MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling

Paweł Budzianowski[1], **Tsung-Hsien Wen**[2*], **Bo-Hsiang Tseng**[1],
Iñigo Casanueva[2*], **Stefan Ultes**[1], **Osman Ramadan**[1] and **Milica Gašić**[1]
[1]Department of Engineering, University of Cambridge, UK,
[2]PolyAI, London, UK
{pfb30,mg436}@cam.ac.uk

arXiv:1810.00278v3 [cs.CL] 20 Apr 2020

Abstract

Even though machine learning has become the major scene in dialogue research community, the real breakthrough has been blocked by the scale of data available. To address this fundamental obstacle, we introduce the Multi-Domain Wizard-of-Oz dataset (MultiWOZ), a fully-labeled collection of human-human written conversations spanning over multiple domains and topics. At a size of 10k dialogues, it is at least one order of magnitude larger than all previous annotated task-oriented corpora. The contribution of this work apart from the open-sourced dataset labelled with dialogue belief states and dialogue actions is two-fold: firstly, a detailed description of the data collection procedure along with a summary of data structure and analysis is provided. The proposed data-collection pipeline is entirely based on crowd-sourcing without the need of hiring professional annotators; secondly, a set of benchmark results of belief tracking, dialogue act and response generation is reported, which shows the usability of the data and sets a baseline for future studies.

1 Introduction

Conversational Artificial Intelligence (Conversational AI) is one of the long-standing challenges in computer science and artificial intelligence since the Dartmouth Proposal (McCarthy et al., 1955). As human conversation is inherently complex and ambiguous, learning an open-domain conversational AI that can carry on arbitrary tasks is still very far-off (Vinyals and Le, 2015). As a consequence, instead of focusing on creating ambitious conversational agents that can reach human-level intelligence, industrial practice has focused on building task-oriented dialogue systems (Young et al., 2013) that can help with specific tasks such as flight reservation (Seneff and Polifroni, 2000) or bus information (Raux et al., 2005). As the need of hands-free use cases continues to grow, building a conversational agent that can handle tasks across different application domains has become more and more prominent (Ram et al., 2018).

Dialogues systems are inherently hard to build because there are several layers of complexity: the noise and uncertainty in speech recognition (Black et al., 2011); the ambiguity when understanding human language (Williams et al., 2013); the need to integrate third-party services and dialogue context in the decision-making (Traum and Larsson, 2003; Paek and Pieraccini, 2008); and finally, the ability to generate natural and engaging responses (Stent et al., 2005). These difficulties have led to the same solution of using statistical framework and machine learning for various system components, such as natural language understanding (Henderson et al., 2013; Mesnil et al., 2015; Mrkšić et al., 2017a), dialogue management (Gašić and Young, 2014; Tegho et al., 2018), language generation (Wen et al., 2015; Kiddon et al., 2016), and even end-to-end dialogue modelling (Zhao and Eskenazi, 2016; Wen et al., 2017; Eric et al., 2017).

To drive the progress of building dialogue systems using data-driven approaches, a number of conversational corpora have been released in the past. Based on whether a structured annotation scheme is used to label the semantics, these corpora can be roughly divided into two categories: corpora with structured semantic labels (Hemphill et al., 1990; Williams et al., 2013; Asri et al., 2017; Wen et al., 2017; Eric et al., 2017; Shah et al., 2018); and corpora without semantic labels but with an implicit user goal in mind (Ritter et al., 2010; Lowe et al., 2015). Despite these efforts, aforementioned datasets are usually constrained in one or more dimensions such as missing proper annotations, only available in a limited capacity,

*The work was done while at the University of Cambridge.

Metric	DSTC2	SFX	WOZ2.0	FRAMES	KVRET	M2M	MultiWOZ
# Dialogues	1,612	1,006	600	1,369	2,425	1,500	**8,438**
Total # turns	23,354	12,396	4,472	19,986	12,732	14,796	**113, 556**
Total # tokens	199,431	108,975	50,264	251,867	102,077	121,977	**1,490,615**
Avg. turns per dialogue	14.49	12.32	7.45	**14.60**	5.25	9.86	13.46
Avg. tokens per turn	8.54	8.79	11.24	12.60	8.02	8.24	**13.13**
Total unique tokens	986	1,473	2,142	12,043	2,842	1,008	**23689**
# Slots	8	14	4	**61**	13	14	24
# Values	212	1847	99	3871	1363	138	**4510**

Table 1: Comparison of our corpus to similar data sets. Numbers in bold indicate best value for the respective metric. The numbers are provided for the training part of data except for FRAMES data-set were such division was not defined.

lacking multi-domain use cases, or having a negligible linguistic variability.

This paper introduces the Multi-Domain Wizard-of-Oz (MultiWOZ) dataset, a large-scale multi-turn conversational corpus with dialogues spanning across several domains and topics. Each dialogue is annotated with a sequence of dialogue states and corresponding system dialogue acts (Traum, 1999). Hence, MultiWOZ can be used to develop individual system modules as separate classification tasks and serve as a benchmark for existing modular-based approaches. On the other hand, MultiWOZ has around 10k dialogues, which is at least one order of magnitude larger than any structured corpus currently available. This significant size of the corpus allows researchers to carry on end-to-end based dialogue modelling experiments, which may facilitate a lot of exciting ongoing research in the area.

This work presents the data collection approach, a summary of the data structure, as well as a series of analyses of the data statistics. To show the potential and usefulness of the proposed MultiWOZ corpus, benchmarking baselines of belief tracking, natural language generation and end-to-end response generation have been conducted and reported. The dataset and baseline models will be freely available online.[1]

2 Related Work

Existing datasets can be roughly grouped into three categories: machine-to-machine, human-to-machine, and human-to-human conversations. A detailed review of these categories is presented below.

Machine-to-Machine Creating an environment with a simulated user enables to exhaustively generate dialogue templates. These templates can be mapped to a natural language by either pre-defined rules (Bordes et al., 2017) or crowd workers (Shah et al., 2018). Such approach ensures a diversity and full coverage of all possible dialogue outcomes within a certain domain. However, the naturalness of the dialogue flows relies entirely on the engineered set-up of the user and system bots. This poses a risk of a mismatch between training data and real interactions harming the interaction quality. Moreover, these datasets do not take into account noisy conditions often experienced in real interactions (Black et al., 2011).

Human-to-Machine Since collecting dialogue corpus for a task-specific application from scratch is difficult, most of the task-oriented dialogue corpora are fostered based on an existing dialogue system. One famous example of this kind is the Let's Go Bus Information System which offers live bus schedule information over the phone (Raux et al., 2005) leading to the first Dialogue State Tracking Challenge (Williams et al., 2013). Taking the idea of the Let's Go system forward, the second and third DSTCs (Henderson et al., 2014b,c) have produced bootstrapped human-machine datasets for a restaurant search domain in the Cambridge area, UK. Since then, DSTCs have become one of the central research topics in the dialogue community (Kim et al., 2016, 2017).

While human-to-machine data collection is an

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

obvious solution for dialogue system development, it is only possible with a provision of an existing working system. Therefore, this chicken (system)-and-egg (data) problem limits the use of this type of data collection to existing system improvement instead of developing systems in a completely new domain. What is even worse is that the capability of the initial system introduces additional biases to the collected data, which may result in a mismatch between the training and testing sets (Wen et al., 2016). The limited understanding capability of the initial system may prompt the users to adapt to simpler input examples that the system can understand but are not necessarily natural in conversations.

Human-to-Human Arguably, the best strategy to build a natural conversational system may be to have a system that can directly mimic human behaviors through learning from a large amount of real human-human conversations. With this idea in mind, several large-scale dialogue corpora have been released in the past, such as the Twitter (Ritter et al., 2010) dataset, the Reddit conversations (Schrading et al., 2015), and the Ubuntu technical support corpus (Lowe et al., 2015). Although previous work (Vinyals and Le, 2015) has shown that a large learning system can learn to generate interesting responses from these corpora, the lack of grounding conversations onto an existing knowledge base or APIs limits the usability of developed systems. Due to the lack of an explicit goal in the conversation, recent studies have shown that systems trained with this type of corpus not only struggle in generating consistent and diverse responses (Li et al., 2016) but are also extremely hard to evaluate (Liu et al., 2016).

In this paper, we focus on a particular type of human-to-human data collection. The Wizard-of-Oz framework (WOZ) (Kelley, 1984) was first proposed as an iterative approach to improve user experiences when designing a conversational system. The goal of WOZ data collection is to log down the conversation for future system development. One of the earliest dataset collected in this fashion is the ATIS corpus (Hemphill et al., 1990), where conversations between a client and an airline help-desk operator were recorded.

More recently, Wen et al. (2017) have shown that the WOZ approach can be applied to collect high-quality typed conversations where a machine learning-based system can learn from. By modifying the original WOZ framework to make it suitable for crowd-sourcing, a total of 676 dialogues was collected via Amazon Mechanical Turk. The corpus was later extended to additional two languages for cross-lingual research (Mrkšić et al., 2017b). Subsequently, this approach is followed by Asri et al. (2017) to collect the Frame corpus in a more complex travel booking domain, and Eric et al. (2017) to collect a corpus of conversations for in-car navigation. Despite the fact that all these datasets contain highly natural conversations comparing to other human-machine collected datasets, they are usually small in size with only a limited domain coverage.

- You are traveling to Cambridge and looking forward to try local restaurants.
- You are looking for a **place to stay**. The hotel should be in the type of **hotel** and should be in the **centre**.
- The hotel should **include free wifi** and should have **a star of 4**.
- Once you find the **hotel** you want to book it for **3 people** and **5 nights** starting from **monday**.
- Make sure you get the **reference number**.
- You are also looking for a **restaurant**. The restaurant should serve **australasian** food and should be in the **moderate** price range.
- The restaurant should be **in the same area as the hotel**.
- If there is no such restaurant, how about one that serves **british** food.
- Once you find the **restaurant** you want to book a table for **the same group of people** at **18:30** on **the same day**.
- Make sure you get the **reference number**

Figure 1: A sample task template spanning over three domains - hotels, restaurants and booking.

3 Data Collection Set-up

Following the Wizard-of-Oz set-up (Kelley, 1984), corpora of annotated dialogues can be gathered at relatively low costs and with a small time effort. This is in contrast to previous approaches (Henderson et al., 2014a) and such WOZ set-up has been successfully validated by Wen et al. (2017) and Asri et al. (2017).

Therefore, we follow the same process to create a large-scale corpus of natural human-human conversations. Our goal was to collect multi-domain dialogues. To overcome the need of relying the data collection to a small set of trusted workers[2],

[2]Excluding annotation phase.

Table 2: Full ontology for all domains in our data-set. The upper script indicates which domains it belongs to. *: universal, 1: restaurant, 2: hotel, 3: attraction, 4: taxi, 5: train, 6: hospital, 7: police.

act type	inform* / request* / select^{123} / recommend/^{123} / not found^{123} request booking info^{123} / offer booking^{1235} / inform booked^{1235} / decline booking^{1235} welcome* /greet* / bye* / reqmore*
slots	address* / postcode* / phone* / name^{1234} / no of choices^{1235} / area^{123} / pricerange^{123} / type^{123} / internet^{2} / parking^{2} / stars^{2} / open hours^{3} / departure^{45} destination^{45} / leave after^{45} / arrive by^{45} / no of people^{1235} / reference no.^{1235} / trainID^{5} / ticket price^{5} / travel time^{5} / department^{7} / day^{1235} / no of days^{123}

the collection set-up was designed to provide an easy-to-operate system interface for the Wizards and easy-to-follow goals for the users. This resulted in a bigger diversity and semantical richness of the collected data (see Section 4.3). Moreover, having a large set of workers mitigates the problem of artificial encouragement of a variety of behavior from users. A detailed explanation of the data-gathering process from both sides is provided below. Subsequently, we show how the crowd-sourcing scheme can also be employed to annotate the collected dialogues with dialogue acts.

3.1 Dialogue Task

The domain of a task-oriented dialogue system is often defined by an ontology, a structured representation of the back-end database. The ontology defines all entity attributes called slots and all possible values for each slot. In general, the slots may be divided into *informable* slots and *requestable* slots. *Informable* slots are attributes that allow the user to constrain the search (e.g., area or price range). *Requestable* slots represent additional information the users can request about a given entity (e.g., phone number). Based on a given ontology spanning several domains, a task template was created for each task through random sampling. This results in single and multi-domain dialogue scenarios and domain specific constraints were generated. In domains that allowed for that, an additional booking requirement was sampled with some probability.

To model more realistic conversations, goal changes are encouraged. With a certain probability, the initial constraints of a task may be set to values so that no matching database entry exists. Once informed about that situation by the system, the users only needed to follow the goal which provided alternative values.

3.2 User Side

To provide information to the users, each task template is mapped to natural language. Using heuristic rules, the task is then gradually introduced to the user to prevent an overflow of information. The goal description presented to the user is dependent on the number of turns already performed. Moreover, if the user is required to perform a sub-task (for example - booking a venue), these sub-goals are shown straight-away along with the main goal in the given domain. This makes the dialogues more similar to spoken conversations.[3] Figure 1 shows a sampled task description spanning over two domains with booking requirement. Natural incorporation of co-referencing and lexical entailment into the dialogue was achieved through implicit mentioning of some slots in the goal.

3.3 System Side

The wizard is asked to perform a role of a clerk by providing information required by the user. He is given an easy-to-operate graphical user interface to the back-end database. The wizard conveys the information provided by the current user input through a web form. This information is persistent across turns and is used to query the database. Thus, the annotation of a belief state is performed implicitly while the wizard is allowed to fully focus on providing the required information. Given the result of the query (a list of entities satisfying current constraints), the wizard either requests more details or provides the user with the adequate information. At each system turn, the wizard starts with the results of the query from the previous turn.

To ensure coherence and consistency, the wizard and the user alike first need to go through the

[3] However, the length of turns are significantly longer than with spoken interaction (Section 4.3).

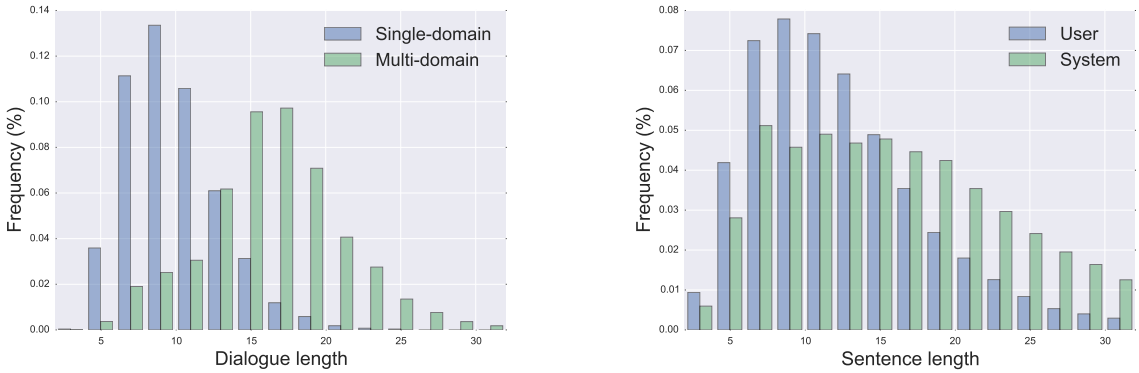

Figure 2: Dialogue length distribution (left) and distribution of number of tokens per turn (right).

dialogue history to establish the respective context. We found that even though multiple workers contributed to one dialogue, only a small margin of dialogues were incoherent.

3.4 Annotation of Dialogue Acts

Arguably, the most challenging and time-consuming part of any dialogue data collection is the process of annotating dialogue acts. One of the major challenges of this task is the definition of a set and structure of dialogue acts (Traum and Hinkelman, 1992; Bunt, 2006). In general, a dialogue act consists of the intent (such as request or inform) and slot-value pairs. For example, the act `inform(domain=hotel,price=expensive)` has the intent *inform*, where the user is informing the system to constrain the search to expensive hotels.

Expecting a big discrepancy in annotations between annotators, we initially ran three trial tests over a subset of dialogues using Amazon Mechanical Turk. Three annotations per dialogue were gathered resulting in around 750 turns. As this requires a multi-annotator metric over a multi-label task, we used Fleiss' kappa metric (Fleiss, 1971) per single dialogue act. Although the weighted kappa value averaged over dialogue acts was at a high level of 0.704, we have observed many cases of very poor annotations and an unsatisfactory coverage of dialogue acts. Initial errors in annotations and suggestions from crowd workers gradually helped us to expand and improve the final set of dialogue acts from 8 to 13 - see Table 2.

The variation in annotations made us change the initial approach. We ran a two-phase trial to first identify set of workers that perform well. Turkers were asked to annotate an illustrative, long dialogue which covered many problematic examples that we have observed in the initial run described above. All submissions that were of high quality were inspected and corrections were reported to annotators. Workers were asked to re-run a new trial dialogue. Having passed the second test, they were allowed to start annotating real dialogues. This procedure resulted in a restricted set of annotators performing high quality annotations. Appendix A contains a demonstration of a created system.

3.5 Data Quality

Data collection was performed in a two-step process. First, all dialogues were collected and then the annotation process was launched. This setup allowed the dialogue act annotators to also report errors (e.g., not following the task or confusing utterances) found in the collected dialogues. As a result, many errors could be corrected. Finally, additional tests were performed to ensure that the provided information in the dialogues match the pre-defined goals.

To estimate the inter-annotator agreement, the averaged weighted kappa value for all dialogue acts was computed over 291 turns. With $\kappa = 0.884$, an improvement in agreement between annotators was achieved although the size of action set was significantly larger.

4 MultiWOZ Dialogue Corpus

The main goal of the data collection was to acquire highly natural conversations between a tourist and a clerk from an information center in a touristic

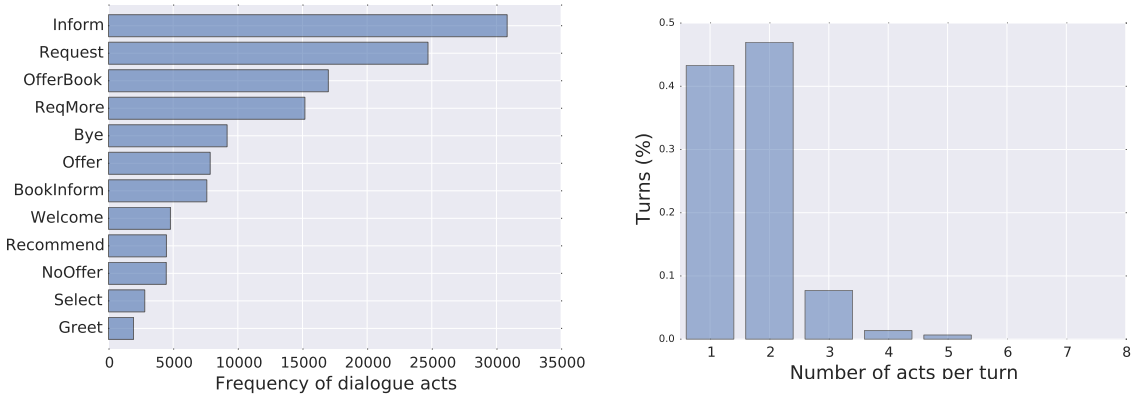

Figure 3: Dialogue acts frequency (left) and number of dialogue acts per turn (right) in the collected corpus.

city. We considered various possible dialogue scenarios ranging from requesting basic information about attractions through booking a hotel room or travelling between cities. In total, the presented corpus consists of 7 domains - *Attraction, Hospital, Police, Hotel, Restaurant, Taxi, Train*. The latter four are extended domains which include the sub-task *Booking*. Through a task sampling procedure (Section 3.1), the dialogues cover between 1 and 5 domains per dialogue thus greatly varying in length and complexity. This broad range of domains allows to create scenarios where domains are naturally connected. For example, a tourist needs to find a hotel, to get the list of attractions and to book a taxi to travel between both places. Table 2 presents the global ontology with the list of considered dialogue acts.

4.1 Data Statistics

Following data collection process from the previous section, a total of $10,438$ dialogues were collected. Figure 2 (left) shows the dialogue length distribution grouped by single and multi domain dialogues. Around 70% of dialogues have more than 10 turns which shows the complexity of the corpus. The average number of turns are 8.93 and 15.39 for single and multi-domain dialogues respectively with $115,434$ turns in total. Figure 2 (right) presents a distribution over the turn lengths. As expected, the wizard replies are much longer - the average sentence lengths are 11.75 and 15.12 for users and wizards respectively. The responses are also more diverse thus enabling the training of more complex generation models.

Figure 3 (left) shows the distribution of dialogue acts annotated in the corpus. We present here a summarized list where different types of actions like *inform* are grouped together. The right graph in the Figure 3 presents the distribution of number of acts per turn. Almost 60% of dialogues turns have more than one dialogue act showing again the richness of system utterances. These create a new challenge for reinforcement learning-based models requiring them to operate on concurrent actions.

In total, $1,249$ workers contributed to the corpus creation with only few instances of intentional wrongdoing. Additional restrictions were added to automatically discover instances of very short utterances, short dialogues or missing single turns during annotations. All such cases were corrected or deleted from the corpus.

4.2 Data Structure

There are $3,406$ single-domain dialogues that include booking if the domain allows for that and $7,032$ multi-domain dialogues consisting of at least 2 up to 5 domains. To enforce reproducibility of results, the corpus was randomly split into a train, test and development set. The test and development sets contain 1k examples each. Even though all dialogues are coherent, some of them were not finished in terms of task description. Therefore, the validation and test sets only contain fully successful dialogues thus enabling a fair comparison of models.

Each dialogue consists of a goal, multiple user and system utterances as well as a belief state and

set of dialogue acts with slots per turn. Additionally, the task description in natural language presented to turkers working from the visitor's side is added.

4.3 Comparison to Other Structured Corpora

To illustrate the contribution of the new corpus, we compare it on several important statistics with the DSTC2 corpus (Henderson et al., 2014a), the SFX corpus (Gašić et al., 2014), the WOZ2.0 corpus (Wen et al., 2017), the FRAMES corpus (Asri et al., 2017), the KVRET corpus (Eric et al., 2017), and the M2M corpus (Shah et al., 2018). Figure 1 clearly shows that our corpus compares favorably to all other data sets on most of the metrics with the number of total dialogues, the average number of tokens per turn and the total number of unique tokens as the most prominent ones. Especially the latter is important as it is directly linked to linguistic richness.

5 MultiWOZ as a New Benchmark

The complexity and the rich linguistic variation in the collected MultiWOZ dataset makes it a great benchmark for a range of dialogue tasks. To show the potential usefulness of the MultiWOZ corpus, we break down the dialogue modelling task into three sub-tasks and report a benchmark result for each of them: dialogue state tracking, dialogue-act-to-text generation, and dialogue-context-to-text generation. These results illustrate new challenges introduced by the MultiWOZ dataset for different dialogue modelling problems.

5.1 Dialogue State Tracking

A robust natural language understanding and dialogue state tracking is the first step towards building a good conversational system. Since multi-domain dialogue state tracking is still in its infancy and there are not many comparable approaches available (Rastogi et al., 2017), we instead report our state-of-the-art result on the restaurant subset of the MultiWOZ corpus as the reference baseline. The proposed method (Ramadan et al., 2018) exploits the semantic similarity between dialogue utterances and the ontology terms which allows the information to be shared across domains. Furthermore, the model parameters are independent of the ontology and belief states, therefore the number of the parameters does not increase with the size of the domain itself.[4]

Slot	WOZ 2.0	MultiWOZ (restaurant)
Overall accuracy	96.5	89.7
Joint goals	85.5	80.9

Table 3: The test set accuracies *overall* and for *joint goals* in the restaurant sub-domain.

The same model was trained on both the WOZ2.0 and the proposed MultiWOZ datasets, where the WOZ2.0 corpus consists of 1200 single domain dialogues in the restaurant domain. Although not directly comparable, Table 3 shows that the performance of the model is consecutively poorer on the new dataset compared to WOZ2.0. These results demonstrate how demanding is the new dataset as the conversations are richer and much longer.

5.2 Dialogue-Context-to-Text Generation

After a robust dialogue state tracking module is built, the next challenge becomes the dialogue management and response generation components. These problems can either be addressed separately (Young et al., 2013), or jointly in an end-to-end fashion (Bordes et al., 2017; Wen et al., 2017; Li et al., 2017). In order to establish a clear benchmark where the performance of the composite of dialogue management and response generation is completely independent of the belief tracking, we experimented with a baseline neural response generation model with an *oracle* belief-state obtained from the wizard annotations as discussed in Section 3.3.[5]

Following Wen et al. (2017) which frames the dialogue as a context to response mapping problem, a sequence-to-sequence model (Sutskever et al., 2014) is augmented with a belief tracker and a discrete database accessing component as additional features to inform the word decisions in the decoder. Note, in the original paper the belief tracker was pre-trained while in this work the annotations of the dialogue state are used as an oracle tracker. Figure 4 presents the architecture of the system (Budzianowski et al., 2018).

[4]The model is publicly available at https://github.com/osmanio2/multi-domain-belief-tracking

[5]The model is publicly available at https://github.com/budzianowski/multiwoz

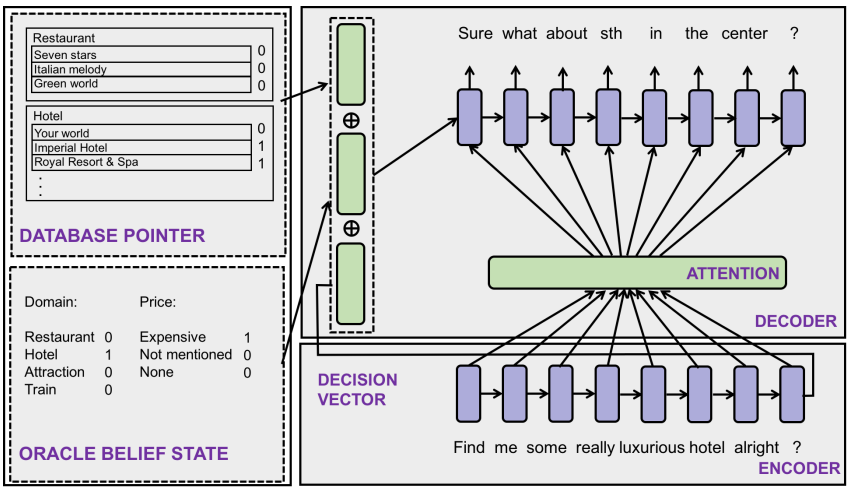

Figure 4: Architecture of the multi-domain response generator. The attention is conditioned on the oracle belief state and the database pointer.

Training and Evaluation Since often times the evaluation of a dialogue system without a direct interaction with the real users can be misleading (Liu et al., 2016), three different automatic metrics are included to ensure the result is better interpreted. Among them, the first two metrics relate to the dialogue task completion - whether the system has provided an appropriate entity (Inform rate) and then answered all the requested attributes (Success rate); while fluency is measured via BLEU score (Papineni et al., 2002). The best models for both datasets were found through a grid search over a set of hyper-parameters such as the size of embeddings, learning rate and different recurrent architectures.

We trained the same neural architecture (taking into account different number of domains) on both MultiWOZ and Cam676 datasets. The best results on the Cam676 corpus were obtained with bidirectional GRU cell. In the case of MultiWOZ dataset, the LSTM cell serving as a decoder and an encoder achieved the highest score with the global type of attention (Bahdanau et al., 2014). Table 4 presents the results of a various of model architectures and shows several challenges. As expected, the model achieves almost perfect score on the Inform metric on the Cam676 dataset taking the advantage of an oracle belief state signal. However, even with the perfect dialogue state tracking of the user intent, the baseline models obtain almost 30% lower score on the Inform metric on the new corpus. The addition of the attention improves the score on the Success metric on the new dataset by less than 1%. Nevertheless, as expected, the best model on MultiWOZ is still falling behind by a large margin in comparison to the results on the Cam676 corpus taking into account both Inform and Success metrics. As most of dialogues span over at least two domains, the model has to be much more effective in order to execute a successful dialogue. Moreover, the BLEU score on the MultiWOZ is lower than the one reported on the Cam676 dataset. This is mainly caused by the much more diverse linguistic expressions observed in the MultiWOZ dataset.

5.3 Dialogue-Act-to-Text Generation

Natural Language Generation from a structured meaning representation (Oh and Rudnicky, 2000; Bohus and Rudnicky, 2005) has been a very popular research topic in the community, and the lack of data has been a long standing block for the field to adopt more machine learning methods. Due to the additional annotation of the system acts, the MultiWOZ dataset serves as a new benchmark for studying natural language generation from a structured meaning representation. In order to verify the difficulty of the collected dataset for the language generation task, we compare it to the SFX dataset (see Table 1), which consists of around 5k dialogue act and natural language sentence pairs. We trained the same Semantically Conditioned Long Short-term Memory network (SC-LSTM) proposed by Wen et al. (2015) on both datasets and used the metrics as a proxy to estimate the difficulty of the two corpora. To make a fair comparison, we constrained our dataset to only the restaurant sub-domain which contains around 25k dia-

	Cam676		MultiWOZ	
	w/o attention	w/ attention	w/o attention	w/ attention
Inform (%)	99.17	99.58	71.29	71.33
Success (%)	75.08	73.75	60.29	60.96
BLEU	0.219	0.204	0.188	0.189

Table 4: Performance comparison of two different model architectures using a corpus-based evaluation.

logue turns. To give more statistics about the two datasets: the SFX corpus has 9 different act types with 12 slots comparing to 12 acts and 14 slots in our corpus. The best model for both datasets was found through a grid search over a set of hyper-parameters such as the size of embeddings, learning rate, and number of LSTM layers.[6]

Table 6 presents the results on two metrics: BLEU score (Papineni et al., 2002) and slot error rate (SER) (Wen et al., 2015). The significantly lower metrics on the MultiWOZ corpus showed that it is much more challenging than the SFX restaurant dataset. This is probably due to the fact that more than 60% of the dialogue turns are composed of at least two system acts, which greatly harms the performance of the existing model.

Metric	SFX	MultiWOZ (restaurant)
SER (%)	0.46	4.378
BLEU	0.731	0.616

Table 5: The test set slot error rate (SER) and BLEU on the SFX dataset and the MultiWOZ restaurant subset.

	Single	Multi
# of dialogues	3,406	7,032
# of domains	1-2	2-6

Table 6: The test set slot error rate (SER) and BLEU on the SFX dataset and the MultiWOZ restaurant subset.

6 Conclusions

As more and more speech oriented applications are commercially deployed, the necessity of building an entirely data-driven conversational agent becomes more apparent. Various corpora were gathered to enable data-driven approaches to dialogue modelling. To date, however, the available datasets were usually constrained in linguistic variability or lacking multi-domain use cases. In this paper, we established a data-collection pipeline entirely based on crowd-sourcing enabling to gather a large scale, linguistically rich corpus of human-human conversations. We hope that MultiWOZ offers valuable training data and a new challenging testbed for existing modular-based approaches ranging from belief tracking to dialogue acts generation. Moreover, the scale of the data should help push forward research in the end-to-end dialogue modelling.

Acknowledgments

This work was funded by a Google Faculty Research Award (RG91111), an EPSRC studentship (RG80792), an EPSRC grant (EP/M018946/1) and by Toshiba Research Europe Ltd, Cambridge Research Laboratory (RG85875). The authors thank many excellent Mechanical Turk contributors for building this dataset. The authors would also like to thank Thang Minh Luong for his support for this project and Nikola Mrkšić and anonymous reviewers for their constructive feedback. The data is available at https://github.com/budzianowski/multiwoz.

[6]The model is publicly available at https://github.com/andy194673/nlg-sclstm-multiwoz

References

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A corpus for adding memory to goal-oriented dialogue systems. *Proceedings of SigDial*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *ICLR*.

Alan W Black, Susanne Burger, Alistair Conkie, Helen Hastie, Simon Keizer, Oliver Lemon, Nicolas Merigaud, Gabriel Parent, Gabriel Schubiner, Blaise Thomson, et al. 2011. Spoken dialog challenge 2010: Comparison of live and control test results. In *Proceedings of the SIGDIAL 2011 Conference*, pages 2–7. Association for Computational Linguistics.

Dan Bohus and Alexander I Rudnicky. 2005. Sorry, i didn't catch that! - an investigation of non-understanding errors and recovery strategies. In *6th SIGdial workshop on discourse and dialogue*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. *Proceedings of ICLR*.

Paweł Budzianowski, Iñigo Casanueva, Bo-Hsiang Tseng, and Milica Gašić. 2018. Towards end-to-end multi-domain dialogue modelling. *Tech. Rep. CUED/F-INFENG/TR.706, University of Cambridge, Engineering Department*.

Harry Bunt. 2006. Dimensions in dialogue act annotation. In *Proc. of LREC*, volume 6, pages 919–924.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

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

Milica Gašić, Dongho Kim, Pirros Tsiakoulis, Catherine Breslin, Matthew Henderson, Martin Szummer, Blaise Thomson, and Steve Young. 2014. Incremental on-line adaptation of pomdp-based dialogue managers to extended domains. In *Interspeech*.

Milica Gašić and Steve Young. 2014. Gaussian processes for pomdp-based dialogue manager optimization. *TASLP*, 22(1):28–40.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania*.

M. Henderson, B. Thomson, and J. Williams. 2014a. The second dialog state tracking challenge. In *Proceedings of SIGdial*.

M. Henderson, B. Thomson, and S. J. Young. 2014b. Word-based Dialog State Tracking with Recurrent Neural Networks. In *Proceedings of SIGdial*.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014c. The third dialog state tracking challenge. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 324–329. IEEE.

Matthew Henderson, Blaise Thomson, and Steve Young. 2013. Deep neural network approach for the dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 467–471.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339.

Seokhwan Kim, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, Matthew Henderson, and Koichiro Yoshino. 2016. The fifth dialog state tracking challenge. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, pages 511–517. IEEE.

Seokhwan Kim, Luis Fernando DHaro, Rafael E Banchs, Jason D Williams, and Matthew Henderson. 2017. The fourth dialog state tracking challenge. In *Dialogues with Social Robots*, pages 435–449. Springer.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 733–743.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Ryan Lowe, Nissan Pow, Iulian V Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *16th Annual Meeting of the*

Special Interest Group on Discourse and Dialogue, page 285.

J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon. 1955. A proposal for the dartmouth summer research project on artificial intelligence.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017a. 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)*, volume 1, pages 1777–1788.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017b. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association of Computational Linguistics*, 5(1):309–324.

Alice H Oh and Alexander I Rudnicky. 2000. Stochastic language generation for spoken dialogue systems. In *Proceedings of the 2000 ANLP/NAACL Workshop on Conversational systems-Volume 3*, pages 27–32. Association for Computational Linguistics.

Tim Paek and Roberto Pieraccini. 2008. Automating spoken dialogue management design using machine learning: An industry perspective. *Speech communication*, 50(8-9):716–729.

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

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Osman Ramadan, Paweł Budzianowski, and Milica Gašić. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 432–437.

Abhinav Rastogi, Dilek Hakkani-Tur, and Larry Heck. 2017. Scalable multi-domain dialogue state tracking. *arXiv preprint arXiv:1712.10224*.

Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Let's go public! taking a spoken dialog system to the real world. In *Ninth European Conference on Speech Communication and Technology*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180.

Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583.

Stephanie Seneff and Joseph Polifroni. 2000. Dialogue management in the mercury flight reservation system. In *Proceedings of the 2000 ANLP/NAACL Workshop on Conversational Systems - Volume 3*, ANLP/NAACL-ConvSyst '00, pages 11–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

P Shah, D Hakkani-Tur, G Tur, A Rastogi, A Bapna, N Nayak, and L Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 341–351. Springer.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Christopher Tegho, Paweł Budzianowski, and Milica Gašić. 2018. Benchmarking uncertainty estimates with deep reinforcement learning for dialogue policy optimisation. In *IEEE ICASSP 2018*.

David R. Traum. 1999. *Foundations of Rational Agency*, chapter Speech Acts for Dialogue Agents. Springer.

David R Traum and Elizabeth A Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Computational intelligence*, 8(3):575–599.

David R Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In *Current and new directions in discourse and dialogue*, pages 325–353. Springer.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2016. Multi-domain neural network language generation for spoken dialogue systems. *ACL*.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gašić, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. *EACL*.

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.

Steve Young, Milica Gašić, Blaise Thomson, and Jason Williams. 2013. POMDP-based Statistical Spoken Dialogue Systems: a Review. In *Proc of IEEE*, volume 99, pages 1–20.

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 1.

A MTurk Website Set-up

Figure A1 presents the user side interface where the worker needs to properly respond given the task description and the dialogue history. Figure A2 shows the wizard page with the GUI over all domains. Finally, Figure A3 shows the set-up for annotation of the system acts with Restaurant domain being turned on.

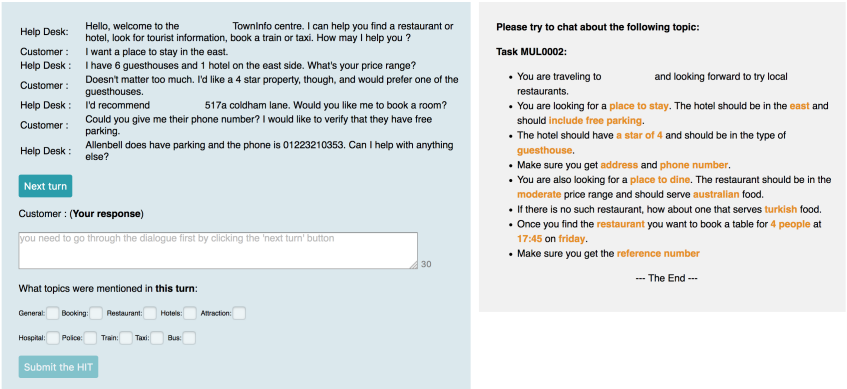

Figure A1: Interface from the User side

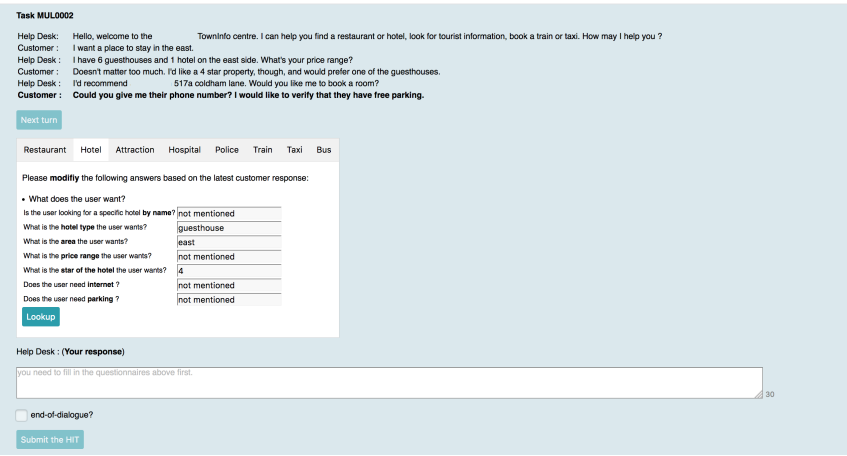

Figure A2: Interface from the Wizard side

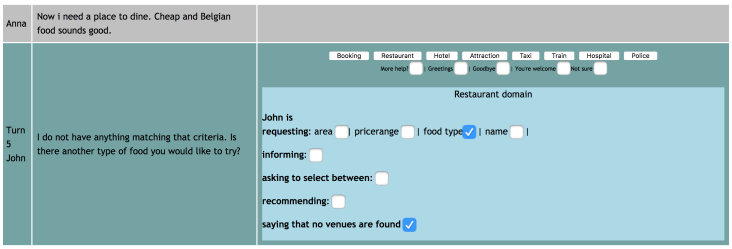

Figure A3: Interface for the annotation.,