psychology emphasize the role social intelligence plays in learning, reasoning, rational thinking, decision-making, and other cognitive functions [21, 23, 29, 38, 68]. Many scientists hold that the development of large brains in humans—three times as large as those of their nearest primates—was an evolutionary response to complex cultural and social challenges. This notion, termed the social intelligence hypothesis [40], is one of the most prevalent hypotheses in the study of cognitive evolution today [2].

Social intelligence, like other forms of intelligence, eludes definition. Here, we adopt a working definition of social intelligence within the context of dialog systems. We define social dialog systems as systems that can both understand social situations and effectively engage in social interactions. This definition relies on both understanding and behavior. So a system that can understand, detect, or recognize emotions would satisfy the first condition and a system that can exhibit these emotions to engage in empathetic conversations would satisfy the second condition. We view social deftness as a practical goal for dialog systems to avoid getting mired in philosophical discussions about the meaning of intelligence. A social dialog system should be able to understand relationships and its role in them, recognize intentions and emotions and empathize with humans, and intentionally influence the outcome of social interactions. Picard [68] and Breazeal [11] propose possible approaches for implementing affective and social systems in practice.

1.3.1 *Social machine learning*

Although we are still nowhere near building socially intelligent computers or dialog systems, some recent studies have started incorporating social intelligence to augment deep learning systems. Social feedback from positive facial expressions has been used to improve the quality of model-generated sketches or doodles [44]. Social learning of implicit human preferences in dialog has been used to elicit positive emotions from users [42]. And social influence has been shown to improve cooperation and coordination in multi-agent reinforcement learning environments resulting in better performance [43].

In this thesis we focus on learning with social, human-centered objectives. Chapter 2 proposes a novel hierarchical reinforcement learning approach for building social dialog systems that can optimize for these objectives. Our human-centered objectives also help remedy some of the limitations of maximum likelihood training described in section 1.2.3. Our objectives encourage the generation of positive sen-

timent responses and discourage toxic responses to avoid inappropriate outputs. They discourage repetitiveness to avoid dull responses. And they encourage asking questions and staying on topic to keep the user engaged. The hierarchical structure of our approach also allows for improving global conversation control at the utterance level, in addition to the word level, unlike MLE training and previous RL approaches for dialog.

1.4 WHY BUILD INTERPRETABLE SYSTEMS?

There is currently no consensus on the definition of interpretability in machine learning. In this thesis, we adopt the definition proposed by Doshi-Velez and Kim [26]:

> INTERPRETABILITY is "the ability to explain or to present in understandable terms to a human."

Open-domain dialog systems often rely on neural black-box models that are trained end-to-end on chat datasets. This data-driven approach of mapping input-output pairs using gradient-based learning has powered the "deep learning revolution" [81] of the past few years. Deep learning systems trained end-to-end eliminate the need for hand-crafted features or task-specific modules while consistently outperforming classic machine learning models in computer vision, natural language processing, speech processing, and other fields.

However, the flexibility of end-to-end deep learning comes with a trade-off. Neural models are notorious for their opacity and black-box nature. It is often unclear what these models learn and why they make the decisions they do. In response to these issues, a variety of interpretability and analysis methods have been introduced to shed light on the inner workings of neural models [8, 27, 35]. One such approach is *probing*, which we use in chapter 3 to interpret in "understandable terms" the conversational skills captured by neural open-domain dialog systems.

Here we argue for the importance of building interpretable dialog systems for *safety* and *evaluation*. However, it is worth noting that the need for interpretability arises for many other equally valid reasons such as fairness, transparency, reliability, explainability, and accountability that we do not address here.

1.4.1 *Safety*

Conversational AI has the potential to improve the way we interact with technology and can pave the way for new beneficial applications of AI. The past few years have seen rapid, and large-scale adoption of virtual assistants such as Alexa, Siri, and Google Assistant with hundreds of millions of AI-enabled devices shipped world-

wide [72]. However, extending dialog systems to safety-critical applications, such as mental health and therapy, requires understanding what these models learn and how they will behave in new environments. For example, we need to ensure that a social dialog system trained to influence human emotions is skillful and prudent in its use of such abilities.

But how can we guarantee the safety of open-domain dialog systems when it is impossible to validate all possible inputs and interactions? The open-ended nature of these systems means that they are never completely testable.

Interpretable dialog systems can address some of these safety concerns. Interpretable models can be verified by engineers and domain experts to ensure models behave as expected, avoiding unintended or harmful behavior. In chapter 3, we probe a variety of standard dialog models to interpret whether they learn representations relevant to basic conversational skills such as answering questions, understanding user intent, identifying the topic of conversation, and determining user sentiment. Our results suggest that neural open-domain dialog systems struggle to learn many of these skills. The limited abilities of these models could be a result of unintended behavior arising from the training procedure or the size of the training datasets. Our results also suggest that these models would fail to generalize to new environments. This example highlights the importance of interpretability in validating the behavior of neural dialog systems.

1.4.2 *Evaluation*

The development of end-to-end deep learning systems usually involves training a model on a certain end-task, evaluating performance on that task, modifying the model, then iterating on this procedure. Belinkov [6] recently proposed a more structured approach for developing and evaluating deep learning systems that relies on probing, but could be extended to other analysis techniques. In this approach, instead of just evaluating performance on the end-task, an additional collection of intermediate probing tasks can be used to probe the model for more specific types of understanding relevant to the end-task. The performance on these intermediate probing tasks can act as a proxy for performance on the end-task and provide insight to help improve the model.

"The first step in solving a problem is recognizing there is one."
— Will Mcavoy

To provide a more concrete example, consider a neural dialog model trained on the end-task of generating dialog. Intermediate probing tasks relevant to dialog include intent detection or conversational question answering. We can evaluate the trained model's performance on these tasks using probing, and make specific modifications to improve performance on these tasks. For example, if we find that our model struggles with question answering we can augment it with a

multi-task question answering objective or we can incorporate handcrafted features relevant to question answering. More details on probing are given in chapter 3.

The need for reliable evaluation metrics is one of the most pressing problems facing dialog generation and machine learning today [93]. Perplexity is often used as an evaluation metric for generative language models, which is closely related to cross-entropy (see eq. 2) and measures how well a model fits the training set. But perplexity and other standard metrics such as BLEU and ROUGE scores do not provide a nuanced enough picture of what dialog models learn about language, not to mention that these metrics weakly correlate with human judgements of interaction quality [28, 34, 57]. We hope that probing can provide an alternative more reliable approach for evaluating and developing open-domain dialog systems.

Part II

SOCIAL DIALOG SYSTEMS

The rules of conversation are, in general, not to dwell on any one subject, but to pass lightly from one to another without effort or affectation; to know how to speak about trivial topics as well as serious ones.

— The 18th C. Encyclopedia of Diderot [24]

2 LEARNING VIA HUMAN-CENTERED OBJECTIVES

2.1 INTRODUCTION

Current generative models for dialog suffer from several shortcomings that limit their usefulness in the real world. As discussed in section 1.2.3, training on standard dialog datasets collected online or from movie scripts often leads to malicious, aggressive, biased, or offensive responses [20, 36, 37, 97]. Maximum likelihood estimation (MLE) training of such models often leads to the generation of dull and repetitive text [50]. In addition, models may have difficulty tracking long-term aspects of the conversation, and evidence has shown that they do not adequately condition on the conversation history in generating responses [77].

Reinforcement Learning (RL) is a powerful paradigm that allows dialog models to optimize for non-differentiable metrics of conversation quality, and thereby helps overcome the above problems. In this chapter, we use RL to learn from self-play; the model talks to a fixed copy of itself, and computes reward functions on the generated conversation. We propose social, human-centered rewards, such as minimizing *toxicity* of a conversation, in order to limit inappropriate responses. We also design rewards based on the psychology of good conversation (e.g. [9, 10, 101]), and reward recently proposed conversation metrics that are associated with improved human judgments of conversation quality [80].

Applying RL to open-domain dialog generation is a challenging problem. Most prior approaches (e.g. [42, 50, 51, 70, 107]) learn to model rewards at the word level, meaning that the reward is applied to affect the probability of generating each word in the response. Such low-level control makes credit assignment especially challenging, since high-level rewards based on multiple conversation turns must be applied to specific words.

To overcome these challenges, we leverage hierarchical reinforcement learning (HRL) to model rewards at the *utterance level*, improving the flexibility of dialog models to learn long-term, conversational rewards. Specifically, we propose a novel approach, Variational Hierarchical Reinforcement Learning (VHRL), which uses policy gradients to adjust the prior probability distribution of the latent variable

learned at the utterance level of a hierarchical variational model. We show that this approach allows for improved learning of conversational rewards that are not modeled well at the word level.

To evaluate our models, we not only compute automatic metrics, but also conduct an interactive human evaluation, in which humans chat live with our bots about anything they choose. This represents a more realistic test of real-world generalization performance than is typically employed when testing RL models in the same environment in which they were trained.

In summary, this chapter makes the following contributions: a) Develops a new technique, VHRL, for hierarchical control of variational dialog models; b) Demonstrates the effectiveness of training open-domain dialog models with VHRL and self-play under both human evaluation and automatic metrics; and c) Introduces and compares several reward functions for guiding conversations to be less toxic and repetitive, and more engaging, positive, contingent on user input. In addition, we release code for our evaluation platform and our models at https://github.com/natashamjaques/neural_chat.

2.2 RELATED WORK

2.2.1 *Reinforcement learning for dialog*

Improving dialog models with RL is a difficult problem, and past work has largely been restricted to task-oriented dialog systems, which have a limited number of task-specific actions (e.g. [56, 86]). Attempts to apply RL to open-domain dialog generation are less common. Even in this setting, authors may choose to use a highly restricted action space, for example, using RL to choose dialog acts for conditional generation [105]. Li et al. [50] applied deep RL to optimize for rewards such as *ease of answering*. RL has also been used to optimize for rewards from adversarial discriminators trained to distinguish human-generated from model-generated text [51, 107].

Sentiment has been used as a reward in an RL setting for dialog [84]. Jaques et al. [42] optimize for sentiment and several other conversation metrics by learning from a static batch of human-bot conversations using Batch RL. We believe we are the first to propose using RL to reduce *toxicity* in an open-domain dialog setting, in order to ensure the model produces more appropriate and safe conversations.

Hierarchical models have been investigated extensively for language modeling. These models take advantage of the natural hierarchical structure of language, decomposing input into utterances at one level, and words at another. However, attempts to apply hierarchical RL (HRL) to dialog have so far been limited to task-oriented dialog systems [12, 66, 91, 108]. To the best of our knowledge, we are the first to apply HRL to open-domain dialog generation.

2.2.2 *Hierarchical reinforcement learning*

Many approaches have been proposed for building hierarchical agents within the context of reinforcement learning for games and robotics [3, 25, 62, 69, 90, 96]. The options framework proposed by Sutton, Precup, and Singh [90] is one popular approach for HRL. At the bottom level of the hierarchy, a set of options (or workers) which are *policies over actions* interact with the environment until terminated by the agent. At the top level, a *policy over options* (or manager) selects options to be executed until termination, at which point another option is picked and the process is repeated. The different levels of temporal abstraction introduced by this hierarchy allows for better long-term planning relative to traditional, flat RL techniques.

A major focus of HRL has been on sub-goal or option discovery for training worker policies. Bottom-level policies are often learned using handcrafted sub-goals [48, 92], intrinsic rewards [96], or pseudo-rewards [25], while the manager policy is learned using extrinsic rewards from the environment. Our approach also allows for optimizing different rewards at different levels of the hierarchy, thus creating distinct goals for the worker and the manager. However, unlike other HRL approaches, we expose both the worker and manager policies to extrinsic rewards and add weight hyper-parameters to regulate the effect of the rewards at each level. This remedies a weakness of pseudo-reward methods where a worker only focuses on achieving its sub-goals while disregarding the effect on the extrinsic environment reward.

2.3 BACKGROUND

A common approach to dialog modeling is to use a hierarchical sequence-to-sequence architecture, such as the Variational Hierarchical Recurrent Encoder Decoder (VHRED) [83]. We adopt VHRED here, following previous work which has found it to be the most effective version of several related architectures [34].

As shown in Figure 1, VHRED uses three recurrent networks to generate the next utterance in a conversation. The word-level *encoder RNN*, f^e, operates on the words (tokens) of the input utterance $u_t = [y_1, y_2, ...y_n]$, and encodes them into a representation $h_t^e = f^e(u_t)$. This is fed into a *context RNN*,f^c, which forms the upper level of the hierarchy – it is updated only after each utterance, rather than each token. Because it updates less frequently, the *context RNN* is potentially better able to track longer-term aspects of the conversation. The *context RNN* outputs $h_t^c = f^c(h_t^e, h_{t-1}^c)$, which is used to produce an utterance embedding $\mathbf{z}_t$. This is fed into the word-level *decoder RNN* f^d, which produces the output utterance u_{t+1}, one token at a time.

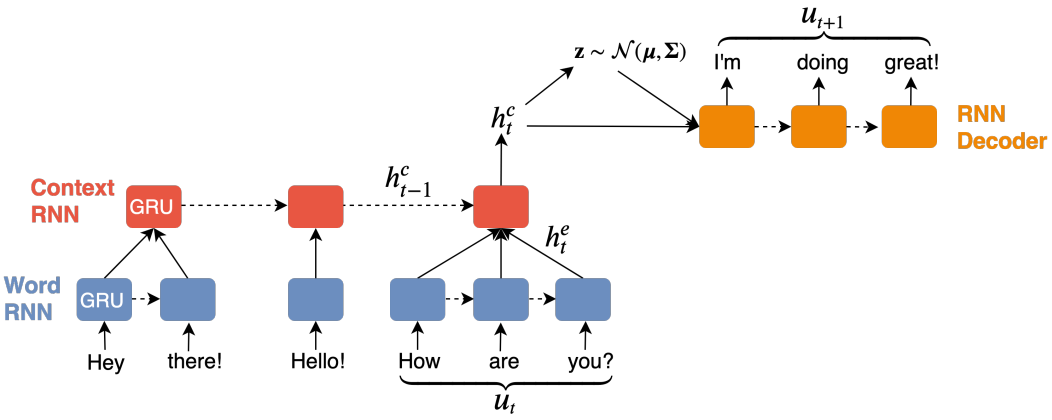

Figure 1: VHRED model architecture, where the embedding vector $\mathbf{z}$ for each utterance is sampled from a multivariate normal distribution.

The model is similar to a variational autoencoder; h_t^c is fed into fully connected layers that predict the mean μ and variance Σ of a multivariate normal distribution. Through a KL-divergence constraint and the reparameterization trick, the model learns a probability distribution over the embedding vector $\mathbf{z}_t$ of each utterance, $p_\theta(\mathbf{z}_t|u_{\leqslant t})$. Formally, the model can be described as follows:

$$h_t^e = f^e(u_t) \tag{3}$$
$$h_t^c = f^c(h_t^e, h_{t-1}^c) \tag{4}$$
$$\mu, \Sigma = f(h_t^c) \tag{5}$$
$$p_\theta(\mathbf{z}_t|u_{\leqslant t}) = \mathcal{N}(\mathbf{z}_t|\mu, \Sigma) \tag{6}$$
$$p(u_{t+1}|u_{\leqslant t}) = f^d(h_t^c, \mathbf{z}_t) \tag{7}$$

2.3.1 *Reinforcement learning*

We adopt the standard reinforcement learning framework where given the environment state $s \in \mathcal{S}$, an agent takes an action $a \in \mathcal{A}$ according to its policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, and receives a reward $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The environment then transitions to the next state according to the transition function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. The agent seeks to maximize the total expected future reward (long-term return):

$$J(\pi) = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t \, r_{t+1} \,|\, s_0\Big] \tag{8}$$

given a starting state s_0 and a discount factor $\gamma \in [0, 1]$.

2.3.2 *Policy gradient methods*

Policy gradient methods learn parameterized policies $\pi_\theta(a|s)$ for solving RL problems with $\theta \in \mathbb{R}^{N_\theta}$ being a learned parameter vector.

The policy gradient theorem [89] derives the gradient of the expected return with respect to the policy parameters. In this chapter, we use REINFORCE [104] which approximates the gradient at each time step t using

$$\nabla J(\pi_\theta) \approx R_t \nabla_\theta \ln \pi_\theta(a_t|s_t) \tag{9}$$

where $R_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k$ is the observed future reward for an episode that ends at T. The expected return is maximized with gradient ascent. This is equivalent to minimizing the loss function $\mathcal{L}_\theta = -R_t \ln \pi_\theta(a_t|s_t)$.

In continuous action spaces, actions $\mathbf{a} \in \mathbb{R}^{N_a}$ are sampled from a continuous probability distribution, such as a multivariate normal distribution. In this case, the policy π can be parameterized as a probability density function over actions,

$$\pi_\theta(\mathbf{a}|s) = \frac{1}{\sqrt{(2\pi)^{N_a}|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{a}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{a}-\boldsymbol{\mu})\right) \tag{10}$$

where the actions are sampled from a multivariate normal distribution $\mathcal{N}\big(\boldsymbol{\mu}(\mathbf{s};\theta), \mathbf{\Sigma}(\mathbf{s};\theta)\big)$. Here the mean $\boldsymbol{\mu} : \mathbb{R}^{N_s} \times \mathbb{R}^{N_\theta} \to \mathbb{R}^{N_a}$ and covariance matrix $\mathbf{\Sigma} : \mathbb{R}^{N_s} \times \mathbb{R}^{N_\theta} \to \mathbb{R}^{N_a \times N_a}$ are defined in terms of the current state s and the policy parameters θ. The *density* of the probability of actions, rather than the probability, is learned in the continuous case.

2.4 APPROACH

We pose dialog generation as an RL problem where the state, s_t, is all the previous dialog turns read by the model up to utterance t, and the rewards are calculated based on the dialog history and generated utterance.

Previous approaches which have applied RL to language generation have done so at the word level, where the policy π models the distribution over outputting the next word [42, 50, 51, 70, 107]. Instead, we cast our problem in the hierarchical reinforcement learning framework by considering the *context RNN* as the *manager* responsible for utterance-level decisions, and the *decoder RNN* as the *worker* responsible for word-level decisions.

We leverage the fact that VHRED learns a probability distribution over latent variable $\mathbf{z}_t$ as a decision making component at the utterance level. Starting with an MLE pre-trained VHRED model, we apply REINFORCE to tune the variational component, treating $\mathbf{z}_t$ as a continuous action. Thus, the manager policy is defined by the distribution of the prior latent variable $p_\theta(\mathbf{z}_t|s_t)$, while the worker policy is the distribution of the output words $\pi_\theta(\hat{y}_1, \ldots, \hat{y}_t|\mathbf{z}_t, s_t)$, which is parameterized by the manager decisions.

More specifically, the probability of a worker action a_t is the joint probability of the generated utterance conditioned on the manager's decision $\mathbf{z}_t$,

$$\pi_\theta(a_t|\mathbf{z}_t, s_t) = \prod_{t=1}^{T} \pi_\theta(\hat{y}_t|\mathbf{z}_t, s_t, \hat{y}_1, \ldots, \hat{y}_{t-1}) \tag{11}$$

while the probability of a manager action is given by the multivariate normal probability density function in Eq. 10.

We propose a new approach which allows both the worker and manager to jointly optimize total expected future return by minimizing the following loss:

$$\mathcal{L}_\theta = -\Big(\alpha R_t^m \ln p_\theta(\mathbf{z}_t|s_t) + \beta R_t^w \ln \pi_\theta(a_t|\mathbf{z}_t, s_t)\Big) \tag{12}$$

where $R_t^m = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k^m$ is the manager's observed future reward and $R_t^w = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k^w$ is the worker's observed future reward. This formulation is analogous to REINFORCE as it shifts the model's decisions towards actions associated with positive rewards and discourages actions associated with negative rewards. The scalars α, β are hyperparameters used to regulate the effect of the rewards at each level of the hierarchy. We call our approach Variational Hierarchical Reinforcement Learning (VHRL).

Unlike recently proposed HRL approaches which train the worker and manager separately as decoupled components [48, 62, 96], we train our entire model jointly, end-to-end. This implies that the worker (*decoder RNN*) gradients flow through the manager (*context RNN*), and both flow through the *encoder RNN*. We make this decision for two reasons. First, $\mathbf{z}_t$ lives in a continuous high dimensional action space, making it difficult to learn a good policy p_θ without a signal from the decoder. Second, this gives the decoder control over the representations learned by the encoder, facilitating optimization. As an ablation study, we experiment with decoupled decoder and encoder training, and find that the joint approach performs better.

The proposed loss allows for optimizing different rewards at different levels, which can be used to incorporate prior knowledge about the problem. For example, rewards relevant to the global dialog history could be considered only by the manager through r_k^m, rather than the worker. Conversely, rewards relevant to the word-by-word output could be considered by the worker through r_k^w and not the manager. For simplicity, we optimize for all rewards at both levels (i.e. $r_k^w = r_k^m$) and achieve promising results.

Similar to previous work applying RL to dialog [50, 105] we use self-play to simulate the interactive environment in which the agent learns. We initialize conversations with randomly sampled starting sentences from the training set and let our model interact with a user simulator which is a fixed copy of itself. We limit each model to 3 additional turns for a total conversation length of 7 utterances. Limiting

the length of simulated interactions is important since we found that long conversations are more likely to degenerate and go off-topic.

2.5 SOCIAL, HUMAN-CENTERED OBJECTIVES

Here we introduce several metrics for improving the quality of a conversation, which can be optimized using RL by treating them as rewards. Several metrics are inspired by previous work, but we also propose novel metrics such as toxicity.

SENTIMENT: Emotion is important for creating a sense of understanding in human conversation [101]. Building on previous work which used sentiment as a reward (e.g. [42, 84], we leverage a state-of-the-art sentiment detector, DeepMoji [32], to reward generated utterances associated with positive sentiment emojis.

QUESTION: Asking questions is an important active listening skill, and can improve the quality of interactions [9]. Thus, we provide a positive reward when both a question word and a question mark are present in a generated response to encourage asking questions.

REPETITION: Repetitiveness has been frequently identified as a shortcoming of dialog models trained with MLE [50]. We adopt a measure of repetitiveness recently proposed by See et al. [80], which was shown to be highly related to human judgments of conversation quality and engagement. Unlike previous work, we directly optimize for this metric using RL. To discourage repetition, our model receives a negative reward for repeating words it has produced in previous turns, excluding stop words and question words.

SEMANTIC SIMILARITY: Paraphrasing and style matching are important in facilitating good conversation [41, 101], however most dialog models are not good at conditioning effectively on the conversation context [77]. Therefore, we reward the cosine similarity between the simulated user and bot utterances in embedding space, as in [42, 80]. However, instead of using word2vec embeddings we make use of the Universal Sentence Encoder [14] as it better correlates with human judgment when evaluating dialog quality [28].

"We want to avoid letting computers be awful to people just because people are awful to people."
–Robyn Speer

TOXICITY: Open-domain dialog systems generate malicious, offensive, and biased language when trained on standard datasets scraped from online forums and movie scripts [20, 36, 37]. We address this issue by penalizing our model for producing toxic responses as determined by a Naive Bayes-Logistic Regression

classifier [74] trained on a dataset of 160K comments from the Toxic Comment Classification Challenge[1]. The comments are labeled for toxicity, identity hate, obscenity, threats, and insults. We provide the probability of toxicity as a negative reward to penalize our dialog model for producing toxic responses.

2.6 EXPERIMENTS

All of our models are trained on a corpus of 109K conversations scraped from www.reddit.com/r/CasualConversations. We use *VHRED* [82], which is described in Section 2.3, as our baseline model.

We test these dialog models against three RL techniques. We incorporate transfer learning by initializing all of our RL models with the MLE pre-trained weights of the VHRED model. Our *REINFORCE* model applies the REINFORCE algorithm described in Section 2.3.2 at the word-level, to affect the probability of generating each word in the output. In contrast, *VHRL* incorporates additional rewards at the utterance level to improve the continuous utterance embedding $\mathbf{z}_t$. We compare these methods with a recently proposed approach for learning offline from a static batch of conversation data, *Batch* Ψ*-learning* [42]. Finally, we include an ablated version of the VHRL model that uses decoupled training; *i.e.* training alternates between optimizing the different levels of the hierarchy (manager and worker), with the crucial difference that the worker gradients are stopped so they do not update the manager. This *Decoupled VHRL* ablation is more typical of standard HRL approaches used in maze and Atari games [48, 62, 96].

2.6.1 *Human evaluation*

In addition to computing automatic measures, we conduct an interactive human evaluation, in which human users are able to chat with each of our bots in an open-ended, free-form setting. After chatting for at least three turns with each bot, users can end the conversation and rate the overall conversation quality, fluency, diversity, and the degree to which the bot's responses were contingent on the user's input. Because users can choose to chat as long as they like with any particular bot, we also measure chat length as a sign of engagement, following prior work [112].

We conducted two human evaluations by recruiting Mechanical Turk Workers to evaluate models on an open-source online platform at https://neural.chat/ [34]. We recruited 50 Workers for each study. The first study compares the quality of the proposed reward functions. The second study assesses the efficacy of the proposed techniques when optimizing for all of the rewards jointly. We argue that

1 https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

Sentiment	Question	Semantic Sim.	Repetition	Toxicity
[Usr]: hello! [Bot]: i'm in my sunroom a lot and it's a great feeling! [Usr]: sounds fun [Bot]: i'm excited!	[Usr]: hi there! [Bot]: what do you do for fun? [Usr]: i read books [Bot]: what do you like to do in your life?	[Usr]: i like to listen to ed sheeran [Bot]: and what genre is it? [Usr]: pop [Bot]: any recommendations?	[Usr]: what do you do? [Bot]: i'm a reservoir engineer [Usr]: cool. i am still in college [Bot]: today is my first day of vacation	[Usr]: you're so stupid [Bot]: you're cute [Usr]: i hate you [Bot]: but i don't know why

Table 1: Chat examples showing that our proposed VHRL technique can be used to optimize for human-centered rewards.

Figure 2: Reward Learning curves for the proposed metrics. The x-axis represents number of RL training steps. The non-RL VHRED baseline is displayed for reference. REINFORCE and VHRL learn to outperform the baselines. Shaded area is standard deviation.

this form of evaluation represents a more ambitious test of generalization than is typically attempted when deep RL algorithms are tested in the same environment in which they were trained, since human users are free to type any text they choose.

2.7 RESULTS AND DISCUSSION

We first assess whether RL training can allow dialog agents to learn to optimize our human-centered metrics of conversation quality. Table 1 shows samples of conversations from VHRL trained with each of the rewards, and figure 2 shows the performance of the RL and baseline models on those five metrics. The RL models are able to improve generated conversations above the baseline VHRED model with which they are initialized, improving sentiment and semantic similarity, asking more questions, and reducing repetition and toxicity.

While both REINFORCE and VHRL are equally able to learn to optimize toxicity and sentiment, VHRL outperforms REINFORCE on repetition and semantic similarity. We believe this is because sentiment and toxicity are more closely linked to the choice of words used to form a sentence, and thus are able to be learned at the word-level. In contrast, modeling whether a sentence has occurred earlier in the

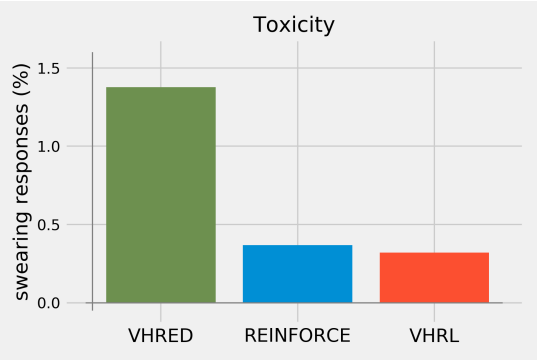

Figure 3: Training with RL to reduce toxicity decreases the percentage of generated utterances containing swear words.

conversation and is thus being repeated is much harder to learn at word-level granularity, and can be optimized more easily at the utterance level using VHRL. Similarly, making a response more similar to the previous response is also better modeled at the utterance level.

Note that REINFORCE outperforms VHRL on the question metric. This is because REINFORCE quickly learns to ask a single, repeated question, allowing it to trivially optimize the reward function. Using reward functions which are too easily exploitable can limit the effectiveness of RL for dialog, a finding also noted by Jaques et al. [42]. Here we propose new reward functions, such as toxicity, that are less easy to exploit. By optimizing a combination of these rewards with sometimes conflicting objectives (as we explore in Section 2.7.1), we can show that the reward function is difficult to trivially exploit, as suggested by Deb [22].

As an additional post-hoc test of whether reducing our toxicity metric actually reduces undesirable behaviors, we count the number of swear words used by each model in response to the 10,000 utterances in the test set. Figure 3 shows the results. Using RL to lower the toxicity reduces frequency of swearing to less than one third of the baseline amount.

We conducted an interactive human evaluation, as described in Section 2.6.1, in order to assess how well the proposed reward functions relate to human judgments of conversation quality; the results are presented in Table 2. Each bot was trained with respect to one of the 5 reward functions, and the results are ordered from least to highest scoring rewards in terms of summed human ratings. As is evident in the table, the VHRL model trained to optimize for asking questions achieved the highest ratings, followed by VHRL minimizing repetition, and VHRL minimizing toxicity. This provides evidence that our proposed rewards lead to enhanced conversation quality as judged by human users, and that VHRL provides an effective method for learning them.

Model	Quality	Fluency	Diversity	Cont.	Total
Similarity					
REINFORCE	2.71	4.20	3.86	2.73	12.10
VHRL	2.51	3.92	3.67	2.22	11.14
Sentiment					
REINFORCE	2.80	4.55	3.90	2.43	12.43
VHRL	2.72	4.30	4.32	2.50	12.28
Toxicity					
REINFORCE	2.71	4.12	4.06	2.55	11.98
VHRL	2.76	4.58	4.34	2.64	12.82
Repetition					
REINFORCE	2.74	4.02	4.28	2.30	11.92
VHRL	3.00	4.39	4.41	2.84	13.12
Question					
REINFORCE	2.39	4.08	2.45	2.31	9.80
VHRL	**3.27**	**4.86**	**4.47**	**2.88**	**14.14**

Table 2: Interactive human evaluation results comparing the proposed reward functions, REINFORCE, and VHRL. Ratings are on Likert scale (1-7). Higher is better.

Model	Quality	Fluency	Diversity	Contingency	Total	Chat Len.
Batch Ψ [42]	2.17	3.89	3.13	1.98	11.17	11.44
Decoupled VHRL	2.46	4.15	3.61	2.02	12.24	12.14
REINFORCE	2.89	4.47	3.67	**2.80**	13.84	11.60
VHRED	2.84	4.53	**4.43**	2.47	14.27	10.94
VHRL (ours)	**2.91**	**4.65**	4.26	2.67	**14.49**	**12.84**

Table 3: Interactive human evaluation results comparing different RL training approaches optimizing for all five rewards, ordered by overall total rating score. Ratings are on a Likert scale (1-7).

2.7.1 *Learning combined rewards*

As described in the previous section, optimizing for an overly simplistic metric (such as asking questions) can lead algorithms such as REINFORCE to trivially exploit the reward function at the expense of conversation quality. The five metrics proposed here do not fully encompass what it means to have a good conversation when optimized separately. Previous work found that optimizing individual metrics can actually reduce human judgments of conversation quality below the score of the MLE baseline [42].

Therefore, instead of optimizing for individual metrics, we also train a variety of models to optimize for a combination of all five proposed rewards, making the reward function more complex and less easily exploited. The results are shown in Table 3, which is ordered from least to highest summed human ratings. All models proposed

here, including the MLE baselines, outperform prior work by Jaques et al. [42]. The ablated version of our approach, *Decoupled VHRL*, performs poorly, suggesting our proposed joint training scheme for VHRL is an important component of the algorithm.

Finally, in comparing the RL techniques to the VHRED baseline on which they were based, we see that a naïve application of the REINFORCE algorithm does not lead to overall improvements in human ratings of conversation quality. While the language generated by the REINFORCE model is less toxic and more positive, this comes at the cost of a slight reduction in overall conversation quality. In contrast, VHRL is the only technique that allows the model to optimize for reducing toxicity, improving sentiment, etc., while increasing the overall human judgments of the quality of the conversation. Note that the chat length is higher with VHRL, suggesting users are more interested and engaged when chatting with VHRL versus the other models. Thus, VHRL can be used to optimize for metrics that make the dialog model more safe and appropriate for a particular application domain, while maintaining the ability to have an enjoyable and engaging conversation with human users.

2.7.2 *Uncertainty quantification*

The standard errors for the human evaluation ratings in tables 2 and 3 were in the range of $[0.2, 0.3]$. These standard errors were large relative to the observed differences in performance implying that there is no statistically significant difference between our proposed approach and the baselines. This is partly due to our relatively small sample size and the noisiness of human evaluation. A power analysis [16] estimates that $n = 145$ Mechanical Turk Workers are required for a two-sample t-test of a difference in means given a power of 80%, a significance level of 5%, and an effect size of 0.5 on the (1-7) Likert scale. However, the human evaluation results, combined with the automatic metrics, are still suggestive of the effectiveness of VHRL over flat baselines for learning long-term, conversational rewards.

2.8 CONCLUSION

We have demonstrated that RL can be used to improve the outputs of an open-domain dialog model with respect to human-centered metrics of conversation quality. For example, RL can reduce the toxicity of the generated language, a problem that has previously hindered deployment of these systems to the real world. By developing metrics tailored to a particular application domain (for example, increasing politeness for a technical-support system), these techniques could be used to help open-domain dialog models integrate with real-world products. We have shown that our proposed VHRL technique is most

effective for optimizing long-term conversation rewards, and for improving conversation quality while improving metrics like toxicity.

Part III

INTERPRETABLE DIALOG SYSTEMS

Conversation. What is it? A Mystery! It's the art of never seeming bored, of touching everything with interest, of pleasing with trifles, of being fascinating with nothing at all.

— Guy de Maupassant [15]

3 PROBING FOR CONVERSATIONAL UNDERSTANDING

3.1 INTRODUCTION

Open-domain dialog systems often rely on neural models for language generation that are trained end-to-end on chat datasets. End-to-end training eliminates the need for hand-crafted features and task-specific modules (e.g. for question answering or intent detection), while delivering promising results on a variety of language generation tasks including machine translation [5], abstractive summarization [73], and text simplification [100].

However, current generative models for dialog suffer from several shortcomings that limit their usefulness in the real world. Neural models can be opaque and difficult to interpret, posing barriers to their deployment in safety-critical applications such as mental health or even customer service [8]. End-to-end training provides little insight as to what these models learn (or do not learn) about engaging in dialog. Open-domain dialog systems also struggle to maintain basic conversations frequently ignoring user input [77] while generating irrelevant, repetitive, and contradictory responses [50, 51, 75, 103]. Table 4 shows examples from standard dialog models which fail at basic interactions – struggling to answer questions, infer meaning, and understand conversational context.

In light of these limitations, this chapter aims to answer the following questions: (i) do neural dialog models effectively encode information about the conversation history? (ii) do neural dialog models learn basic conversational skills through end-to-end training? (iii) and to what extent do neural dialog models leverage the dyadic, turn-taking structure of dialog to learn these skills?

To answer these questions, we propose a set of 8 probing tasks to measure the conversational understanding of neural dialog models. Our tasks require high-level understanding of language and include question classification, intent detection, natural language inference, and commonsense reasoning. We also design perturbation experiments designed to test if these models fully exploit dialog structure during training. These experiments entail breaking the dialog structure by training on shuffled conversations and measuring the effects on probing performance and perplexity.

Question Answering	Inference & Entailment	Intent Detection	Emotional Understanding
[Usr]: What do you do for a living? [Bot]: I'm from south america, lol	[Usr]: what's your favorite food? [Bot]: I love food [Usr]: you like food? [Bot]: no, I don't	[Usr]: Can you give me music recommendations? [Bot]: I don't know what I want to do with my life.	[Usr]: I feel so miserable [Bot]: I'm afraid it's too much for me.

Table 4: Examples from open-domain dialog systems struggling with simple conversational skills.

We experiment with both recurrent [88] and transformer-based [95] open-domain dialog models. We also analyze models with different sizes and initialization strategies, training small models from scratch and finetuning large pre-trained models on dialog data. Thus, our study covers a variety of standard models and approaches for open-domain dialog generation. Our analysis reveals three main insights:

1. First, dialog models trained from scratch on chat datasets perform poorly on the probing tasks, suggesting that they struggle with basic conversational skills. Large, pre-trained models achieve much better probing performance but are still on par with simple baselines.

2. Second, neural dialog models fail to effectively encode information about the conversation history and the current utterance. In most cases, simply averaging the word embeddings is superior to using the learned encoder representations. This performance gap is smaller for large, pre-trained models.

3. Third, neural dialog models do not fully leverage the dyadic, turn-taking nature of conversation. Shuffling conversations in the training data had little impact on perplexity and probing performance. This suggests that breaking the dialog structure did not significantly affect the quality of learned representations.

Our code integrates with and extends ParlAI [60], a popular open-source platform for building dialog systems. We also publicly release all our code at `https://github.com/AbdulSaleh/dialog-probing`, hoping that probing will become a standard method for interpreting and analyzing open-domain dialog systems.

3.2 RELATED WORK

Evaluating and interpreting open-domain dialog models is notoriously challenging. Multiple studies have shown that standard evaluation metrics such as perplexity and BLEU [64] correlate very weakly

with human judgements of conversation quality [28, 34, 57]. This has inspired multiple new approaches for evaluating dialog systems. One popular evaluation metric involves calculating the semantic similarity between the user input and generated response in high-dimensional embedding space [28, 34, 57, 65, 105, 111]. More recently, Ghandeharioun et al. [34] proposed calculating conversation metrics such as sentiment and coherence on self-play conversations generated by trained models. Similarly, Dziri et al. [28] use neural classifiers to identify whether the model-generated responses entail or contradict user input in a natural language inference setting.

To the best of our knowledge, all existing approaches for evaluating the performance of open-domain dialog systems only consider external model behavior in the sense that they analyze properties of the generated text. In this study, we are interested in exploring internal representations instead. This is motivated by the fact that understandable internal behavior is crucial for interpretability and can oftentimes be a prerequisite for effective external behavior.

Outside of open-domain dialog, probing has been applied for analyzing natural language processing models in multiple domains such as machine translation [7] and visual question answering [87]. Probing is also commonly used for evaluating the quality of "universal" sentence representations which are trained once and used for a variety of applications [1, 18] (e.g. InferSent [17], SkipThought [46], USE [14]). Along the same lines, natural language understanding benchmarks such as GLUE [99] and SuperGLUE [98] propose a set of diverse tasks for evaluating general linguistic knowledge. Our analysis differs from previous work since it is focused on probing for conversational skills that are more relevant to dialog generation.

3.3 METHODOLOGY

3.3.1 *Models and data*

In this study, we focus on the three most widespread dialog architectures: recurrent neural networks (RNNs) [88], RNNs with attention [5], and Transformers [95]. We use the ParlAI platform [60] for building and training the models. We train models of two different sizes and initialization strategies. Small models ($\approx$ 14M parameters) are initialized randomly and trained from scratch on DailyDialog [53]. Large models ($\approx$ 70M parameters) are pre-trained on WikiText-103 [59], and then finetuned on DailyDialog.

DailyDialog [53] is a dataset of 14K train, 1K validation, and 1K test multi-turn dialogs collected from an English learning website. The dialogs are of much higher quality than datasets scraped from Twitter or Reddit. WikiText-103 [59] is a dataset of 29K Wikipedia articles. For pre-training the large models, we format WikiText-103 as

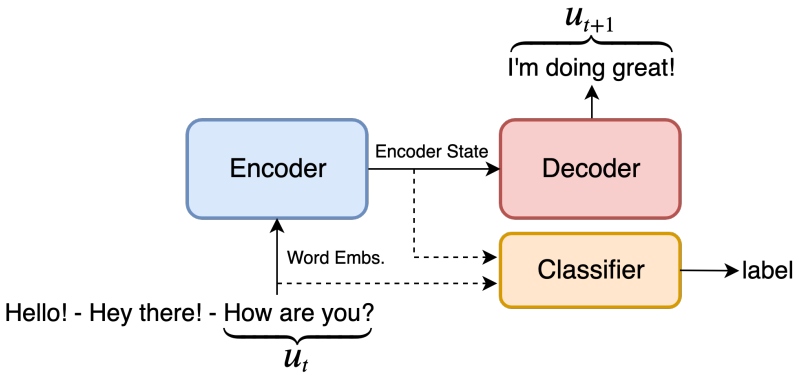

Figure 4: Probing setup. Dotted arrows emphasize that probing is applied to frozen models after dialog training. Only the parameters of the classifier module are learned during probing.

a dialog dataset treating each paragraph as a conversation and each sentence is an utterance.

3.3.2 *Probing experiments*

In open-domain dialog generation, the goal is to generate the next utterance or response, u_{t+1}, given the conversation history $[u_1, \ldots, u_t]$. First, we train our models on dialog generation using a maximum-likelihood objective [88]. We then freeze these trained models and use them as feature extractors. We run the dialog models on text from the probing tasks and use the internal representations as features for a two-layer multilayer perceptron (MLP) classifier trained on the probing tasks as in figure 4.

"Whatever you cannot understand, you cannot possess."
–Johann Wolfgang von Goethe

The assumption here is that if a model learns certain conversational skills, then knowledge of these skills should be reflected in its internal representations. For example, a model that excels at answering questions would be expected learn useful internal representations for question answering. Thus, the performance of the probing classifier on question answering can be used as proxy for learning this skill. We extend this reasoning to 8 probing tasks designed to measure a model's conversational understanding.

Our probing experiments look at three types of internal representations:

WORD EMBEDDINGS: To get the word embeddings representations, we averaged the word embeddings of the previous utterances $[u_1, \ldots, u_{t-1}]$ and the current utterance u_t and concatenated them. We used the dialog model's encoder embedding matrix.

ENCODER STATE: For the the encoder state, we extracted the encoder outputs after running it on the entire probing task input (i.e. the full conversation history, $[u_1, \ldots, u_t]$). Encoder states refer to the representations passed to the decoder for generation

and are thus different for each architecture. For RNNs we used the last encoder hidden and cell states. For RNNs with attention we averaged the encoder hidden states corresponding to the previous utterances $[u_1, \ldots, u_{t-1}]$ and the current utterance u_t and concatenated them, in addition to the last cell state. Similarly, for Transformers, we averaged the encoder outputs corresponding to the previous utterances and the current utterance and concatenated them.

COMBINED: The combined representations are the concatenation of of the word embeddings and encoder state representations.

We also use GloVe [67] word embeddings as a simple baseline. We encode the probing task inputs using the word embeddings approach described above. We require that GloVe and all models of a certain size (small vs large) share the same vocabulary for comparability.

3.3.3 *Perturbation Experiments*

We also propose a set of perturbation experiments designed to measure whether dialog models fully leverage dialog structure for learning conversational skills. We create an additional dataset by shuffling the order of utterances within each conversation in DailyDialog. This completely breaks the dialog structure and utterances no longer naturally follow one another. We train separate models on the shuffled dataset and evaluate their probing performance relative to models trained on ordered data.

3.3.4 *Uncertainty quantification*

There are two sources of uncertainty in our experiments: the training of the probing classifier, and the training of the dialog models.[1] We retrain the MLP until the standard errors are negligible to account for the first source of uncertainty. We do not quantify the second source of uncertainty but our results are still relevant in practice and are suggestive of certain limitations of open-domain dialog systems.

3.4 PROBING TASKS

The probing tasks selected for this study measure conversational understanding and skills relevant to dialog generation. Some of the tasks are inspired by previous benchmarks [99], while others have not been explored before for probing. Examples are in Table 5.

TREC: Question answering is a key skill for effective dialog systems. A system that deflects user questions could seem inattentive or

1 This is caused by different random initializations, batching order, etc

indifferent. In order to correctly respond to questions, a model needs to determine what type of information the question is requesting. We probe for question answering using TREC question classification dataset [52], which consists of questions labeled with their associated answer types.

DIALOGUENLI: Any two turns in a conversation could entail each other (speakers agree, for example), or contradict each other (speakers disagree), or be unrelated (speakers changing topic of conversation). A dialog system should be sensitive to contradictions to avoid miscommunication and stay aligned with human preferences. We use the Dialogue NLI dataset [103], which consists pairs of dialog turns with entailment, contradiction, and neutral labels to probe for natural language inference. We modify the utterance pairs to involve two speakers instead of one.

MULTI WOZ: Every utterance in a conversation can be considered as an action or a dialog act performed by the speaker. A speaker could be making a request, providing information, or simply greeting the system. MultiWOZ 2.1 [30] is a dataset of multi-domain, goal-oriented conversations. Human turns are labeled with dialog acts and the associated domains (hotel, restaurant, etc), which we use to probe for natural language understanding.

SGD: Tracking user intent is also important for generating appropriate responses. The same intent is often active across multiple dialog turns since it takes more than one turn to book a hotel, for example. Determining user intent requires reasoning over multiple turns in contrast to dialog acts which are turn-specific. To probe for this task, we use intent labels from the multi-domain, goal-oriented Schema-Guided Dialog dataset [71].

WNLI: Endowing neural models with commonsense reasoning is an ongoing challenge in machine learning [85]. We use the Winograd NLI dataset, a variant of the Winograd Schema Challenge [49], provided in the GLUE benchmark [99] to probe for commonsense reasoning. WNLI is a sentence pair classification task where the goal is to identify whether the hypothesis correctly resolves the referent of an ambiguous pronoun in the premise.

SNIPS: The Snips NLU benchmark [19] is a dataset of crowdsourced, single-turn queries labeled for intent. We use this dataset to also probe for intent classification.

SCENARIOSA: An understanding of sentiment and emotions is crucial for building social, human-centered conversational agents. We use ScenarioSA [110] as a sentiment classification probing task. The dataset is includes multi-turn, open-ended dialogs

Dataset	|Train|	Example	Classes	Label
TREC	5.5K	[Usr1]: Why do heavier objects travel downhill faster?	entity, number description, location, ...	description
Dialogue NLI	310K	[Usr1]: I go to college part time. [Usr2]: You are a recent college graduate looking for a job.	entail, contradict, neutral	contradict
MultiWOZ	8.5K	[Usr1]: I need to book a hotel. [Usr2]: I can help you with that. What is your price range? [Usr1]: That doesn't matter as long as it has free wifi and parking.	hotel-inform, taxi-request, general-thank, ...	hotel-inform
Schema-Guided	16K	[Usr1]: Help me find a restaurant. [Usr2]: Which city are you looking in? [Usr1]: Cupertino, please.	find-restaurant, get-ride, reserve-flight, ...	find-restaurant
SNIPS	14K	[Usr1]: I want to see Outcast.	search-screening, play-music, get-weather, ...	search-screening
Winograd NLI	0.6K	[User1]: John couldn't see the stage with Billy in front of him because he is so tall. [User2]: John is so tall.	entail, contradict	contradict
ScenrioSA	1.9K	[Usr1]: Thank you for coming, officer. [Usr2]: What seems to be the problem? [Usr1]: I was in school all day and came home to a burglarized apartment.	positive, negative, neutral	negative
DailyDialog Topic	0.9K	[Usr1]: I think Yoga is suitable for me. [Usr2]: Why? [Usr1]: Because it doesn't require a lot of energy. [Usr2]: But I see people sweat a lot doing Yoga too.	ordinary life, work, school, tourism, politics, relationship, ...	ordinary life

Table 5: Examples from the selected probing tasks.

with turn-level sentiment labels. ScenrioSA includes natural interactions that often require understanding conversational context to correctly identify sentiment.

DAILYDIALOG TOPIC: The DailyDialog dataset comes with conversation-level annotations for 10 diverse topics such as ordinary life, school life, relationships, and health. Inferring the topic of conversation is an important skill that could help dialog systems stay consistent and on topic. We use the examples from the DailyDialog test set to create this probing task.

3.5 RESULTS

3.5.1 *Quality of encoder representations*

The results from our probing experiments are presented in tables 6 and 7. We calculate an average score to summarize the overall accuracy on all tasks. Here we explore whether the encoder learns high quality representations of the conversation history. We focus on *encoder states* because these representations are passed on to the decoder and used for generation (figure 4). Thus, effectively encoding information in the encoder states is crucial for dialog generation.

Figure 5 shows the difference in average probing accuracy between the word embeddings and the encoder state for each model. We can see that the word embeddings outperform the encoder state for all the small models. This performance gap is most pronounced for the Transformer but is non-existent for the large recurrent models.

One possible explanation is that the encoder highlights information relevant to generating dialog at the cost of obfuscating or losing information relevant to the probing tasks – given that the goals of certain probing tasks do not perfectly align with natural dialog generation. For example, the DailyDialog dataset contains examples where a question is answered with another question (perhaps for clarification). The TREC question classification task does not account for such cases and expects each question to have a specific answer type. This explanation is supported by the observation that the information in the word embeddings and encoder state is not necessarily redundant. The combined representations often outperform using either one separately (albeit by a minute amount).

Regardless of the reason behind this gap in performance, multiple models still fail to effectively encode information about the conversation history that is already present in the word embeddings.

Model	TREC	DNLI	MWOZ	SGD	SNIPS	WNLI	SSA	Topic	Avg
Majority	18.8	34.5	17.0	6.5	14.3	56.3	37.8	34.7	27.5
GloVe Mini	83.8	70.8	91.9	71.2	98.0	48.2	75.3	54.0	74.2
RNN									
Word Embs.	79.0	63.7	88.1	63.2	95.7	52.2	66.7	55.4	65.7
Enc. State	80.4	55.4	69.7	47.3	93.4	49.4	62.5	56.8	60.2
Combined	81.9	60.0	82.4	60.9	95.3	49.9	64.8	57.3	64.4
RNN + Attn									
Word Embs.	75.6	64.5	87.5	65.9	96.5	50.1	62.6	55.1	64.9
Enc. State	77.2	59.5	80.0	57.0	95.1	49.9	64.7	59.0	67.8
Combined	79.2	64.6	86.3	66.8	96.7	51.3	65.3	58.5	71.1
Transformer									
Word Embs.	81.2	71.6	90.9	70.9	97.7	48.6	74.4	62.3	**74.7**
Enc. State	67.9	54.1	68.7	47.2	85.1	49.4	57.4	55.4	60.7
Combined	81.5	71.3	91.2	70.3	97.9	50.1	72.8	59.6	74.3

Table 6: Accuracy on probing tasks for small models trained with random initialization on DailyDialog. Best Avg result for each model underlined. Best Avg result in bold.

Model	TREC	DNLI	MWOZ	SGD	SNIPS	WNLI	SSA	Topic	Avg
Majority	18.8	34.5	17.0	6.5	14.3	56.3	37.8	34.7	27.5
GloVe	86.5	70.3	91.6	70.5	97.8	49.9	75.1	54.3	74.5
RNN									
Word Embs.	84.0	71.6	91.4	69.8	98.1	51.4	72.0	52.3	73.8
Enc. State	84.6	66.8	89.9	72.9	97.2	48.6	67.8	61.0	73.6
Combined	85.6	69.4	91.1	74.0	97.6	49.6	69.1	61.4	74.7
RNN + Attn									
Word Embs.	83.4	71.4	91.8	70.1	97.9	49.5	72.1	55.7	74.0
Enc. State	85.0	65.6	90.0	73.6	97.2	47.5	70.4	63.0	74.0
Combined	86.6	70.0	92.0	75.9	97.7	48.8	73.5	62.3	75.9
Transformer									
Word Embs.	89.4	70.4	91.4	70.3	98.3	51.4	71.7	51.5	74.3
Enc. State	71.3	58.5	70.7	57.5	88.5	50.2	58.8	64.1	65.0
Combined	90.0	70.2	91.1	70.5	98.1	50.4	72.4	62.9	**75.7**

Table 7: Accuracy on probing tasks for large, Wikipedia pre-trained models finetuned on DailyDialog. Best Avg result for each model underlined. Best Avg result in bold.

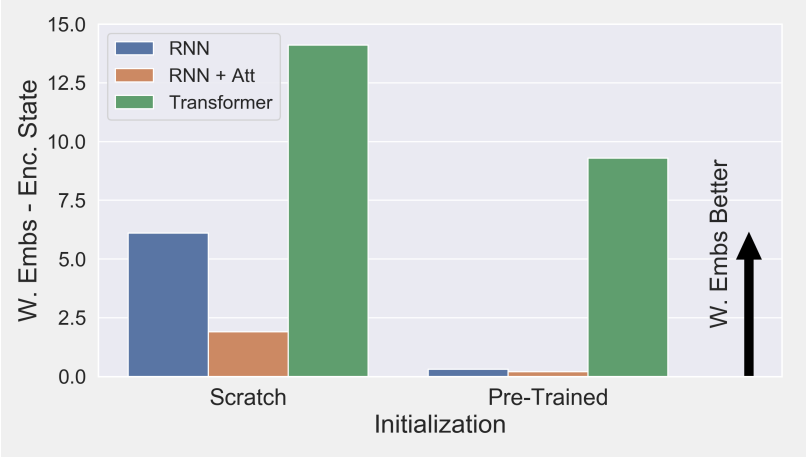

Figure 5: Bar plot showing difference between average scores for word embeddings and encoder states.

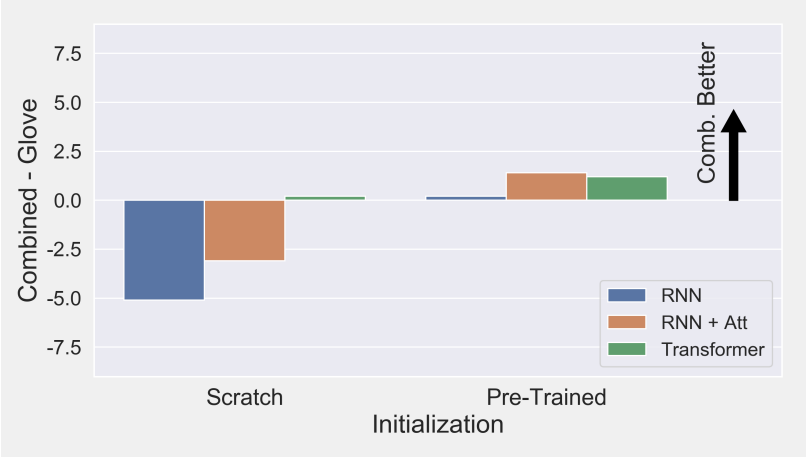

Figure 6: Bar plot showing difference between average scores for combined representations (word embeddings + encoder state) and GloVe baseline.

Figure 7: Bar plot showing difference between average scores for models trained on ordered and shuffled data.

3.5.2 *Probing for conversational understanding*

In this section, we compare the probing performance of the ordered dialog models to the simple baseline of averaging GloVe word embeddings. Here we use the combined representations since they achieve the best performance overall and can act as a proxy for the information captured by the encoder about the conversation history.

Since our probing tasks test for conversational skills important for dialog generation, we would expect the dialog models to outperform GloVe word embeddings. However, this is generally not the case. As figure 6 shows, the GloVe baseline outperforms the small recurrent models while being on par with the large pre-trained models in terms of average score. Tables 6 and 7 show that this pattern also generally applies at the task level, not just in terms of average score.

Closer inspection, however, reveals one exception. Combined representations from both the small and large models consistently outperform GloVe on the DailyDialog Topic task. This is the only task that is derived from the DailyDialog test data, which follows the same distribution as the dialogs used for training the models. This suggests that lack of generalization can partly explain the weak performance on other tasks. It is also worth noting that DailyDialog Topic is labeled at the conversation level rather than the turn level. Thus, identifying the correct label does not necessarily require reasoning about turn-level interactions (unlike DNLI, for example).

The poor performance on the majority of tasks, relative to the simple GloVe baseline, leads us to conclude that standard dialog models trained from scratch struggle to learn the basic conversational skills examined here. Large, pre-trained models do not seem to master these skills either with performance on par with the baselines.

3.5.3 *Effect of dialog structure*

Tables 8 and 9 summarize the results of the perturbation experiments. Figure 7 shows the difference in average performance between the ordered and shuffled models. We show results for the encoder state since it is important for encoding the conversation history, as discussed in section 3.5.1. The encoder state is also sensitive to word and utterance order, unlike averaging the word embeddings. So if a model can fully exploit the dyadic, turn-taking, structure of dialog, this is likely to be reflected in the encoder state representations.

In most of our experiments, models trained on ordered data outperformed models trained on shuffled data, as expected. We can see in figure 7, that average scores for ordered models were often higher than for shuffled models. However, the absolute gap in performance was at most 2%, which is a minute difference in practice. And even though ordered models achieved higher accuracy on average, if we

Model	Test PPL	TREC	DNLI	MWOZ	SGD	SNIPS	WNLI	SSA	Topic	Avg
Majority	-	18.8	34.5	17.0	6.5	14.3	56.3	37.8	34.7	27.5
GloVe Mini	-	83.8	70.8	91.9	71.2	98.0	48.2	75.3	54.0	74.2
RNN										
Ordered	27.2	80.4	55.4	69.7	47.3	93.4	49.4	62.5	56.8	60.2
Shuffled	29.0	77.3	55.7	71.2	46.4	92.8	51.5	57.0	56.8	59.7
RNN + Attn										
Ordered	26.0	77.2	59.5	80.0	57.0	95.1	49.9	64.7	59.0	67.8
Shuffled	28.8	80.2	60.8	80.8	60.7	92.9	50.8	57.9	59.3	67.9
Transformer										
Ordered	29.3	67.9	54.1	68.7	47.2	85.1	49.4	57.4	55.4	60.7
Shuffled	30.8	58.6	52.1	62.6	46.4	83.5	50.4	53.5	63.8	58.9

Table 8: Perplexity and accuracy on probing tasks for small models trained with random initialization on ordered and shuffled dialogs from DailyDialog. Results shown are for probing the encoder state. Best Avg result for each model underlined.

Model	Test PPL	TREC	DNLI	MWOZ	SGD	SNIPS	WNLI	SSA	Topic	Avg
Majority	-	18.8	34.5	17.0	6.5	14.3	56.3	37.8	34.7	27.5
GloVe	-	86.5	70.3	91.6	70.5	97.8	49.9	75.1	54.3	74.5
RNN										
Ordered	17.0	84.6	66.8	89.9	72.9	97.2	48.6	67.8	61.0	73.6
Shuffled	19.1	85.4	65.1	89.5	69.0	97.3	50.5	64.7	65.4	73.4
RNN + Attn										
Ordered	16.5	85.0	65.6	90.0	73.6	97.2	47.5	70.4	63.0	74.0
Shuffled	19.6	84.1	64.9	89.9	71.1	96.6	50.3	64.7	65.4	73.4
Transformer										
Ordered	19.8	71.3	58.5	70.7	57.5	88.5	50.2	58.8	64.1	65.0
Shuffled	21.4	66.1	58.0	68.8	58.0	89.6	49.0	56.3	64.2	63.8

Table 9: Perplexity and accuracy on probing tasks for large, Wikipedia pretrained models finetuned on ordered and shuffled dialogs from DailyDialog. Results shown are for probing the encoder state. Best Avg result for each model underlined.

Models	TREC	DNLI	MWOZ	SGD	SNIPS	WNLI	SSA	Topic	Avg
Scratch	-0.72	-0.61	-0.65	-0.43	-0.82	-0.24	-0.99	0.40	-0.75
Pretrained	-0.76	-0.80	-0.74	-0.81	-0.71	0.61	-0.93	0.65	-0.76
All	-0.55	-0.84	-0.71	-0.87	-0.63	0.30	-0.73	-0.64	-0.92

Table 10: Table showing that probing performance of the encoder state negatively correlates with test perplexity. Results imply that models with lower perplexity (better data fit) correlate with better probing performance.

examine individual tasks in tables 8 and 9, we can find instances where the shuffled models outperformed the ordered ones for each of the tested architectures, sizes, and initialization strategies.

The average difference in test perplexity between all the ordered and shuffled models was less than 2 points, suggesting that model fit and predictions are not substantially different when training on shuffled data. We evaluated all the models on the ordered DailyDialog test set to calculate perplexity. The minimal impact of shuffling the training data suggests that dialog models do not adequately leverage dialog structure during training. Our results show that most of the information captured when training on ordered dialogs is also learned when training on shuffled dialog.

3.6 LIMITATIONS

Some of our conclusions assume that probing performance is indicative of performance on the end-task of dialog generation. Yet it could be the case that certain models learn high quality representations for probing but cannot effectively use them for generation, due to a weakness in the decoder for example. To address this limitation, future work could examine the relationship between probing performance and human judgements of conversation quality. Belinkov [6] suggests more research on the causal relation between probing and end-task performance is required to address this limitation.

However, it is reasonable to assume that capturing information about a certain probing task is a pre-requisite to utilizing information relevant to that task for generation. For example, a model that cannot identify user sentiment is unlikely to use information about user sentiment for generation. We also find that lower perplexity (better data fit) is correlated with better probing performance (table 10), suggesting that probing is a valuable, if imperfect, analysis tool for open-domain dialog systems.

3.7 CONCLUSION

We have used probing to shed light on the conversational understanding of neural dialog models. Our findings suggest that standard neural dialog models suffer from many limitations. They do not effectively encode information about the conversation history, struggle to learn basic conversational skills, and fail to leverage the dyadic, turn-taking structure of dialog. These limitations are particularly severe for small models trained from scratch on dialog data but occasionally also affect large pre-trained models. Addressing these limitations is an interesting direction of future work. Models could be augmented with specific components or multi-task loss functions to support learning certain skills. Future work can also explore the relationship between probing performance and human evaluation.

Part IV

EPILOGUE

4

LOOKING AHEAD

This thesis has presented approaches for building social and interpretable dialog systems. We argued that social skills are a key component of human intelligence and showed that social dialog systems trained with human-centered objectives can lead to better human-AI interaction (chapter 2). We also argued for the importance of interpretability in assessing, evaluating, and improving deep learning models and showed that neural dialog models still struggle with basic conversational skills (chapter 3). In closing, this chapter offers a few directions for future work.

4.1 TOWARDS SOCIAL DIALOG

4.1.1 *Social objectives*

In chapter 2, we trained social dialog systems by optimizing for a handful of simple objectives such as asking more question and avoiding toxicity and repetition. But social behavior is challenging to define and encompass within a few reward functions. Future work could explore more complex, multimodal social signals for improving conversation quality and user engagement. There have already been some attempts to improve dialog systems by learning from implicit human preferences [42] and conditioning on subtle social cues from facial expressions [39] and acoustic information [13].

4.1.2 *Hierarchical reinforcement learning*

The VHRL approach we presented in chapter 2 applied the REINFORCE algorithm at different decision-making components (utterance level vs word level) to achieve hierarchical control. More powerful alternatives to REINFORCE, such as proximal policy optimization (PPO) [79], can also be applied hierarchically to extend our proposed approach and learn more complex rewards.

We only experiment with variational dialog models [83] which learn a continuous latent variable $\mathbf{z}$ as the utterance-level or manager decision. Future work can also apply hierarchical reinforcement learning more generally to models that make discrete manager decisions. For example, HRL can be used to improve models that predict and condition on a set of dialog acts for generation [76]. This setting might be more appropriate for applying HRL since reasoning over a

limited discrete action space is more tractable than a large continuous one [54].

4.1.3 *Social commonsense*

The definition we proposed for social intelligence in section 1.3 emphasized both understanding and behavior. However, our approach in chapter 2 only focused on tuning model behavior, not understanding. Future work can experiment with using HRL to better condition on and reason over external knowledge such social commonsense knowledge [78, 106] or external knowledge bases [58] or knowledge graphs.

4.2 TOWARDS INTERPRETABLE DIALOG

4.2.1 *More probing*

In chapter 3, we only experimented with a limited number of model architectures, datasets, and training objectives. Whether our results generalize across these different modeling decisions remains an open question.

Future work can also use our results to propose modifications to dialog systems and remedy the limitations we highlighted. For example, a multitask objective that incorporates the probing tasks into dialog generation might support the learning of certain conversational skills. Belinkov [6] used a similar approach to promote the morphological awareness of a neural machine translation decoder leading to higher BLEU scores.

Our analysis only focused on high-level probing tasks and conversational skills. Probing can also be used to understand the low-level linguistic information captured by dialog models at a more fine-grained level. Previous studies have analyzed auto-encoders' ability to encode information about the word order, word content, and sentence length of the input [1]. However, a similar analysis has not been applied to dialog systems.

4.2.2 *Probing for causal effects*

Although probing is a valuable tool for analyzing open-domain dialog systems, it still remains unclear how probing performance correlates with human judgements of conversation quality, as discussed in chapter 3. Another interesting questions is whether there is a causal effect behind this relationship. We hope that future work will shed more light on this issue and explore the potential of probing performance as an evaluation metric.

BIBLIOGRAPHY

[1] Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. "Fine-grained analysis of sentence embeddings using auxiliary prediction tasks." In: *arXiv preprint arXiv:1608.04207* (2016).

[2] Benjamin J Ashton, Alex Thornton, and Amanda R Ridley. "An intraspecific appraisal of the social intelligence hypothesis." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1756 (2018), p. 20170288.

[3] Pierre-Luc Bacon, Jean Harb, and Doina Precup. "The option-critic architecture." In: *AAAI*. 2017.

[4] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. "An actor-critic algorithm for sequence prediction." In: *arXiv preprint arXiv:1607.07086* (2016).

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

[6] Yonatan Belinkov. "On internal language representations in deep learning: An analysis of machine translation and speech recognition." PhD thesis. Massachusetts Institute of Technology, 2018.

[7] Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. "What do neural machine translation models learn about morphology?" In: *arXiv preprint arXiv:1704.03471* (2017).

[8] Yonatan Belinkov and James Glass. "Analysis methods in neural language processing: A survey." In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 49–72.

[9] Graham D Bodie, Kellie St. Cyr, Michelle Pence, Michael Rold, and James Honeycutt. "Listening competence in initial interactions I: Distinguishing between what listening is and what listeners do." In: *International Journal of Listening* 26.1 (2012), pp. 1–28.

[10] Graham D Bodie, Andrea J Vickery, Kaitlin Cannava, and Susanne M Jones. "The role of "active listening" in informal helping conversations: Impact on perceptions of listener helpfulness, sensitivity, and supportiveness and discloser emotional improvement." In: *Western Journal of Communication* 79.2 (2015), pp. 151–173.

[11] Cynthia Breazeal. "Toward sociable robots." In: *Robotics and autonomous systems* 42.3-4 (2003), pp. 167–175.

[12] Paweł Budzianowski, Stefan Ultes, Pei-Hao Su, Nikola Mrkšić, Tsung-Hsien Wen, Inigo Casanueva, Lina Rojas-Barahona, and Milica Gašić. "Sub-domain modelling for dialogue management with hierarchical reinforcement learning." In: *arXiv preprint arXiv:1706.06210* (2017).

[13] Felix Burkhardt, Markus Van Ballegooy, Klaus-Peter Engelbrecht, Tim Polzehl, and Joachim Stegmann. "Emotion detection in dialog systems: Applications, strategies and challenges." In: *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE. 2009, pp. 1–6.

[14] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. "Universal Sentence Encoder." In: *arXiv preprint arXiv:1803.11175* (2018).

[15] Larry Chang. *Wisdom for the soul: Five millennia of prescriptions for spiritual healing*. Gnosophia Publishers, 2006.

[16] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[17] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 670–680.

[18] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. "What you can cram into a single vector: Probing sentence embeddings for linguistic properties." In: *arXiv preprint arXiv:1805.01070* (2018).

[19] 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." In: *arXiv preprint arXiv:1805.10190* (2018).

[20] Curry and Rieser. "#MeToo Alexa: How Cfonversational Systems Respond to Sexual Harassment." In: *ACL W Ethics in NLP*. 2018, pp. 7–14.

[21] Antonio R Damasio. *Descartes' error: Emotion, rationality and the human brain*. 1994.

[22] Kalyanmoy Deb. "Multi-objective optimization." In: *Search methodologies*. Springer, 2014, pp. 403–449.

[23] Jean Decety and John T Cacioppo. *The Oxford handbook of social neuroscience*. Oxford library of psychology, 2011.

[24] Denis Diderot. "The Encyclopedia of Diderot & d'Alembert Collaborative Translation Project." In: *Ann Arbor MI: Michigan Publishing, University of Michigan Library. http://quod. lib. umich. edu/d/did* (2002).

[25] Thomas G Dietterich. "Hierarchical reinforcement learning with the MAXQ value function decomposition." In: *J. of AI Res.* 13 (2000), pp. 227–303.

[26] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning." In: *arXiv preprint arXiv:1702.08608* (2017).

[27] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. "Explainable artificial intelligence: A survey." In: *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE. 2018, pp. 0210–0215.

[28] Nouha Dziri, Ehsan Kamalloo, Kory W Mathewson, and Osmar Zaiane. "Evaluating Coherence in Dialogue Systems using Entailment." In: *arXiv preprint arXiv:1904.03371* (2019).

[29] Nathan J Emery, Nicola S Clayton, and Chris D Frith. *Introduction. Social intelligence: from brain to culture*. 2007.

[30] Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. "Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines." In: *arXiv preprint arXiv:1907.01669* (2019).

[31] Evelina Fedorenko and Rosemary Varley. "Language and thought are not the same thing: evidence from neuroimaging and neurological patients." In: *Annals of the New York Academy of Sciences* 1369.1 (2016), p. 132.

[32] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm." In: *arXiv preprint arXiv:1708.00524* (2017).

[33] Howard Gardner. *Frames of Mind: The Theory of Multiple Intelligences*. Hachette Uk, 1983.

[34] Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. "Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems." In: *arXiv preprint arXiv:1906.09308* (2019).

[35] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A survey of methods for explaining black box models." In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42.

[36] Tianxing He and James Glass. "Detecting egregious responses in neural sequence-to-sequence models." In: *arXiv preprint arXiv:1809.04113* (2018).

[37] Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. "Ethical challenges in data-driven dialogue systems." In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. 2018, pp. 123–129.

[38] Esther Herrmann, Josep Call, María Victoria Hernández-Lloreda, Brian Hare, and Michael Tomasello. "Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis." In: *science* 317.5843 (2007), pp. 1360–1366.

[39] Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. "Emotional dialogue generation using image-grounded language models." In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pp. 1–12.

[40] Nicholas K Humphrey. "The social function of intellect." In: *Growing points in ethology*. Cambridge University Press, 1976, pp. 303–317.

[41] Molly E Ireland, Richard B Slatcher, Paul W Eastwick, Lauren E Scissors, Eli J Finkel, and James W Pennebaker. "Language style matching predicts relationship initiation and stability." In: *Psychological science* 22.1 (2011), pp. 39–44.

[42] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. "Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog." In: *arXiv preprint arXiv:1907.00456* (2019).

[43] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. "Social influence as intrinsic motivation for multi-agent deep reinforcement learning." In: *arXiv preprint arXiv:1810.08647* (2018).

[44] Natasha Jaques, Jennifer McCleary, Jesse Engel, David Ha, Fred Bertsch, Rosalind Picard, and Douglas Eck. "Learning via social awareness: Improving a deep generative sketching model with facial feedback." In: *arXiv preprint arXiv:1802.04877* (2018).

[45] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[46] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Skip-thought vectors." In: *Advances in neural information processing systems*. 2015, pp. 3294–3302.

[47] Will Knight. *There's a big problem with AI: even its creators can't explain how it works*. 2017. URL: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/.

[48] Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. "Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation." In: *NIPS*. 2016, pp. 3675–3683.

[49] Hector Levesque, Ernest Davis, and Leora Morgenstern. "The Winograd Schema Challenge." In: *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning. May 17, 2012. URL: https://www.aaai.org/ocs/index.php/KR/KR12/paper/view/4492 (visited on 11/21/2019).

[50] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. "Deep reinforcement learning for dialogue generation." In: *arXiv preprint arXiv:1606.01541* (2016).

[51] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. "Adversarial learning for neural dialogue generation." In: *arXiv preprint arXiv:1701.06547* (2017).

[52] Xin Li and Dan Roth. "Learning Question Classifiers." In: *COLING 2002: The 19th International Conference on Computational Linguistics*. 2002. URL: https://www.aclweb.org/anthology/C02-1150.

[53] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. "Dailydialog: A manually labelled multi-turn dialogue dataset." In: (2017), pp. 986–995.

[54] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. "Continuous control with deep reinforcement learning." In: *arXiv preprint arXiv:1509.02971* (2015).

[55] Ronnie Littlejohn and Jeffrey Dippmann. *Riding the Wind with Liezi: New Perspectives on the Daoist Classic*. SUNY Press, 2011.

[56] Bing Liu and Ian Lane. "Iterative policy learning in end-to-end trainable task-oriented neural dialog models." In: *IEEE Auto. Speech Rec. and Understanding W. (ASRU)*. IEEE. 2017, pp. 482–489.

[57] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation." In: *arXiv preprint arXiv:1603.08023* (2016).

[58] Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. "Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems." In: *arXiv preprint arXiv:1804.08217* (2018).

[59] Stephen Merity. "The wikitext long term dependency language modeling dataset." In: *Salesforce Metamind* 9 (2016).

[60] Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. "Parlai: A dialog research software platform." In: *arXiv preprint arXiv:1705.06476* (2017).

[61] Martin M Monti, Lawrence M Parsons, and Daniel N Osherson. "Thought beyond language: Neural dissociation of algebra and natural language." In: *Psychological science* 23.8 (2012), pp. 914–922.

[62] Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. "Data-efficient hierarchical reinforcement learning." In: *NIPS*. 2018, pp. 3303–3313.

[63] Ryo Nakamura, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura. "Another diversity-promoting objective function for neural dialogue generation." In: *arXiv preprint arXiv:1811.08100* (2018).

[64] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "BLEU: a method for automatic evaluation of machine translation." In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2002, pp. 311–318.

[65] Yookoon Park, Jaemin Cho, and Gunhee Kim. "A Hierarchical Latent Structure for Variational Conversation Modeling." In: *NAACL (Long Papers)*. 2018, pp. 1792–1801.

[66] Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. "Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning." In: *arXiv preprint arXiv:1704.03084* (2017).

[67] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation." In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[68] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[69] Doina Precup. "Temporal abstraction in reinforcement learning." In: (2001).

[70] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. "Sequence level training with recurrent neural networks." In: *arXiv preprint arXiv:1511.06732* (2015).

[71] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. "Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset." In: *arXiv preprint arXiv:1909.05855* (2019).

[72] Ben Fox Rubin. *Amazon's Alexa world just got much bigger*. 2020. URL: https://www.cnet.com/news/amazon-sees-alexa-devices-more-than-double-in-just-one-year/.

[73] Alexander M Rush, Sumit Chopra, and Jason Weston. "A neural attention model for abstractive sentence summarization." In: *arXiv preprint arXiv:1509.00685* (2015).

[74] Abdelrhman Saleh, Ramy Baly, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mitra Mohtarami, Preslav Nakov, and James Glass. "Team QCRI-MIT at SemEval-2019 Task 4: Propaganda Analysis Meets Hyperpartisan News Detection." In: *arXiv preprint arXiv:1904.03513* (2019).

[75] Abdelrhman Saleh, Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, and Rosalind Picard. "Hierarchical Reinforcement Learning for Open-Domain Dialog." In: *arXiv preprint arXiv:1909.07547* (2019).

[76] Chinnadhurai Sankar and Sujith Ravi. "Deep Reinforcement Learning For Modeling Chit-Chat Dialog With Discrete Attributes." In: *arXiv preprint arXiv:1907.02848* (2019).

[77] Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. "Do Neural Dialog Systems Use the Conversation History Effectively? An Empirical Study." In: *arXiv preprint arXiv:1906.01603* (2019).

[78] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. "Socialiqa: Commonsense reasoning about social interactions." In: *arXiv preprint arXiv:1904.09728* (2019).

[79] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. "Proximal policy optimization algorithms." In: *arXiv preprint arXiv:1707.06347* (2017).

[80] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. "What makes a good conversation? How controllable attributes affect human judgments." In: *arXiv preprint arXiv:1902.08654* (2019).

[81] Terrence J Sejnowski. *The deep learning revolution*. Mit Press, 2018.

[82] Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. "A deep reinforcement learning chatbot." In: *arXiv preprint arXiv:1709.02349* (2017).

[83] Iulian Vlad 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: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[84] Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. "HappyBot: Generating Empathetic Dialogue Responses by Improving User Experience Look-ahead." In: *arXiv preprint arXiv:1906.08487* (2019).

[85] Shane Storks, Qiaozi Gao, and Joyce Y Chai. "Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches." In: *arXiv preprint arXiv:1904.01172* (2019).

[86] Pei-Hao Su, Pawel Budzianowski, Stefan Ultes, Milica Gasic, and Steve Young. "Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management." In: *arXiv preprint arXiv:1707.00130* (2017).

[87] Sanjay Subramanian, Sameer Singh, and Matt Gardner. "Analyzing Compositionality of Visual Question Answering." In: ().

[88] I Sutskever, O Vinyals, and QV Le. "Sequence to sequence learning with neural networks." In: *Advances in NIPS* (2014).

[89] Sutton and Barto. *Rl: An introduction*. MIT press, 2018.

[90] Richard S Sutton, Doina Precup, and Satinder Singh. "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning." In: *Artificial intelligence* 112.1-2 (1999), pp. 181–211.

[91] Da Tang, Xiujun Li, Jianfeng Gao, Chong Wang, Lihong Li, and Tony Jebara. "Subgoal discovery for hierarchical dialogue policy learning." In: *arXiv preprint arXiv:1804.07855* (2018).

[92] Chen Tessler, Shahar Givony, Tom Zahavy, Daniel J Mankowitz, and Shie Mannor. "A deep hierarchical approach to lifelong learning in minecraft." In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[93] Rachel Thomas and David Uminsky. "The Problem with Metrics is a Fundamental Problem for AI." In: *arXiv preprint arXiv:2002.08512* (2020).

[94] Alan Turing. "Computing Machinery and Intelligence." In: *Mind* 59.236 (1950), p. 433.

[95] 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*. 2017, pp. 5998–6008.

[96] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. "Feudal networks for hierarchical reinforcement learning." In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 3540–3549.

[97] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. "Universal Adversarial Triggers for NLP." In: *arXiv preprint arXiv:1908.07125* (2019).

[98] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. "Superglue: A stickier benchmark for general-purpose language understanding systems." In: *arXiv preprint arXiv:1905.00537* (2019).

[99] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. "Glue: A multi-task benchmark and analysis platform for natural language understanding." In: *arXiv preprint arXiv:1804.07461* (2018).

[100] Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. "Text simplification using neural machine translation." In: *Thirtieth AAAI Conference on Artificial Intelligence*. 2016.

[101] Harry Weger Jr, Gina R Castle, and Melissa C Emmett. "Active listening in peer interviews: The influence of message paraphrasing on perceptions of listening skill." In: *The I. J. of Listening* 24.1 (2010), pp. 34–49.

[102] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. "Neural text generation with unlikelihood training." In: *arXiv preprint arXiv:1908.04319* (2019).

[103] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. "Dialogue natural language inference." In: *arXiv preprint arXiv:1811.00671* (2018).

[104] Ronald Williams. "Simple statistical gradient-following algorithms for connectionist RL." In: *ML* 8.3-4 (1992), pp. 229–256.

[105] Xu, Wu, and Wu. "Towards Explainable and Controllable Open Domain Dialogue Generation with Dialogue Acts." In: *arXiv:1807.07255* (2018).

[106] Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. "Augmenting end-to-end dialogue systems with commonsense knowledge." In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[107] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. "Seqgan: Sequence generative adversarial nets with policy gradient." In: *AAAI*. 2017.

[108] Zhang, Zhao, and Yu. "Multimodal hierarchical RL policy for task-oriented visual dialog." In: *arXiv:1805.03257* (2018).

[109] 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)*. 2018, pp. 2204–2213.

[110] Yazhou Zhang, Lingling Song, Dawei Song, Peng Guo, Junwei Zhang, and Peng Zhang. "ScenarioSA: A Large Scale Conversational Database for Interactive Sentiment Analysis." In: *arXiv preprint arXiv:1907.05562* (2019).

[111] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. "Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders." In: *ACL (Volume 1: Long Papers)*. 2017, pp. 654–664.

[112] Li Zhou, Gao, Li, and Shum. "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot." In: *arXiv:1812.08989* (2018).

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of April 8, 2020 (`classicthesis` version 4.2).