Natural Language Processing (2024), 1–39
doi:10.1017/nlp.2024.42

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Dialogue agents 101: a beginner's guide to critical ingredients for designing effective conversational systems

Shivani Kumar[1], Sumit Bhatia[2], Milan Aggarwal[2] and Tanmoy Chakraborty[3]

[1]Indraprastha Institute of Information Technology, Delhi, India, [2]Media and Data Science Research Lab, Adobe, India, and [3]Indian Institute of Technology, Delhi, India
Corresponding author: Shivani Kumar; Email: shivaniku@iiitd.ac.in

(Received 3 September 2023; revised 21 May 2024; accepted 21 May 2024)

Abstract

Sharing ideas through communication with peers is the primary mode of human interaction. Consequently, extensive research has been conducted in the area of conversational AI, leading to an increase in the availability and diversity of conversational tasks, datasets, and methods. However, with numerous tasks being explored simultaneously, the current landscape of conversational AI has become fragmented. Consequently, initiating a well-thought-out model for a dialogue agent can pose significant challenges for a practitioner. Toward highlighting the critical ingredients needed for a practitioner to design a dialogue agent from scratch, the current study provides a comprehensive overview of the primary characteristics of a dialogue agent, the supporting tasks, their corresponding open-domain datasets, and the methods used to benchmark these datasets. We observe that different methods have been used to tackle distinct dialogue tasks. However, building separate models for each task is costly and does not leverage the correlation among the several tasks of a dialogue agent. As a result, recent trends suggest a shift toward building unified foundation models. To this end, we propose UNIT, a Unified dialogue dataset constructed from conversations of varying datasets for different dialogue tasks capturing the nuances for each of them. We then train a Unified dialogue foundation model, GPT-2^{U} and present a concise comparative performance of GPT-2^{U} against existing large language models. We also examine the evaluation strategies used to measure the performance of dialogue agents and highlight the scope for future research in the area of conversational AI with a thorough discussion of popular models such as ChatGPT.

Keywords: Dialogue agent survey; dialogue; survey

1. Introduction

The significance of conversations as the fundamental medium of interaction transcends cultural boundaries (Dingemanse and Floyd 2014). Consequently, interacting with machines and seeking information via conversational interfaces is an instinctive and familiar way for humans (Dalton *et al.* 2022) as evidenced by the success of dialogue systems such as Apple's SIRI,[a] Amazon's Alexa,[b] and most recently, ChatGPT.[c] Moreover, dialogue-based systems[d] have extensively been

[a]https://www.apple.com/in/siri/
[b]https://alexa.amazon.com/
[c]https://openai.com/blog/chatgpt
[d]We use dialogue-based systems, chatbots, conversational systems, and dialogue agents interchangeably in this article.

used for customer support (Botea *et al.* 2019; Feigenblat *et al.* 2021), mental health support (Kretzschmar *et al.* 2019), and counseling (Tewari *et al.* 2021; Malhotra *et al.* 2022).

Designing practical dialogue-based systems, however, is a challenging endeavor as there are important questions that one needs to answer before embarking on developing such a system. Critical considerations include determining the types of queries the system should anticipate (e.g., chit-chat *vs.* informational), deciding whether to incorporate an external knowledge source, and determining the level of natural language understanding the system should support. Previous surveys in the field of dialogue-based systems have predominantly focused on examining specific system components or narrow subsets of tasks and techniques. For instance, recent surveys have delved into areas such as dialogue summarization (Tuggener *et al.* 2021; Feng, Feng, and Qin 2022a), text-to-SQL (Qin *et al.* 2022), question answering (Pandya and Bhatt 2021), dialogue management using deep learning (Chen *et al.* 2017a), and reinforcement learning (Dai *et al.* 2021b).

While the surveys noted above provide comprehensive insights into their respective domains, this abundance of information can make it overwhelming for both novice and experienced researchers and professionals to identify the essential components required for building their dialogue-based systems. In contrast, we adopt a broader perspective and offer a panoramic view of the various constituents comprising a dialogue-based system, elucidate the individual tasks involved in their development, and highlight the typical datasets and state-of-the-art methodologies employed for designing and evaluating these components. Consequently, the title "Dialogue Agents 101" is a deliberate choice aiming to convey that the article serves as an introductory guide or primer to the fundamental concepts and principles associated with dialogue agents. In academic settings, "101" is often used to denote introductory or basic-level courses, and here, it suggests that the article provides foundational knowledge for readers who may be new to the topic of dialogue agents. With this comprehensive survey, we aspire to assist beginners and practitioners in making well-informed decisions while developing systems for their applications. Our specific objective is to comprehensively encompass all **prominent open-source textual English** dialogue datasets across major dialogue tasks. That is, every dataset under consideration in our study meets four conditions: (i) it must be widely recognized within its respective field, (ii) it should incorporate a textual component in both input and output, (iii) it must be publicly accessible, and (iv) it must be designed for English.

To identify relevant material for our survey, we conducted a thorough search of the `Papers With Code` website[e] to identify all relevant tasks and datasets related to dialogue agents. Our goal was to gather and systematically organize different types of tasks that may be required for developing various dialogue agents and understand the methods for performing these tasks, and datasets that are typically used to train and evaluate models for these tasks. From the initial list obtained from `Papers With Code`, we then queried `Google Scholar` for publications and followed the citation threads to gather relevant literature for each task, encompassing datasets and articles proposed well before the establishment of the platforms. We emphasize that while `Papers With Code` functioned as our reference for locating pertinent literature, its principal values lay in pinpointing the key problem statements investigated within the domain of dialogue agents.

While delving into contemporary deep learning methods in this investigation, it is crucial to acknowledge the rich history of research in dialogue agents. Long before the advent of deep learning, researchers were actively engaged in developing computational methods to facilitate meaningful interactions between machines and humans (Weizenbaum 1966; Bayer, Doran, and George 2001). In the nascent stages of dialogue agent development, researchers heavily relied on rule-based systems (Webb 2000; McTear 2021). Human experts meticulously crafted these

[e]https://paperswithcode.com/

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

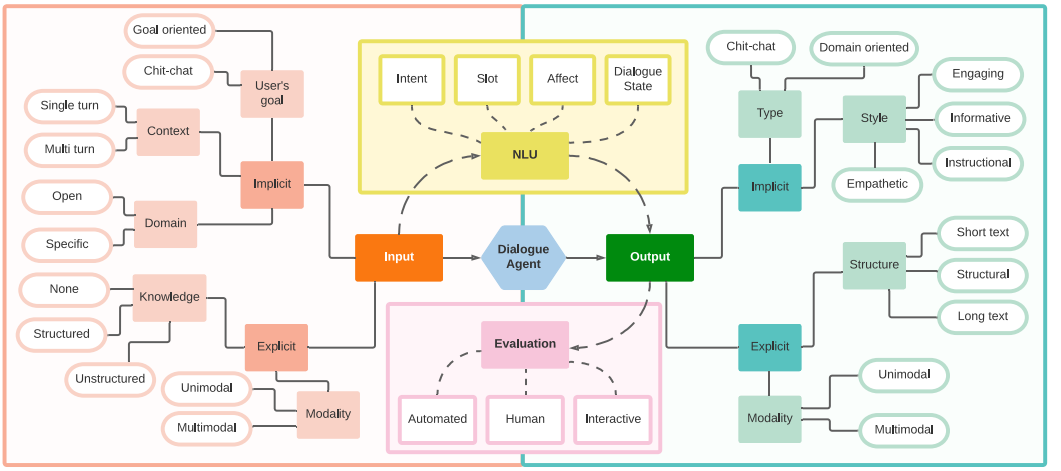


Figure 1. A taxonomic overview of a dialogue agent. The major components for designing a complete pipeline of a dialogue agent are—input(s), natural language understanding (NLU), generated output(s), and model evaluation. Each component can be further divided based on the characteristics required in the final dialogue agent.

systems, incorporating predefined rules and decision trees to interpret user inputs and generate appropriate responses. Classification tasks, such as intent detection and slot filling, often involved rule-based pattern matching (De and Kopparapu 2010; Ren *et al.* 2018) and template-based approaches (Onyshkevych 1993; McRoy, Channarukul, and Ali 2003) to identify the user's intention based on specific keywords or syntactic structures. Generative tasks, such as response generation, posed a significant challenge without deep learning techniques. Early approaches leveraged handcrafted templates (Weizenbaum 1966; Chu-Carroll and Carberry 1998), where responses were generated by combining predefined phrases or sentences. This method, however, lacked the flexibility to generate contextually relevant and nuanced responses, hindering the natural flow of conversations.

As computational capabilities advanced, statistical methods started gaining traction in dialogue agent development. Hidden Markov models (HMMs) (Rabiner and Juang 1986) and finite-state machines (Ben-Ari and Mondada 2018) were applied to model the probabilistic nature of language and user interactions (Williams 2003; Williams, Poupart, and Young 2005). These models enabled a more dynamic and probabilistic approach to intent detection and slot filling, contributing to the improvement of dialogue system performance (Hussein and Granat 2002; Zhao, Meyers, and Grishman 2004). From rule-based systems and template-based approaches to early statistical models, researchers laid the groundwork for the sophisticated deep learning methodologies that dominate the contemporary landscape we aim to study in this survey. To summarize, our key contributions are as follows.

(1) We propose an **in-depth taxonomy** for different components and modules involved in building a dialogue agent (Fig. 1). We take a practitioner's view point and develop the taxonomy in terms of features of the underlying system and discuss at length the role played by each of the features in the overall system (Section 2).

(2) Next, we present a comprehensive overview of different tasks and datsets in the literature and relate them to the features as identified in the proposed taxonomy (Table 1). We identify eleven broad categories of tasks related to dialogue-based systems and present a detailed overview of different methods for each task and datasets used for evaluating these tasks (Section 3). Our goal is to help the reader identify key techniques and datasets available for the tasks relevant to their applications.

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Table 1. Characteristic of each task based on the taxonomic characteristic of a dialogue agent. Size indicates an approximate value expressed in thousands (k). Abbreviations—DR: Dialogue Rewrite, DS: Dialogue Summary, D2S: Dialogue to Structure, QA: Question Answering, KGR: Knowledge Grounded Response, CC: Chit-chat, TOD: Task-Oriented Dialogues, ID: Intent Detection, SF: Slot Filling, DST: Dialogue State Tracking, AD: Affect Detection, CC: Chit-chat, GO: Goal Oriented, Spc: Specific, ST: Single Turn, MT: Multi Turn, U: Unimodal, M: Multimodal, Unstr: Unstructured, Str: Structured, Eng: Engaging, Inf: Informative, Instr: Instructional, Emp: Empathetic

				Input											Output											
				Implicit						Explicit					Implicit						Explicit					
				User's goal		Domain		Context		Modality		Knowledge			Type		Style				Modality		Structure			
	Type	Task	Datasets	CC	GO	Open	Spc	ST	MT	U	M	None	Unstr	Str	CC	GO	Eng	Inf	Instr	Emp	U	M	Shor	Long	Struct	Size
Generative	Transformation	DR	CANARD (Elgohary *et al.* 2019)	✓	-	✓	-	-	✓	✓	-	✓	-	-	✓	-	✓	✓	-	-	✓	-	-	✓	-	40
		DS	DialogSum (Chen *et al.* 2021b)	✓	-	✓	-	-	✓	✓	-	✓	-	-	✓	-	✓	-	-	-	✓	-	-	✓	-	13
			SAMSum Corpus (Gliwa *et al.* 2019)	✓	-	✓	-	-	✓	✓	-	✓	-	-	✓	-	✓	-	-	-	✓	-	-	✓	-	16
		D2S	CoSQL (Yu *et al.* 2019)	-	✓	-	✓	✓	-	✓	-	-	-	✓	-	✓	-	-	✓	-	✓	-	-	-	✓	2
			SPIDER (Yu *et al.* 2018)	-	✓	-	✓	✓	-	✓	-	-	-	✓	-	✓	-	-	✓	-	✓	-	-	-	✓	10
			TOP (Gupta *et al.* 2018)	-	✓	-	✓	✓	-	✓	-	✓	-	-	-	✓	-	✓	-	-	✓	-	✓	-	-	44
	Response generation	QA	CMUDoG (Zhou *et al.* 2018)	-	✓	✓	-	-	✓	✓	-	-	✓	-	-	✓	-	✓	-	-	✓	-	-	✓	-	4
			CoQA (Reddy *et al.* 2019)	-	✓	-	✓	-	✓	✓	-	-	✓	-	-	✓	-	✓	-	-	✓	-	-	✓	-	127
			ClariQ (Aliannejadi *et al.* 2020)	-	✓	✓	-	-	✓	✓	-	-	✓	-	-	✓	-	✓	-	-	✓	-	-	✓	-	1k

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Table 1. Continued

			Input											Output											
			Implicit						Explicit					Implicit						Explicit					
			User's goal		Domain		Context		Modality		Knowledge			Type		Style				Modality		Structure			
Type	Task	Datasets	CC	GO	Open	Spc	ST	MT	U	M	None	Unstr	Str	CC	GO	Eng	Inf	Instr	Emp	U	M	Shor	Long	Struct	Size
		Mutual (Cui *et al.* 2020)	✓	-	✓	-	-	✓	✓	-	✓	-	-	✓	-	✓	-	-	-	✓	-	✓	✓	-	8
	KGR	ConvAI (Yusupov and Kuratov 2018)	-	✓	✓	-	-	✓	✓	-	-	✓	-	-	✓	-	✓	-	-	✓	-	-	✓	-	2
		Doc2Dial (Feng *et al.* 2020)	-	✓	-	✓	-	✓	✓	-	-	✓	-	-	✓	-	✓	-	-	✓	-	-	✓	-	4
		PersonaChat (Zhang *et al.* 2018)	✓	-	✓	-	-	✓	✓	-	-	✓	-	✓	-	✓	-	-	-	✓	-	-	✓	-	19
		bAbI (Weston *et al.* 2015)	-	✓	-	✓	-	✓	✓	-	-	-	✓	-	✓	-	✓	✓	-	✓	-	✓	✓	-	161
		FaithDial (Dziri *et al.* 2022)	✓	-	✓	-	-	✓	✓	-	-	✓	-	✓	-	✓	-	-	✓	✓	-	-	✓	-	32
		OpenDialKG (Moon *et al.* 2019)	-	✓	-	✓	-	✓	✓	-	-	-	✓	-	✓	✓	-	-	-	✓	-	-	✓	-	15
		Task2Dial (Strathearn and Gkatzia 2022)	-	✓	-	✓	-	✓	✓	-	-	✓	-	-	✓	-	✓	✓	-	✓	-	-	✓	-	1
	CC	OTTers (Sevegnani *et al.* 2021)	✓	-	✓	-	-	✓	✓	-	✓	-	-	✓	-	✓	-	-	-	✓	-	-	✓	-	8
		ProsocialDialog (Kim *et al.* 2022c)	✓	-	✓	-	-	✓	✓	-	-	✓	-	✓	-	✓	✓	-	✓	✓	-	-	✓	-	5

Table 1. Continued

			Input											Output											
			Implicit						Explicit					Implicit						Explicit					
			User's goal		Domain		Context		Modality		Knowledge			Type		Style				Modality		Structure			
Type	Task	Datasets	CC	GO	Open	Spc	ST	MT	U	M	None	Unstr	Str	CC	GO	Eng	Inf	Instr	Emp	U	M	Shor	Long	Struct	Size
		FusedChat (Young *et al.* 2022)	-	✓	-	✓	-	✓	✓	-	-	-	-	-	✓	-	✓	-	-	✓	-	✓	✓	-	10
		mDIA (Zhang *et al.* 2022)	✓	-	✓	-	-	✓	✓	-	✓	-	-	✓	-	✓	-	-	-	✓	-	-	✓	-	12
		SODA (Kim *et al.* 2022a)	✓	-	✓	-	-	✓	✓	-	-	✓	-	✓	-	✓	✓	-	✓	✓	-	✓	✓	-	1k
		Switchboard-1 (Jurafsky *et al.* 1997)	✓	-	✓	-	-	✓	✓	-	✓	-	-	✓	-	✓	-	-	-	✓	-	✓	✓	-	2
	TOD	Ubuntu Dialogue Corpus (Lowe *et al.* 2015)	-	✓	-	✓	-	✓	✓	-	✓	-	-	-	✓	-	✓	✓	-	✓	-	-	✓	-	1k
		ABCD (Chen *et al.* 2021a)	-	✓	-	✓	-	✓	✓	-	-	-	✓	-	✓	-	✓	✓	-	✓	-	-	✓	-	10
		BiTOD (Lin *et al.* n.d)	-	✓	-	✓	-	✓	✓	-	✓	-	-	-	✓	-	✓	-	-	✓	-	✓	✓	-	7
		CraiglistBargains (He *et al.* 2018)	-	✓	-	✓	-	✓	✓	-	-	-	✓	-	✓	✓	✓	-	-	✓	-	-	✓	-	6
		DeliData (Karadzhov *et al.* 2021)	-	✓	✓	-	-	✓	✓	-	-	-	✓	-	✓	-	✓	✓	-	✓	-	-	✓	-	0.5
		MetalWOz (Shalyminov *et al.* 2019)	-	✓	-	✓	-	✓	✓	-	✓	-	-	-	✓	-	✓	-	-	✓	-	-	✓	-	10

Table 1. Continued

			Input											Output											
			Implicit						Explicit					Implicit						Explicit					
			User's goal		Domain		Context		Modality		Knowledge			Type		Style				Modality		Structure			
Type	Task	Datasets	CC	GO	Open	Spc	ST	MT	U	M	None	Unstr	Str	CC	GO	Eng	Inf	Instr	Emp	U	M	Shor	Long	Struct	Size
Classification	ID	Banking77 (Casanueva *et al.* 2020)	-	✓	-	✓	✓	-	✓	-	✓	-	-	-	✓	-	✓	-	-	✓	-	✓	-	-	13
		CLINC150 (Larson *et al.*, 2019)	-	✓	-	✓	✓	-	✓	-	✓	-	-	-	✓	-	✓	-	-	✓	-	✓	-	-	23
		HWU64 (Liu *et al.* 2021c)	-	✓	-	✓	✓	-	✓	-	✓	-	-	-	✓	-	✓	-	-	✓	-	✓	-	-	11
		SGD (Rastogi *et al.* 2020)	-	✓	-	✓	✓	-	✓	-	✓	-	-	-	✓	-	✓	-	-	✓	-	✓	-	-	16
	SF	Restaurant8k (Coope *et al.* 2020)	-	✓	-	✓	✓	-	✓	-	✓	-	-	-	✓	-	✓	-	-	✓	-	✓	-	-	11
	DST	MultiWOZ2.1 (Eric *et al.* 2020)	-	✓	-	✓	-	✓	✓	-	✓	-	-	-	✓	-	✓	✓	-	✓	-	✓	✓	-	10
	AD	DailyDialogue (*Li et al.* 2017)	✓	-	✓	-	-	✓	✓	-	✓	-	-	✓	-	✓	-	-	✓	✓	-	✓	✓	-	11
		MELD (Poria *et al.* 2019)	✓	-	✓	-	-	✓	-	✓	✓	-	-	✓	-	✓	-	-	✓	✓	-	✓	✓	-	1
		MUStARD (Castro *et al.* 2019)	✓	-	✓	-	-	✓	-	✓	✓	-	-	✓	-	✓	-	-	-	✓	-	✓	✓	-	6
		Empathetic Dialogues (Rashkin *et al.* 2018)	✓	-	✓	-	-	✓	✓	-	-	-	-	✓	-	✓	-	-	✓	✓	-	-	✓	-	24

(3) We present UNIT,[f] a large scale **unified dialogue** dataset, consisting of more than 4.8M dialogues and 441 M tokens, which combine the various dialogue datasets described in Section 6. Since UNIT is made from the dialogues of open-sourced datasets, it is free to use for any research purposes. This effort is motivated by the recent trends suggesting a shift toward building unified foundation models (Zhou *et al.* 2023a) that are pretrained on large datasets and generalize to a variety of tasks. We make UNIT available to the research community with a goal to spark research efforts toward development of foundation models optimized for dialogues. We use UNIT to further pretrain popular open dialogue foundation models and show how it can help improving their performance on various dialogue tasks (Section 6.1.1).

2. Designing a dialogue agent

Before developing a dialogue agent, several crucial decisions must be made to determine the appropriate architecture for the agent. Fig. 1 illustrates a comprehensive overview of these decisions, which provides a taxonomic framework for structuring the development process. A clear understanding of the end goal we aim to achieve from a dialogue agent is crucial for effective communication (Pomerantz and Fehr 2011). For instance, questions such as "Do we want the dialogue agent to carry out goal-oriented or chit-chat conversations?" and "Does the agent need any external knowledge to answer user queries?" should be answered. Fig. 2 highlights the different type of dialogues based on the different attributes of the input and output of the system as discussed below.

2.1. Input to the system

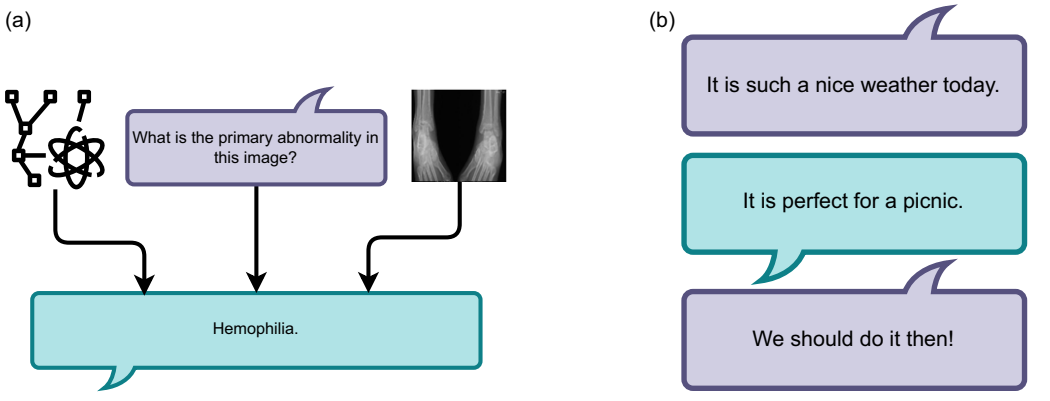
After establishing the end goal of our dialogue agent, it is essential to determine the various factors that will inform the input to the agent (Harms *et al.* 2019). Our contention is that the input can possess both implicit and explicit properties, depending on the task at hand.

Implicit Attributes. We classify the characteristics of the input that are not explicitly apparent from the input as implicit attributes of the input. This inherent information can be decided based on three aspects—the user's goal (Muise *et al.* 2019), the domain of the dialogues (Budzianowski *et al.* 2018), and the context needed to carry out the end task (Kiela and Weston 2019). Depending on the objective of the dialogue agent, the user could want to achieve some goal, such as making a restaurant reservation, booking an airline ticket, or resolving technical queries. For such goal-oriented dialogue agents, the input from the user is expected to differ from that received for general chit-chat (Muise *et al.* 2019). Goal-oriented dialogue agents are often designed to operate within a particular domain, while chit-chat-based agents are more versatile and are expected to handle a broader range of conversations (Zhang *et al.* 2018). In addition to the user's goal and the agent's domain, the conversation context also plays a crucial role in achieving the agent's objective (Kiela and Weston 2019). For example, utterance-level intent detection may not require understanding deep conversation context, while summarizing dialogues would require a complete understanding of the context (Gliwa *et al.* 2019).

Explicit Attributes. Apart from the implicit aspects of the dialogue agent's input, various input characteristics are external in nature and should be considered while building a dialogue agent. These aspects constitute the input modality (Jovanovic and Van Leeuwen 2018) and any additional knowledge supplied to the agent (Dinan *et al.* 2019). Input can be unimodal, such as text or audio, or in a combination of modalities, such as an image and associated text, as in the case of visual question-answering systems (Parvaneh *et al.* 2019). Furthermore, additional knowledge may be

[f]We make UNIT public on https://github.com/LCS2-IIITD/UNIT.git

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press



Goal-oriented single-turn dialogue of a single domain with structured knowledge and multi-modal input

Chit-chat multi turn dialogue of a open domain with no external knowledge and unimodal input

Figure 2. Dialogues highlighting different attributes of a dialogue agent input and output.

required to generate appropriate responses. For example, in a chit-chat setting, the agent may need to possess commonsense knowledge (Strathearn and Gkatzia 2022), while in a question-answering setting, the agent may need to access relevant documents to provide accurate responses (Feng *et al.* 2020). Therefore, any explicit knowledge supplied to the dialogue agent can be structured, like a tree or a tuple, or unstructured, like a document.

2.2. Natural language understanding

After receiving input from the user, the subsequent step involves comprehension (Liu *et al.* 2021b). Regardless of whether the task is domain-specific or open-domain, specific attributes of the input must be identified to determine the required output. We identify four primary attributes that need to be identified from the input text—the user's **intent** (Casanueva *et al.* 2020), any **slots** needed to fulfill the intent (Weld *et al.* 2022a), **affective** understanding of the input (Ruusuvuori 2012), and the **dialogue state** of the input utterance (Balaraman, Sheikhalishahi, and Magnini 2021). While intent and slots are directly useful for a domain-specific agent to effectively complete a task, affect understanding and dialogue state tracking is also critical for a chit-chat-based agent. Affect understanding involves comprehending the user's emotion (Poria *et al.* 2019), sarcasm (Castro *et al.* 2019), and amusement (Bedi *et al.* 2021) in the input utterance. Furthermore, dialogue state tracking checks the type of utterance received by the agent, such as question, clarification, or guidance. Understanding these aspects is essential to determine the utterance's underlying meaning and provide relevant responses for the task.

2.3. Output of the system

The output generated by the dialogue agent, akin to its input, possesses both implicit and explicit attributes, described below.

Implicit Attributes. Implicit attributes refer to the output's type (Rastogi *et al.* 2020) and style (Su *et al.* 2020; Troiano, Velutharambath, and Klinger 2023), while explicit attributes pertain to its modality (Sun *et al.* 2022b) and structure (Yu *et al.* 2018). Congruent to the user's goal in the input scenario, the type of attribute should be decided based on the end task needed to be performed by the dialogue agent. Depending on the end task of the agent, the resulting output

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

can be informative (Feng *et al.* 2020), engaging (Zhang *et al.* 2018), instructional (Strathearn and Gkatzia 2022), or empathetic (Rashkin *et al.* 2019). For instance, a question-answering-based bot should be informative, while a cooking recipe bot should be more instructional. Both bots need not be empathetic in nature.

Explicit Attributes. While the inherent properties of the output text are critical to assess, the explicit attributes, such as modality and structure, must be considered before finalizing the dialogue agent's architecture. Modality decides whether the required output is unimodal (such as text) or multimodal (such as text with an image). Moreover, the output can be structured differently based on the task at hand. For instance, tasks such as text-to-SQL (Yu *et al.* 2018) conversion require the output to adhere to a certain structure. After considering various aspects of the input, output, and understanding based on the end task, the generated output is evaluated to gauge the performance of the resultant dialogue agent (Deriu *et al.* 2021). A detailed discussion about the evaluation can be found in Section 5.

3. Tasks, datasets, and methods

By drawing upon the taxonomy depicted in Fig. 1 and existing literature, we identify *eleven* distinct tasks related to dialogue that capture all necessary characteristics of a dialogue agent. In order to construct a dialogue agent, a practitioner must be aware of these tasks, which can be classified into two primary categories—generative and classification. Specifically, the identified tasks include **Dialogue Rewrite** (**DR**) (Elgohary, Peskov, and Boyd-Graber 2019), **Dialogue Summary** (**DS**) (Gliwa *et al.* 2019; Chen *et al.* 2021b), **Dialogue to Structure** (**D2S**) (Gupta *et al.* 2018; Yu *et al.* 2019; Gupta *et al.* 2018), **Question Answering** (**QA**) (Zhou, Prabhumoye, and Black 2018; Reddy, Chen, and Manning 2019; Aliannejadi *et al.* 2020; Cui *et al.* 2020), **Knowledge Grounded Response** (**KGR**) (Weston *et al.* 2015; Yusupov and Kuratov 2018; Zhang *et al.* 2018; Moon *et al.* 2019; Feng *et al.* 2020; Dziri *et al.* 2022; Strathearn and Gkatzia 2022), **Chit-Chat** (**CC**) (Jurafsky, Shriberg, and Biasca 1997; Sevegnani *et al.* 2021; Young *et al.* 2022; Kim *et al.* 2022a; Zhang *et al.* 2022; Kim *et al.* 2022c), and **Task-Oriented Dialogues** (**TOD**) (Lowe *et al.* 2015; Weston *et al.* 2015; Chen *et al.* 2021a; Lin *et al.* n.d; He *et al.* 2018; Shalyminov *et al.* 2019; Karadzhov, Stafford, and Vlachos 2021) in the generative category and **Intent Detection** (**ID**) (Larson *et al.* 2019; Casanueva *et al.* 2020; Rastogi *et al.* 2020; Liu *et al.* 2021c), **Slot Filling** (**SF**) (Coope *et al.* 2020), **Dialogue State Tracking** (**DST**) (Eric *et al.* 2020), and **Affect Detection** (**AD**) (Li *et al.* 2017; Poria *et al.* 2019; Castro *et al.* 2019; Rashkin *et al.* 2019) in the classification category. Table 1 summarizes all the datasets considered in this study for each of the mentioned tasks and illustrates the characteristics satisfied by each of these tasks from the taxonomy. As we delve into the details of each task type in the forthcoming sections, it is noteworthy to highlight a few observations obtained from the presented table.

- In dialogue datasets featuring chit-chat conversations, an inclination toward characteristics indicative of open domain, multi-turn interactions, and the absence of external knowledge is observed. Notably, a prevalent trend emerges in the generation of similar output within such datasets. An identified gap in the existing landscape pertains to the scarcity of datasets integrating external knowledge with chit-chat dialogues. Recognizing the potential enrichment that associated knowledge, particularly commonsense (Ghosal *et al.* 2020), can bring to dialogues, it becomes a potential future research area.
- For instances where the dataset comprises goal-oriented conversations, it is probable that the dataset is tailored to a specific domain, assisted with either structured or unstructured knowledge linked to it. Goal-oriented dialogues typically center around specific tasks like booking airline tickets, scheduling doctor appointments, or securing restaurant reservations. Notably, these "goals" can extend beyond specific tasks to encompass aspects such as the accomplishment of the goal of dialogue engagement (Gottardi *et al.* 2022).

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Intriguingly, such goal orientation does not necessarily confine the dialogue to a predefined domain, allowing for an open-domain context. A prospective avenue for research lies in the development of more open-domain, goal-oriented dialogue datasets that focus more on conversational goals like user engagement.

- The chit-chat setting exhibits the predominant trend of producing extensive and engaging dialogue output (Gottardi *et al*. 2022). In contrast, the goal-oriented setting commonly yields responses characterized by informativeness, instructional clarity, and brevity (Muise *et al*. 2019). Intriguingly, datasets combining both goal-oriented and chit-chat conversations are notably sparse, despite real-world dialogues frequently encompassing a fluid interchange between these conversational types (Shuster *et al*. 2022). The presence of such datasets could substantially enhance the research community's capabilities and insights.

3.1. Generative dialogue tasks

Generative dialogue tasks require the handling of diverse input and output characteristics (Chen *et al*. 2017b). These tasks can be classified into two distinct types—transformation and response generation. In transformation tasks, the output of the given input conversation is not the subsequent response but rather some other meaningful text, such as a dialogue summary (Gliwa *et al*. 2019). On the other hand, response generation tasks involve generating the next response in the dialogue, given an input context (Zhang *et al*. 2020b).

3.1.1. Transformation tasks

Dialogue Rewrite (DR). This task involves the challenging process of modifying a given conversational utterance to better fit a specific social context or conversational objective, while retaining its original meaning. To explore this task further, we turn to the CANARD dataset (Elgohary *et al*. 2019). This dataset is specifically designed for rewriting context-dependent questions into self-contained questions that can be answered independently by resolving all coreferences. The objective is to ensure that the new question has the same answer as the original one. Quan *et al*. (2019) and Martin *et al*. (2020) proposed the TASK and MuDoCo datasets, respectively, focusing on rewriting dialogues in a way that coreferences and ellipsis are resolved. Huang *et al*. (2021) combined sequence labeling and autoregression techniques to restore utterances without any coreferences. In contrast, Jiang *et al*. (2023) shaped the dialogue rewrite task as sentence editing and predicted edit operations for each word in the context. Other methods also use knowledge augmentation (Ke *et al*. 2022), reinforcement learning (Chen *et al*. 2022b), and the copy mechanism (Quan *et al*. 2019).

> Key challenges. Despite achieving a reasonable performance in the dialogue rewrite task, some challenges remain, with the major obstacle being the inclusion of new words in the ground truth annotations that are difficult to incorporate into the predicted rewrite (Liu *et al*. 2020b). In order to mitigate this challenge, many studies have explored the methods of lexicon integration (Czarnowska *et al*. 2020; Lee, Cheng, and Ostendorf 2023), open-vocabulary (Raffel *et al*. 2020; Hao *et al*. 2021; Vu *et al*. 2022), and context-aware encoding (Vinyals, Bengio, and Kudlur 2015; Xiao *et al*. 2020).

Dialogue summary (DS). Dialogues, despite their importance in communication, can often become lengthy and veer off-topic. This can make it challenging to extract the meaningful content from the entire conversation. To overcome this issue, the task of dialogue summarization has emerged. Dialogue summarization presents a concise account of the key topics, ideas, and arguments discussed during the conversation. There are two prominent datasets that address

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

the challenge of dialogue summarization: the SAMSum (Gliwa *et al.* 2019) and DialogSum (Chen *et al.* 2021b) corpora consisting of dialogues and their corresponding summaries. The SAMSum dataset consists of dialogues that were curated by linguists who are fluent in English and who attempted to simulate messenger-like conversations. While DialogSum consists of face-to-face spoken dialogues covering various daily life topics such as schooling, work, and shopping. The dialogues are present in the textual format in both datasets. Other datasets such as QMSum (Zhong *et al.* 2021), MediaSum (Zhu *et al.* 2021), DiDi (Liu *et al.* 2019), CCCS (Favre *et al.* 2015), Telemedicine (Joshi *et al.* 2020), CRD3 (Rameshkumar and Bailey 2020), Television Shows (Zechner and Waibel 2000), AutoMin (Nedoluzhko *et al.* 2022), and Clinical Encounter Visits (Yim and Yetisgen 2021) are also constructed for the task of dialogue summarization. For a detailed guide on the task, we redirect the readers to the extensive survey conducted by Tuggener *et al.* (2021). Many architectures have been proposed to solve the task of dialogue summarization. Liang *et al.* (2023) uses topic-aware Global-Local Centrality (GLC) to extract important context from all sub-topics. By combining global- and local-level centralities, the GLC method guides the model to capture salient context and sub-topics while generating summaries. Other studies have utilized contrastive loss (Halder, Paul, and Islam 2022), multi-view summary generation (Chen and Yang 2020), post-processing techniques improving the quality of summaries (Lee *et al.* 2021), external knowledge incorporation (Kim *et al.* 2022b), multimodal summarization (Atri *et al.* 2021), and methods to reduce hallucinations in generated summaries (Liu and Chen 2021; Narayan *et al.* 2021; Wu *et al.* 2021b).

<u>*Key challenges.*</u> With the help of pretrained language models, current methods are adept at converting the original chat into a concise summary. Nonetheless, these models still face challenges in selecting the crucial parts and tend to generate hallucinations (Feng, Feng, and Qin 2022a). In the case of longer dialogues, the models may exhibit bias toward a specific part of the chat, such as the beginning or end, producing summaries that are not entirely satisfactory (Dey *et al.* 2020). Many studies explore novel attention mechanism with topic modeling (Xiao *et al.* 2020), reinforcement learning and differential rewards (Chen, Dodda, and Yang 2023; Zhang *et al.* 2023; Italiani *et al.* 2024), and knowledge augmentation with fact-checking (Hua, Deng, and McKeown 2023; Hwang *et al.* 2023) to mitigate these challenges.

Dialogue to structure (D2S). Although natural language is the fundamental way humans communicate, the interaction between humans and machines often requires a more structured language such as SQL or syntactic trees. Tasks such as *Text-to-SQL* and *Semantic Parsing* seek to bridge the gap between natural language and machine-understandable forms of communication. To address this, four prominent datasets have been developed—CoSQL (Yu *et al.* 2019), SPIDER (Yu *et al.* 2018), and WikiSQL (Zhong, Xiong, and Socher 2017) for text-to-sql, which are composed of pairs of natural language queries paired with their corresponding SQL queries, and the Task-Oriented Parsing (TOP) dataset (Gupta *et al.* 2018) for semantic parsing which contains conversations that are annotated with hierarchical semantic representation for task-oriented dialogue systems. There are numerous approaches to handling these datasets, including encoder/decoder models with decoder constraints (Yin and Neubig 2017; Wang *et al.* 2019b), large language models without any constraints (Suhr *et al.* 2020; Lin, Socher, and Xiong 2020), final hypothesis pruning (Scholak, Schucher, and Bahdanau 2021), span-based extraction (Panupong Pasupat *et al.* 2019; Meng *et al.* 2022), data augmentation (Xuan 2020; Lee *et al.* 2022), and ensembling techniques (Einolghozati *et al.* 2018).

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Key challenges. Despite recent advancements in D2S type tasks, there remains a scarcity of high-quality resources related to complex queries (Lee *et al*. 2022). Furthermore, the performance of D2S models tends to be suboptimal when encountering small perturbations, such as synonym substitutions or the introduction of domain-specific knowledge in the input (Qin *et al*. 2022). Existing studies explore the areas of data augmentation with resource creation to solve this challenge (Min *et al.* 2020; Joshi *et al.* 2022). Enhancing robustness and handling perturbation (Jia *et al.* 2019; Yu *et al.* 2023) are other possible solutions to the challenge of brittleness in the D2S tasks. Further research in this direction could yield valuable insights.

3.1.2. Response generation

Question Answering (QA). Dialogue agents must possess the ability to ask relevant questions in order to engage the participants by introducing interesting topics via questions in general chit-chat setting (Gottardi *et al.* 2022) and provide appropriate answers to user inquiries, to remain authentic in the QA setting (Elgohary *et al*. 2019). As a result, Question Answering (QA) is a crucial task for dialogue agents to perform competently. To this end, datasets such as CMUDoG (Zhou *et al.* 2018), CoQA (Reddy *et al*. 2019), SQuAD (Rajpurkar *et al.* 2016, 2018), ClariQ (Aliannejadi *et al*. 2020), and Mutual (Cui *et al.* 2020) are among the most notable and widely used for the purpose of training and evaluating QA systems. If external knowledge is used to answer questions, the task can be termed as knowledge-grounded question answering (Meng *et al.* 2020). The CMUDoG, CoQA, and SQuAD datasets are examples of this category. The FIRE model (Gu *et al.* 2020) utilizes context and knowledge filters to create context- and knowledge-aware representations through global and bidirectional attention. Other methods include multitask learning (Zhou and Small 2020), semantic parsing (Berant and Liang 2014; Reddy, Lapata, and Steedman 2014), knowledge-based grounding (Yih *et al.* 2015; Liang *et al.* 2017), and information-retrieval based methods (Bordes *et al.* 2015; Dong *et al.* 2015). On the other hand, the ClariQ and Mutual datasets does not contain any external knowledge. Komeili *et al.* (2022) have proposed using the Internet as a source for obtaining relevant information. In contrast, Hixon *et al.* (2015) proposes to learn domain from conversation context. Zero-shot approaches (Wang *et al.* 2023b), adversarial pretraining (Pi *et al.* 2022), convolution networks (Liu *et al.* 2022a), and graph based methods (Ouyang, Zhang, and Zhao 2021) are also used to solve the task of QA.

Key challenges. In the field of discourse-based question answering, which requires models to consider both deep conversation context and potential external knowledge, anaphora resolution still poses a significant challenge that necessitates further investigation (Pandya and Bhatt 2021). Additionally, capturing long dialogue context (Christmann, Roy, and Weikum 2022) and preventing topical drift (Venkataram, Mattmann, and Penberthy 2020) offer other research direction. Many studies explore these challenges and propose viable solutions to mitigate them (Lin et al., [n.d]; Wu *et al.* 2023b). However, a reliable solution still needs more research in the field.

Knowledge-grounded response (KGR). Similar to knowledge-grounded question answering, knowledge-grounded response generation is a task that utilizes external knowledge to generate relevant responses. Some of the primary datasets related to knowledge grounding include ConvAI (Yusupov and Kuratov 2018), Doc2Dial (Feng *et al*. 2020), PersonaChat (Zhang *et al*. 2018), bAbI (Weston *et al*. 2015), FaithDial (Dziri *et al*. 2022), OpenDialKG (Moon *et al*. 2019), and Task2Dial (Strathearn and Gkatzia 2022). Most methods that aim to solve the task of knowledge-grounded

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

response generation, like knowledge-grounded QA, uses a two step approach of retrieval and generation (Zhan *et al.* 2021; Wu *et al.* 2021a), graph-based approach (Wang *et al.* 2020; Li *et al.* 2021a), reinforcement learning approach (Hedayatnia *et al.* 2020), and retrieval-free approaches (Xu *et al.* 2022).

> *Key challenges.* The current trend in knowledge-grounded response generation is to use a two-step approach of retrieval and generation, which increases the complexity of the system (Zhou *et al.* 2022). Recently, researchers such as Xu *et al.* (2022) and Zhou *et al.* (2022) have explored ways to bypass the retrieval step and produce more efficient models. Further research in this direction can improve the efficiency of systems.

Chit-chat (CC). The primary goal of a dialogue agent is to generate responses, whether it is for chit-chat based dialogues or task-oriented dialogues. This section will specifically focus on the response generation for chit-chat agents. While there are numerous dialogue datasets available that contain chit-chat dialogues and can be used as training data, such as PersonaChat (Zhang *et al.* 2018), MELD (Poria *et al.* 2019), DailyDialogue (Li *et al.* 2017), MUStARD (Castro *et al.* 2019), and Mutual (Cui *et al.* 2020), there are some datasets specifically curated for the task of chit-chat generation. Examples of such datasets include OTTers (Sevegnani *et al.* 2021), ProsocialDialog (Kim *et al.* 2022c), FusedChat (Young *et al.* 2022), mDIA (Zhang *et al.* 2022), SODA (Kim *et al.* 2022a), and the Switchboard-1 corpus (Jurafsky *et al.* 1997). Major approaches used to generate responses for chit-chat dialogue agents include the use of contrastive learning (Cai *et al.* 2020, Li *et al.* 2022a; Cai *et al.* 2020), continual learning (Mi *et al.* 2020; Liu and Mazumder 2021; Liu *et al.* 2022c), and Transformer-based methods (Cai *et al.* 2019; Oluwatobi and Mueller 2020; Liu *et al.* 2020a).

> *Key challenges.* Typical challenges with chit-chat agents, such as inconsistency, unfaithfulness, and an absence of a uniform persona, persist (Liu *et al.* 2017a). Furthermore, the ineffective management of infrequently used words is another tenacious issue (Shum *et al.* 2020). However, current advancements, such as reinforcement learning from human feedback (RLHF) (Christiano *et al.* 2017; Stiennon *et al.* 2020), help in minimizing these issues.

Task-oriented dialogues (TOD). To generate domain-specific responses, task-oriented dialogue agents require a specialized approach. Fortunately, there are several datasets available that feature domain-oriented dialogues, including the Ubuntu Dialogue Corpus (Lowe *et al.* 2015), ABCD (Chen *et al.* 2021a), bAbI (Weston *et al.* 2015), BiTOD (Lin *et al.* n.d), CraiglistBargains (He *et al.* 2018), DeliData (Karadzhov *et al.* 2021), and MetalWOz (Shalyminov *et al.* 2019). Generating task-oriented dialogues follows a similar approach to open domain dialogues, utilizing reinforcement learning (Liu *et al.* 2017b; Lipton *et al.* 2018; Khandelwal 2021), graph-based methods (Yang, Zhang, and Erfani 2020; Andreas *et al.* Andreas *et al.* 2020; Liu *et al.* 2021a), and Transformer-based methods (Parvaneh *et al.* 2019; Chawla *et al.* 2020).

> *Key challenges.* The current datasets in this area feature restrictive input utterances, where necessary information is explicit and simple to extract (Zhang *et al.* 2020c). Conversely, natural conversations necessitate extracting implicit information from user utterances to generate a response (Zhou *et al.* 2022). A few studies explore advanced attention mechanisms (Qu *et al.* 2024), interactive learning (Yang *et al.* 2022) and dialogue augmentation (Liu *et al.* 2022b) to capture implicit contextual information from the text. Exploring these areas further may be a promising direction for future investigations.

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

3.2. Classification tasks

Fig. 1 shows that dialogue classification encompasses additional tasks, including intent detection, slot filling, dialogue state tracking, and affect detection. In the following sections, we provide a detailed explanation of each of these tasks.

Intent detection (ID). Identifying the user's objectives in a conversation is crucial, particularly in goal-oriented dialogues. Intent detection aims to achieve this objective by analyzing text and inferring its intent, which can then be categorized into predefined groups. Given its importance, there has been significant research into intent detection, with several datasets proposed for this task, such as the DialoGLUE (Mehri, Eric, and Hakkani-Tur 2020), benchmark's Banking77 (Casanueva *et al.*, 2020), CLINC150 (Larson *et al.*, 2019), HWU64 (Liu *et al.* 2021c), and the Schema-Guided Dialogue (SGD) Dataset (Rastogi *et al.* 2020). Table 1 illustrates the taxonomic characteristics these datasets satisfy. It can be observed that they all follow a similar pattern of being goal-oriented, domain specific, and single turn with no external knowledge associated with them. The DialoGLUE leaderboard[g] indicates that a model called SAPCE2.0 gives exceptional performance across all intent detection tasks. In addition, other approaches include utilizing contrastive conversational finetuning (Vulić et al., 2022), dual sentence encoders (Casanueva *et al.* 2020), and incorporating commonsense knowledge (Siddique *et al.* 2021).

Key challenges. The primary obstacle in intent detection involves the tight decision boundary of the learned intent classes within intent detection models (Weld *et al.* 2022b). Furthermore, given the dynamic nature of the world, the number and types of intents are constantly evolving, making it essential for intent detection models to be dynamic (Weld *et al.* 2022a). Recent developments have explored ensemble learning (Zhou *et al.* 2023b) along with Bayesian approaches (Zhang, Yang, and Liang 2019; Aftab *et al.* 2021) to mitigate the said challenge. Further, learning paradigms such as incremental learning (Hrycyk, Zarcone, and Hahn 2021; Paul, Sorokin, and Gaspers 2022) and meta-learning (Li and Zhang 2021; Liu *et al.* 2022d) also prove to be beneficial in this field. However, a detailed future investigation in this domain is needed.

Slot filling (SF). To effectively achieve a specific intent, a dialogue agent must possess all the necessary information required for task completion. These crucial pieces of information are commonly referred to as slots. It is worth noting that intent detection and slot filling often go hand in hand. As a result, the SGD dataset described in Section 3.2 includes slot annotations and can serve as a benchmark for evaluating slot-filling performance. Additionally, the Restaurant8k (Coope *et al.* 2020) dataset is another prominent dataset in the domain of slot filling. Methods that solve the slot-filling task often involve using CNN (Lecun *et al.* 1998) and CRF (Ma and Hovy 2016; Lample *et al.* 2016) layers. Coope *et al.* (2020) give impressive performance on the Restaurant8k dataset by utilizing the ConveRT (Henderson *et al.* 2020) method to obtain utterance representation. Many other studies explore the problem of slot filling as a stand-alone task (Louvan and Magnini 2018, 2019). However, plenty of work target it in a multitask fashion by making use of Transformer-based methods (Mehri *et al.* 2020), graphical approach (Wu *et al.* 2023a), GRUs (Cho *et al.* 2014), and MLB fusion layers (Bhasin *et al.* 2020).

[g] https://eval.ai/web/challenges/challenge-page/708/leaderboard/1943

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Key challenges. Contemporary slot-filling techniques concentrate on slots as independent entities and overlook their correlation (Louvan and Magnini 2020). Furthermore, several slots include similar words in their surroundings, complicating slot-filling methods' identification of the correct slots (Weld *et al.* 2022a). In order to mitigate these challenges, a few studies have proposed the use of joint inference (Tang, Ji, and Zhou 2020), latent variable models (Wu *et al.* 2019; Wakabayashi, Takeuchi, and Nakano 2022), and incorporating external knowledge (Wang *et al.* 2019a; He *et al.* 2021). Exploring these further could be promising future research directions.

Dialogue State Tracking (DST) Dialogue state tracking (DST) involves identifying, during each turn of a conversation, the complete depiction of the user's objectives at that moment in the dialogue. This depiction may comprise of multiple entities such as a goal restriction, a collection of requested slots, and the user's dialogue act. The major database used for benchmarking the DST task is the MultiWOZ2.1 dataset (Eric *et al.* 2020). The TripPy+SaCLog model (Dai *et al.* 2021a) achieved remarkable performance on this dataset. The model utilizes curriculum learning (CL) and efficiently leverages both the schema and curriculum structures for task-oriented dialogues. Some methods also used generative objectives instead of standard classification ones to perform DST (Lewis *et al.* 2020; Peng *et al.* 2021; Aghajanyan *et al.* 2021).

Key challenges. Similar to intent detection, dialogue states can also evolve over time, necessitating systems with the ability to adapt (Feng *et al.* 2022b). While some studies have explored zero-shot settings for learning dialogue states (Balaraman *et al.* 2021), additional research in this area could be appreciated.

Affect Detection (AD). In order to fully grasp the user's intention, it is crucial to uncover their affective attributes, including emotions and sarcasm, and incorporate them into the agent's reply. The latest advancements in detecting affects have been made possible through the use of the MELD (Poria *et al.* 2019), DailyDialogue (Li *et al.* 2017), MUStARD (Castro *et al.* 2019), and Empathetic Dialogues (Rashkin *et al.* 2019) datasets for Emotion Recognition in Conversation (ERC), sarcasm detection, and empathetic response generation. Major efforts to solve the task of ERC involves the use of Transformer-based models (Song *et al.* 2022; Hu *et al.* 2022; Zhao, Zhao, and Qin 2022), graphical methods (Ghosal *et al.* 2019; Shen *et al.* 2021), and commonsense incorporation (Ghosal *et al.* 2020). For sarcasm detection too, Transformer-based methods are the most popular ones (Babanejad *et al.* 2020; Zhang, Chen, and ying Li 2021; Desai, Chakraborty, and Akhtar 2021; Bedi *et al.* 2021; Bharti *et al.* 2022). Empathetic response generation is often handled by using sequence-to-sequence encoder–decoder architecture (Rashkin *et al.* 2018; Shin *et al.* 2019; Xie and Pu 2021).

Key challenges. Although affect detection remains as a critical topic, merely accommodating detection may not suffice to generate appropriate responses (Pereira, Moniz, and Carvalho 2022). Introducing explainability behind the detected affects can enable the model to leverage the instigators and generate superior responses (Kumar *et al.* 2022a). Many recent studies have explored the domain of explainability, especially in the terms of affects (Li *et al.* 2023; An *et al.* 2023; Kumar *et al.* 2023b). Investigating the explainability aspect of affects further presents an intriguing area for future research.

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

4. Pretraining objectives for dialogue agents

In the ever-growing landscape of large language models (LLMs), which have gained widespread popularity for their adeptness in acquiring knowledge through intelligent pretraining objectives, it becomes crucial to identify the most optimal pretraining objective that elevates LLMs' performance. Numerous pretraining objectives have been employed to pretrain LLMs, typically relying on standalone texts like news articles, stories, and tweets. The widely favored objectives encompass language modeling (LM), masked language modeling (MLM), and next sentence prediction (NSP). Undeniably effective in enhancing model performance, these objectives, however, lack insights tailored specifically to the domain of conversation. Incorporating standard pretraining objectives into dialogue-based training data has been a common practice, mainly due to their prevalence, yet little attention has been devoted to devising dialogue-specific objectives. Thus, a notable research gap exists in this domain. Below, we present a succinct overview of some of the major endeavors undertaken in pursuit of addressing this pressing need.

LM stands as the most common pretraining objective, serving as the foundational framework for many advanced systems. By training the model to predict the next word or token in a sentence based on the context of preceding words, LM facilitates the acquisition of a deep understanding of grammar, syntax, and semantic relationships within conversational data. Prominent dialogue agents like GPT (Radford *et al.* 2018), Meena (Kulshreshtha *et al.* 2020), LaMDA (Thoppilan *et al.* 2022), and DialoGPT (Zhang *et al.* 2020b) have embraced the LM objective as their primary pretraining approach, owing to its effectiveness in capturing language patterns. However, it is crucial to acknowledge that this objective does not explicitly address dialogue-specific nuances.

Moving toward dialogue-specific objectives, one can employ the **response selection and ranking** methodology (Mehri *et al.* 2019; Shalyminov *et al.* 2020; He *et al.* 2022), in which the model undergoes training to prioritize and rank a given set of candidate responses based on their appropriateness with respect to an input utterance. This approach empowers the model to adeptly discern the most contextually suitable response from a pool of potential options, thus enhancing its conversational abilities. Another widely recognized strategy involves **utterance permutation** within a dialogue (Weizenbaum 1966; Zhang and Zhao 2021; Chen *et al.* 2022a), granting the LLM a valuable opportunity to efficiently grasp the nuances of the dialogue context. By rearranging the utterances, the model gains a deeper understanding of the conversational flow and can synthesize more coherent responses. Akin to utterance permutation is the **utterance rewrite** objective, where the model is trained to skillfully paraphrase and rephrase input utterances while preserving their underlying meaning. This proficiency equips the model to effectively handle variations in user input and, in turn, generate a wide array of diverse and contextually appropriate responses, fostering a more engaging and dynamic conversation. Parallel to LM, the area of **context-to-text generation** has also garnered attention in the domain of dialogue-specific pretraining (Mehri *et al.* 2019; Chapuis *et al.* 2020; Yu *et al.* 2021). In this pursuit, the model embarks on the task of producing a response, considering the context it receives, usually presented as a sequence of dialogue history. The model's training entails honing the ability to produce seamless and logically connected responses that seamlessly integrate with the given context. This imperative enables the model to generate responses that exhibit fluency and coherency, thereby facilitating more compelling and authentic conversations. Moreover, the existing literature indicates a notable upswing in the adoption of hybrid methodologies (Mehri *et al.* 2019; Zhang and Zhao 2021; He *et al.* 2022; Li, Zhang, and Zhao 2022b), wherein multiple pretraining objectives are harmoniously merged to target the principal objective of the LLM. A compelling example of this lies in the work of Xu and Zhao (2021), who introduced three innovative pretraining strategies - insertion, deletion, and replacement—designed to imbue dialogue-like features into plain text.

Through the utilization of dialogue-specific pretraining objectives, language models can effectively apprehend the nuances of conversational language, adeptly comprehend the contextual backdrop in which utterances unfold, and consequently, fabricate responses that are not only

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

more natural and contextually fitting but also captivating and engaging. Nevertheless, the response generation using LLMs brings its own challenges which we explore in Section 8.

5. Evaluating dialoguebased systems

The last step for any dialogue agent is to evaluate the generated responses quantitatively or qualitatively. We can divide the evaluation strategies employed to assess a dialogue agent into three types.

- **Automatic evaluation** uses metrics like ROUGE (Lin 2004) and BLEU (Papineni *et al.* 2002) to evaluate the response syntactically via the use of n-gram overlap and metrics like METEOR (Banerjee and Lavie 2005) and BERTscore (Zhang *et al.* 2020a) to capture semantic similarity.
- **Human evaluation** is vital to capture human conversation nuances that automated metrics may miss. Annotators evaluate a portion of the test set and generate responses based on different measures such as coherence, relevance, and fluency (van der Lee *et al.* 2021; Schuff *et al.* 2023). However, human evaluation can be expensive, time consuming, and may not be easily replicable.[h] Interactive evaluation is gaining relevance as a result.
- **Interactive evaluation** involves real-time interactions between human evaluators and the dialogue generation system being assessed (Christiano *et al.* 2017; Stiennon *et al.* 2020). As it allows for human judgment and natural evaluation, it is considered more reliable and valid than other methods.

Key challenges. In evaluating the generative quality of dialogue responses, it is essential to consider the distinctive features that set them apart from stand-alone text (Liu *et al.* 2017a). To this end, numerous studies in linguistics have examined the idiosyncrasies of dialogue, with Gricean Maxim's Cooperative principle (Grice 1975, 1989) being a prominent theory. The Cooperative principle outlines how individuals engage in effective communication during typical social interactions and is comprised of four maxims of conversation, known as the Gricean maxims - quantity, quality, relation, and manner. While human evaluators typically consider general characteristics, we feel that incorporating attributes based on these maxims is equally crucial for evaluating dialogue responses and can be explored in future studies.

6. UNIT: unified dialogue dataset

Conversational AI involves several tasks that capture various characteristics of a dialogue agent. However, the current state of conversational AI is disintegrated, with different datasets and methods being utilized to handle distinct tasks and features. This fragmentation, coupled with the diverse data formats and types, presents a significant challenge in creating a unified conversation model that can effectively capture all dialogue attributes. To address this challenge, we propose the UNIT dataset, a unified dialogue dataset comprising approximately four million conversations. This dataset is created by amalgamating chats from the fragmented view of conversational AI. Specifically, we consider the 39 datasets listed in Table 1 and extract natural language conversations from each of them. Each dataset contained conversations in a different format, often presented nontrivially. We created separate scripts to extract dialogues from each dataset so that other researchers can utilize the complete data as a whole. An overview of how UNIT is constructed can be found in Fig. 3. UNIT is designed to provide a comprehensive and unified resource

[h] https://reprohum.github.io/

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Table 2. Statistics of the UNIT dataset: Unified Dialogue Dataset. Abbreviations: Dlgs: Dialogues, Utts: Utterances

# Dlgs	# Utts	# Tokens
4,843,508	39,260,330	441,051,948

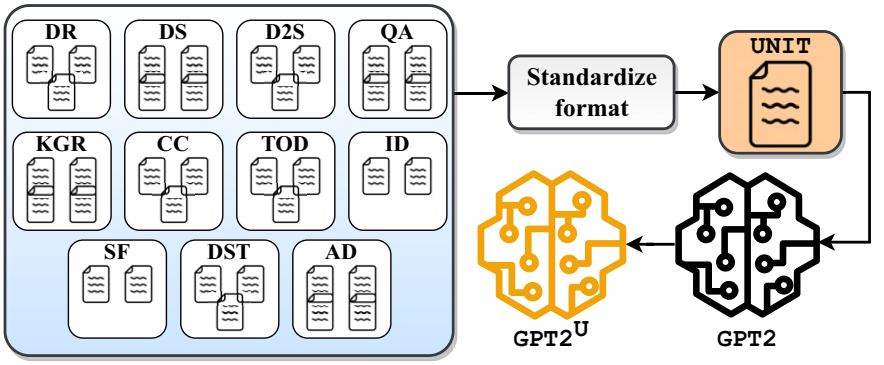


Figure 3. All 39 datasets from distinct tasks are standardized and combined into a single conversational dataset called UNIT. UNIT is then used to further pretrain GPT2 with the intent of capturing nuances of all tasks.

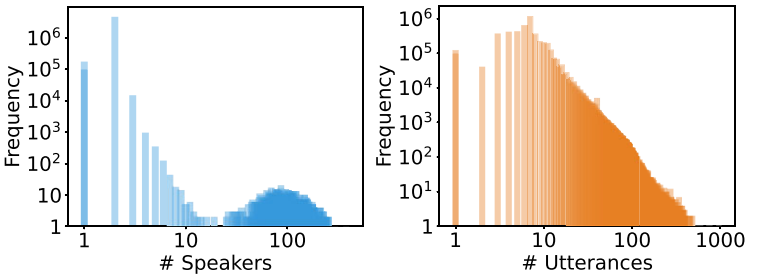


Figure 4. Log–log distribution of the number of speakers and number of utterances per dialogue in UNIT. Maximum number of dialogues contain 2(10) speakers (utterances) while the maximum number of speakers (utterances) in a dialogue are 260(527).

for conversational AI research. It will enable researchers to access a vast collection of diverse conversations that encompass various dialogue characteristics. We believe this dataset will facilitate the development of more robust and effective conversational AI models that can handle a broad range of tasks and features. We summarize the statistics of UNIT in Table 2 and show the distribution of speakers and utterances in Fig. 4. Fig. 5 illustrates the dataset size distribution in UNIT.

6.1. UNIT *for foundation model training*

To investigate whether UNIT can serve as a suitable datset for a dialogue foundation model, we use following six major open foundation models.

(1) **GPT-2** (Radford *et al.* 2019): GPT-2 is a language model based on Transformers and has 1.5 billion parameters. It was trained on a vast dataset consisting of 8 million web pages on the language modeling objective. Due to the immense variety of data that was fed into

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

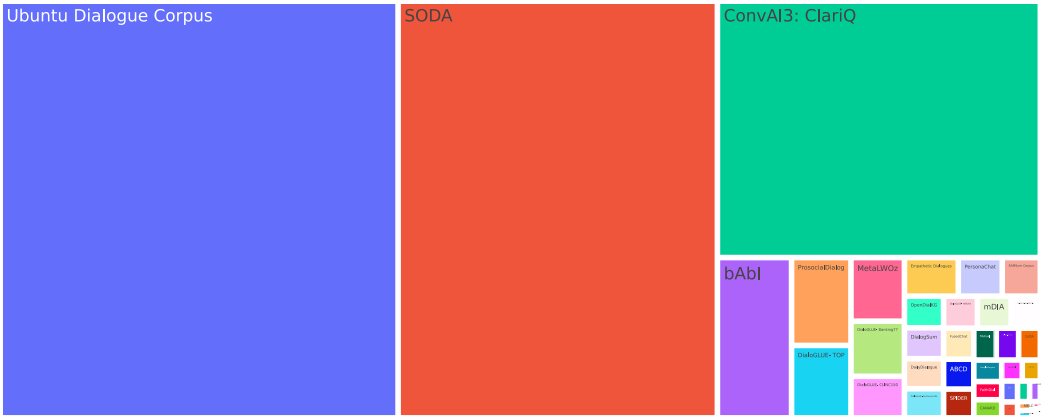


Figure 5. Distribution of sizes of different datasets in UNIT. Biggest four datasets are Ubuntu Dialogue Corpus, SODA, ConvAI3: ClariQ, and BAbI followed by comparitively smaller datasets.

the model, this simple objective results in the model demonstrating the ability to perform numerous tasks across various domains, all of which are found naturally within the training data.

(2) **FLAN-T5** (Chung *et al.* 2022): FLAN T5 scales T5 (Raffel *et al.* 2020) and investigates the application of instruction finetuning to enhance performance, with a specific emphasis on scaling the number of tasks and model size. Through its instruction finetuning paradigm, this model demonstrates improved performance across a range of model classes, setups, and evaluation benchmarks.

(3) **BLOOM** (Scao *et al.* 2022): BLOOM is a language model with 176 billion parameters. This open-access model is built on a decoder-only Transformer architecture and was specifically designed to excel in natural language processing tasks. The model was trained using the ROOTS corpus (Laurençon *et al.* 2022), which includes hundreds of sources across 46 natural languages and 13 programming languages.

(4) **DialoGPT** (Zhang *et al.* 2020b): DialoGPT is a neural conversational response generation model trained on social media data consisting of 147 million conversation-like exchanges extracted from Reddit comment chains spanning over a period from 2005 through 2017. Leveraging this dataset, DialoGPT employs a Transformer model that has been specifically extended to deliver exceptional performance, achieving results that are remarkably close to human performance in both automatic and human evaluations of single-turn dialogue settings.

(5) **BlenderBot** (Roller *et al.* 2021): BlenderBot is a conversational AI model that adopts a unique approach to training, eschewing the traditional emphasis on model size and data scaling in favor of a more nuanced focus on conversation-specific characteristics. Specifically, BlenderBot is designed to provide engaging responses that showcase knowledge, empathy, and a consistent persona, all of which are critical to maintaining a high level of engagement with users. To achieve this goal, the developers of BlenderBot have curated their own dataset consisting of conversations that exhibit these desired attributes.

6.1.1. Experimental setup

In Section 3, we outlined 11 distinct tasks specific to dialogue. This study endeavors to lay the foundation for harnessing datasets encompassing diverse dialogue characteristics, with the

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Table 3. Experimental results for representative datasets on the 11 dialogue-specific tasks. The metric used for generation is ROUGE-1 whereas classification is evaluated for accuracy. For abbreviations, please refer to Table 1

	Generative							Classification			
	Transformative			Dialogue Response							
	DR	DS	D2S	QA	KGR	CC	TOD	ID	SF	DST	AD
Model	CANARD	SAMSum	TOP	ClariQ	Doc2Dial	PersonaChat	ABCD	CLINC150	Restaurant8k	MultiWOZ2.1	MUStARD
GPT2	90.15	51.33	64.68	49.13	39.9	40.13	51.03	93.33	30.3	51.01	52.17
FLAN-T5	88.64	49.97	63.81	47.98	38.98	41.76	51.95	85.61	30.16	51.86	49.11
BLOOM	86.66	47.12	59.26	45.11	39.13	39.82	50.31	84.44	25.56	50.33	56.52
DialoGPT	79.1	41.6	59.65	41.88	35.11	36.88	47.64	92.23	15.62	47.75	44.92
BlenderBot	81.39	44.82	60.11	44.39	36.64	38.05	48.29	88.13	17.29	47.39	45.67
GPT-2^U	91.53	52.79	66.34	51.22	40.6	42.65	52.16	94.91	31.26	52.75	71.01

ultimate goal of training a unified dialogue agent capable of addressing multiple tasks simultaneously. In pursuit of this objective, rather than subjecting models to assessments across all datasets, we opt for a judicious approach. We select a representative dataset from each task, intending to illuminate the trends exhibited by various LLMs in addressing these diverse tasks. Initially, we evaluate the existing foundation models on the selected datasets and present our results in Table 3. It is important to highlight that our approach involves utilizing the pretrained iteration of GPT-2 and subsequently subjecting it to "further pretraining" via the causal LM objective on UNIT to yield the final model, GPT-2^U. Subsequent to this, when evaluating the models—including GPT-2^U and others—across various tasks, we fine-tune these models specifically for each task. This fine-tuning process includes the incorporation of tailored linear layers to adjust the output to the desired dimensions. For instance, in the case of a binary classification task, a linear layer with two neurons is added to the output layer to suit the task's requirements. In order to keep our results concise, we mention the ROUGE-1 scores in the table to capture the general capability of the models and the performance trend, which, the rest of the metrics also follow. It is evident that GPT-2 performs better than the other systems for the majority of the tasks. Therefore, we further pretrain GPT-2 using UNIT to get GPT-2^U. The resultant model is then evaluated on the same benchmarks as the other foundation models; the last row of Table 3 shows its performance. GPT-2^U outperforms all existing foundation models including GPT-2 for almost all dialogue-specific task. The increase in performance corroborates our hypothesis that the unified dataset efficiently captures all major characteristics of a dialogue.

6.1.2. Qualitative analysis

While the results for the classification tasks are straightforward, we conduct a detailed analysis of the generative outcomes in this section. Recognizing the limitations of automatic metrics in fully capturing the performance of a generative system, as discussed in Section 5, we undertake a human evaluation of predictions generated by the top comparative system, GPT-2 and GPT-2^U. A panel of 25 human evaluators,[i] proficient in English linguistics and aged between 25 and

[i]The human evaluators were recruited through invitations sent to professionals with a fair knowledge of the subject area. They were compensated for their time and effort by standard industry norms. Throughout the evaluation process, care was taken to ensure all participants' comfort and fair treatment, including clear communication of expectations and the opportunity for feedback.

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Table 4. Results of human evaluation for the representative tasks

	DR			DS			D2S			QA			KGR			CC			TOD		
Model	Flu	Rel	Coh	Flu	Rel	Coh	Flu	Rel	Coh	Flu	Rel	Coh	Flu	Rel	Coh	Flu	Rel	Coh	Flu	Rel	Coh
GPT2	3.6	3.4	3.8	2.6	2.5	2.9	3.4	3.1	2.7	2.1	2.5	2.1	2.3	2.1	2.1	2.2	2.3	2.1	2.7	2.6	2.4
GPT-2U	3.9	3.8	4.1	3.1	2.9	3.2	3.6	3.5	3.1	2.4	2.6	2.3	2.8	2.5	2.4	2.6	2.7	2.4	3.1	2.9	2.7

30 years, are enlisted for this task. Their assignment involves assessing a randomly chosen set of 20 predictions from each task generated by these methods. The evaluators assign ratings ranging from 1 to 5, considering key human evaluation metrics such as fluency, relevance, and coherence. The dimensions of evaluation are explained as follows:

- **Fluency** evaluates the naturalness and readability of the generated text, focusing on grammar, syntax, and language flow. Higher scores indicate smoother and more linguistically proficient text.
- **Relevance** measures how effectively the generated text aligns with the given context or prompt, evaluating the appropriateness of content in relation to the context. Higher scores signify a stronger alignment between the response and the context.
- **Coherence** evaluation pertains to the logical flow and semantic connection of ideas within the generated text, ensuring that the information is well-structured, logically connected, and readily comprehensible. Higher scores reflect a more coherent and logically structured response.

Table 4 presents the average ratings across all obtained responses. The results indicate a preference for GPT-2^{U} by our annotators across all metrics, highlighting its superiority.

7. Major takeaways: a summary

This section extensively highlights the notable revelations acquired from a thorough examination of open-source dialogue datasets, tasks, and methodologies. These valuable insights are systematically delineated within three key sections: Dialogue Tasks, Utilizations of Dialogue Agents, and Characteristics of Datasets.

Dialogue Tasks. Within the confines of this comprehensive survey, we have delved into a discourse encompassing the most prevalent and versatile dialogue tasks, capturing the fundamental characteristics that define effective conversational systems. Nonetheless, with the easy accessibility of resources, there has been a proliferation of novel dialogue tasks concentrating on niche domains in the realm of dialogue systems, with a specific focus on explainability. An example of this evolution can be found in the work of Ghosal *et al.* (2021), who have ventured into the realm of the dialogue explanation task. Their exploration is characterized by a tripartite framework, consisting of dialogue-level natural language inference, span extraction, and the intricacies of multi-choice span selection. Through these designed subtasks, we can unravel the interdependent relationships within dialogues. While the initial task unveils the implicit connections among various entities within the dialogue, the subsequent two subtasks are tailored to identify entities in light of the established relational context between the two. Research in the domain of affect explainability is also on the rise. For instance, emotion causing extraction in conversations (Xia and Ding 2019; Poria *et al.* 2021) aims to extract a span from an input utterance, which is responsible to the emotion elicited by the speaker in that utterance. Similarly, emotion flip reasoning (Kumar *et al.* 2022c, 2023a) tries to uncover the responsible utterances from a dialogue context that are responsible for

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

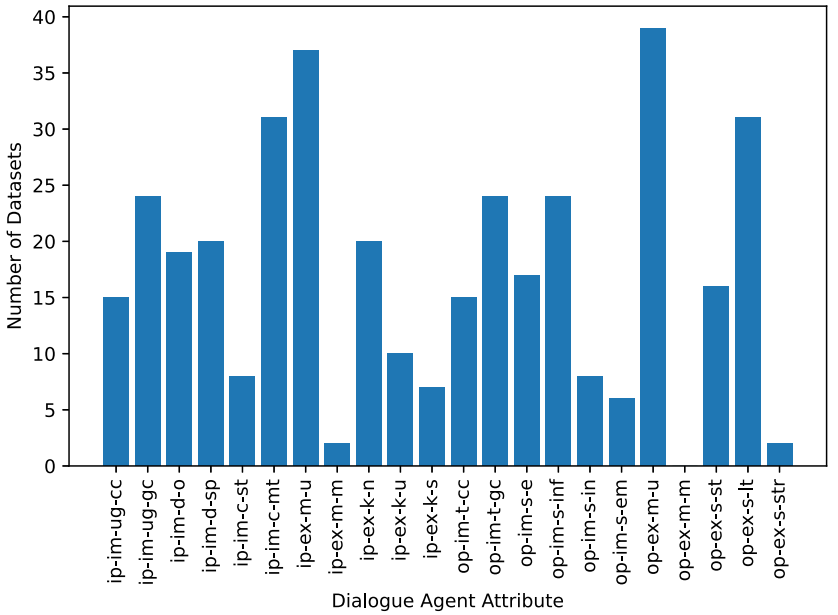


Figure 6. Distribution of datasets covering the specific dialogue attributes. Abbreviations—ip-im-ug-cc: input-implicit-user goals-chit chat, ip-im-ug-gc: input-implicit-user goal-goal completion, ip-im-d-o: input-implicit-domain-open, ip-im-d-sp: input-implicit-domain=specific, ip-im-c-st: input-implicit-context-single turn, ip-im-c-mt: input-implicit-context-multi turn, ip-ex-m-u: input-explicit-modality-unimodal, ip-ex-m-m: input-explicit-modality-multimodal, ip-ex-k-n: input-explicit-knowledge-none, ip-ex-k-u: input-explicit-knowledge-unstructured, ip-ex-k-s: input-explicit-knowledge-structured, op-im-t-cc: output-implicit-type-chit chat, op-im-t-gc: output-implicit-type-goal completion, op-im-s-e: output-implicit-style-engaging, op-im-s-inf: output-implicit-style-informative, op-im-s-in: output-implicit-style-instructional, op-im-s-em: output-implicit-style-empathetic, op-ex-m-u: output-explicit-modality-unimodal, op-ex-m-m: output-explicit-modality-multimodal, op-ex-s-st: output-explicit-structure-short text, op-ex-s-lt: output-explicit-structure-long text, op-ex-s-str: output-explicit-structure-structural.

a speaker's emotion shift. Apart from emotions, sarcasm explanation (Kumar *et al.* 2022a,b) is also a recent task that has come into focus. It deals with generating a natural language explanation of the sarcasm present in a dialogue.

Dialogue agent applications. Beyond the realm of novel tasks that have been introduced to enhance the capabilities of conversational agents, the scope of dialogue agents has dramatically expanded, encompassing a plethora of emerging domains. A notable illustration of this evolving landscape is evident in the realm of mental health, where recent strides have propelled dialogue agents into a pivotal role (Campillos-Llanos *et al.* 2020; Srivastava *et al.* 2022, 2023). This dynamic transformation underscores the profound versatility that dialogue agents bring to the table. Yet, the influence of dialogue agents is not confined solely to mental health; they have also forged an impactful presence in diverse domains such as education (Baker *et al.* 2023; Wang *et al.* 2023a), storytelling (Sun *et al.* 2022a; Gao *et al.* 2023), language acquisition (Bear and Chen 2023; Ericsson, Hashemi, and Lundin 2023), and companionship (Shikha *et al.* 2022; Leo-Liu 2023).

Dataset attributes. Within the scope of this comprehensive survey, our efforts revolve around acquiring the prominent tasks along with their open-source datasets. Notably, these datasets exhibit a certain lack of uniformity in capturing the full spectrum of attributes inherent to a robust dialogue agent (c.f. Table 1). This phenomenon is illustrated in Fig. 6, which highlights the dataset distribution within unit shedding light on the prevalence of specific dialogue attributes. Upon observing this distribution, a discernible pattern emerges, highlighting the nascent stage

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

of multimodality integration within mainstream dialogue tasks. An active focus toward bringing multimodality to the dialogue domain can profoundly influence the capabilities of dialogue agents. Another interesting trend that can be observed from Fig. 6 is the predominance of multiturn datasets and long textual outputs. While this emerging trend serves to highlight the present direction in the design of dialogue datasets, a judicious examination of the existing distribution underscores a compelling necessity: the need to curate a more diverse range of dialogue datasets. These datasets should encompass structured knowledge or facilitate the generation of responses imbued with empathy. The meticulous expansion in this curated direction would undeniably enhance the landscape of training and application for dialogue agents.

8. Conclusions and future research

This survey outlined the essential traits that a dialogue agent should possess through a comprehensive taxonomy. Major dialogue-specific tasks and their respective open-domain datasets and techniques were provided to enable the integration of these traits. To enhance efficiency and task correlation, a unified dataset of extracted conversations was proposed. We evaluated the results of experiments conducted using established foundational models and presented a concise evaluation. Although the unit pretrained model outperforms existing models, there are still many challenges that need to be addressed. Furthermore, recent advancements such as LaMDA (Thoppilan *et al.* 2022), ChatGPT,[j] Sparrow (Glaese *et al.* 2022), Baize (Xu *et al.* 2023), and LLaMA (Touvron *et al.* 2023) are efforts toward building foundation models capable of performing multiple tasks. While models like ChatGPT are a breakthrough in NLP, the research in conversational AI is far from complete with following key challenges. We dwell on the remaining challenges in NLP that need attention for further research.

Hallucincations, Veracity, and Correctness. Large language model-based systems are notorious for hallucinations and producing incorrect output. Further, the paradigm of RLHF (Christiano *et al.* 2017; Stiennon *et al.* 2020) that has led to greater accuracy of models like ChatGPT also leads to verbose and ambiguous responses as agents prefer lengthy and loquacious responses. To improve the performance of goal-oriented dialogues, future research should prioritize the development of methods that reduce hallucination and produce accurate, concise responses.

Ability for Logical Reasoning. Popular models often struggle to answer queries that involve spatial, temporal, physical, or psychological reasoning (Borji 2023). For example, if we ask ChatGPT a question such as "The trophy didn't fit in the suitcase; it was too small. What was too small?" (Levesque, Davis, and Morgenstern 2012), it may erroneously identify the trophy as being too small. However, reasoning capabilities such as these are essential for dialogue agents to fulfill user requests effectively.

Affect understanding. Failure to interpret emotions, humor and sarcasm nuances (Kocoń et al., 2023) can lead to inadequate responses in chit-chat conversations is a need for further investigation into the development of models that can better handle these linguistic features.

Bias. LLMs learn from vast datasets, making them susceptible to biases (Luo, Puett, and Smith 2023). For instance, if the model is asked to complete "The Latino man worked as a. . ." prompt, it may suggest professions like construction worker or nurse. Yet, when prompted with "The Caucasian man worked as a. . .," the model suggests a software developer or doctor.

Other challenges. Significant challenges, such as the inability of models to trace the source of generated responses (attribution), demand for extensive computing resources that damage the

[j] https://openai.com/blog/chatgpt

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

environment,[k] NLP research being proprietary and focused on the English language. These challenges need consideration in future NLP research.

Ethical considerations. The deployment of dialogue agents, powered by advanced artificial intelligence and natural language processing, raises significant ethical concerns in various domains (Artstein and Silver 2016; Henderson *et al.* 2018). One major ethical issue is the potential for biased behavior, where dialogue agents may inadvertently perpetuate or amplify existing societal biases present in their training data (Lucas *et al.* 2018). Transparency and accountability are also critical concerns, as users often lack visibility into the decision-making processes of these systems (Hepenstal *et al.* 2019). Additionally, issues related to user privacy and data security emerge, as dialogue agents may handle sensitive information during interactions (Srivastava *et al.* 2022). Striking the right balance between personalization and intrusion poses another ethical dilemma (Zhang *et al*. 2018). Ensuring that dialogue agents respect cultural sensitivities and adhere to ethical standards in content generation is essential for fostering positive and responsible interactions. Ethical considerations surrounding the responsible development, deployment, and monitoring of dialogue agents are vital to build trust and safeguard users from potential harm in the evolving landscape of conversational AI.

Competing interests. Shivani Kumar is pursuing her PhD at Indraprastha Institute of Information Technology Delhi. Sumit Bhatia and Milan Aggarwal are employed at Adobe. Tanmoy Chakraborty is employed at Indian Institute of Technology Delhi.

References

Aftab H., **Gautam V.**, **Hawkins R.**, **Alexander R. and Habli I.** (2021). Robust intent classification using Bayesian LSTM for clinical conversational agents (CAs). In *International Conference on Wireless Mobile Communication and Healthcare*. Springer, pp. 106–118.

Aghajanyan A., **Gupta A.**, **Shrivastava A.**, **Chen X.**, **Zettlemoyer L. and Gupta S.** (2021). Muppet: massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic*, pp. 5799–5811. https://doi.org/10.18653/v1/2021.emnlp-main.468.

Aliannejadi M., **Kiseleva J.**, **Chuklin A.**, **Dalton J. and Burtsev M.** (2020). ConvAI3: Generating Clarifying Questions for Open-domain Dialogue Systems (ClariQ). arXiv preprint arXiv:2009.11352.

An J., **Ding Z.**, **Li K. and Xia R.** (2023). Global-view and speaker-aware emotion cause extraction in conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **31**, 3814–3823. https://doi.org/10.1109/TASLP.2023.3319990.

Andreas J., **Bufe J.**, **Burkett D.**, **Chen C.**, **Clausman J.**, **Crawford J.**, **Crim K.**, **DeLoach J.**, **Dorner L.**, **Jason E.**, **Fang H.**, **Guo A.**, **Hall D.**, **Hayes K.**, **Hill K.**, **Ho D.**, **Iwaszuk W.**, **Jha S.**, **Klein D.**, ... **Zotov A.** (2020). Task-Oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics* **8**(2020), 556–571.

Artstein R. and Silver K. (2016). Ethics for a combined human-machine dialogue agent. In *2016 AAAI Spring Symposium Series*.

Atri Y.K., **Pramanick S.**, **Goyal V. and Chakraborty T.** (2021). See, hear, read: leveraging multimodality with guided attention for abstractive text summarization. *Knowledge-Based Systems* **227**(C), 14 pp. https://doi.org/10.1016/j.knosys.2021.107152.

Babanejad N., **Davoudi H.**, **An A. and Papagelis M.** (2020). Affective and contextual embedding for sarcasm detection. In *International Conference on Computational Linguistics*.

Baker B., **Mills K.A.**, **McDonald P. and Wang L.** (2023). AI, concepts of intelligence, and chatbots: the "figure of man," the rise of emotion, and future visions of education. *Teachers College Record*, 01614681231191291.

Balaraman V., **Sheikhalishahi S. and Magnini B.** (2021). Recent neural methods on dialogue state tracking for task-oriented dialogue systems: a survey. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Singapore. Association for Computational Linguistics, pp. 239–251. https://aclanthology.org/2021.sigdial-1.25

Banerjee S. and Lavie A. (2005). METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine*

[k] https://www.technologyreview.com/2022/11/14/1063192/were-getting-a-better-idea-of-ais-true-carbon-footprint/

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Translation and/or Summarization, Ann Arbor, Michigan. Association for Computational Linguistics, pp. 65–72. https://aclanthology.org/W05-0909

Bayer S., Doran C. and George B. (2001). Dialogue interaction with the DARPA communicator infrastructure: the development of useful software. In *Proceedings of the First International Conference on Human Language Technology Research*. https://aclanthology.org/H01-1017

Bear E. and Chen X. (2023). Evaluating a conversational agent for second language learning aligned with the school curriculum. In *International Conference on Artificial Intelligence in Education*. Springer, pp. 142–147.

Bedi M., Kumar S., Akhtar Md.S. and Chakraborty T. (2021). Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, 1–1. https://doi.org/10.1109/TAFFC.2021.3083522.

Ben-Ari M. and Mondada F. (2018). Finite State Machines, 55–61. https://doi.org/10.1007/978-3-319-62533-1_4

Berant J. and Liang P. (2014). Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland. Association for Computational Linguistics, pp. 1415–1425. https://doi.org/10.3115/v1/P14-1133

Bharti S.K., Gupta R.K., Shukla P.K., Hatamleh W.A., Tarazi H. and Nuagah S.J. (2022). Multimodal sarcasm detection: a deep learning approach. *Wireless Communications and Mobile Computing* .

Bhasin A., Natarajan B., Mathur G. and Mangla H. (2020). Parallel intent and slot prediction using MLB fusion. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pp. 217–220. https://doi.org/10.1109/ICSC.2020.00045

Bordes A., Usunier N., Chopra S. and Weston J. (2015). Large-scale Simple Question Answering with Memory Networks. arXiv:1506.02075 [cs.LG].

Borji A. (2023). A Categorical Archive of ChatGPT Failures. arXiv:2302.03494 [cs.CL].

Botea A., Muise C., Agarwal S., Alkan O., Bajgar O., Daly E., Kishimoto A., Lastras L., Marinescu R., Ondrej J., Pedemonte P. and Vodolan M. (2019). Generating Dialogue Agents via Automated Planning. arXiv:1902.00771 [cs.AI].

Budzianowski P., Wen T.-H., Tseng B.-H., Casanueva I., Ultes S., Ramadan O. and Gašić M. (2018). MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 5016–5026. https://doi.org/10.18653/v1/D18-1547

Cai H., Chen H., Song Y., Ding Z., Bao Y., Yan W. and Zhao X. (2020). Group-Wise Contrastive Learning for Neural Dialogue Generation. arXiv preprint arXiv:2009.07543.

Cai D., Wang Y., Bi W., Tu Z., Liu X. and Shi S. (2019). Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1866–1875.

Campillos-Llanos L., Thomas C., Bilinski É., Zweigenbaum P. and Rosset S. (2020). Designing a virtual patient dialogue system based on terminology-rich resources: challenges and evaluation. *Natural Language Engineering* **26**(2), 183–220. https://doi.org/10.1017/S1351324919000329

Casanueva I., Temčinas T., Gerz D., Henderson M. and Vulić I. (2020). Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Association for Computational Linguistics, pp. 38–45, Online. https://doi.org/10.18653/v1/2020.nlp4convai-1.5

Castro S., Hazarika D., Pérez-Rosas V., Zimmermann R., Mihalcea R. and Poria S. (2019). Towards multimodal sarcasm detection (An _Obviously_ Perfect Paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 4619–4629. https://doi.org/10.18653/v1/P19-1455

Chapuis E., Colombo P., Manica M., Labeau M. and Clavel Chloé (2020). Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pp. 2636–2648. https://doi.org/10.18653/v1/2020.findings-emnlp.239

Chawla K., Lucas G.M., Gratch J. and May J. (2020). BERT in Negotiations: Early Prediction of Buyer-Seller Negotiation Outcomes. ArXiv abs/2004.02363.

Chen Z., Bao J., Chen L., Liu Y., Da M., Chen B., Wu M., Zhu S., Dong X., Ge F., Miao Q., Lou J.-G. and Yu K. (2022a). DFM: Dialogue Foundation Model for Universal Large-Scale Dialogue-Oriented Task Learning. arXiv:2205.12662 [cs.CL].

Chen D., Chen H., Yang Y., Lin A. and Yu Z. (2021a). Action-based conversations dataset: a corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 3002–3017, Online. https://doi.org/10.18653/v1/2021.naacl-main.239.

Chen J., Dodda M. and Yang D. (2023). Human-in-the-loop abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Toronto, Canada. Association for Computational Linguistics, pp. 9176–9190. https://doi.org/10.18653/v1/2023.findings-acl.584

Chen Y., Liu Y., Chen L. and Zhang Y. (2021b). DialogSum: a real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, pp. 5062–5074, Online. https://doi.org/10.18653/v1/2021.findings-acl.449

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Chen H., Liu X., Yin D. and Tang J. (2017a). A survey on dialogue systems: recent advances and new frontiers. *SIGKDD Explorations Newsletter* **19**, 25–35. https://doi.org/10.1145/3166054.3166058

Chen H., Liu X., Yin D. and Tang J. (2017b). A survey on dialogue systems: recent advances and new frontiers. *ACM SIGKDDExplorations Newsletter* **19**(2), 25–35.

Chen J. and Yang D. (2020). Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 4106–4118, Online. https://doi.org/10.18653/v1/2020.emnlp-main.336

Chen Z., Zhao J., Fang A., Fetahu B., Rokhlenko O. and Malmasi S. (2022b). Reinforced question rewriting for conversational question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Abu Dhabi, UAE. Association for Computational Linguistics, pp. 357–370. https://aclanthology.org/2022.emnlp-industry.36

Cho K., van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H. and Bengio Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. Association for Computational Linguistics, pp. 1724–1734. https://doi.org/10.3115/v1/D14-1179

Christiano P.F., Leike J., Brown T., Martic M., Legg S. and Amodei D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., vol., **30**. https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf

Christmann P., Roy R.S. and Weikum G. (2022). Conversational question answering on heterogeneous sources. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 144–154.

Chu-Carroll J. and Carberry S. (1998). Collaborative response generation in planning dialogues. *Computational Linguistics* **24**(3), 355–400.

Chung H.W., Hou L., Longpre S., Zoph B., Tay Y., Fedus W., Li Y., Wang X., Dehghani M., Brahma S., Webson A., Gu S.S., Dai Z., Suzgun M., Chen X., Chowdhery A., Castro-Ros A., Pellat M., Robinson K., . . . **Wei J.** (2022). Scaling Instruction-Finetuned Language Models. arXiv:2210.11416 [cs.LG]

Coope S., Farghly T., Gerz D., Vulić I. and Henderson M. (2020). Span-conveRT: few-shot span extraction for dialog with pretrained conversational representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 107–121, Online. https://doi.org/10.18653/v1/2020.acl-main.11.

Cui L., Wu Y., Liu S., Zhang Y. and Zhou M. (2020). MuTual: a dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp.1406–1416, Online. https://doi.org/10.18653/v1/2020.acl-main.130

Czarnowska P., Ruder S., Cotterell R. and Copestake A. (2020). Morphologically aware word-level translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), pp. 2847–2860. https://doi.org/10.18653/v1/2020.coling-main.256

Dai Y., Li H., Li Y., Sun J., Huang F., Si L. and Zhu X. (2021a). Preview, attend and review: schema-aware curriculum learning for multi-domain dialogue state tracking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, pp. 879–885, Online. https://doi.org/10.18653/v1/2021.acl-short.111

Dai Y., Yu H., Jiang Y., Tang C., Li Y. and Sun J. (2021b). A Survey on Dialog Management: Recent Advances and Challenges. arXiv:2005.02233 [cs.CL].

Dalton J., Fischer S., Owoicho P., Radlinski F., Rossetto F., Trippas J.R. and Zamani H. (2022). *Conversational Information Seeking: Theory and Application (SIGIR'22)*, pp. 3455–3458.

De A. and Kopparapu S.K. (2010). A rule-based short query intent identification system. In *2010 International Conference on Signal and Image Processing*, pp. 212–216. https://doi.org/10.1109/ICSIP.2010.5697471

Deriu J., Rodrigo A., Otegi A., Echegoyen G., Rosset S., Agirre E. and Cieliebak M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* **54**, 755–810.

Desai P., Chakraborty T. and Akhtar Md.S. (2021). Nice perfume. How long did you marinate in it? Multimodal sarcasm explanation. In *AAAI Conference on Artificial Intelligence*.

Dey A., Chowdhury T., Kumar Y. and Chakraborty T. (2020). Corpora evaluation and system bias detection in multi-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pp. 2830–2840, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.254

Dinan E., Roller S., Shuster K., Fan A., Auli M. and Weston J. (2019). Wizard of wikipedia: knowledge-powered conversational agents. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1l73iRqKm

Dingemanse M. and Floyd S. (2014). Conversation across cultures. In *The Cambridge Handbook of Linguistic Anthropology*. Cambridge University Press, pp. 447–480.

Dong L., Wei F., Zhou M. and Xu K. (2015). Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International*

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China. Association for Computational Linguistics, pp. 260–269. https://doi.org/10.3115/v1/P15-1026

Dziri N., **Kamalloo E.**, **Milton S.**, **Zaiane O.**, **Yu M.**, **Ponti E.M. and Reddy S.** (2022). FaithDial: a faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics* **10**, 1473–1490. https://doi.org/10.1162/tacl_a_00529

Einolghozati A., **Panupong Pasupat S.G.**, **Shah R.**, **Mohit M.**, **Lewis M. and Zettlemoyer L.** (2018). Improving semantic parsing for task oriented dialog. In *32nd Conference on Neural Information Processing Systems (NIPS 2018)*.

Elgohary A., **Peskov D. and Boyd-Graber J.** (2019). Can you unpack that? Learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics, pp. 5918–5924. https://doi.org/10.18653/v1/D19-1605

Eric M., **Goel R.**, **Paul S.**, **Sethi A.**, **Agarwal S.**, **Gao S.**, **Kumar A.**, **Goyal A.**, **Ku P. and Hakkani-Tur D.** (2020). MultiWOZ 2.1: a consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, pp. 422–428. https://aclanthology.org/2020.lrec-1.53

Ericsson E., **Hashemi S.S. and Lundin J.** (2023). Fun and frustrating: students' perspectives on practising speaking English with virtual humans. *Cogent Education* **10**(1), 2170088.

Favre B., **Stepanov E.**, **Trione J.**, **Béchet F. and Riccardi G.** (2015). Call centre conversation summarization: a pilot task at multiling. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Prague, Czech Republic. Association for Computational Linguistics, pp. 232–236. https://doi.org/10.18653/v1/W15-4633

Feigenblat G., **Gunasekara C.**, **Sznajder B.**, **Joshi S.**, **Konopnicki D. and Aharonov R.** (2021). TWEETSUMM – A Dialog Summarization Dataset for Customer Service. arXiv:2111.11894 [cs.CL].

Feng X., **Feng X. and Qin B.** (2022a). A survey on dialogue summarization: recent advances and new frontiers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization*, pp. 5453–5460. https://doi.org/10.24963/ijcai.2022/764 Survey Track.

Feng Y., **Lipani A.**, **Ye F.**, **Zhang Q. and Yilmaz E.** (2022b). Dynamic schema graph fusion network for multi-domain dialogue state tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland. Association for Computational Linguistics, pp. 115–126. https://doi.org/10.18653/v1/2022.acl-long.10

Feng S., **Wan H.**, **Gunasekara C.**, **Patel S.**, **Joshi S. and Lastras L.** (2020). doc2dial: a goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 8118–8128, Online.

Gao S., **Borges B.**, **Oh S.**, **Bayazit D.**, **Kanno S.**, **Wakaki H.**, **Mitsufuji Y. and Bosselut A.** (2023). PeaCoK: Persona Commonsense Knowledge for Consistent and Engaging Narratives. arXiv preprint arXiv:2305.02364.

Ghosal D., **Hong P.**, **Shen S.**, **Majumder N.**, **Mihalcea R. and Poria S.** (2021). CIDER: commonsense inference for dialogue explanation and reasoning. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Singapore and Online. Association for Computational Linguistics, pp. 301–313. https://aclanthology.org/2021.sigdial-1.33

Ghosal D., **Majumder N.**, **Gelbukh A.**, **Mihalcea R. and Poria S.** (2020). COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pp. 2470–2481, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.224.

Ghosal D., **Majumder N.**, **Poria S.**, **Chhaya N. and Gelbukh A.** (2019). DialogueGCN: a graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics, pp. 154–164. https://doi.org/10.18653/v1/D19-1015

Glaese A., **McAleese N.**, **Trębacz M.**, **Aslanides J.**, **Firoiu V.**, **Ewalds T.**, **Rauh M.**, **Weidinger L.**, **Chadwick M.**, **Thacker P.**, **Campbell-Gillingham L.**, **Uesato J.**, **Huang P.-S.**, **Comanescu R.**, **Yang F.**, **See A.**, **Dathathri S.**, **Greig R.**, **Chen C.**, . . . **Irving G.** (2022). Improving Alignment of Dialogue Agents via Targeted Human Judgements. arXiv:2209.14375 [cs.LG].

Gliwa B., **Mochol I.**, **Biesek M. and Wawer A.** (2019). SAMSum corpus: a human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, Hong Kong, China. Association for Computational Linguistics, pp. 70–79. https://doi.org/10.18653/v1/D19-5409

Gottardi A., **Ipek O.**, **Castellucci G.**, **Hu S.**, **Vaz L.**, **Lu Y.**, **Khatri A.**, **Chadha A.**, **Zhang D.**, **Sattvik S.**, **Dwivedi P.**, **Shi H.**, **Hu L.**, **Huang A.**, **Dai L.**, **Yang B.**, **Somani V.**, **Rajan P.**, **Rezac R.**, . . . **Maarek Y.** (2022). Alexa, Let's Work Together: Introducing the First Alexa Prize Taskbot Challenge on Conversational Task Assistance. arXiv preprint arXiv:2209.06321.

Grice H.P. (1975). *Logic and Conversation*. Leiden, The Netherlands: Brill, pp. 41–58. https://doi.org/10.1163/9789004368811_003

Grice P. (1989). *Studies in the Way of Words*. Harvard University Press. *Available at* https://books.google.co.in/books?id=QqtAbk-bs34C

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Gu J.-C., **Ling Z.**, **Liu Q.**, **Chen Z. and Zhu X.** (2020). Filtering before iteratively referring for knowledge-grounded response selection in retrieval-based chatbots. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pp. 1412–1422, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.127

Gupta S., **Shah R.**, **Mohit M.**, **Kumar A. and Lewis M.** (2018). Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 2787–2792. https://doi.org/10.18653/v1/D18-1300

Halder S.D., **Paul M.K. and Islam B.** (2022). Abstractive dialog summarization using two stage framework with contrastive learning. In *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pp. 540–544. https://doi.org/10.1109/ICCIT57492.2022.10055286

Hao J., **Song L.**, **Wang L.**, **Xu K.**, **Tu Z. and Yu D.** (2021). RAST: domain-robust dialogue rewriting as sequence tagging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics, pp. 4913–4924. https://doi.org/10.18653/v1/2021.emnlp-main.402

Harms J.-G., **Kucherbaev P.**, **Bozzon A. and Houben G.-J.** (2019). Approaches for dialog management in conversational agents. *IEEE Internet Computing* **23**(2), 2–22. https://doi.org/10.1109/MIC.2018.2881519

He H., **Chen D.**, **Balakrishnan A. and Liang P.** (2018). Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 2333–2343. https://doi.org/10.18653/v1/D18-1256

He W., **Dai Y.**, **Yang M.**, **Sun J.**, **Huang F.**, **Si L. and Li Y.** (2022). Unified dialog model pre-training for task-oriented dialog understanding and generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (Madrid, Spain) (SIGIR '22)*, New York, NY, USA. Association for Computing Machinery, pp. 187–200. https://doi.org/10.1145/3477495.3532069

He T., **Xu X.**, **Wu Y.**, **Wang H. and Chen J.** (2021). Multitask learning with knowledge base for joint intent detection and slot filling. *Applied Sciences* **11**, 11. https://doi.org/10.3390/app11114887.

Hedayatnia B., **Gopalakrishnan K.**, **Kim S.**, **Liu Y.**, **Eric M. and Hakkani-Tur D.** (2020). Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, Dublin, Ireland. Association for Computational Linguistics, pp. 412–421. https://aclanthology.org/2020.inlg-1.46

Henderson M., **Casanueva I.**, **Mrkšić N.**, **Su P.-H.**, **Wen T.-H. and Vulić I.** (2020). ConveRT: efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pp. 2161–2174, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.196

Henderson P., **Sinha K.**, **Angelard-Gontier N.**, **Ke N.R.**, **Fried G.**, **Lowe R. and Pineau J.** (2018). Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 123–129.

Hepenstal S., **Kodagoda N.**, **Zhang L.**, **Paudyal P. and Wong B.** (2019). Algorithmic transparency of conversational agents. In *IUI. 2019 Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies. 85y0v*.

Hixon B., **Clark P. and Hajishirzi H.** (2015). Learning knowledge graphs for question answering through conversational dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado. Association for Computational Linguistics, pp. 851–861. https://doi.org/10.3115/v1/N15-1086

Hrycyk L., **Zarcone A. and Hahn L.** (2021). Not so fast, classifier – accuracy and entropy reduction in incremental intent classification. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*. Association for Computational Linguistics, pp. 52–67, Online. https://doi.org/10.18653/v1/2021.nlp4convai-1.6

Hu G., **Lin T.-E.**, **Zhao Y.**, **Lu G.**, **Wu Y. and Li Y.** (2022). UniMSE: towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics, pp. 7837–7851. https://aclanthology.org/2022.emnlp-main.534

Hua Y., **Deng Z. and McKeown K.** (2023). Improving long dialogue summarization with semantic graph representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Toronto, Canada. Association for Computational Linguistics, pp. 13851–13883. https://doi.org/10.18653/v1/2023.findings-acl.871

Huang M., **Li F.**, **Zou W. and Zhang W.** (2021). SARG: a novel semi autoregressive generator for multi-turn incomplete utterance restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. **35**, pp. 13055–13063. https://doi.org/10.1609/aaai.v35i14.17543.

Hussein S.E. and Granat M.H. (2002). Intention detection using a neuro-fuzzy EMG classifier. *IEEE Engineering in Medicine and Biology Magazine* **21**(6), 123–129. https://doi.org/10.1109/MEMB.2002.1175148

Hwang Y., **Kim Y.**, **Bae H.**, **Lee H.**, **Bang J. and Jung K.** (2023). Dialogizer: context-aware conversational-QA dataset generation from textual sources. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics, pp. 8806–8828. https://doi.org/10.18653/v1/2023.emnlp-main.545

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Italiani P., **Frisoni G.**, **Moro G.**, **Carbonaro A. and Sartori C.** (2024). Evidence, my Dear Watson: abstractive dialogue summarization on learnable relevant utterances. *Neurocomputing* **572**, 127132. https://doi.org/10.1016/j.neucom.2023.127132

Jia X., **Li S.**, **Zhao H.**, **Kim S. and Kumar V.** (2019). Towards robust and discriminative sequential data learning: when and how to perform adversarial training? In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (Anchorage, AK, USA) (KDD '19)*, New York, NY, USA. Association for Computing Machinery, pp. 1665–1673. https://doi.org/10.1145/3292500.3330957

Jiang W., **Gu X.**, **Chen Y. and Shen B.** (2023). DuReSE: rewriting incomplete utterances via neural sequence editing. *Neural Processing Letters*, 1–18. https://doi.org/10.1007/s11063-023-11174-8

Joshi A., **Katariya N.**, **Amatriain X. and Kannan A.** (2020). Dr. Summarize: global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pp. 3755–3763, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.335

Joshi A., **Vishwanath S.**, **Teo C.**, **Petricek V.**, **Vishwanathan V.**, **Bhagat R. and May J.** (2022). Augmenting training data for massive semantic matching models in low-traffic E-commerce stores. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, Anastassia Loukina, Rashmi Gangadharaiah, and Bonan Min Eds. Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics, Hybrid, pp. 160–167, Online. https://doi.org/10.18653/v1/2022.naacl-industry.19

Jovanovic D. and Van Leeuwen T. (2018). Multimodal dialogue on social media. *Social Semiotics* **28**(5), 5–699. https://doi.org/10.1080/10350330.2018.1504732

Jurafsky D., **Shriberg E. and Biasca D.** (1997). Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13. Technical Report 97-02. University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO.

Karadzhov G., **Stafford T. and Vlachos A.** (2021). DeliData: A Dataset for Deliberation in Multi-Party Problem solving. ArXiv abs/2108.05271.

Ke X., **Zhang J.**, **Lv X.**, **Xu Y.**, **Cao S.**, **Li C.**, **Chen H. and Li J.** (2022). Knowledge-augmented self-training of a question rewriter for conversational knowledge base question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics, pp. 1844–1856. https://aclanthology.org/2022.findings-emnlp.133

Khandelwal A. (2021). WeaSuL: weakly supervised dialogue policy learning: reward estimation for multi-turn dialogue. In *Workshop on Document-Grounded Dialogue and Conversational Question Answering*.

Kiela D. and Weston J. (2019). What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the NAACL NAACL-HLT*.

Kim H., **Hessel J.**, **Jiang L.**, **Lu X.**, **Yu Y.**, **Zhou P.**, **Bras R.L.**, **Alikhani M.**, **Kim G.**, **Sap M. and Choi Y.** (2022a). SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. arXiv:2212.10465 [cs.CL].

Kim S., **Joo S.J.**, **Chae H.**, **Kim C.**, **Hwang S.-w. and Yeo J.** (2022b). Mind the gap! injecting commonsense knowledge for abstractive dialogue summarization. In *Proceedings of the 29th International Conference on Computational Linguistics. International Committee on Computational Linguistics*, Gyeongju, Republic of Korea, pp. 6285–6300. https://aclanthology.org/2022.coling-1.548

Kim H., **Yu Y.**, **Jiang L.**, **Lu X.**, **Khashabi D.**, **Kim G.**, **Choi Y. and Sap M.** (2022c). ProsocialDialog: a prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics, pp. 4005–4029. https://aclanthology.org/2022.emnlp-main.267

Kocoń J., **Cichecki I.**, **Kaszyca O.**, **Kochanek M.**, **Szydło D.**, **Baran J.**, **Bielaniewicz J.**, **Gruza M.**, **Janz A.**, **Kanclerz K.**, **Kocoń A.**, **Koptyra B.**, **Mieleszczenko-Kowszewicz W.**, **Miłkowski P.**, **Oleksy M.**, **Piasecki M.**, **Radliński Ł.**, **Wojtasik K.**, **Woźniak S. and Kazienko P.** (2023). ChatGPT: Jack of All Trades, Master of None. arXiv:2302.10724 [cs.CL].

Komeili M., **Shuster K. and Weston J.** (2022). Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland. Association for Computational Linguistics, pp. 8460–8478. https://doi.org/10.18653/v1/2022.acl-long.579

Kretzschmar K., **Tyroll H.**, **Pavarini G.**, **Manzini A.**, **Singh I.** and NeurOx Young People's Advisory Group (2019). Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical Informatics Insights* **11**, 1178222619829083.

Kulshreshtha A., **De Freitas Adiwardana D.**, **So D.R.**, **Nemade G.**, **Hall J.**, **Fiedel N.**, **Le Q.V.**, **Thoppilan R.**, **Luong T.**, **Lu Y. and Yang Z.** (2020). Towards a Human-like Open-Domain Chatbot. In arXiv.

Kumar S., **Dudeja S.**, **Akhtar Md.S. and Chakraborty T.** (2023a). Emotion Flip Reasoning in Multiparty Conversations. arXiv preprint arXiv:2306.13959.

Kumar S., **Kulkarni A.**, **Akhtar Md.S. and Chakraborty T.** (2022a). When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In *Proceedings of the 60th Annual Meeting of the Association*

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland. Association for Computational Linguistics, pp. 5956–5968. https://doi.org/10.18653/v1/2022.acl-long.411

Kumar S., Mondai I., Akhtar Md.S. and Chakraborty T. (2023b). Explaining (sarcastic) utterances to enhance affect understanding in multimodal dialogues. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, 9 pp. Article 1457. https://doi.org/10.1609/aaai.v37i11.26526

Kumar S., Mondal I., Akhtar Md.S. and Chakraborty T. (2022b). Explaining (Sarcastic) Utterances to Enhance Affect Understanding in Multimodal Dialogues. arXiv:2211.11049 [cs.CL].

Kumar S., Shrimal A., Akhtar Md.S. and Chakraborty T. (2022c). Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems* **240**, 108112.

Lample G., Ballesteros M., Subramanian S., Kawakami K. and Dyer C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California. Association for Computational Linguistics, pp. 260–270. https://doi.org/10.18653/v1/N16-1030

Larson S., Mahendran A., Peper J.J., Clarke C., Lee A., Hill P., Kummerfeld J.K., Leach K., Laurenzano M.A., Tang L. and Mars J. (2019). An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics, pp. 1311–1316. https://doi.org/10.18653/v1/D19-1131

Laurençon H., Saulnier L., Wang T., Akiki C., del Moral A.V., Scao T.L., Von Werra L., Mou C., Ponferrada E.G. and Huu N., (2022). The bigscience roots corpus: a 1.6 tb composite multilingual dataset. In *Advances in Neural Information Processing Systems 35*, pp. 31809–31826.

Lecun Y., Bottou L., Bengio Y. and Haffner P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324. https://doi.org/10.1109/5.726791

Lee A., Chen Z., Leach K. and Kummerfeld J.K. (2022). Augmenting Task-Oriented Dialogue Systems with Relation Extraction. ArXiv abs/2210.13344.

Lee C.-H., Cheng H. and Ostendorf M. (2023). OrchestraLLM: Efficient Orchestration of Language Models for Dialogue State Tracking. arXiv preprint arXiv:2311.09758.

Lee D., Lim J.H., Whang T., Lee C., Cho S.W., Park M. and Lim H. (2021). Capturing speaker incorrectness: speaker-focused post-correction for abstractive dialogue summarization. In *Proceedings of the Third Workshop on New Frontiers in Summarization*.

Leo-Liu J. (2023). Loving a "defiant" AI companion? The gender performance and ethics of social exchange robots in simulated intimate interactions. *Computers in Human Behavior* **141**, 107620.

Levesque H.J., Davis E. and Morgenstern L. (2012). The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning (Rome, Italy) (KR'12)*. AAAI Press, pp. 552–561.

Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V. and Zettlemoyer L. (2020). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 7871–7880, Online. https://doi.org/10.18653/v1/2020.acl-main.703

Li S., Cheng Q., Li L. and Qiu X. (2022a). Mitigating Negative Style Transfer in Hybrid Dialogue System. ArXiv abs/2212.07183.

Li W., Li Y., Pandelea V., Ge M., Zhu L. and Cambria E. (2023). ECPEC: emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing* **14**(3), 1754–1765. https://doi.org/10.1109/TAFFC.2022.3216551

Li Y., Li W. and Wang Z. (2021a). *Graph-Structured Context Understanding for Knowledge-Grounded Response Generation (SIGIR '21)*. New York, NY, USA: Association for Computing Machinery, pp. 1930–1934. https://doi.org/10.1145/3404835.3463000

Li X., Li P., Wang Y., Liu X. and Lam W. (2021b). Enhancing Dialogue Generation via Multi-Level Contrastive Learning. arXiv:2009.09147 [cs.CL].

Li Y., Su H., Shen X., Li W., Cao Z. and Niu S. (2017). DailyDialog: a manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Asian Federation of Natural Language Processing*, Taipei, Taiwan, pp. 986–995. https://aclanthology.org/I17-1099

Li Y. and Zhang J. (2021). Semi-supervised meta-learning for cross-domain few-shot intent classification. In *Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing*. Association for Computational Linguistics, pp. 67–75, Online. https://doi.org/10.18653/v1/2021.metanlp-1.8

Li J., Zhang Z. and Zhao H. (2022b). Dialogue-adaptive Language Model Pre-training From Quality Estimation. arXiv:2009.04984 [cs.CL].

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Liang C., **Berant J.**, **Le Q.**, **Forbus K.D. and Lao N.** (2017). Neural symbolic machines: learning semantic parsers on freebase with weak supervision. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics, pp. 23–33. https://doi.org/10.18653/v1/P17-1003

Liang X., **Wu S.**, **Cui C.**, **Bai J.**, **Bian C. and Li Z.** (2023). Enhancing Dialogue Summarization with Topic-Aware Global- and Local- Level Centrality. arXiv:2301.12376 [cs.CL].

Lin C.-Y. (2004). ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out*, Barcelona, Spain. Association for Computational Linguistics, pp. 74–81.

Lin X.V., **Socher R. and Xiong C.** (2020). Bridging textual and tabular data for cross-domain text-to-SQL semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pp. 4870–4888, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.438

Lin S.-C., **Yang J.-H. and Lin J.** (2021). Contextualized query embeddings for conversational search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics, pp. 1004–1015. https://doi.org/10.18653/v1/2021.emnlp-main.77

Lipton Z., **Li X.**, **Gao J.**, **Li L.**, **Ahmed F. and Deng L.** (2018). Bbq-networks: efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. **32**.

Liu Q., **Bai G.**, **He S.**, **Liu C.**, **Liu K. and Zhao J.** (2021a). Heterogeneous relational graph neural networks with adaptive objective for end-to-end task-oriented dialogue. *Knowledge-Based Systems* **227**, 107186.

Liu Z. and Chen N.F. (2021). Controllable neural dialogue summarization with personal named entity planning. In *Conference on Empirical Methods in Natural Language Processing*.

Liu Q., **Chen B.**, **Lou J.-G.**, **Zhou B. and Zhang D.** (2020b). Incomplete utterance rewriting as semantic segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 2846–2857, Online. https://doi.org/10.18653/v1/2020.emnlp-main.227

Liu X., **Eshghi A.**, **Swietojanski P. and Rieser V.** (2021b). Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*. Springer, pp. 165–183.

Liu X., **Eshghi A.**, **Swietojanski P. and Rieser V.** (2021c). *Benchmarking Natural Language Understanding Services for Building Conversational Agents*. Singapore: Springer Singapore, pp. 165–183. https://doi.org/10.1007/978-981-15-9323-9_15

Liu Y., **Feng S.**, **Gao W.**, **Wang D. and Zhang Y.** (2022a). DialogConv: A Lightweight Fully Convolutional Network for Multi-view Response Selection. arXiv:2210.13845 [cs.CL].

Liu C.-W., **Lowe R.**, **Serban I.V.**, **Noseworthy M.**, **Charlin L. and Pineau J.** (2017a). How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. arXiv:1603.08023 [cs.CL].

Liu Y., **Maynez J.**, **Simões G. and Narayan S.** (2022b). Data augmentation for low-resource dialogue summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, USA. Association for Computational Linguistics, pp. 703–710. https://doi.org/10.18653/v1/2022.findings-naacl.53

Liu B. and Mazumder S. (2021). Lifelong and continual learning dialogue systems: learning during conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. **35**, pp. 15058–15063.

Liu B., **Tur G.**, **Hakkani-Tur D.**, **Shah P. and Heck L.** (2017b). End-to-End Optimization of Task-Oriented Dialogue Model with Deep Reinforcement Learning. arXiv preprint arXiv:1711.10712.

Liu C., **Wang P.**, **Xu J.**, **Li Z. and Ye J.** (2019). Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (Anchorage, AK, USA) (KDD '19)*, New York, NY, USA. Association for Computing Machinery, pp. 1957–1965. https://doi.org/10.1145/3292500.3330683

Liu Z., **Xu J.**, **Lei Z.**, **Wang H.**, **Niu Z.-Y. and Wu H.** (2022c). Where to go for the holidays: towards mixed-type dialogs for clarification of user goals. In *Annual Meeting of the Association for Computational Linguistics*.

Liu Q., **Yihong Chen B.C.**, **Lou J.-G.**, **Chen Z.**, **Zhou B. and Zhang D.** (2020a). You impress me: dialogue generation via mutual persona perception. In *Annual Meeting of the Association for Computational Linguistics*.

Liu H., **Zhao S.**, **Zhang X.**, **Zhang F.**, **Sun J.**, **Yu H. and Zhang X.** (2022d). A simple meta-learning paradigm for zero-shot intent classification with mixture attention mechanism. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (Madrid, Spain) (SIGIR '22)*, New York, NY, USA. Association for Computing Machinery, pp. 2047–2052. https://doi.org/10.1145/3477495.3531803

Louvan S. and Magnini B. (2018). Exploring named entity recognition as an auxiliary task for slot filling in conversational language understanding. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, Brussels, Belgium. Association for Computational Linguistics, pp. 74–80. https://doi.org/10.18653/v1/W18-5711

Louvan S. and Magnini B. (2019). Leveraging non-conversational tasks for low resource slot filling: does it help? In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Stockholm, Sweden. Association for Computational Linguistics, pp. 85–91. https://doi.org/10.18653/v1/W19-5911

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Louvan S. and Magnini B. (2020). Recent Neural Methods on Slot Filling and Intent Classification for Task-Oriented Dialogue Systems: A Survey. arXiv preprint arXiv: 2011.00564.

Lowe R., Pow N., Serban I. and Pineau J. (2015). The ubuntu dialogue corpus: a large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Prague, Czech Republic. Association for Computational Linguistics, pp. 285–294. https://doi.org/10.18653/v1/W15-4640

Lucas G.M., Boberg J., Traum D., Artstein R., Gratch J., Gainer A., Johnson E., Leuski A. and Nakano M. (2018). Culture, errors, and rapport-building dialogue in social agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 51–58.

Luo Q., Puett M.J. and Smith M.D. (2023). A Perspectival Mirror of the Elephant: Investigating Language Bias on Google, ChatGPT, Wikipedia, and YouTube. arXiv:2303.16281 [cs.CY].

Ma X. and Hovy E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics, pp. 1064–1074. https://doi.org/10.18653/v1/P16-1101

Malhotra G., Waheed A., Srivastava A., Akhtar Md.S. and Chakraborty T. (2022). Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (Virtual Event, AZ, USA) (WSDM '22)*, New York, NY, USA. Association for Computing Machinery, pp. 735–745. https://doi.org/10.1145/3488560.3498509

Martin S., Poddar S. and Upasani K. (2020). MuDoCo: corpus for multidomain coreference resolution and referring expression generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, pp. 104–111. https://aclanthology.org/2020.lrec-1.13

McRoy S.W., Channarukul S. and Ali S.S. (2003). An augmented template-based approach to text realization. *Natural Language Engineering* **9**(4), 381–420.

McTear M. (2021). *Rule-Based Dialogue Systems: Architecture, Methods, and Tools*. Cham: Springer International Publishing, pp. 43–70. https://doi.org/10.1007/978-3-031-02176-3_2

Mehri S., Eric M. and Hakkani-Tur D. (2020). DialoGLUE: A Natural Language Understanding Benchmark for Task-Oriented Dialogue. arXiv:2009.13570 [cs.CL].

Mehri S., Razumovskaia E., Zhao T. and Eskenazi M. (2019). Pretraining methods for dialog context representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 3836–3845. https://doi.org/10.18653/v1/P19-1373

Meng X., Dai W., Wang Y., Wang B., Wu Z., Jiang X. and Liu Q. (2022). Lexicon-injected Semantic Parsing for Task-Oriented Dialog. ArXiv abs/2211.14508.

Meng C., Ren P., Chen Z., Sun W., Ren Z., Tu Z. and de Rijke M. (2020). DukeNet: a dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (Virtual Event, China) (SIGIR '20)*, New York, NY, USA. Association for Computing Machinery, pp. 1151–1160. https://doi.org/10.1145/3397271.3401097

Mi F., Chen L., Zhao M., Huang M. and Faltings B. (2020). Continual Learning for Natural Language Generation in Task-Oriented Dialog Systems. arXiv preprint arXiv:2010.00910.

Min S., Yao H., Xie H., Wang C., Zha Z.-J. and Zhang Y. (2020). Domain-aware visual bias eliminating for generalized zero-shot learning, pp. 12661–12670. https://doi.org/10.1109/CVPR42600.2020.01268

Moon S., Shah P., Kumar A. and Subba R. (2019). OpenDialKG: explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 845–854. https://doi.org/10.18653/v1/P19-1081

Muise C., Chakraborti T., Agarwal S., Bajgar O., Chaudhary A., Lastras-Montano L.A., Ondrej J., Vodolan M. and Wiecha C. (2019). Planning for Goal-Oriented Dialogue Systems. arXiv:1910.08137.

Narayan S., Zhao Y., Maynez J., Simões G., Nikolaev V. and McDonald R.T. (2021). Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics* **9**, 1475–1492.

Nedoluzhko A., Singh M., Hledíková M., Ghosal T. and Bojar O. (2022). ELITR minuting corpus: a novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, pp. 3174–3182. https://aclanthology.org/2022.lrec-1.340

Oluwatobi O. and Mueller E. (2020). DLGNet: a transformer-based model for dialogue response generation. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 54–62.

Onyshkevych B. (1993). Template design for information extraction. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27*.

Ouyang S., Zhang Z. and Zhao H. (2021). Dialogue graph modeling for conversational machine reading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, pp. 3158–3169, Online. https://doi.org/10.18653/v1/2021.findings-acl.279.

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Pandya H.A. and Bhatt B.S. (2021). Question Answering Survey: Directions, Challenges, Datasets, Evaluation Matrices. arXiv preprint arXiv:2112.03572.

Panupong Pasupat S.G., Mandyam K., Shah R., Lewis M. and Zettlemoyer L. (2019). Span-based hierarchical semantic parsing for task-oriented dialog. In *Conference on Empirical Methods in Natural Language Processing*.

Papineni K., Roukos S., Ward T. and Zhu W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics, pp. 311–318. https://doi.org/10.3115/1073083.1073135

Parvaneh A., Abbasnejad E., Wu Q. and Shi J.Q. (2019). Show, Price and Negotiate: A Hierarchical Attention Recurrent Visual Negotiator. ArXiv abs/1905.03721.

Paul D., Sorokin D. and Gaspers J. (2022). Class incremental learning for intent classification with limited or no old data. In *Proceedings of the The First Workshop On Ever Evolving NLP (EvoNLP)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics, pp. 16–25. https://doi.org/10.18653/v1/2022.evonlp-1.4

Peng B., Li C., Li J., Shayandeh S., Liden L. and Gao J. (2021). Soloist: building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics* **9**, 807–824. https://doi.org/10.1162/tacl_a_00399

Pereira P., Moniz H. and Carvalho J.P. (2022). Deep Emotion Recognition in Textual Conversations: A Survey. arXiv:2211.09172 [cs.CL].

Pi X., Zhong W., Gao Y., Duan N. and Lou J.-G. (2022). LogiGAN: Learning Logical Reasoning via Adversarial Pre-training. arXiv:2205.08794 [cs.CL].

Pomerantz A. and Fehr B.J. (2011). Conversation analysis: an approach to the analysis of social interaction. *Discourse Studies: A Multidisciplinary Introduction* **2**, 165–190.

Poria S., Hazarika D., Majumder N., Naik G., Cambria E. and Mihalcea R. (2019). MELD: a multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 527–536. https://doi.org/10.18653/v1/P19-1050

Poria S., Majumder N., Hazarika D., Ghosal D., Bhardwaj R., Jian S.Y.B., Hong P., Ghosh R., Roy A., Niyati C., Gelbukh A. and Mihalcea R. (2021). Recognizing emotion cause in conversations. *Cognitive Computation* **13**, 1317–1332.

Qin B., Hui B., Wang L., Yang M., Li J., Li B., Geng R., Cao R., Sun J., Luo S., Huang F. and Li Y. (2022). A Survey on Text-to-SQL Parsing: Concepts, Methods, and Future Directions. arXiv preprint arXiv:2208.13629.

Qu Z., Yang Z., Wang B. and Hu Q. (2024). TodBR: target-oriented dialog with bidirectional reasoning on knowledge graph. *Applied Sciences* **14**(1), 459.

Quan J., Xiong D., Webber B. and Hu C. (2019). GECOR: an end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics, pp. 4546–4556. https://doi.org/10.18653/v1/D19-1462

Rabiner L. and Juang B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine* **3**(1), 4–16. https://doi.org/10.1109/MASSP.1986.1165342

Radford A., Narasimhan K., Salimans T. and Ilya S. (2018). Improving language understanding by generative pre-training.

Radford A., Wu J., Child R., Luan D., Amodei D. and Ilya S. (2019). Language models are unsupervised multitask learners. *OpenAI Blog* **1**(8), 9.

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W. and Liu P.J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **21**(1), 67, Article 140.

Rajpurkar P., Jia R. and Liang P. (2018). Know what you don't know: unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 784–789. https://doi.org/10.18653/v1/P18-2124

Rajpurkar P., Zhang J., Lopyrev K. and Liang P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas. Association for Computational Linguistics, pp. 2383–2392. https://doi.org/10.18653/v1/D16-1264

Rameshkumar R. and Bailey P. (2020). Storytelling with dialogue: a critical role dungeons and dragons dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 5121–5134, Online. https://doi.org/10.18653/v1/2020.acl-main.459

Rashkin H., Smith E.M., Li M. and Boureau Y.-L. (2018). Towards empathetic open-domain conversation models: a new benchmark and dataset. In *Annual Meeting of the Association for Computational Linguistics*.

Rashkin H., Smith E.M., Li M. and Boureau Y.-L. (2019). Towards empathetic open-domain conversation models: a new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 5370–5381. https://doi.org/10.18653/v1/P19-1534

Rastogi A., Zang X., Sunkara S., Gupta R. and Khaitan P. (2020). Towards scalable multi-domain conversational agents: the schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. **34**, pp. 8689–8696. https://doi.org/10.1609/aaai.v34i05.6394

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Reddy S., **Chen D. and Manning C.D.** (2019). CoQA: a conversational question answering challenge. *Transactions of the Association for Computational Linguistics* **7**, 249–266. https://doi.org/10.1162/tacl_a_00266.

Reddy S., **Lapata M. and Steedman M.** (2014). Large-scale semantic parsing without question-answer pairs. *Transactions of the Association for Computational Linguistics* **2**, 377–392. https://doi.org/10.1162/tacl_a_00190.

Ren S., **Wang H.**, **Yu D.**, **Li Y.**, **Zhixing Li S.H. and Zou L.** (2018). Joint intent detection and slot filling with rules. *CCKS Tasks* **2242**, 34–40.

Roller S., **Dinan E.**, **Goyal N.**, **Da J.**, **Williamson M.**, **Liu Y.**, **Xu J.**, **Ott M.**, **Eric Michael Smith Y.-L.B. and Weston J.** (2021). Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, pp. 300–325, Online. https://doi.org/10.18653/v1/2021.eacl-main.24

Ruusuvuori J. (2012). Emotion, affect and conversation. In *The Handbook of Conversation Analysis*, pp. 330–349.

Scao T.L., **Fan A.**, **Akiki C.**, **Pavlick E.**, **Ilić S.**, **Hesslow D.**, **Castagné R.**, **Luccioni A.S.**, **Yvon F.**, **Matthias G.**, **Tow J.**, **Rush A.M.**, **Biderman S.**, **Webson A.**, **Ammanamanchi P.S.**, **Wang T.**, **Sagot B.**, **Muennighoff N.**, **Villanova del Moral A.**, . . . **Wolf T.** (2022). Bloom: A 176b-Parameter Open-Access Multilingual Language Model. arXiv preprint arXiv:2211.05100.

Scholak T., **Schucher N. and Bahdanau D.** (2021). PICARD: parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. pp. 9895–9901. https://doi.org/10.18653/v1/2021.emnlp-main.779

Schuff H., **Vanderlyn L.**, **Adel H. and Vu N.T.** (2023). How to do human evaluation: a brief introduction to user studies in NLP. *Natural Language Engineering*, 1–24. https://doi.org/10.1017/S1351324922000535

Sevegnani K., **Howcroft D.M.**, **Konstas I. and Rieser V.** (2021). OTTers: one-turn topic transitions for open-domain dialogue. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 2492–2504, Online. https://doi.org/10.18653/v1/2021.acl-long.194

Shalyminov I., **Lee S.**, **Eshghi A. and Lemon O.** (2019). Few-shot dialogue generation without annotated data: a transfer learning approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Stockholm, Sweden. Association for Computational Linguistics, pp. 32–39. https://doi.org/10.18653/v1/W19-5904

Shalyminov I., **Sordoni A.**, **Atkinson A. and Schulz H.** (2020). Fast domain adaptation for goal-oriented dialogue using a hybrid generative-retrieval transformer. In *ICASSP. 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8039–8043. https://doi.org/10.1109/ICASSP40776.2020.9053599

Shen W., **Wu S.**, **Yang Y. and Quan X.** (2021). Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 1551–1560, Online. https://doi.org/10.18653/v1/2021.acl-long.123

Shikha N., **Naidu K.**, **Choudhury A.R. and Kayarvizhy N.** (2022). Smart memory companion for elderly. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. IEEE, pp. 1497–1502.

Shin J., **Xu P.**, **Madotto A. and Fung P.** (2019). HappyBot: Generating Empathetic Dialogue Responses by Improving User Experience Look-ahead. ArXiv abs/1906.08487.

Shum M., **Zheng S.**, **Kryscinski W.**, **Xiong C. and Socher R.** (2020). Sketch-fill-A-R: a persona-grounded chit-chat generation framework. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Association for Computational Linguistics, pp. 118–131, Online. https://doi.org/10.18653/v1/2020.nlp4convai-1.14

Shuster K., **Xu J.**, **Komeili M.**, **Da J.**, **Smith E.M.**, **Roller S.**, **Ung M.**, **Chen M.**, **Arora K.**, **Joshua L.**, **Behrooz M.**, **Ngan W.**, **Poff S.**, **Goyal N.**, **Szlam A.**, **Boureau Y.-L.**, **Kambadur M. and Weston J.** (2022). Blenderbot 3: A Deployed Conversational Agent that Continually Learns to Responsibly Engage. arXiv preprint arXiv:2208.03188.

Siddique A.B., **Jamour F.**, **Xu L. and Hristidis V.** (2021). *Generalized Zero-Shot Intent Detection via Commonsense Knowledge (SIGIR '21)*, New York, NY, USA, Association for Computing Machinery, pp. 1925–1929. https://doi.org/10.1145/3404835.3462985

Song X., **Huang L.**, **Xue H. and Hu S.** (2022). Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics, pp. 5197–5206. https://aclanthology.org/2022.emnlp-main.347

Srivastava A., **Pandey I.**, **Akhtar Md.S. and Chakraborty T.** (2023). Response-act guided reinforced dialogue generation for mental health counseling. In *Proceedings of the ACM Web Conference 2023 (Austin, TX, USA) (WWW'23)*, New York, NY, USA. Association for Computing Machinery, pp. 1118–1129. https://doi.org/10.1145/3543507.3583380

Srivastava A., **Suresh T.**, **Lord S.P.**, **Akhtar Md.S. and Chakraborty T.** (2022). Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Washington DC, USA) (KDD '22)*, New York, NY, USA. Association for Computing Machinery, pp. 3920–3930. https://doi.org/10.1145/3534678.3539187

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Stiennon N., Ouyang L., Wu J., Ziegler D., Lowe R., Voss C., Radford A., Amodei D. and Christiano P.F. (2020). Learning to summarize with human feedback. In Larochelle H., Ranzato M., Hadsell R., Balcan M. F. and Lin H. (eds), *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc., pp. 3008–3021. https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf

Strathearn C. and Gkatzia D. (2022). Task2Dial: a novel task and dataset for commonsense-enhanced task-based dialogue grounded in documents. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering. Association for Computational Linguistics*, Dublin, Ireland, pp. 187–196. https://doi.org/10.18653/v1/2022.dialdoc-1.21

Su Y., Cai D., Wang Y., Baker S., Korhonen A., Collier N. and Liu X. (2020). Stylistic Dialogue Generation via Information-Guided Reinforcement Learning Strategy. arXiv preprint arXiv:2004.02202.

Suhr A., Chang M.-W., Shaw P. and Lee K. (2020). Exploring unexplored generalization challenges for cross-database semantic parsing. In *Annual Meeting of the Association for Computational Linguistics*.

Sun Y., Ni X., Feng H., Ray L.C., Lee C.H. and Asadipour A. (2022a). Bringing stories to life in 1001 nights: a co-creative text adventure game using a story generation model. In *International Conference on Interactive Digital Storytelling*. Springer, pp. 651–672.

Sun Q., Wang Y., Xu C., Zheng K., Yang Y., Hu H., Xu F., Zhang J., Geng X. and Jiang D. (2022b). Multimodal dialogue response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland. Association for Computational Linguistics, pp. 2854–2866. https://doi.org/10.18653/v1/2022.acl-long.204

Tang H., Ji D. and Zhou Q. (2020). End-to-end masked graph-based CRF for joint slot filling and intent detection. *Neurocomputing* **413**, 348–359. https://doi.org/10.1016/j.neucom.2020.06.113.

Tewari A., Chhabria A., Khalsa A.S., Chaudhary S. and Kanal H. (2021). A survey of mental health chatbots using NLP. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*.

Thoppilan R., De Freitas D., Hall J., Shazeer N., Kulshreshtha A., Cheng H.-T., Jin A., Bos T., Baker L., Du Y., Li Y., Lee H., Zheng H.S., Ghafouri A., Menegali M., Huang Y., Krikun M., Lepikhin D., Qin J., . . . **Le Q.** (2022). LaMDA: Language Models for Dialog Applications. arXiv:2201.08239 [cs.CL].

Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M.-A., Lacroix T., Rozière B., Goyal N., Hambro E., Azhar F., Rodriguez A., Joulin A., Grave E. and Lample G. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL].

Troiano E., Velutharambath A. and Klinger R. (2023). From theories on styles to their transfer in text: bridging the gap with a hierarchical survey. *Natural Language Engineering* **29**(4), 849–908. https://doi.org/10.1017/S1351324922000407

Tuggener D., Mieskes M., Deriu J. and Cieliebak M. (2021). Are we summarizing the right way? A survey of dialogue summarization data sets. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, Online and in Dominican Republic. Association for Computational Linguistics, pp. 107–118. https://doi.org/10.18653/v1/2021.newsum-1.12

van der Lee C., Gatt A., van Miltenburg E. and Krahmer E. (2021). Human evaluation of automatically generated text: current trends and best practice guidelines. *Computer Speech & Language* **67**, 101151. https://doi.org/10.1016/j.csl.2020.101151

Venkataram H.S., Mattmann C.A. and Penberthy S. (2020). TopiQAL: topic-aware question answering using scalable domain-specific supercomputers. In *2020 IEEE/ACM Fourth Workshop on Deep Learning on Supercomputers (DLS)*. IEEE, pp. 48–55.

Vinyals O., Bengio S. and Kudlur M. (2015). Order Matters: Sequence to Sequence for Sets. arXiv preprint arXiv:1511.06391.

Vu T., Barua A., Lester B., Cer D., Iyyer M. and Constant N. (2022). Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics, pp. 9279–9300. https://doi.org/10.18653/v1/2022.emnlp-main.630

Vulić I., Casanueva I., Spithourakis G., Mondal A., Wen T.-H. and Budzianowski P. (2022). Multi-label intent detection via contrastive task specialization of sentence encoders. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics, pp. 7544–7559. https://aclanthology.org/2022.emnlp-main.512

Wakabayashi K., Takeuchi J. and Nakano M. (2022). Robust slot filling modeling for incomplete annotations using segmentation-based formulation. *Transactions of the Japanese Society for Artificial Intelligence* **37**(3), IDS–E_1-12. https://doi.org/10.1527/tjsai.37-3_IDS-E

Wang Y., He T., Fan R., Zhou W. and Tu X. (2019a). Effective utilization of external knowledge and history context in multi-turn spoken language understanding model. In *2019 IEEE International Conference On Big Data (Big Data)*, pp. 960–967. https://doi.org/10.1109/BigData47090.2019.9006162.

Wang J., Kang D., AbuHussein A. and Collen L.A. (2023a). Designing a Conversational Agent for Education: A Personality-based Approach.

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Wang Y., Rong W., Zhang J., Ouyang Y. and Xiong Z. (2020). Knowledge grounded pre-trained model for dialogue response generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207054

Wang B., Shin R., Liu X., Polozov O. and Richardson M. (2019b). RAT-SQL: relation-aware schema encoding and linking for text-to-SQL parsers. In *Annual Meeting of the Association for Computational Linguistics*.

Wang Z., Tu Y., Rosset C., Craswell N., Wu M. and Ai Q. (2023b). Zero-shot Clarifying Question Generation for Conversational Search. arXiv:2301.12660 [cs.IR].

Webb N. (2000). Rule-based dialogue management systems. In *Proceedings of the 3rd International Workshop on Human-Computer Conversation*, Bellagio, Italy, pp. 3–5.

Weizenbaum J. (1966). ELIZA–a computer program for the study of natural language communication between man and machine. *Communications of The ACM* **9**(1), 36–45. https://doi.org/10.1145/365153.365168

Weld H., Huang X., Long S., Poon J. and Han S.C. (2022a). A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys* **55**(8), 38 pp, Article 156. https://doi.org/10.1145/3547138

Weld H., Huang X., Long S., Poon J. and Han S.C. (2022b). A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys* **55**(8), 1–38.

Weston J., Bordes A., Chopra S., Rush A.M., van Merriënboer B., Joulin A. and Mikolov T. (2015). Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. arXiv:1502.05698 [cs.AI].

Williams J.D. (2003). A probabilistic model of human/computer dialogue with application to a partially observable Markov decision process. PhD first year report. Department of Engineering, University of Cambridge.

Williams J.D., Poupart P. and Young S. (2005). Factored partially observable Markov decision processes for dialogue management. In *Proceedings of IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Citeseer, pp. 76–82.

Wu Z., Galley M., Brockett C., Zhang Y., Gao X., Quirk C., Koncel-Kedziorski R., Gao J., Hajishirzi H., Ostendorf M. and Dolan B. (2021a). A controllable model of grounded response generation. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**(16), 14085–14093. https://doi.org/10.1609/aaai.v35i16.17658

Wu J., Harris I.G., Zhao H. and Ling G. (2023a). A graph-to-sequence model for joint intent detection and slot filling. In *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, pp. 131–138. https://doi.org/10.1109/ICSC56153.2023.00028

Wu C.-S., Liu L., Liu W., Stenetorp P. and Xiong C. (2021b). Controllable abstractive dialogue summarization with sketch supervision. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, pp. 5108–5122, Online. https://doi.org/10.18653/v1/2021.findings-acl.454

Wu H., Shen X., Lan M., Mao S., Bai X. and Wu Y. (2023b). A multi-task dataset for assessing discourse coherence in chinese essays: structure, theme, and logic analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics, pp. 6673–6688. https://doi.org/10.18653/v1/2023.emnlp-main.412

Wu T., Wang M., Gao H., Qi G. and Li W. (2019). Zero-shot slot filling via latent question representation and reading comprehension. In *PRICAI 2019: Trends in Artificial Intelligence*. Cham:, Springer International Publishing, pp. 123–136.

Xia R. and Ding Z. (2019). Emotion-cause pair extraction: a new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 1003–1012. https://doi.org/10.18653/v1/P19-1096

Xiao S., Zhao Z., Zhang Z., Yan X. and Yang M. (2020). Convolutional hierarchical attention network for query-focused video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. **34**, pp. 12426–12433.

Xie Y. and Pu P. (2021). Generating Empathetic Responses with a Large Scale Dialog Dataset. ArXiv abs/2105.06829.

Xu C., Guo D., Duan N. and McAuley J. (2023). Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data. arXiv:2304.01196 [cs.CL].

Xu Y., Ishii E., Cahyawijaya S., Liu Z., Winata G.D., Madotto A., Su D. and Fung P. (2022). Retrieval-free knowledge-grounded dialogue response generation with adapters. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, Dublin, Ireland. Association for Computational Linguistics, pp. 93–107. https://doi.org/10.18653/v1/2022.dialdoc-1.10

Xu Y. and Zhao H. (2021). Dialogue-oriented pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, pp. 2663–2673, Online. https://doi.org/10.18653/v1/2021.findings-acl.235

Xuan C. (2020). Improving sequence-to-sequence semantic parser for task oriented dialog. In *Proceedings of the First Workshop on Interactive and Executable Semantic Parsing*.

Yang S., Huang X., Lau J.H. and Erfani S. (2022). Robust task-oriented dialogue generation with contrastive pre-training and adversarial filtering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics, pp. 1220–1234. https://doi.org/10.18653/v1/2022.findings-emnlp.88

Yang S., Zhang R. and Erfani S. (2020). Graphdialog: Integrating Graph Knowledge into End-to-End Task-Oriented Dialogue Systems. arXiv preprint arXiv:2010.01447.

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Yih W.-T., Chang M.-W., He X. and Gao J. (2015). Semantic parsing via staged query graph generation: question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China. Association for Computational Linguistics, pp. 1321–1331. https://doi.org/10.3115/v1/P15-1128

Yim W.-W. and Yetisgen M. (2021). Towards automating medical scribing : clinic visit Dialogue2Note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop On Natural Language Processing for Medical Conversations*. Association for Computational Linguistics, pp. 10–20, Online. https://doi.org/10.18653/v1/2021.nlpmc-1.2

Yin P. and Neubig G. (2017). A syntactic neural model for general-purpose code generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics, pp. 440–450. https://doi.org/10.18653/v1/P17-1041

Young T., Xing F., Pandelea V., Ni J. and Cambria E. (2022). Fusing task-oriented and open-domain dialogues in conversational agents. *Proceedings of the AAAI Conference on Artificial Intelligence* **36**(10), 11622–11629. https://doi.org/10.1609/aaai.v36i10.21416

Yu T., Zhang R., Er H., Li S., Xue E., Pang B., Lin X.V., Tan Y.C., Shi T., Li Z., Jiang Y., Yasunaga M., Shim S., Chen T., Fabbri A., Li Z., Chen L., Zhang Y., Dixit S., . . . **Radev D.** (2019). CoSQL: a conversational text-to-SQL challenge towards cross-domain natural language interfaces to databases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics, pp. 1962–1979. https://doi.org/10.18653/v1/D19-1204

Yu W., Zhang H., Pan X., Ma K., Wang H. and Yu D. (2023). Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models. arXiv preprint arXiv:2311.09210.

Yu T., Zhang R., Polozov A., Meek C. and Awadallah A.H. (2021). {SC}oRe: pre-training for context representation in conversational semantic parsing. In *International Conference on Learning Representations*. https://openreview.net/forum?id=oyZxhRI2RiE

Yu T., Zhang R., Yang K., Yasunaga M., Wang D., Li Z., Ma J., Li I., Yao Q., Roman S., Zhang Z. and Radev D. (2018). Spider: a large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 3911–3921. https://doi.org/10.18653/v1/D18-1425

Yusupov I. and Kuratov Y. (2018). NIPS conversational intelligence challenge 2017 winner system: skill-based conversational agent with supervised dialog manager. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. Association for Computational Linguistics, pp. 3681–3692. https://aclanthology.org/C18-1312

Zechner K. and Waibel A. (2000). DIASUMM: flexible summarization of spontaneous dialogues in unrestricted domains. In *COLING. 2000 Volume 2: The 18th International Conference on Computational Linguistics*. https://aclanthology.org/C00-2140

Zhan H., Zhang H., Chen H., Ding Z., Bao Y. and Lan Y. (2021). Augmenting knowledge-grounded conversations with sequential knowledge transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 5621–5630, Online. https://doi.org/10.18653/v1/2021.naacl-main.446

Zhang X., Chen Y. and ying Li G. (2021). Multi-modal sarcasm detection based on contrastive attention mechanism. In *Natural Language Processing and Chinese Computing*

Zhang S., Dinan E., Urbanek J., Szlam A., Kiela D. and Weston J. (2018). Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 2204–2213. https://doi.org/10.18653/v1/P18-1205

Zhang T., Kishore V., Wu F., Weinberger K.Q. and Artzi Y. (2020a). BERTScore: evaluating text generation with BERT. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

Zhang Y., Liu Y., Yang Z., Fang Y., Chen Y., Radev D., Zhu C., Zeng M. and Zhang R. (2023). MACSum: controllable summarization with mixed attributes. *Transactions of the Association for Computational Linguistics* **11**, 787–803. https://doi.org/10.1162/tacl_a_00575 2023.

Zhang Q., Shen X., Chang E., Ge J. and Chen P. (2022). MDIA: A Benchmark for Multilingual Dialogue Generation in 46 Languages. arXiv:2208.13078 [cs.CL].

Zhang Y., Sun S., Galley M., Chen Y.-C., Brockett C., Gao X., Gao J., Liu J. and Dolan B. (2020b). DIALOGPT : large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, pp. 270–278, Online. https://doi.org/10.18653/v1/2020.acl-demos.30

Zhang Z., Takanobu R., Zhu Q., Huang M. and Zhu X. (2020c). Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences* **63**(10), 2011–2027.

Zhang W., Yang F. and Liang Y. (2019). A Bayesian framework for joint target tracking, classification, and intent inference. *IEEE Access* **7**, 66148–66156. https://doi.org/10.1109/ACCESS.2019.2917541

https://doi.org/10.1017/nlp.2024.42 Published online by Cambridge University Press

Zhang Z. and Zhao H. (2021). Structural pre-training for dialogue comprehension. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 5134–5145, Online. https://doi.org/10.18653/v1/2021.acl-long.399

Zhao S., Meyers A. and Grishman R. (2004). Discriminative slot detection using kernel methods. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics. COLING*, Geneva, Switzerland, pp. 757–763. https://aclanthology.org/C04-1109

Zhao W., Zhao Y. and Qin B. (2022). MuCDN: mutual conversational detachment network for emotion recognition in multi-party conversations. In *Proceedings of the 29th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Gyeongju, Republic of Korea*, pp. 7020–7030. https://aclanthology.org/2022.coling-1.612

Zhaojiang L., Andrea M., Genta I.W., Peng X., Feijun J., Yuxiang H., Chen S. and Pascale F. (n.d). BiToD: a bilingual multi-domain dataset for task-oriented dialogue modeling. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Zhong M., Da Y., Yu T., Zaidi A., Mutuma M., Jha R., Awadallah A.H., Celikyilmaz A., Liu Y., Qiu X. and Radev D. (2021). QMSum: a new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Kristina Toutanova, Anna Rumshisky*. Association for Computational Linguistics, pp. 5905–5921, Online. https://doi.org/10.18653/v1/2021.naacl-main.472

Zhong V., Xiong C. and Socher R. (2017). Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. CoRR abs/1709.00103.

Zhou P., Gopalakrishnan K., Hedayatnia B., Kim S., Pujara J., Ren X., Liu Y. and Hakkani-Tur D. (2022). Think before you speak: explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1237–1252.

Zhou C., Li Q., Li C., Yu J., Liu Y., Wang G., Zhang K., Ji C., Yan Q., Lifang H., He L., Peng H., Li J., Wu J., Liu Z., Xie P., Xiong C., Pei J., Yu P.S. and Sun L. (2023a). A Comprehensive Survey on Pretrained Foundation Models: A History from Bert to Chatgpt. arXiv preprint arXiv: 2302.09419.

Zhou K., Prabhumoye S. and Black A.W. (2018). A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 708–713. https://doi.org/10.18653/v1/D18-1076

Zhou L. and Small K. (2020). Multi-domain Dialogue State Tracking as Dynamic Knowledge Graph Enhanced Question Answering. arXiv:1911.06192 [cs.CL].

Zhou Y., Yang J., Wang P. and Qiu X. (2023b). Two birds one stone: dynamic ensemble for OOD intent classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10659–10673.

Zhu C., Liu Y., Mei J. and Zeng M. (2021). MediaSum: a large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 5927–5934, Online. https://doi.org/10.18653/v1/2021.naacl-main.474

Cite this article: Kumar S, Bhatia S, Aggarwal M and Chakraborty T. Dialogue agents 101: a beginner's guide to critical ingredients for designing effective conversational systems. *Natural Language Processing* https://doi.org/10.1017/nlp.2024.42