"Do you follow me?":
A Survey of Recent Approaches in Dialogue State Tracking

Léo Jacqmin[1,2]   Lina M. Rojas-Barahona[1]   Benoit Favre[2]

[1]Orange Innovation, Lannion, France
[2]Aix-Marseille Université / CNRS / LIS, Marseille, France
`{leo.jacqmin, linamaria.rojasbarahona}@orange.com`
`benoit.favre@lis-lab.fr`

Abstract

While communicating with a user, a task-oriented dialogue system has to track the user's needs at each turn according to the conversation history. This process called dialogue state tracking (DST) is crucial because it directly informs the downstream dialogue policy. DST has received a lot of interest in recent years with the text-to-text paradigm emerging as the favored approach. In this review paper, we first present the task and its associated datasets. Then, considering a large number of recent publications, we identify highlights and advances of research in 2021-2022. Although neural approaches have enabled significant progress, we argue that some critical aspects of dialogue systems such as generalizability are still underexplored. To motivate future studies, we propose several research avenues.

1 Introduction

Since human conversation is inherently complex and ambiguous, creating an open-domain conversational agent capable of performing arbitrary tasks is an open problem. Therefore, the practice has focused on building task-oriented dialogue (TOD) systems limited to specific domains such as flight booking. These systems are typically implemented through a modular architecture that provides more control and allows interaction with a database, desirable features for a commercial application. Among other components, dialogue state tracking (DST) seeks to update a representation of the user's needs at each turn, taking into account the dialogue history. It is a key component of the system: the downstream dialogue policy uses this representation to predict the next action to be performed (e.g. asking for clarification). A response is then generated based on this action. Beyond processing isolated turns, DST must be able to accurately accumulate information during a conversation and adjust its prediction according to the observations to provide a summary of the dialogue so far.
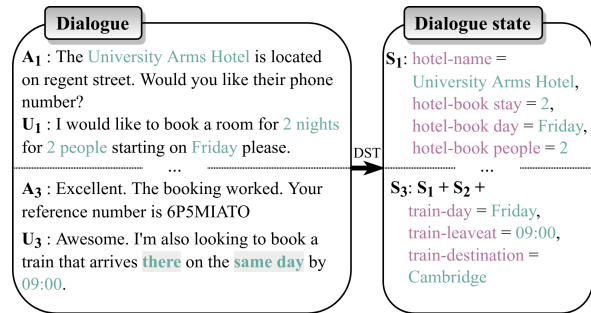

Figure 1: Dialogue state tracking (DST) example taken from MultiWOZ. DST seeks to update the dialogue state at each turn as (slot, value) pairs. Current DST models struggle to parse U_3, which requires world knowledge and long-term context awareness.

Several papers have surveyed DST from the perspective of a specific paradigm or period. Young et al. (2013) give an overview of POMDP-based statistical dialogue systems. These generative approaches model the dialogue in the form of a dynamic bayesian network to track a belief state. They made it possible to account for the uncertainty of the input related to speech recognition errors and offered an alternative to conventional deterministic systems which were expensive to implement and often brittle in operation. However, POMDP-based systems presuppose certain independencies to be tractable and were gradually abandoned in favor of discriminative approaches that directly model the dialogue state distribution. These methods, surveyed by Henderson (2015), rely on annotated dialogue corpora such as those from the Dialogue State Tracking Challenge, the first standardized benchmark for this task (Williams et al., 2016). The past years have been marked by the use of neural models that have allowed significant advances, covered by Balaraman et al. (2021). Since then, many works dealing with key issues such as adaptability have been published. The first contribution of this paper is a synthesis of recent advances in order to identify the major achievements in the field.

The recent development of neural models has addressed fundamental issues in dialogue systems. DST models can now be decoupled from a given domain and shared across similar domains (Wu et al., 2019). They even achieve promising results in zero-shot scenarios (Lin et al., 2021b). Despite these advances, modern approaches are still limited to a specific scenario as dialogues in most DST datasets consist of filling in a form: the system asks for constraints until it can query a database and returns the results to the user. Moreover, these data-driven approaches are rigid and do not correspond to the need for fine-grained control of conversational agents in an application context. In a real-world scenario, access to annotated data is limited but recent DST models appear to have poor generalization capacities (Li et al., 2021a). These limitations show that there remain major challenges to the development of versatile conversational agents that can be adopted by the public. The second contribution of this paper is to propose several research directions to address these challenges.

2 Dialogue State Tracking

A successful TOD system makes it possible to automate the execution of a given task in interaction with a user. The last few years have seen an increase in research in this field, which goes hand in hand with the growing interest of companies in implementing solutions based on TOD systems to reduce their customer support costs.

A typical modular architecture for these systems is composed of the following components: natural language understanding (NLU), DST, dialogue policy, and natural language generation (NLG). NLU consists of two main subtasks, namely, intent detection which consists in identifying the user's intent, such as booking a hotel, and slot filling which consists in identifying relevant semantic concepts, such as price and location. DST aims to update the dialogue state, a representation of the user's needs expressed during the conversation. The dialogue policy predicts the system's next action based on the current state. Along with DST, it forms the dialogue manager which interfaces with the ontology, a structured representation of the back-end database containing the information needed to perform the task. Finally, NLG converts the system action into natural language.

In the case of a spoken dialogue system, automatic speech recognition (ASR) and speech synthe-

sis components are integrated to move from speech to text and back. NLU then operates on the ASR hypotheses and is denoted SLU for spoken language understanding.[1] Traditionally, DST operates on the output of SLU to update the dialogue state by processing noise from ASR and SLU. For example, SLU may provide a list of possible semantic representations based on the top ASR hypotheses. DST manages all these uncertainties to update the dialogue state. However, recent DST datasets are collected in text format with no consideration for noisy speech inputs, which has caused slot filling and DST to be studied separately.

Note that such modular architecture is one possible approach to conversational AI and others have been considered. Another approach is end-to-end systems which generate a response from the user's utterance using a continuous representation. Such systems have been studied extensively for open-domain dialogue (Huang et al., 2020) and have also been applied to TOD (Bordes et al., 2017).

2.1 Dialogue State

Consider the example dialogue shown in Figure 1. Here, the agent acts as a travel clerk and interacts with the user to plan a trip according to the user's preferences. The dialogue state is a formal representation of these constraints and consists of a set of predefined slots, each of which can take a possible value, e.g. book day = Friday. At a given turn t, the goal of DST is to extract values from the dialogue context and accumulate information from previous turns as the dialogue state S_t, which is then used to guide the system in choosing its next action towards accomplishing the task, e.g. finding a suitable hotel. Traditionally, informable slots (constraints provided by the user, e.g. price range) are distinguished from requestable slots (information that the user can request, e.g. phone number). Informable slots can take a special value: don't care when the user has no preference and none when the user has not specified a goal for the slot. The current dialogue state can be predicted either from the previous turn using the previous dialogue state S_{t-1} as context, or from the entire dialogue history, predicting a new dialogue state S_t at each turn.

The dialogue state representation as (*slot, value*) pairs is used to deal with a single domain. When

---

[1]Another possibility is end-to-end approaches which seek to obtain a semantic representation directly from speech.

dealing with a multidomain corpus, this approach is usually extended by combining domains and slots to extract (*domain-slot*, *value*) pairs.[2]

2.2 Datasets

Many public datasets have been published to advance machine learning approaches for DST. The evolution of these datasets is marked by increasing dialogue complexity, especially with the advent of multidomain corpora. We distinguish two main approaches for collecting dialogue datasets: (i) Wizard-of-Oz (WOZ or H2H for *human-to-human*) approaches where two humans (asynchronously) play the roles of user and agent according to a task description. This approach allows for natural and varied dialogues, but the subsequent annotation can be a source of errors. (ii) Simulation-based approaches (M2M for *machine-to-machine*) where two systems play the roles of user and agent and interact with each other to generate conversation templates that are then paraphrased by humans. The advantage of this method is that the annotations are obtained automatically. However, the complexity of the task and linguistic diversity are often limited because the dialogue is simulated.

Table 1 lists the main datasets as well as recent datasets relevant to the problems discussed in Section 4. In addition to the datasets mentioned, there are other less recent corpora listed in the survey of Balaraman et al. (2021) that we do not include here for the sake of clarity. What follows is a description of the main datasets. Recent datasets, namely SMCalFlow[3] (Andreas et al., 2020), ABCD[4] (Chen et al., 2021), SIMMC 2.0[5] (Kottur et al., 2021a), BiToD[6] (Lin et al., 2021d) and DSTC10[7] (Kim et al., 2021), will be discussed in Section 4.

DSTC2[8] (**Henderson et al., 2014a**) Early editions of the Dialogue State Tracking Challenge (DSTC) introduced the first shared datasets and evaluation metrics for DST and thus catalyzed research in this area. The corpus of the second edition remains a reference today. It includes dialogues between paid participants and various telephone dialogue systems (H2M collection). The user needs to find a restaurant by specifying constraints such as the type of cuisine and can request specific information such as the phone number.

MultiWOZ[9] (**Budzianowski et al., 2018**) The first large-scale multidomain corpus and currently the main benchmark for DST. It contains dialogues between a tourist and a travel clerk that can span several domains. A major problem related to the way the data was collected is the inconsistency and errors of annotation, which was crowdsourced. Four more versions were later released to try and fix these errors. (Eric et al., 2019; Zang et al., 2020; Han et al., 2021; Ye et al., 2021a). Multilingual versions[10] were obtained by a process of machine translation followed by manual correction (Gunasekara et al., 2020; Zuo et al., 2021).

SGD[11] (**Rastogi et al., 2020b**) The Schema-Guided Dataset was created to elicit research on domain independence through the use of schemas. Schemas describe domains, slots, and intents in natural language and can be used to handle unseen domains. The test set includes unseen schemas to encourage model generalization. SGD-X is an extension designed to study model robustness to different schema wordings (Lee et al., 2021b).

2.3 Evaluation Metrics

As there is a strict correspondence between dialogue history and dialogue state, DST models' performance is measured with accuracy metrics. Two metrics introduced by Williams et al. (2013) are commonly used for joint and individual evaluation.

Joint goal accuracy (JGA) Main metric referring to the set of user goals. JGA indicates the performance of the model in correctly predicting the dialogue state at a given turn. It is equivalent to the proportion of turns where all predicted values for all slots exactly match the reference values. A similar metric is the average goal accuracy which only evaluates predictions for slots present in the reference dialogue state (Rastogi et al., 2020a).

---

[2]For simplicity, in the rest of this paper, we use the term "slot" to refer to "domain-slots"

[3]https://microsoft.github.io/task_oriented_dialogue_as_dataflow_synthesis/

[4]https://github.com/asappresearch/abcd

[5]https://github.com/facebookresearch/simmc2

[6]https://github.com/HLTCHKUST/BiToD

[7]https://github.com/alexa/alexa-with-dstc10-track2-dataset

[8]https://github.com/matthen/dstc

[9]https://github.com/budzianowski/multiwoz

[10]Mandarin, Korean, Vietnamese, Hindi, French, Portuguese, and Thai.

[11]https://github.com/google-research-datasets/dstc8-schema-guided-dialogue

	DSTC2	MultiWOZ	SGD	SMCalFlow	ABCD	SIMMC2.0	BiToD	DSTC10
Language(s)	en	en, 7 lang.*	en, ru, ar, id, sw	en	en	en	en, zh	en
Collection	H2M	H2H	M2M	H2H	H2H	M2M	M2M	H2H
Modality	speech	text	text	text	text	text, image	text	speech‡
No. domains	1	7	16	4	30	1	5	3
No. dialogues	1612	8438	16,142	41,517	8034	11,244	5787	107
No. turns	23,354	115,424	329,964	155,923	177,407	117,236	115,638	2292
Avg. turns / dial.	14.5	13.7	20.4	8.2	22.1	10.4	19.9	12.6
Avg. tokens / turn	8.5	13.8†	9.7†	10.2	9.2	12.8	12.2†	17.8
Dialogue state	slot-val	slot-val	slot-val	program	action	slot-val	slot-val	slot-val
Slots	8	25	214	-	231	12	68	18
Values	212	4510	14,139	-	12,047	64	8206	327

Table 1: Characteristics of the main (left) and recent (right) datasets available for dialogue state tracking. *Machine translated. † Average for English. ‡ In the form of ASR hypotheses.

Slot accuracy   Unlike JGA, this metric evaluates the predicted value of each slot individually at each turn. It is computed as a macro-averaged accuracy: $SA = \frac{\sum_i^n acc_i}{n}$, where n represents the number of slots. For a more detailed evaluation, it can be decomposed according to slot type (e.g. requested slots, typically measured with the F1 score).

Alternative metrics   Though these two metrics are widely used, they can be difficult to interpret. Slot accuracy tends to overestimate performance as most slots are not mentioned in a given turn while JGA can be too penalizing as a single wrong value prediction will result in wrong dialogue states for the rest of the conversation when accumulated through subsequent turns. Two new metrics have been recently proposed to complement existing metrics and address these shortcomings: Relative slot accuracy ignores unmentioned slots and rewards the model for correct predictions (Kim et al., 2022). Flexible goal accuracy is a generalized version of JGA which is more tolerant of errors that stem from an earlier turn (Dey et al., 2022).

2.4   Modern Approaches

One feature that categorizes DST models is the way they predict slot values. The prediction can be made either from a predefined set of values (*fixed ontology*) or from an open set of values (*open vocabulary*). What follows is a description of these approaches. A selection of recent models is presented in Table 2 based on this taxonomy.

Fixed ontology   Following the discriminative approaches that preceded them, traditional neural approaches are based on a fixed ontology and treat DST as a multiclass classification problem (Henderson et al., 2014b; Mrkšić et al., 2017). Predictions for a given slot are estimated by a probability distribution over a predefined set of values, restricting the prediction field to a closed vocabulary and thus simplifying the task considerably. The performance of this approach is therefore relatively high (Chen et al., 2020), however, its cost is proportional to the size of the vocabulary as all potential values have to be evaluated. In practice, the number of values can be large and a predefined ontology is rarely available.

Open vocabulary   To overcome these limitations, approaches to predict on an open set of values have been proposed. The first method consists in extracting values directly from the dialogue history, e.g. by formulating DST as a reading comprehension task (Gao et al., 2019). This method depends solely on the dialogue context to extract value spans, however, slot values can be implicit or have different wordings (e.g. the value "expensive" may be expressed as "high-end"). An alternative is to generate slot values using an encoder-decoder architecture. For instance, TRADE uses a copy mechanism to generate a value for each slot based on a representation of the dialogue history (Wu et al., 2019). A common current approach is to decode the dialogue state using a pretrained autoregressive language model (Hosseini-Asl et al., 2020).

Hybrid methods   A trade-off seems to exist between the level of value independence in a model and DST performance. Some works have sought to combine fixed ontology approaches with open vocabulary prediction to benefit from the advantages of both methods. This approach is based on the distinction between categorical slots for which a set of values is predefined, and non-categorical slots with an open set of values (Goel et al., 2019; Zhang et al., 2020; Heck et al., 2020).

Model	Decoder	Context	Extra supervision	Ontology	MWOZ2.1
TRADE (Wu et al., 2019)	Generative	Full history	-	✗	45.60
TOD-BERT (Wu et al., 2020)	Classifier	Full history	Pretraining	✓	48.00
NADST (Le et al., 2020)	Generative	Full history	-	✗	49.04
DS-DST (Zhang et al., 2020)	Extract. + classif.	Previous turn	-	Cat. slots	51.21
SOM-DST (Kim et al., 2020)	Generative	History + prev. state	-	✗	52.57
MinTL (Lin et al., 2020)	Generative	History + prev. state	Response generation	✗	53.62
SST (Chen et al., 2020)	Classifier	Prev. turn and state	Schema graph	✓	55.23
TripPy (Heck et al., 2020)	Extractive	Full history	-	✗	55.30
SimpleTOD (Hosseini-Asl et al., 2020)	Generative	Full history	TOD tasks	✗	55.76
Seq2Seq-DU (Feng et al., 2021)	Generative	Full history	Schema	Cat. slots	56.10
SGP-DST (Lee et al., 2021a)	Generative	Full history	Schema	Cat. slots	56.66
SOLOIST (Peng et al., 2021a)	Generative	Full history	Pretraining	✗	56.85
PPTOD (Su et al., 2022)	Generative	Full history	Pretrain + TOD tasks	✗	57.45
D3ST (Zhao et al., 2022)	Generative	Full history	Schema	✗	57.80
TripPy + SCoRe (Yu et al., 2021)	Extractive	Full history	Pretraining	✗	60.48
TripPy + CoCo (Li et al., 2021a)	Extractive	Full history	Data augmentation	✗	60.53
TripPy + SaCLog (Dai et al., 2021)	Extractive	Full history	Curriculum learning	✗	60.61
DiCoS-DST (Guo et al., 2022)	Extract. + classif.	Relevant turns	Schema graph	✓	61.02

Table 2: Characteristics of recent DST models and performance in terms of joint goal accuracy on MultiWOZ 2.1 (Eric et al., 2019). "Ontology" denotes access to a predefined set of values for each slot.

3 Recent Advances

In recent years, several important problems have been addressed, notably through the use of pre-trained language models (PLMs) trained on large amounts of unannotated text. This section summarizes the advances in 2021-2022.

3.1 Modeling Slot Relationships

The methods mentioned so far treat slots individually without taking into account their relations. However, slots are not conditionally independent, for instance, slot values can be correlated, e.g. hotel stars and price range. An alternative is to explicitly consider these relations using self-attention (Ye et al., 2021b). In a similar vein, Lin et al. (2021a) adopt a hybrid architecture to enable sequential value prediction from a GPT-2 model while modeling the relationships between slots and values with a graph attention network (GAT).

Rather than learning these relationships automatically, another line of work uses the knowledge available from the domain ontology, for example by taking advantage of its hierarchical structure (Li et al., 2021c). These relationships can also be represented in a graph with slots and domains as nodes. Chen et al. (2020) first build a schema graph based on the ontology and then use a GAT to merge information from the dialogue history and the schema graph. Feng et al. (2022) extend this approach by dynamically updating slot relations in the schema graph based on the dialogue context. Guo et al. (2022) incorporate both dialogue turns

and slot-value pairs as nodes to consider relevant turns only and solve implicit mentions.

3.2 Adapting PLMs to Dialogues

Though now commonly used for DST, existing PLMs are pretrained on free-form text using language modeling objectives. Their ability to model dialogue context and multi-turn dynamics is therefore limited. It has been shown that adapting a PLM to the target domain or task by continuing self-supervised learning can lead to performance gains (Gururangan et al., 2020). This method has been applied to TOD systems and DST.

There are two underlying questions with this approach: the selection of adaptation data and the formulation of self-supervised training objectives to learn better dialogue representations for the downstream task. Wu et al. (2020) gather nine TOD corpora and continue BERT's pretraining with masked language modeling and next response selection. The obtained model TOD-BERT provides an improvement over a standard BERT model on several TOD tasks including DST. With a similar setup, Zhu et al. (2021) contrast these results and find that such adaptation is most beneficial when little annotated data is available. Based on TOD-BERT, Hung et al. (2022) show that it is advantageous not only to adapt a PLM to dialogues but also to the target domain. To do so, they use conversational data from Reddit filtered to contain terms specific to the target domain. Finally, Yu et al. (2021) introduce two objective functions designed to inject inductive biases into a PLM in order to jointly represent

dynamic dialogue utterances and ontology structure. They evaluate their method on conversational semantic parsing tasks including DST.

3.3 Mitigating Annotated Data Scarcity

The lack of annotated data hinders the development of efficient and robust DST models. However, the data collection process is costly and time-consuming. One approach to address this problem is to train a model on resource-rich domains and apply it to an unseen domain with little or no annotated data (cross-domain transfer; Wu et al., 2019). Dingliwal et al. (2021) adopt meta-learning and use the source domains to meta-learn the model's parameters and initialize fine-tuning for the target domain. Works around schema-based datasets (Rastogi et al., 2020a) use slot descriptions to handle unseen domains and slots (Lin et al., 2021c; Zhao et al., 2022). A drawback of these approaches is that they rely on the similarity between the unseen domain and the initial fine-tuning domains.

Another set of approaches tries to exploit external knowledge from other tasks with more abundant resources. Hudeček et al. (2021) use FrameNet semantic analysis as weak supervision to identify potential slots. Gao et al. (2020); Li et al. (2021b); Lin et al. (2021b) propose different methods to pretrain a model on reading comprehension data before applying it to DST. Similarly, (Shin et al., 2022) reformulate DST as a dialogue summarization task based on templates and leverage external annotated data.

Note that the PLM adaptation approaches seen above allow for more efficient learning when little data is available and are also a potential solution to the data scarcity problem. Along this line, Mi et al. (2021) present a self-learning method complementary to TOD-BERT for few-shot DST.

3.4 Prompting Generative Models to Address Unseen Domains

A recent paradigm tackles all text-based language tasks with a single model by converting them into a text-to-text format (Raffel et al., 2020). This approach relies on textual instructions called prompts that are prepended to the input to condition the model to perform a task. It has been successfully applied to DST, not only closing the performance gap between classification and generation methods but also opening opportunities to address unseen domains using schemas (cfr. Section 2). In addition to slot and domain names, Lee et al. (2021a) include slot descriptions in a prompt to independently generate values for each slot. They also add possible values for categorical slots. Zhao et al. (2021, 2022) expand on this approach by generating the entire dialogue state sequentially, as illustrated in Figure 2. In an analysis of prompt formulation, Cao and Zhang (2021) find that a question format yields better results for prompt-based DST.

Hand-crafting textual prompts may lead to suboptimal conditioning of a PLM. Instead of using discrete tokens, we can optimize these instructions as continuous embeddings, a method called prompt tuning (Lester et al., 2021), which has been applied to DST for continual learning, adding new domains through time (Zhu et al., 2022). Alternatively, Yang et al. (2022) reverse description-based prompts and formulate a prompt based on values extracted from the utterance to generate their respective slot. They argue that slots that appear in a small corpus do not represent all potential requirements whereas values are often explicitly mentioned. Rather than using descriptions that indirectly convey the semantics of a schema, others have sought to prompt a model with instructions, i.e. in-context learning, in which the prompt consists of a few example input-output pairs (Hu et al., 2022; Gupta et al., 2022).

```
Input:
[user] I'm looking for somewhere in Fisherman's
Wharf to have a meal and a drink. [system] What
kind of food do you want to eat? [user] I feel
like a burger. [state] area: the place of the
restaurant, cuisine: the type of food served,
serves_alcohol: whether alcohol is served
Output:
area: Fisherman's Wharf, cuisine: burger,
serves_alcohol: True
```
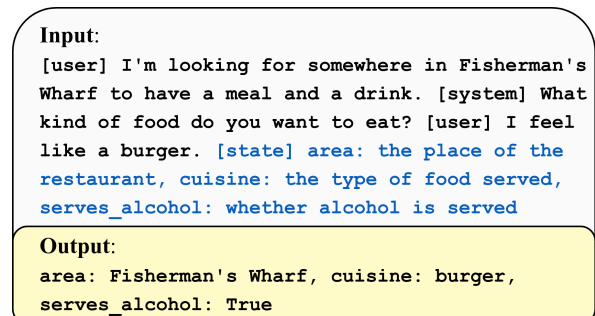
Figure 2: An example of text-to-text DST from SGD (Rastogi et al., 2020b). Including slot descriptions as a prompt makes it possible to address unseen domains.

3.5 Going Beyond English

Until recently, most works around DST were confined to English due to the lack of data in other languages, preventing the creation of truly multilingual models. In recent years, several works have addressed this issue. A DSTC9 track studied cross-lingual DST. For this purpose, a Chinese version of MultiWOZ and an English version of CrossWOZ were obtained by a process of machine translation followed by manual correction (Gunasekara et al., 2020). Zuo et al. (2021) used the same method to

translate MultiWOZ in seven languages.

A problem with machine translations is that they lack naturalness and are not localized. Two Chinese datasets were obtained by H2H collection: CrossWOZ and RiSAWOZ (Zhu et al., 2020; Quan et al., 2020), however, this type of collection is expensive. The M2M approach makes it possible to obtain adequate multilingual corpora by adapting conversation templates according to the target language. Lin et al. (2021d) took advantage of this method to create BiToD, a bilingual English-Chinese corpus. Majewska et al. (2022) used schemas from SGD to create conversation templates that were then adapted into Russian, Arabic, Indonesian and Swahili. Similarly, Ding et al. (2022) translated templates from the MultiWOZ test set in Chinese, Spanish and Indonesian and localized them. Lastly, to overcome the lack of multilingual data, Moghe et al. (2021) leverage parallel and conversational movie subtitle data to pretrain a cross-lingual DST model.

4 Challenges and Future Directions

Despite recent advances, there are still many challenges in developing DST models that can accurately capture user needs dynamically and in a variety of scenarios. There are many interesting paths, this section articulates the needs around three axes: generalization, robustness and relevance of models.

4.1 Generalizability

TOD systems are intended to be deployed in dynamic environments that may involve different settings. In practice, the application domains of these systems are numerous and varied (e.g. customer service in telecommunications, banking, technical support, etc.), which makes manual annotation of corpora for each domain difficult or impossible. This reduces the effectiveness of traditional supervised learning and is one of the reasons why most systems in production are rule-based.

Learning with little or no new annotated data offers an alternative to take advantage of the capabilities of neural networks to guarantee the flexibility of the systems. The importance of this aspect is reflected by the numerous recent works that address this problem, as presented in Section 3. Although significant progress has been made, these approaches remain limited. Models that rely on existing DST resources are unable to handle domains whose distribution deviates from that of the train-

ing data (Dingliwal et al., 2021). Models that use external knowledge offer a more generic approach but achieve relatively poor performance (Lin et al., 2021b). Learning generalizable models thus remains an open problem. An interesting avenue is continual learning, which allows new skills to be added to a system over time after deployment. Without retraining with all the data, the model must be able to accumulate knowledge (Madotto et al., 2021; Liu et al., 2021; Zhu et al., 2022).

In a real scenario, entities that are not observed during training are bound to appear. A DST model must be able to extrapolate from similar entities that have been seen. However, MultiWOZ has been shown to exhibit an entity bias, i.e. the slot values distribution is unbalanced. When a generative DST model was evaluated on a new test set with unseen entities, its performance dropped sharply, hinting at severe memorization (Qian et al., 2021). Similarly, by its constrained nature, a TOD system may be perturbed by out-of-domain utterances. It is desirable to be able to recognize such utterances in order to provide an appropriate response. The ability of a model to generalize to new scenarios is therefore related to its robustness.

4.2 Robustness

Since users may express similar requirements in different ways, a DST model must be able to interpret different formulations of a request in a consistent manner. In other words, it must be robust to variations in the input. Analytical work has shown that models' performance drops when they are confronted with realistic examples that deviate from the test set distribution (Huang et al., 2021; Li et al., 2021a). Recent work has tried to address this issue through regularization techniques (Heck et al., 2022). Related to this, an understudied aspect is dealing with utterances that deviate from the norm in the case of a written dialogue system.

A DST model must be able to take into account all the history and adjust its predictions of the dialogue state using all available information. Many works have found that performance degrades rapidly as the dialogue length increases. Another critical aspect is therefore efficiently processing long dialogues (Zhang et al., 2021). The dialogue state condenses important information, but correcting an error made in an earlier turn may be difficult. To overcome this error propagation issue, Tian et al. (2021a) use a two-pass dialogue state generation

to correct potential errors, while Manotumruksa et al. (2021) propose a turn-based objective function to penalize the model for incorrect prediction in early turns. Despite this need, Jakobovits et al. (2022) have shown that popular DST datasets are not conversational: most utterances can be parsed in isolation. Some works simulate longer and more realistic dialogues by inserting chit-chat into TODs (Kottur et al., 2021b; Sun et al., 2021).

Another point that has not been studied much in recent approaches is robustness to speech inputs (Faruqui and Hakkani-Tür, 2021). For spoken dialogue systems, new challenges arise such as ASR errors or verbal disfluencies. Early editions of DSTC provided spoken corpora that included a transcript and ASR hypotheses. Since then, DST datasets have been primarily text-based. A DSTC10 track considered this aspect again and proposed a DST task with validation and test sets containing ASR hypotheses (Kim et al., 2021).

Learning robust models requires diverse datasets that represent real-world challenges. In this sense, several evaluation benchmarks have been published to study TOD systems' robustness (Lee et al., 2021b; Peng et al., 2021b; Cho et al., 2021). Data augmentation is a potential solution to the lack of variety in datasets (Campagna et al., 2020; Li et al., 2021a; Aksu et al., 2022), especially for simulating ASR errors (Wang et al., 2020; Tian et al., 2021b).

4.3 Relevance

As we have seen, existing datasets do not really reflect real-world conditions resulting in a rather artificial task. It is important to keep a holistic view of the development of DST models to ensure the relevance of their application in a dialogue system.

Since the first TOD systems, the dialogue state has been considered as a form to be filled in as slot-value pairs. This fixed representation is suitable for simple tasks like flight booking but is limited in domains with rich relational structures and a variable number of entities. Indeed, composition is not possible (e.g. "itinerary for my next meeting") and knowledge is not directly shared between slots. Section 3 presented approaches that attempt to address the latter point by using graphs for dialogue state representation. To promote work on more realistic scenarios, some have proposed richer representations with an associated corpus. Andreas et al. (2020) encode the dialogue state as a dataflow graph and introduce the SMCallFlow corpus.

Cheng et al. (2020) propose a tree structure along with the TreeDST corpus. ThingTalk is another alternative representation of the dialogue state which was successfully applied to MultiWOZ (Lam et al., 2022; Campagna et al., 2022). In the ABCD corpus, Chen et al. (2021) adopt a representation of the procedures that a customer service employee must follow in accordance with company policies.

These approaches still rely on specific database schemas and are limited to one modality. For more capable virtual agents, we can extend the scope of dialogues to a multimodal world. Emerging efforts on multimodal DST seek to track the information of visual objects based on a multimodal context. With SIMMC 2.0, Kottur et al. (2021a) introduced a multimodal TOD dataset based on virtual reality rendered shopping scenes. Similarly, Le et al. (2022) proposed a synthetic dataset of dialogues consisting of multiple question-answer pairs about a grounding video input. The questions are based on 3D objects. In both these datasets, the system has to track visual objects in the dialogue state in the form of slot-value pairs. For a broader background on multimodal conversational AI, we refer the reader to Sundar and Heck (2022)'s survey.

Dialogue is dynamic: in a real scenario, an erroneous prediction of the dialogue state would have deviated the course of the conversation from the reference dialogue. However, most studies evaluate models in isolation, assuming that it is always possible to assemble a set of well-performing components to build a good TOD system. The overall performance of a system is rarely taken into account as evaluating the system as a whole is complicated and requires human evaluation. Moreover, it can be difficult to identify which component of the system is problematic and needs to be improved. Despite these hurdles, it is important to consider the impact that DST can have on the dialogue system as a whole. Takanobu et al. (2020) conducted automatic and human evaluations of dialogue systems with a wide variety of configurations and settings on MultiWOZ. They found a drop in task success rate using DST rather than NLU followed by a rule-based dialogue state update. They explain this result by the fact that NLU extracts the user's intentions in addition to the slot-value pairs. Another work showed how uncertainty estimates in belief tracking can lead to a more robust downstream policy (van Niekerk et al., 2021). These studies are rare in their kind and call for more similar work.

5 Conclusion

Dialogue state tracking is a crucial component of a conversational agent to identify the user's needs at each turn of the conversation. A growing body of work is addressing this task and we have outlined the latest developments. After giving an overview of the task and the different datasets available, we have categorized modern neural approaches according to the inference of the dialogue state. Despite encouraging results on benchmarks such as MultiWOZ, these systems lack flexibility and robustness, which are critical skills for a dialogue system. In recent years, many works have sought to address these limitations and we have summarized the advances. However, there are still significant challenges to be addressed in the future. There are many interesting avenues and we have proposed three key features of DST models to guide future research: generalizability, robustness and relevance.

Acknowledgements

References

Ibrahim Aksu, Zhengyuan Liu, Min-Yen Kan, and Nancy Chen. 2022. N-shot learning for augmenting task-oriented dialogue state tracking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1659–1671, Dublin, Ireland. Association for Computational Linguistics.

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. 2020. Task-Oriented Dialogue as Dataflow Synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Vevake Balaraman, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. 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*, pages 239–251, Singapore and Online. Association for Computational Linguistics.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

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

Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica Lam. 2020. Zero-Shot Transfer Learning with Synthesized Data for Multi-Domain Dialogue State Tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 122–132, Online. Association for Computational Linguistics.

Giovanni Campagna, Sina Semnani, Ryan Kearns, Lucas Jun Koba Sato, Silei Xu, and Monica Lam. 2022. A Few-Shot Semantic Parser for Wizard-of-Oz Dialogues with the Precise ThingTalk Representation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4021–4034, Dublin, Ireland. Association for Computational Linguistics.

Jie Cao and Yi Zhang. 2021. A Comparative Study on Schema-Guided Dialogue State Tracking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 782–796, Online. Association for Computational Linguistics.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. Action-Based Conversations Dataset: A Corpus for Building More In-Depth Task-Oriented Dialogue Systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3002–3017, Online. Association for Computational Linguistics.

Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-Guided Multi-Domain Dialogue State Tracking with Graph Attention Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7521–7528.

Jianpeng Cheng, Devang Agrawal, Héctor Martínez Alonso, Shruti Bhargava, Joris Driesen, Federico Flego, Dain Kaplan, Dimitri Kartsaklis, Lin Li, Dhivya Praviperumal, Jason D. Williams, Hong Yu, Diarmuid Ó Séaghdha, and Anders Johannsen. 2020. Conversational Semantic Parsing

for Dialog State Tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8107–8117, Online. Association for Computational Linguistics.

Hyundong Cho, Chinnadhurai Sankar, Christopher Lin, Kaushik Ram Sadagopan, Shahin Shayandeh, Asli Celikyilmaz, Jonathan May, and Ahmad Beirami. 2021. CheckDST: Measuring Real-World Generalization of Dialogue State Tracking Performance. *arXiv:2112.08321 [cs]*.

Yinpei Dai, Hangyu Li, Yongbin Li, Jian Sun, Fei Huang, Luo Si, and Xiaodan Zhu. 2021. 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)*, pages 879–885, Online. Association for Computational Linguistics.

Suvodip Dey, Ramamohan Kummara, and Maunendra Desarkar. 2022. Towards Fair Evaluation of Dialogue State Tracking by Flexible Incorporation of Turn-level Performances. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 318–324, Dublin, Ireland. Association for Computational Linguistics.

Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022. GlobalWoZ: Globalizing MultiWoZ to Develop Multilingual Task-Oriented Dialogue Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1639–1657, Dublin, Ireland. Association for Computational Linguistics.

Saket Dingliwal, Shuyang Gao, Sanchit Agarwal, Chien-Wei Lin, Tagyoung Chung, and Dilek Hakkani-Tur. 2021. Few Shot Dialogue State Tracking using Meta-learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1730–1739, Online. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. *arXiv:1907.01669 [cs]*.

Manaal Faruqui and Dilek Hakkani-Tür. 2021. Revisiting the Boundary between ASR and NLU in the Age of Conversational Dialog Systems. *arXiv:2112.05842 [cs, eess]*.

Yue Feng, Aldo Lipani, Fanghua Ye, Qiang Zhang, and Emine Yilmaz. 2022. Dynamic Schema Graph Fusion Network for Multi-Domain Dialogue State Tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 115–126, Dublin, Ireland. Association for Computational Linguistics.

Yue Feng, Yang Wang, and Hang Li. 2021. A Sequence-to-Sequence Approach to Dialogue State Tracking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1714–1725, Online. Association for Computational Linguistics.

Shuyang Gao, Sanchit Agarwal, Di Jin, Tagyoung Chung, and Dilek Hakkani-Tur. 2020. From Machine Reading Comprehension to Dialogue State Tracking: Bridging the Gap. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 79–89, Online. Association for Computational Linguistics.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialog State Tracking: A Neural Reading Comprehension Approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 264–273, Stockholm, Sweden. Association for Computational Linguistics.

Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. 2019. HyST: A Hybrid Approach for Flexible and Accurate Dialogue State Tracking. In *Interspeech 2019*, pages 1458–1462. ISCA.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, Maxine Eskenazi, Ahmad Beirami, Eunjoon, Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba. 2020. Overview of the Ninth Dialog System Technology Challenge: DSTC9. *arXiv:2011.06486 [cs]*.

Jinyu Guo, Kai Shuang, Jijie Li, Zihan Wang, and Yixuan Liu. 2022. Beyond the Granularity: Multi-Perspective Dialogue Collaborative Selection for Dialogue State Tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2320–2332, Dublin, Ireland. Association for Computational Linguistics.

Raghav Gupta, Harrison Lee, Jeffrey Zhao, Abhinav Rastogi, Yuan Cao, and Yonghui Wu. 2022. Show, Don't Tell: Demonstrations Outperform Descriptions for Schema-Guided Task-Oriented Dialogue. *arXiv:2204.04327 [cs]*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey,

345

and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Ting Han, Ximing Liu, Ryuichi Takanobu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang. 2021. MultiWOZ 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. *arXiv:2010.05594 [cs]*.

Michael Heck, Nurul Lubis, Carel van Niekerk, Shutong Feng, Christian Geishauser, Hsien-Chin Lin, and Milica Gašić. 2022. Robust Dialogue State Tracking with Weak Supervision and Sparse Data. *arXiv:2202.03354 [cs]*.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. TripPy: A Triple Copy Strategy for Value Independent Neural Dialog State Tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 35–44, 1st virtual meeting. Association for Computational Linguistics.

Matthew Henderson. 2015. Machine Learning for Dialog State Tracking: A Review. In *Proceedings of The First International Workshop on Machine Learning in Spoken Language Processing*.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014a. The Second Dialog State Tracking Challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014b. Word-Based Dialog State Tracking with Recurrent Neural Networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A Simple Language Model for Task-Oriented Dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2022. In-Context Learning for Few-Shot Dialogue State Tracking. *arXiv:2203.08568 [cs]*.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in Building Intelligent Open-domain Dialog Systems. *ACM Transactions on Information Systems*, 38(3):21:1–21:32.

Yi Huang, Junlan Feng, Xiaoting Wu, and Xiaoyu Du. 2021. Counterfactual Matters: Intrinsic Probing For Dialogue State Tracking. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 1–6, Online. Association for Computational Linguistics.

Vojtěch Hudeček, Ondřej Dušek, and Zhou Yu. 2021. Discovering Dialogue Slots with Weak Supervision. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2430–2442, Online. Association for Computational Linguistics.

Chia-Chien Hung, Anne Lauscher, Simone Ponzetto, and Goran Glavaš. 2022. DS-TOD: Efficient Domain Specialization for Task-Oriented Dialog. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 891–904, Dublin, Ireland. Association for Computational Linguistics.

Alice Shoshana Jakobovits, Francesco Piccinno, and Yasemin Altun. 2022. What Did You Say? Task-Oriented Dialog Datasets Are Not Conversational!? *arXiv:2203.03431 [cs]*.

Seokhwan Kim, Yang Liu, Di Jin, A. Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Z. Hakkani-Tür. 2021. "How Robust R U?": Evaluating Task-Oriented Dialogue Systems on Spoken Conversations. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2020. Efficient Dialogue State Tracking by Selectively Overwriting Memory. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 567–582, Online. Association for Computational Linguistics.

Takyoung Kim, Hoonsang Yoon, Yukyung Lee, Pilsung Kang, and Misuk Kim. 2022. Mismatch between Multi-turn Dialogue and its Evaluation Metric in Dialogue State Tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 297–309, Dublin, Ireland. Association for Computational Linguistics.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021a. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Satwik Kottur, Chinnadhurai Sankar, Zhou Yu, and Alborz Geramifard. 2021b. DialogStitch: Synthetic Deeper and Multi-Context Task-Oriented Dialogs. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*,

pages 21–26, Singapore and Online. Association for Computational Linguistics.

Monica S. Lam, Giovanni Campagna, Mehrad Moradshahi, Sina J. Semnani, and Silei Xu. 2022. ThingTalk: An Extensible, Executable Representation Language for Task-Oriented Dialogues.

Hung Le, Nancy Chen, and Steven Hoi. 2022. Multimodal dialogue state tracking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3394–3415, Seattle, United States. Association for Computational Linguistics.

Hung Le, Richard Socher, and Steven C. H. Hoi. 2020. Non-Autoregressive Dialog State Tracking. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021a. Dialogue State Tracking with a Language Model using Schema-Driven Prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4937–4949, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Harrison Lee, Raghav Gupta, Abhinav Rastogi, Yuan Cao, Bin Zhang, and Yonghui Wu. 2021b. SGD-X: A Benchmark for Robust Generalization in Schema-Guided Dialogue Systems. *arXiv:2110.06800 [cs]*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shiyang Li, Semih Yavuz, Kazuma Hashimoto, Jia Li, Tong Niu, Nazneen Rajani, Xifeng Yan, Yingbo Zhou, and Caiming Xiong. 2021a. CoCo: Controllable Counterfactuals for Evaluating Dialogue State Trackers. *arXiv:2010.12850 [cs]*.

Shuyang Li, Jin Cao, Mukund Sridhar, Henghui Zhu, Shang-Wen Li, Wael Hamza, and Julian McAuley. 2021b. Zero-shot Generalization in Dialog State Tracking through Generative Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1063–1074, Online. Association for Computational Linguistics.

Xinmeng Li, Qian Li, Wansen Wu, and Quanjun Yin. 2021c. Generation and Extraction Combined Dialogue State Tracking with Hierarchical Ontology Integration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2249, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weizhe Lin, Bo-Hsiang Tseng, and Bill Byrne. 2021a. Knowledge-Aware Graph-Enhanced GPT-2 for Dialogue State Tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7871–7881, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhaojiang Lin, Bing Liu, Andrea Madotto, Seungwhan Moon, Zhenpeng Zhou, Paul Crook, Zhiguang Wang, Zhou Yu, Eunjoon Cho, Rajen Subba, and Pascale Fung. 2021b. Zero-Shot Dialogue State Tracking via Cross-Task Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7890–7900, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhaojiang Lin, Bing Liu, Seungwhan Moon, Paul Crook, Zhenpeng Zhou, Zhiguang Wang, Zhou Yu, Andrea Madotto, Eunjoon Cho, and Rajen Subba. 2021c. Leveraging Slot Descriptions for Zero-Shot Cross-Domain Dialogue StateTracking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5640–5648, Online. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist Transfer Learning for Task-Oriented Dialogue Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021d. 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)*.

Qingbin Liu, Pengfei Cao, Cao Liu, Jiansong Chen, Xunliang Cai, Fan Yang, Shizhu He, Kang Liu, and Jun Zhao. 2021. Domain-Lifelong Learning for Dialogue State Tracking via Knowledge Preservation Networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2301–2311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. 2021. Continual Learning in Task-Oriented Dialogue Systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Olga Majewska, Evgeniia Razumovskaia, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2022. Cross-Lingual Dialogue Dataset Creation via Outline-Based Generation. *arXiv:2201.13405 [cs]*.

Jarana Manotumruksa, Jeff Dalton, Edgar Meij, and Emine Yilmaz. 2021. Improving Dialogue State Tracking with Turn-based Loss Function and Sequential Data Augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1674–1683, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fei Mi, Wanhao Zhou, Lingjing Kong, Fengyu Cai, Minlie Huang, and Boi Faltings. 2021. Self-training Improves Pre-training for Few-shot Learning in Task-oriented Dialog Systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1887–1898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikita Moghe, Mark Steedman, and Alexandra Birch. 2021. Cross-lingual Intermediate Fine-tuning improves Dialogue State Tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1137–1150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021a. SOLOIST: Building Task Bots at Scale with Transfer Learning and Machine Teaching. *arXiv:2005.05298 [cs]*.

Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2021b. RADDLE: An Evaluation Benchmark and Analysis Platform for Robust Task-oriented Dialog Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4418–4429, Online. Association for Computational Linguistics.

Kun Qian, Ahmad Beirami, Zhouhan Lin, Ankita De, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2021. Annotation Inconsistency and Entity Bias in MultiWOZ. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 326–337, Singapore and Online. Association for Computational Linguistics.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. RiSAWOZ: A Large-Scale Multi-Domain Wizard-of-Oz Dataset with Rich Semantic Annotations for Task-Oriented Dialogue Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 930–940, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020a. Schema-Guided Dialogue State Tracking Task at DSTC8. *arXiv:2002.01359 [cs]*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020b. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8689–8696. AAAI Press.

Jamin Shin, Hangyeol Yu, Hyeongdon Moon, Andrea Madotto, and Juneyoung Park. 2022. Dialogue Summaries as Dialogue States (DS2), Template-Guided Summarization for Few-shot Dialogue State Tracking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3824–3846, Dublin, Ireland. Association for Computational Linguistics.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.

Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding Chit-Chat to Enhance Task-Oriented Dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1570–1583, Online. Association for Computational Linguistics.

Anirudh Sundar and Larry Heck. 2022. Multimodal conversational AI: A survey of datasets and approaches. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 131–147, Dublin, Ireland. Association for Computational Linguistics.

Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. 2020. Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 297–310, 1st virtual meeting. Association for Computational Linguistics.

Xin Tian, Liankai Huang, Yingzhan Lin, Siqi Bao, Huang He, Yunyi Yang, Hua Wu, Fan Wang, and Shuqi Sun. 2021a. Amendable Generation for Dialogue State Tracking. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 80–92, Online. Association for Computational Linguistics.

Xin Tian, Xinxian Huang, Dongfeng He, Yingzhan Lin, Siqi Bao, Huang He, Liankai Huang, Qiang Ju, Xiyuan Zhang, Jian Xie, Shuqi Sun, Fan Wang, Hua Wu, and Haifeng Wang. 2021b. TOD-DA: Towards Boosting the Robustness of Task-oriented Dialogue Modeling on Spoken Conversations. *arXiv:2112.12441 [cs]*.

Carel van Niekerk, Andrey Malinin, Christian Geishauser, Michael Heck, Hsien-chin Lin, Nurul Lubis, Shutong Feng, and Milica Gasic. 2021. Uncertainty Measures in Neural Belief Tracking and the Effects on Dialogue Policy Performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7914, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Longshaokan Wang, Maryam Fazel-Zarandi, Aditya Tiwari, Spyros Matsoukas, and Lazaros Polymenakos. 2020. Data Augmentation for Training Dialog Models Robust to Speech Recognition Errors. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 63–70, Online. Association for Computational Linguistics.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France. Association for Computational Linguistics.

Jason D. Williams, Antoine Raux, and Matthew Henderson. 2016. The Dialog State Tracking Challenge Series: A Review. *Dialogue & Discourse*, 7(3):4–33.

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

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems. In *Proceedings*

of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.

Yuting Yang, Wenqiang Lei, Juan Cao, Jintao Li, and Tat-Seng Chua. 2022. Prompt Learning for Few-Shot Dialogue State Tracking. *arXiv:2201.05780 [cs]*.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021a. MultiWOZ 2.4: A Multi-Domain Task-Oriented Dialogue Dataset with Essential Annotation Corrections to Improve State Tracking Evaluation. *arXiv:2104.00773 [cs]*.

Fanghua Ye, Jarana Manotumruksa, Qiang Zhang, Shenghui Li, and Emine Yilmaz. 2021b. Slot Self-Attentive Dialogue State Tracking. In *Proceedings of the Web Conference 2021*, WWW '21, pages 1598–1608, New York, NY, USA. Association for Computing Machinery.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE*, 101(5):1160–1179.

Tao Yu, Rui Zhang, Alex Polozov, Chris Meek, and Ahmed H. Awadallah. 2021. SCoRe: Pre-Training for Context Representation in Conversational Semantic Parsing. In *International Conference on Learning Representations*.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, Philip Yu, Richard Socher, and Caiming Xiong. 2020. Find or Classify? Dual Strategy for Slot-Value Predictions on Multi-Domain Dialog State Tracking. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 154–167, Barcelona, Spain (Online). Association for Computational Linguistics.

Ye Zhang, Yuan Cao, Mahdis Mahdieh, Jeffrey Zhao, and Yonghui Wu. 2021. Improving Longer-range Dialogue State Tracking. *arXiv:2103.00109 [cs]*.

Jeffrey Zhao, Raghav Gupta, Yuan Cao, Dian Yu, Mingqiu Wang, Harrison Lee, Abhinav Rastogi, Izhak Shafran, and Yonghui Wu. 2022. Description-Driven Task-Oriented Dialog Modeling. *arXiv:2201.08904 [cs]*.

Jeffrey Zhao, Mahdis Mahdieh, Ye Zhang, Yuan Cao, and Yonghui Wu. 2021. Effective Sequence-to-Sequence Dialogue State Tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7486–7493, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qi Zhu, Yuxian Gu, Lingxiao Luo, Bing Li, Cheng Li, Wei Peng, Minlie Huang, and Xiaoyan Zhu. 2021. When does Further Pre-training MLM Help? An Empirical Study on Task-Oriented Dialog Pre-training. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 54–61, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. CrossWOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

Qi Zhu, Bing Li, Fei Mi, Xiaoyan Zhu, and Minlie Huang. 2022. Continual Prompt Tuning for Dialog State Tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1124–1137, Dublin, Ireland. Association for Computational Linguistics.

Lei Zuo, Kun Qian, Bowen Yang, and Zhou Yu. 2021. AllWOZ: Towards Multilingual Task-Oriented Dialog Systems for All. *arXiv:2112.08333 [cs]*.