Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1286–1296
Marseille, 20-25 June 2022

DIASER: A Unifying View On Task-oriented Dialogue Annotation

Vojtěch Hudeček[1], Léon-Paul Schaub[2,3], Daniel Štancl[1], Patrick Paroubek[2], Ondřej Dušek[1]
hudecek@ufal.mff.cuni.cz, schaub@limsi.fr, stancl@ufal.mff.cuni.cz, pap@limsi.fr, odusek@ufal.mff.cuni.cz
[1]Charles University, Faculty of Mathematics and Physics
Malostranské náměstí 25, 118 00 Prague, Czechia
[2]Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique
91400 Orsay, France
[3]Akio
43 rue de Dunkerque, 75010 Paris, France

Abstract

Every model is only as strong as the data that it is trained on. In this paper, we present a new dataset, obtained by merging four publicly available annotated corpora for task-oriented dialogues in several domains (MultiWOZ 2.2, CamRest676, DSTC2 and Schema-Guided Dialogue Dataset). This way, we assess the feasibility of providing a unified ontology and annotation schema covering several domains with a relatively limited effort. We analyze the characteristics of the resulting dataset along three main dimensions: language, information content and performance. We focus on aspects likely to be pertinent for improving dialogue success, e.g. dialogue consistency. Furthermore, to assess the usability of this new corpus, we thoroughly evaluate dialogue generation performance under various conditions with the help of two prominent recent end-to-end dialogue models: MarCo and GPT-2. These models were selected as popular open implementations representative of the two main dimensions of dialogue modelling. While we did not observe a significant gain for dialogue state tracking performance, we show that using more training data from different sources can improve language modelling capabilities and positively impact dialogue flow (consistency). In addition, we provide the community with one of the largest open dataset for machine learning experiments.

Keywords: task oriented dialog, annotated corpora, resource merging

1. Introduction

Recent research attention in task-oriented dialogue systems focuses on end-to-end neural models. There have been many successful systems, including systems capable of working with multiple domains (Lei et al., 2018; Mehri et al., 2019; Qin et al., 2020). These neural systems are notoriously data-hungry; especially with pretrained language models (LMs), their capacity vastly exceeds the amounts of data available with task-oriented annotation (Budzianowski and Vulić, 2019; Yang et al., 2021). There are several datasets with similar annotation style and similar domains (Serban et al., 2018), but they are often incompatible and training a model on a union of these datasets remains a challenge (Chen et al., 2019a). Despite recent efforts to provide support at the software level (Zhu et al., 2020b), data unification efforts have been limited so far. The aim of this paper is to address this issue and produce a large-scale open dataset with unified dialogue annotation. We call this new corpus **DIASER** (DIAlog System extendER). To the best of our knowledge, no one has yet proposed a corpus as large as the one we have produced by merging annotations from different corpora and domains under a common annotation scheme. We observe the effect of the increased size of training data on the performance of two state-of-the-art end-to-end neural models: MarCo (Wang et al., 2020) and GPT-2 (Budzianowski and Vulić, 2019), considering both in-domain and cross-domain performance. The contributions of this paper can be summarized as follows:

- We unify the annotation of multiple task-oriented datasets as well as their ontologies, producing a standard applicable for many different domains.
- We thoroughly analyze and visualize the data properties using the resulting annotation, providing a direct comparison between the source datasets and showing differences between dialogue styles.
- The resulting dataset is one of the largest with annotated task-oriented dialogues, certainly the largest with this level/granularity of annotation.
- We show a detailed performance comparison of two very different dialogue model architectures on the new dataset, specifically targeting similar tasks (small variations on intents) or cross-domain learning capabilities.

The DIASER dataset is available at:
https://github.com/ufal/diaser

2. Related Work

Unifying annotation The idea of merging/unifying multi-domain datasets has been applied in many areas of natural language processing. Gao and Zhang (2005) merged three datasets to achieve efficient text retrieval. The OPUS project (Tiedemann and Nygaard, 2004) consists of dozens of different corpora merged together and is still active today. In the vision field, data merging has shown some efficiency for autonomous driving (Yang et al., 2018) and object detection in the wild

(Rame et al., 2018). Fortuna et al. (2018) successfully merged different social network datasets to perform aggressive text identification. In the emotion detection field, Bostan and Klinger (2018) achieved state-of-the-art results by merging datasets. Acedo et al. (2018) proposed a model for merging datasets as well as a method to detect similar text datasets (Acedo et al., 2019) and to identify representative words of each dataset (Fernández-Sellers et al., 2019).

Dialogue data unification Despite some theoretical efforts in annotation standardization (Bunt et al., 2020), dialogue data unification attempts have been mostly limited to providing data for different languages under a common scheme, with no attempt to merge datasets (Chen and Kan, 2013; Noh et al., 2015). Recently, ConvLab-2 (Zhu et al., 2020b) offered a dataset of around 106K dialogues that merges four different task-oriented dialogue datasets: CamRest (Wen et al., 2016), MultiWOZ (Budzianowski et al., 2018), DealOrNoDeal (Lewis et al., 2017) and CrossWOZ (Zhu et al., 2020a) in a dialogue toolkit with software support for cross-domain experiments. But contrary to our work, ConvLab-2 does not provide a unified annotation scheme. We can also cite the MetaLWOz (Shalyminov et al., 2020) and Taskmaster (Byrne et al., 2019) subcorpora (over 70k dialogues). However the first has no annotation and the second was built using a crowdsourced Wizard-of-Oz procedure, with the crowd worker playing both roles at the same time. Therefore they are not usable for training an end-to-end task-oriented system with db access.
Merging (or combining) datasets also involve ontology merging. The ontology model represents the concepts referenced in the speaker utterances, such as intents, entities and their relations (Minker, 1998; Ginzburg, 2012). It is not a trivial problem for dialog: Freddo et al. (2007) show the extreme difficulty for autonomous dialogue agents that do not share a common ontology to merge similar concepts together. Reed et al. (2020) successfully combined ontology for dialogue response generation, but for a single domain.

Dialogue models Recently, models based on pre-trained LMs such as GPT-2 took lead in the research of task-oriented dialogue (Budzianowski and Vulić, 2019; Zhang et al., 2019; Peng et al., 2020). These models proved to perform great at fluency of the responses and are also able to capture structured dialogue information well. However, they are typically large in size and rather slow. An interesting alternative are models based on the Memory Networks architecture (Lin et al., 2019; Chen et al., 2019c). Such models contain one or more memory modules that are able to keep and retrieve relevant memories. When combined with hierarchical self-attention models (Chen et al., 2019a; Zhang et al., 2020a), these networks are capable of generalizing on both belief state prediction and response generation in order to achieve transformer-like performance. The varied properties of LM-based and memory-based models motivate us to experiment with both architectures on our dataset (see Section 6).

3. Problem description

Current neural architectures need a lot of data to achieve good performance. However, collecting annotated dialogue data is time-consuming and demanding. Current task-oriented dialogue datasets are thus relatively small and usually cover just a few domains. This results in rather poor benchmarks for dialogue models that are not challenging enough. We aim to create both larger and more diverse new dataset that would represent the real use cases more accurately. We hope that this way we can challenge the current models more.
Therefore, we combine four major task-oriented dialogue datasets spanning several domains to yield one larger multi-domain training corpus: MultiWOZ 2.2 (MW) (Zang et al., 2020),[1] Schema-guided dialogue (SGD) (Rastogi et al., 2020),[2] DSTC2 (DSTC) (Henderson et al., 2014),[3] and CamRest676 (CR) (Wen et al., 2016).[4]
Our aim here is to cover as many domains as possible in a unified corpus. Our source dataset choice is thus mainly based on the level of annotation available – all source datasets include semantic annotation on the turn level as well as explicit database interaction (see Section 3.1). Despite the dataset similarities, important differences need to be resolved (see Section 3.2). We are not aware of other freely available datasets with the same amount of annotation.

3.1. Input datasets description

MultiWOZ 2.2 MW is an established task-oriented dataset introduced by (Budzianowski et al., 2018). MultiWOZ 2.2 is an improved version of the original dataset by (1) fixing some of annotation errors, inconsistencies and ontology issues, (2) adding slot span annotations for utterances. It contains more than 10,000 annotated dialogues and spans several domains. The data were gathered via a crowd-sourcing Wizard-of-Oz scheme which is described in (Wen et al., 2017).

DSTC2 DSTC was introduced as a part of a challenge to improve a state tracking within dialogue systems. It contains over 3,000 dialogues covering a single domain around restaurant reservations. The dialogue corpus was collected using Amazon Mechanical Turk with a POMDP-based spoken dialogue system. It is the only human-machine dataset in our collection.

CamRest676 CR is another crowd-sourced dialogue corpora gathered via the Wizard-of-Oz scheme. CamRest with its 676 conversations is the smallest out of four datasets used in this work, and it is also a

[1] https://github.com/budzianowski/multiwoz
[2] https://github.com/google-research-datasets/dstc8-schema-guided-dialogue
[3] https://github.com/matthen/dstc
[4] https://github.com/shawnwun/NNDIAL

single-domain dataset focused on helping users to find a restaurant in Cambridge, UK.

Schema-guided dialogue SGD is a large (more than 20,000 dialogues) multi-domain (around 20 domains covered) dataset containing in total 45 API services based on a pre-defined schema. First, the data was collected via a simulator that interacts with the API services, and then the dialogues were paraphrased using crowd-sourcing.

All these datasets have several task-oriented dialogue properties in common that define the conversation according to (Young et al., 2013):

1. **Domain(s)** define the topic (or range of topics) which are mentioned in the dialogue. There can be several domains per dialogue.

2. **Task** – the users in each of dialogues are attempting to reach a certain goal (such as book a restaurant or find a tourist attraction).

3. **Success** There are several definitions of dialogue success in the literature. We follow the DSTC2 guidelines (Henderson et al., 2013) arguing that the success of a dialogue can be measured with three indicators: task completion, user satisfaction and dialogue finalisation.

4. **Turns** We consider *turn-taking* dialogues, i.e. the participating sides exchange utterances alternately. One such utterance exchange is called a dialogue turn. Utterance itself is considered the linguistic realisation of the speakers thoughts and will. It can be spoken or written.

   The meaning of each utterance can be represented in a structured way with **Dialog Act(s)** (Weisser, 2016). It is a meta-information that emerges from the respective utterance, and qualifies it. It describes the beliefs, desires and intentions. Dialogue acts can be represented using *Domains*, *Intents* and *Slots*. The intent represents the user intention, i.e. the sub-goal that the user wants to achieve with a particular utterance. Slots represent the attributes that instantiate the dialogue act. Each domain is associated with a certain set of intents, and each intent can be combined with multiple slots. A slot, however, can be used by multiple intents as well.

This abstract annotation scheme can be mapped onto the four datasets to provide a unified corpus. The basic statistics of all the datasets are presented in Table 1.

Data	SGD	MW	DSTC	CR	Total
Domains	18	7	1	1	19*
Slots	145	29	10	7	166*
Dialogues**	22.8	10.4	3.2	0.7	37.1
Turns**	463.3	143.0	51.0	5.5	662.8
Turns/Dial.	20.30	13.71	15.77	8.12	17.83
Avg. utt. length	9.86	13.23	8.47	10.71	10.49
Unique Words**	32.3	23.2	1.3	1.7	49.9
Shannon ent.	8.96	8.54	7.04	7.69	9.01
Cond. ent.	4.76	4.41	2.14	2.95	4.83

Table 1: Composition of our dataset, with basic statistics, overall and for individual sources (number of domains and slots, total numbers of dialogues and turns, average number of turns per dialogue and average utterance length in terms of words. *not a sum since due to ontology merging. ** in thousands.

3.2. Semantic differences across datasets

The main task when merging datasets is to unify the different domain specific ontologies, i.e. the different attributes contained in the dialogue acts. More precisely, the unified dataset ontology contains all the possible domains, with their corresponding slots and associated value sets. We cannot consider the slots independently from the domain they belong to. Indeed, a slot that represents the *price range* will not have the same range of values when pertaining to a restaurant or a flight ticket. We identified two main problems related to this issue:

1. **Name reference ambiguities** We need to design the final ontology so that different slot names refer to different concepts (with due precaution for label choice) and to merge different slot names associated with the same value set under a single label. For example, in MultiWOZ, there are two different slot names *day* and *book-day* for the same value set (week days) and usage contexts. But in SGD, slot names may be misleading since we can find a slot named *start-day* and another one called *day*; the former refers to a calendar date while the latter refers to a week day.

2. **Absence of ontology/database** When SGD dataset was collected, the authors used API calls instead of database lookup. Therefore there is no database-related metadata released with the corpus, which forced us to create an ontology and a database for the data based on values occurring in the conversations.

4. DIASER Creation

Here we present details on how me merged the data described in Section 3 into common format, including the handling of different ontologies. Quantitative statistics of the final dataset are shown in the rightmost column of Table 1.

Full technical documentation can be found in the data repository.[5] An overview of all the required merging steps is listed here:

[5]https://github.com/ufal/diaser

Matching belief representations. In DSTC and CR, belief state annotations are extracted from both the user and the system utterance, whereas in the MW and SGD dataset, the belief state is only extracted from the user utterance. We had to filter automatically the annotations from DSTC and CR until they matched the MW belief state representation.

Adding meta features from the original datasets. DSTC2, CamRest and MultiWOZ contain the *goal* of the dialogue (also called task) as a dialogue act with the constraints (e.g. expensive restaurant south) and the information the user needs to obtain from the system (e.g. phone number and address). They also contain a short text summarizing this goal for crowd workers. We include both versions of the task description with each dialogue.

Unifying annotation structure. The final dataset structure is similar to the structure of SGD and MultiWOZ 2.2. We create a *Turn* object that contains either the user utterance and dialogue acts, or the system utterance. Two consecutive entries for user and system share the same turn number – we consider a *Turn* as an exchange between the user and the system (i.e. a pair or utterances).

Ontology unification. One of the most difficult parts of this work was to unify ontologies[6] of each original dataset because they were not built on the same dimensions. This concerns slot, domain, and intent names, cf. Section 3.2. After indexing all metadata from the different datasets, we merged them manually using MultiWOZ as the pivot. We always checked the meaning in context, so we match slots/intents with the same semantics.
In addition to unifying the naming, we also needed to solve the following problems:

1. SGD does not include any ontology; we thus had to build it from scratch based on values from the included API responses. Since SGD uses several API schemas per domain, each with its own set of slots, often using different naming for the same concepts, we also unified the different schemas, same as we did with different data sources.

2. DSTC2 and CamRest ontologies distinguish between two kinds of slots: informable and requestable. *Informable* (also called *constraints*) are slots for which the value needs to be specified by the user (e.g. price, area); the user cannot specify *requestable* slots but can ask the system for their value (e.g. phone number, address). We use this distinction to also assign slots in the other two datasets into one of these two groups.

[6]Although none of datasets have a formal specification or an RDF representation, the lists of all possible domains, intents, slots, and slot values are generally referred to as ontologies in dialogue systems literature, including the description papers for the source datasets.

Slot co-reference Some of the source corpora (MW and SGD) include co-reference between slots. For example *start-time* can take the value "sooner than that" or the slot *hotel-name* can take the value "event you mentioned earlier". This is a problem if we assume to have a self-contained ontology that captures all the possible values from the corpus. However, as these co-references are impossible to include in the ontology easily, we leave these values unchanged and the slot co-references are carried over to the unified data.

5. Dialogue Time Span Analysis

During our merging process, we realised that human-human and human-machine dialogues had different behaviour in time. To confirm this by hard evidence, we conducted linguistic analysis of DIASER corpora to provide more insight into the data. There are several works (Pasupat et al., 2019; Qin et al., 2021) that attempt to predict temporal sequences in a dialogue. We took inspiration by the work of Papangelis et al. (2017) on the use of several cues to determine the length of a dialogue from a certain turn and the previous context (the dialogue history). While the most discriminative features for predicting dialogue length are acoustic cues (Lykartsis and Kotti, 2019), the use of conditional entropy of the belief state of a given speech turn also proves effective for this task (Papangelis et al., 2017). Therefore, we consider conditional entropy to be an important explaining feature relevant to corpus analysis.

5.1. Measuring conditional entropy

Shannon's entropy (Shannon, 1948) is a well-known measure that represents the amount of information that can be contained in a message. Shannon's entropy considers in its calculation all variables as independent, i.e. each word or each character as independent of its predecessor. However, it is more explainable to take some sentence context into account. This is why we chose to use conditional entropy, in the manner of (Dušek et al., 2020), inspired by (Manning and Schutze, 1999), to compute the information in a dialogue. The conditional entropy for text is similar to Shannon's entropy but we consider pairs of consecutive words (x, y) and their joint probabilities:

$$H_{y|x} = -\sum_{x,y} p(x,y) \cdot \log_2 \frac{p(x,y)}{p(x)} \qquad (1)$$

We compared the conditional entropy growth in the different DIASER sub-corpora and domains. We also studied the particular case of the restaurant domain because it is contained in all the four sub-corpora. The objective of this comparison is to detect irregularities (acceleration or slowing down) in the growth of entropy from one dialogue turn to the next, and to identify the periods where these irregularities occur.

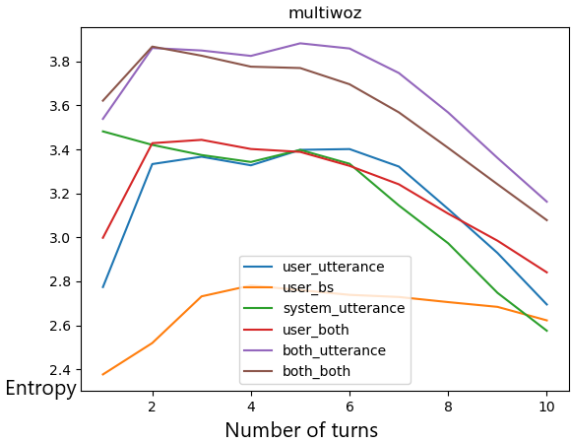

Figure 1: Conditional entropy for the restaurant subset of MultiWOZ (human-human data). Each line corresponds to a feature of the dialogue (utterance, belief state (bs) or both) and the corresponding speaker (user/system/both).

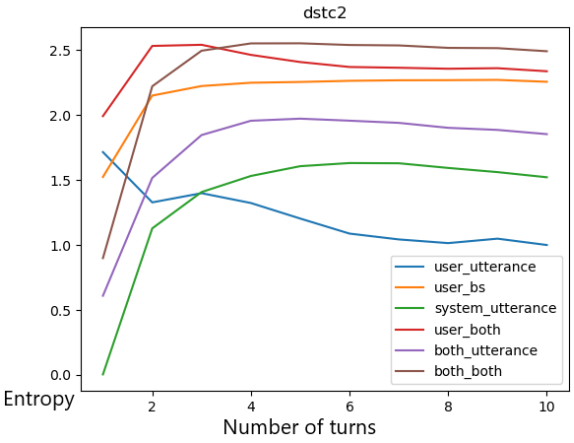

Figure 2: Conditional entropy for the DSTC dataset (human-machine data).

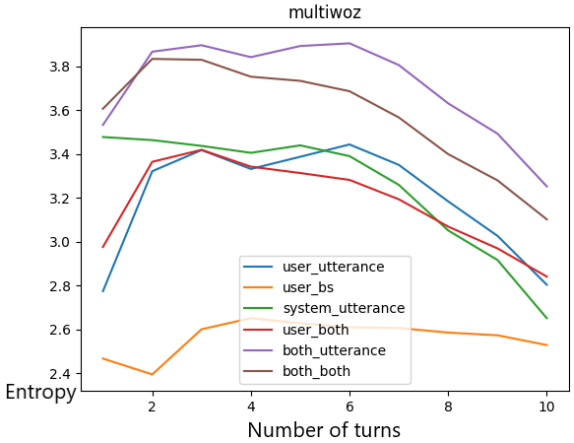

Figure 3: Conditional entropy for the hotel subset of MultiWOZ (human-human data).

5.2. Analysis

Human-machine dialogues: For DSTC2, which was obtained using a human-machine setup, we notice the difference in entropy increase between the user and system utterances (Figure 2), both in Shannon and conditional entropy. We can observe the following general pattern: In the beginning of the dialogue, the user utterances have the highest entropy, then it is caught up by that of the system's utterances and finally it is exceeded by the latter. Therefore, we distinguish three dialogue (information growth, stagnation, information deprecation), which correspond to the phases of the task-oriented dialogue:

1. Presentation of the task by the user to the system
2. Information exchange, confirmations, corrections
3. Resolution of the task by the system by providing the answers expected by the user

Moreover, in the case of DSTC2, the entropy of the belief state is higher than that of the user's utterance, showing the small amount of non-predictable information contained in the latter. We can observe how the curves of the user and system utterances cross (see Figure 2), showing that the system ends up generating more non-predictable information than the user.

Human-human setup: In contrast, the other three corpora show a harmony between user and system entropy (see Figure 1). The belief state is quite predictable with a low entropy. This possibly indicates that the use of human-human corpora to model human-machine dialogues has its limits, because the appearance of unexpected information is not reflected when a human plays the role of the system. The user's utterances become less predictable than the system's as one progresses in the dialogue. We hypothesize that the task assignments for the Wizard-of-Oz schema essentially produce a conversation "script" with several "standard" responses. These can occur multiple times and thus lower the entropy.

Domain differences Just as this phenomenon occurs for each type of interlocutor, it also occurs for each type of domain, but not at the same time. Some domains see the crossover between the user's entropy increase and the system's entropy increase coming earlier, others later (see Figure 3). The restaurant domain in MultiWOZ and SGD resembles the whole dataset, which leads us to think that it can be used as an example for our results and as a domain for the evaluation.

6. Experiments and Results

We choose to compare performance of two dialogue models: MarCo and GPT-2. We evaluate them with respect to both state tracking and language quality metrics. We try out various train-test combinations to understand the influence on each portion of the data on the result. Note, that in our experiments, the CamRest676

training			evaluation			metrics					
DSTC	**MW**	**SGD**	**DSTC**	**MW**	**SGD**	**Slot F1**		**Joint Accuracy**		**BLEU**	
						MarCo†	GPT	MarCo †	GPT	MarCo †	GPT
✗	✓	✗	✓	✗	✗	0.85	0.47	0.45	0.11	10.47	17.90
✗	✗	✓	✓	✗	✗	0.51	0.19	0.05	0.01	3.66	4.11
✓	✓	✗	✓	✗	✗	0.46	0.84	0.05	0.54	10.01	46.31
✓	✗	✓	✓	✗	✗	0.56	0.69	0.18	0.26	23.46	43.47
✗	✓	✓	✓	✗	✗	**0.88**	0.46	**0.57**	0.11	6.33	16.55
✓	✓	✓	✓	✗	✗	0.50	0.85	0.13	0.36	27.31	**46.47**
✗	✓	✗	✗	✓	✗	0.70	0.89	0.48	0.53	17.05	18.61
✗	✗	✓	✗	✓	✗	0.46	0.16	0.11	0.02	2.87	4.01
✓	✓	✗	✗	✓	✗	0.74	0.89	0.49	**0.55**	16.17	19.67
✓	✗	✓	✗	✓	✗	0.51	0.17	0.15	0.03	4.27	5.68
✗	✓	✓	✗	✓	✗	0.64	0.89	0.39	0.52	13.19	19.92
✓	✓	✓	✗	✓	✗	0.65	**0.90**	0.38	0.54	14.59	**21.09**
✗	✓	✗	✗	✗	✓	0.42	0.04	0.04	0.01	2.97	5.63
✗	✗	✓	✗	✗	✓	0.68	0.59	0.19	0.21	9.72	**28.17**
✓	✓	✗	✗	✗	✓	0.51	0.03	0.10	0.01	2.97	5.51
✓	✗	✓	✗	✗	✓	**0.85**	0.58	**0.52**	0.21	7.49	27.96
✗	✓	✓	✗	✗	✓	0.71	0.63	0.17	0.23	6.17	27.54
✓	✓	✓	✗	✗	✓	0.77	0.63	0.32	0.22	1.72	27.72
✓	✓	✗	✓	✓	✓	–	0.28	–	0.12	–	15.30
✓	✗	✓	✓	✓	✓	0.62	0.55	0.23	0.22	3.95	27.28
✗	✓	✓	✓	✓	✓	**0.79**	0.65	**0.40**	0.25	8.70	25.13
✓	✓	✓	✓	✓	✓	0.65	0.70	0.20	0.28	14.49	**29.73**

Table 2: Performance of our models trained and evaluated on various subsets of the unified dataset. †MarCo uses dialogue state as an input feature.

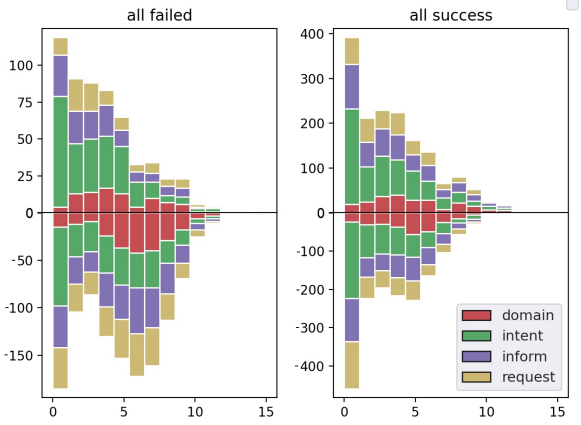

Figure 4: Distribution and types of dialogue failures that happened with the GPT2-based model on the MultiWOZ data (see Section 6.4). Horizontal axis corresponds to turn numbers, positive values represent cases in which the information is captured wrongly while negative values represent cases in which particular information is missing. The source of error (bad domain or intent, missing information or not providing request) is depicted using colors.

dataset is always included in the training set and not used for evaluation as its size is rather small.

6.1. Models description

MarCo (Wang et al., 2020) is a model that jointly generates dialogue acts and system response in the same network. It is treating the generated sequence of dialogue acts as a semantic plan for the final response. MarCo uses belief state as a feature, therefore it relies on an external state tracker, which is its main limitation. The architecture is inspired from the two-stage model called HDSA (Hierarchical disentangled self-attetion) (Chen et al., 2019b). MarCo differs from HDSA because it computes jointly act and response generation instead of a two-stage computation, and improves results compared to HDSA. Both MarCo and HDSA source code are available online.[7]

GPT-2 There was an extensive development of the usage of pretrained Language Models (LMs) in the recent years. These architectures are mostly based on the Transformer architecture (Vaswani et al., 2017) and have a great capacity for learning statistical patterns from large corpora. The GPT architecture (Radford et al., 2019) is among the most used approaches and it was successfully applied on a large variety of tasks across the NLP field. The GPT uses the Transformer's decoder component which essentially represents an auto-regressive LM. The pretrained decoder can be fine-tuned on basically any downstream task that can be expressed as a sequence-to-sequence problem.

[7]https://github.com/InitialBug/MarCo-Dialog

Recently there were works that applied the GPT architecture to the task of dialogue modelling (Peng et al., 2020; Zhang et al., 2020b).

We follow approach taken by previous works (Peng et al., 2020; Kulhánek et al., 2021) and use the pretrained model for both belief state tracking and response generation. Our approach has two stages. First, the belief state is generated based on the dialogue context. The belief state is decoded word by word and we train the model so that the decoded sequence is possible to be parsed in a deterministic way. Next, we perform a database lookup based on the parsed belief state. Then we concatenate the context, belief state and a summary of the database lookup result and generate the system response with the same model. We are working with delexicalized versions of the system utterances.

Training details We use publicly available code for the MarCo model and our own implementation of the GPT-2-based model. We use default hyperparameters for both methods and train on standard GPU cards. Training of each model takes in between several hours to 1 day, depending on the training set size. The GPT architecture is rather robust, however we hypothesize that the MarCo model would need hyperparameter tuning for each data size to achieve better results. Such a tuning is out of our options because of the computational needs of such a process.

6.2. Evaluation Metrics

We evaluate the models' performance with a set of corpus-based metrics. The focus of our dataset is on task-oriented systems, therefore we evaluate accuracy of the generated dialogue states which are essential for database interaction and system's correctness. To evaluate the states, we use common metrics *Joint accuracy* and *F1 score*. Joint accuracy gives exact match percentage over dialogue states, i.e. it reflects how many dialogues have correct state predicted. On the contrary, F1 is computed over slots to provide a more fine-grained view. Additionally, we compute the BLEU score (Papineni et al., 2002) between the generated system utterances and the ground truth to get an approximation of the output fluency. The BLEU score is computed on delexicalized versions of the utterances, which corresponds to the common evaluation scheme. Although the usage of BLEU score for evaluating dialogue systems is controversial (Liu et al., 2016), we decide to use it as it is commonly reported on task-oriented datasets and to measure the output's fluency.

6.3. Results Analysis

GPT model results We can see a general pattern in the results suggesting that the model fails to generalize across different datasets when a subset of data we evaluate on is not included in the training. A significant drop in performance can be observed across all recorded metrics.

On DSTC, this drop is most significant. In terms of F1 slot score and joint accuracy, the explanation for the poor model's accuracy can be found in a significantly different distribution of slots across SGD/MW datasets and DSTC. Training on MW provides us with the most reasonable scores on slot F1 and joint accuracy. Training on both SGD to MW brings only a very marginal improvement. The big difference in performance can be seen in terms of BLEU as well, which suggests a big difference in the language used in each dataset.

The performance of the model evaluated on MW is analogous. Training the model solely on SGD gives us a model which is not able to generalize to MW data. Concatenating MW datasets with SGD, DSTC or both of them during training leads to a little improvement, predominantly in terms of BLEU score. We get the best results when using the whole DIASER dataset for training and obtain the model with better generalizing capabilities supported by the BLEU score of 21.09.

The model evaluation on SGD data mostly follows the same pattern. However, in this case, it is quite interesting to observe that training the model on the whole DIASER seems not to be beneficial and the performance stagnates although the training set is much bigger.

Finally, if we evaluate on the whole DIASER dataset, it is clear that the best performing model is obtained when we train it on the full data. From the results presented in Table 2, we can see that including SGD data is crucial for achieving higher BLEU values, while training on MW helps the model predict slot values.

MarCo model results Results for the MarCo model are harder to interpret. We often see inconsistent behavior with respect to the input data. In general, MarCo often outperforms GPT model in terms of slots detection, but this is mainly due to the use of the belief state representation as a feature from MarCo. Overall, even with the additional information on the input, MarCo often struggles to successfully generate meaningful responses and also seems to be very vulnerable to inconsistencies in DSTC data. We discussed this phenomenon in a previous work (Schaub et al., 2021).

It also might be the case that the MarCo model is much more sensitive to the right hyperparameter choice which therefore has to be made carefully for each dataset (see Section 6.1).

Results summary We observe that each sub-dataset has its specific properties that cannot be substituted completely by other sources. However, combining the training data with additional examples is beneficial. Overall, the GPT model provides more consistent results than MarCo, but it struggles with state tracking on data with complex ontology and lower-quality annotation (i.e., data from SGD).

6.4. Dynamics of Successful and Failed Dialogues

We also perform a detailed quantitative analysis of errors made by the trained models. To do so, we evaluated a GPT model trained on the MW data in terms

of the dialogue success rates.[8] The dialogue is considered successful when it fulfills the following conditions: (a) all constraints introduced by the user were correctly captured by the system (i.e. *inform*) and (b) all the requested information was correctly included in the response (i.e. *request*). "All failed" corresponds to dialogues for which neither inform nor request conditions were met, while "all success" corresponds to dialogues for which both conditions were met and therefore they were successful. It is thus an evaluation concerning the dialogue as a whole, unlike the turn-level methods such as evaluation of the system's response generation.

In Figure 4 we provide a histogram of errors in successful and failed dialogues. We go turn by turn and identify the errors in the dialogue state predicted by the system compared to the expected state. In particular, we show a number of cases where some information is understood wrongly by the system (positive values) or is missing completely (negative values). We can see that in successful dialogues, the system might make some mistake at first but is able to recover eventually. On the other hand, most dialogue failures are actually caused by missing information in turn in the second half of the dialogue, which suggests that the recovery is not certain.

We further identify four reasons that can cause an error:

1. *Domain* is wrongly identified.
2. *Intent* is detected wrongly.
3. Some *inform* value is captured incorrectly.
4. Some *request* was not answered.

We realized some observations:

- For both successful and unsuccessful dialogues, the first turn concentrates the most errors.
- Generally speaking, the failed dialogues show a large amount of superfluous information in the first five turns of speech, especially in terms of the user's intent. On the other hand, from approximately the fourth speech turn onwards, there is a lot of missing information, especially at the level of the dialogue domain. If we add up the missing information with the superfluous information, we see that the peak of errors is between turns three and seven of the dialogue.
- It is interesting to note that for the first four turns, in successful dialogues, we have a significant amount of superfluous information. It matches with the time span where entropy grows the most which interestingly connects the entropy analysis and the model behavior. We hypothesize that there is a correlation between mistakes made by the system and the entropy amount variation.

[8] We only choose the MW data due to additional annotation being available in MultiWOZ.

7. Conclusion and Future Work

In this work we unified a large task-oriented dialogue corpora at both data and annotation levels. which requires complex process of merging ontologies. We called it DIASER. We showed that that using additional data from other sources is helpful for training both GPT-2 and MarCo models when converted to the unified format. Although the new dataset is still far from perfect coverage, it is a step towards wider and more autentic data.

We also showed correlation between entropy growth during dialogue and errors made by our models, demonstrating that future models should focus their attention on certain moments when entropy is at its highest in order to prevent errors. The entropy study reveals that human-human and human-machine dialogues do not share the same pattern of entropy growth over dialogue turns, showing that we still have work to do in order to understand what differs so much in the information exchange between a user and a system. This also means that human-human dialogues, even when built with a WOZ technique, might not be the best fit to train dialogue systems, or at least we should try to chose among the human-human dialogues, which of them are closer to human-machine dialogues.

In our future works, we will focus on the predictions of a dialogue stages and on how important it is for a dialogue model to be aware of them.

8. Acknowledgements

This work was partially supported by AKIO and the ANRT CIFRE #2017/1543, HumanE-AI-Net project / EC Horizon 2020, Grant Agreement H2020-FETFLAG-2018-2020 no. 952026, as the DIASER Miniproject (DIAlog System enhancER), and Charles University projects PRIMUS/19/SCI/10, GA UK No. 302120 and SVV No. 260575. It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth and Sports project No. LM2018101).

9. Bibliographical References

Acedo, J., Lozano-Tello, A., and Fernandez-Sellers, M. (2018). Model of datasets unification from different open data portals. In *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE.

Acedo, J., Fernandez-Sellers, M., and Lozano-Tello, A. (2019). Detection model of similar datasets. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE.

Bostan, L. A. M. and Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119. Association for Computational Linguistics.

Budzianowski, P. and Vulić, I. (2019). Hello, it's gpt-2–how can i help you? towards the use of pretrained

language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*.

Budzianowski, P. and Vulić, I. (2019). Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22, Hong Kong, November. Association for Computational Linguistics. arXiv: 1907.05774.

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*, pages 5016–5026, Brussels, Belgium, October-November. Association for Computational Linguistics.

Bunt, H., Petukhova, V., Gilmartin, E., Pelachaud, C., Fang, A., Keizer, S., and Prévot, L. (2020). The ISO Standard for Dialogue Act Annotation, Second Edition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 549–558, Marseille, France, May. European Language Resources Association.

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Duckworth, D., Yavuz, S., Goodrich, B., Dubey, A., Kim, K.-Y., and Cedilnik, A. (2019). Taskmaster-1: Toward a realistic and diverse dialog dataset.

Chen, T. and Kan, M.-Y. (2013). Creating a live, public short message service corpus: the nus sms corpus. *Language Resources and Evaluation*, 47(2):299–335, Jun.

Chen, W., Chen, J., Qin, P., Yan, X., and Wang, W. Y. (2019a). Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709, Florence, Italy, July. Association for Computational Linguistics.

Chen, W., Chen, J., Qin, P., Yan, X., and Wang, W. Y. (2019b). Semantically conditioned dialog response generation via hierarchical disentangled self-attention. *arXiv preprint arXiv:1905.12866*.

Chen, X., Xu, J., and Xu, B. (2019c). A working memory model for task-oriented dialog response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2687–2693.

Dušek, O., Novikova, J., and Rieser, V. (2020). Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Computer Speech & Language*, 59:123–156, January.

Fernández-Sellers, M., Acedo, J., and Lozano-Tello, A. (2019). Identification of representative terms of datasets. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE.

Fortuna, P., Ferreira, J., Pires, L., Routar, G., and Nunes, S. (2018). Merging datasets for aggressive text identification. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 128–139, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Freddo, A. R., Brito, R. C., Gimenez-Lugo, G., and Tacla, C. A. (2007). Partial and dynamic ontology mapping model in dialogs of agents. In José Neves, et al., editors, *Progress in Artificial Intelligence*, pages 347–356, Berlin, Heidelberg. Springer Berlin Heidelberg.

Gao, J. and Zhang, J. (2005). Clustered svd strategies in latent semantic indexing. *Information Processing & Management*, 41(5):1051–1063.

Ginzburg, J., (2012). *A Semantic Ontology for Dialogue*. The Interactive Stance. Oxford University Press, Oxford.

Henderson, M., Thomson, B., and Williams, J. (2013). Dialog state tracking challenge 2 & 3. Challenge handbook. https://raw.githubusercontent.com/matthen/dstc/master/handbook.pdf.

Kulhánek, J., Hudeček, V., Nekvinda, T., and Dušek, O. (2021). AuGPT: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 198–210, Online, November. Association for Computational Linguistics.

Lei, W., Jin, X., Ren, Z., He, X., Kan, M.-Y., and Yin, D. (2018). Sequicity: Simplifying Task-oriented Dialogue Systems with Single Sequence-to-Sequence Architectures. In *ACL*, pages 1437–1447, Melbourne, Australia, July.

Lewis, M., Yarats, D., Dauphin, Y., Parikh, D., and Batra, D. (2017). Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark, September. Association for Computational Linguistics.

Lin, Z., Huang, X., Ji, F., Chen, H., and Zhang, Y. (2019). Task-oriented conversation generation using heterogeneous memory networks. *arXiv preprint arXiv:1909.11287*.

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Lykartsis, A. and Kotti, M. (2019). Prediction of user emotion and dialogue success using audio spectrograms and convolutional neural networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 336–344, Stockholm,

Sweden, September. Association for Computational Linguistics.

Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Mehri, S., Srinivasan, T., and Eskenazi, M. (2019). Structured Fusion Networks for Dialog. In *SIGdial*, Stockholm, Sweden, September.

Minker, J. (1998). An overview of cooperative answering in databases. In Troels Andreasen, et al., editors, *Flexible Query Answering Systems*, pages 282–285, Berlin, Heidelberg. Springer Berlin Heidelberg.

Noh, T.-G., Padó, S., Shwartz, V., Dagan, I., Nastase, V., Eichler, K., Kotlerman, L., and Adler, M. (2015). Multi-level alignments as an extensible representation basis for textual entailment algorithms. In *Proceedings of the fourth joint conference on lexical and computational semantics*, pages 193–198.

Papangelis, A., Kotti, M., and Stylianou, Y. (2017). Predicting dialogue success, naturalness, and length with acoustic features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5010–5014.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Pasupat, P., Gupta, S., Mandyam, K., Shah, R., Lewis, M., and Zettlemoyer, L. (2019). Span-based hierarchical semantic parsing for task-oriented dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1520–1526, Hong Kong, China, November. Association for Computational Linguistics.

Peng, B., Li, C., Li, J., Shayandeh, S., Liden, L., and Gao, J. (2020). Soloist: Building task bots at scale with transfer learning and machine teaching. *arXiv preprint arXiv:2005.05298*.

Qin, L., Xu, X., Che, W., Zhang, Y., and Liu, T. (2020). Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6344–6354, Online, July. Association for Computational Linguistics.

Qin, L., Gupta, A., Upadhyay, S., He, L., Choi, Y., and Faruqui, M. (2021). TIMEDIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076, Online, August. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rame, A., Garreau, E., Ben-Younes, H., and Ollion, C. (2018). OMNIA Faster R-CNN: Detection in the wild through dataset merging and soft distillation. *arXiv e-prints*, page arXiv:1812.02611, December.

Reed, L., Harrison, V., Oraby, S., Hakkani-Tur, D., and Walker, M. (2020). Learning from mistakes: Combining ontologies via self-training for dialogue generation. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 21–34, 1st virtual meeting, July. Association for Computational Linguistics.

Schaub, L.-P., Hudecek, V., Stancl, D., Dusek, O., and Paroubek, P. (2021). Définition et détection des incohérences du système dans les dialogues orientés tâche. (we present experiments on automatically detecting inconsistent behavior of task-oriented dialogue systems from the context). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 142–152, Lille, France, 6. ATALA.

Serban, I. V., Lowe, R., Charlin, L., and Pineau, J. (2018). A Survey of Available Corpora for Building Data-Driven Dialogue Systems. *Dialogue & Discourse*, 9(1):1–49, May. arXiv: 1512.05742.

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)*, pages 8039–8043.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Tiedemann, J. and Nygaard, L. (2004). The OPUS corpus - parallel and free: http://logos.uio.no/opus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Wang, K., Tian, J., Wang, R., Quan, X., and Yu, J. (2020). Multi-domain dialogue acts and response co-generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7125–7134, Online, July. Association for Computational Linguistics.

Weisser, M. (2016). Dart – the dialogue annotation and research tool. *Corpus Linguistics and Linguistic Theory*, 12(2):355–388.

Wen, T.-H., Gasic, M., Mrkšić, N., Barahona, L. M. R., Su, P.-H., Ultes, S., Vandyke, D., and Young, S. (2016). Conditional generation and snapshot learn-

ing in neural dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2153–2162. Association for Computational Linguistics.

Wen, T.-H., Vandyke, D., Mrkšić, N., Gašić, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., and Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April. Association for Computational Linguistics.

Yang, L., Liang, X., Wang, T., and Xing, E. (2018). Real-to-virtual domain unification for end-to-end autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September.

Yang, Y., Li, Y., and Quan, X. (2021). UBAR: Towards Fully End-to-End Task-Oriented Dialog Systems with GPT-2. In *AAAI*, Online, February. arXiv: 2012.03539.

Young, S., Gašić, M., Thomson, B., and Williams, J. D. (2013). POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhang, Y., Ou, Z., and Yu, Z. (2020a). Task-oriented dialog systems that consider multiple appropriate responses under the same context. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9604–9611, Apr.

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*, pages 270–278, Online, July. Association for Computational Linguistics.

Zhu, Q., Huang, K., Zhang, Z., Zhu, X., and Huang, M. (2020a). CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

Zhu, Q., Zhang, Z., Fang, Y., Li, X., Takanobu, R., Li, J., Peng, B., Gao, J., Zhu, X., and Huang, M. (2020b). ConvLab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 142–149, Online, July. Association for Computational Linguistics.

10. Language Resource References

Henderson, M., Thomson, B., and Williams, J. D. (2014). 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. Association for Computational Linguistics.

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*, volume 34, pages 8689–8696.

Wen, T.-H., Gasic, M., Mrkšić, N., Barahona, L. M. R., Su, P.-H., Ultes, S., Vandyke, D., and Young, S. (2016). Conditional generation and snapshot learning in neural dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2153–2162. Association for Computational Linguistics.

Zang, X., Rastogi, A., and Chen, J. (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.