Multi-Domain Goal-Oriented Dialogues (MultiDoGO): Strategies toward Curating and Annotating Large Scale Dialogue Data

Denis Peskov[1 *†], **Nancy Clarke**[†2], **Jason Krone**[†2],
Brigitta Fodor[*3], **Yi Zhang**[2], **Adel Youssef**[2] **and Mona Diab**[2]
[1]University of Maryland, [2]Amazon AWS AI, [3]Wayfair
dpeskov@cs.umd.edu, bfodor@wayfair.com
{njc, kronej, yizhng, adel, diabmona}@amazon.com

Abstract

The need for high-quality, large-scale, goal-oriented dialogue datasets continues to grow as virtual assistants become increasingly widespread. However, publicly available datasets useful for this area are limited either in their size, linguistic diversity, domain coverage, or annotation granularity. In this paper, we present strategies toward curating and annotating large scale goal oriented dialogue data. We introduce the `MultiDoGO` dataset to overcome these limitations. With a total of over 81K dialogues harvested across six domains, `MultiDoGO` is over 8 times the size of `MultiWOZ`, the other largest comparable dialogue dataset currently available to the public. Over 54K of these harvested conversations are annotated for intent classes and slot labels. We adopt a Wizard-of-Oz approach wherein a crowd-sourced worker (the "customer") is paired with a trained annotator (the "agent"). The data curation process was controlled via biases to ensure a diversity in dialogue flows following variable dialogue policies. We provide distinct class label tags for agents vs. customer utterances, along with applicable slot labels. We also compare and contrast our strategies on annotation granularity, i.e. turn vs. sentence level. Furthermore, we compare and contrast annotations curated by leveraging professional annotators vs the crowd. We believe our strategies for eliciting and annotating such a dialogue dataset scales across modalities and domains and potentially languages in the future. To demonstrate the efficacy of our devised strategies we establish neural baselines for classification on the agent and customer utterances as well as slot labeling for each domain.

1 Introduction

Modern Natural Language Understanding (NLU) frameworks for dialogues are by definition data hungry. They require large amounts of training data representative of goal oriented conversations reflecting both context and diversity. But human responses in goal-oriented dialogues are less predictable than automated systems (Bordes et al., 2016). For example, "Please do this" cannot be interpreted without a broader context. Only by seeing previous utterances, such as requests to book a flight on a specific day to a specific destination, can this task be performed. Additionally, a single intent can be phrased in multiple ways depending on context; "book my flight", "finalize my reservation", "Yes, the 6 pm one" may all be referring to a flight-booking intent. Hence, entire conversations, rather than independent utterances, must be collected. Such data is even more pertinent to modeling NLU and related tasks as they require large, varied, and ideally human-generated datasets. Moreover, recent work (Dong et al., 2015; Devlin et al., 2018) has shown the benefit of applying joint-training and transfer learning techniques to natural language processing tasks. However, these approaches have yet to become widely used in dialogue tasks, due to a lack of large-scale datasets. Furthermore, the latest state of the art end-to-end neural approaches benefit from such training data even more so than past work on goal-oriented dialogues structured around slot filling (Lemon et al., 2006; Wang and Lemon, 2013). One way to simulate data—and not risk releasing personally identifying information—for a domain is to use a Wizard-of-Oz data gathering technique, which requires that participants in a conver-

[*] Work performed while at Amazon AWS AI
[†] Denotes equal contribution

*Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
and the 9th International Joint Conference on Natural Language Processing*, pages 4526–4536,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

Role	Turn	Annotations
A	Hey there! Good morning. You're connected to LMT Airways. How may I help you?	DA = { elicitgoal }
C	Hi, I wonder if you can confirm my seat assignment on my flight tomorrow?	IC = { SeatAssignment }
A	Sure! I'd be glad to help you with that. May I know your last name please?	DA = { elicitslot }

C	My last name is Turker.	IC = { contentonly },
		SL = {Name : Turker }

A	Alright Turker! Could you please share the booking confirmation number?	DA = { elicitslot }

C	I believe it's AMZ685.	IC = { contentonly },
		SL = { Confirmation Number : AMZ685 }

...	...	...

Table 1: A segment of a dialogue from the airline domain annotated at the turn level. This data is annotated with agent dialogue acts (DA), customer intent classes (IC), and slot labels (SL). Roles C and A stand for "Customer" and "Agent", respectively.

sation fulfill a role (Kelley, 1984). This approach has been used in popular public goal-oriented datasets: DSTC and MultiWOZ (Williams et al., 2016; Budzianowski et al., 2018).

Conversations between people and automated systems occur with increasing frequency, especially in customer service. Customers reach out to agents, which could be automated bots or real individuals, to achieve a domain-specific goal. This creates a disparate conversation: agents are incentivized to operate within a set procedure and convey a patient and professional tone. In contrast, customers do not have this incentive. However, to date, the largest available multi-domain goal-oriented dialogue dataset assigns similar dialogue act annotations to both agents and customers (Budzianowski et al., 2018).

To solve the aforementioned challenges, we present our efforts to curate, annotate, and evaluate a large scale multi-domain set of goal oriented dialogues. The dataset is primarily gathered from workers in the crowd paired with professional annotators. The dataset elicited, MultiDoGO, comprises over 86K raw conversations of which 54,818 conversations are annotated at the turn level. We investigate multiple levels of annotation granularity. We annotate a subset of the data on both turn and sentence levels. A turn is defined as a sequence of one or more speech/text sentences by a participant in a conversation. A sentence is a period delimited sequence of words in a turn. A turn may comprise one or more sentences. We do use the term utterance to refer to a unit (turn or sentence, spoken or written by a participant).[1] In our devised annotation strategy, we distinguish between dialogue speech acts for agents vs. customers. In MultiDoGO, the agents' speech acts [DA] are annotated with generic class labels common across all domains, while customer speech acts are labeled with intent classes [IC]. Moreover, we annotate customer utterances with the appropriate slot labels [SL], which consist of the SL span and corresponding tokens with that SL tag. We present the strategies we use to curate and annotate such data given its contextual setting. We furthermore illustrate the efficacy of our devised approaches and annotation decisions against intrinsic metrics and via extrinsic evaluation, namely by applying neural baselines for DA, IC and SL classification leveraging joint models.

2 Existing Dialogue Datasets

There are multiple existing goal-oriented dialogue collections generated by humans through Wizard-of-Oz techniques. The Dialog State Tracking Challenge, *aka* Dialog Systems Technology Challenge, (DSTC) spans 8 iterations and entails the domains of bus timetables, restaurant reservations, and hotel bookings, travel, alarms, movies, etc. (Williams et al., 2016). Frames (Asri et al., 2017) has 1369 dialogues about vacation packages. MultiWOZ contains 10,438 dialogues about Cambridge hotels and restaurants (Budzianowski et al., 2018). There are several dialogue datasets that specialize in a single domain. ATIS (Hemphill et al., 1990) comprises speech data about airlines structured around formal airline flight tables. Similarly, the Google Airlines dataset purportedly contains 400,000 templated dialogues about airline reservations (Wei

[1] We acknowledge that the term utterance is controversial in the literature (Pareti and Lando, 2018)

et al., 2018).[2] The Ubuntu Dialogue Corpus has over a million dialogues about Ubuntu technical support (Lowe et al., 2015).

On the other hand, Chit-chat style dialogues without goals have been popular since ELIZA and have been investigated with neural techniques (Weizenbaum, 1966; Li et al., 2016, 2017). However, these datasets cannot be used for modeling goal-oriented tasks. Related dialogue dataset collections used for Sequential Question Answering rely on dialogue to answer questions, but the task is notably different from our use case of modeling goal oriented conversational AI, hence leading to different evaluation considerations (Reddy et al., 2019; Choi et al., 2018).

3 `MultiDoGO` Dataset Curation

3.1 Data Collection Procedure

We employ both internal data associates, who we train, and crowd-sourced workers from Mechanical Turk (MTurkers) to generate conversational data using a Wizard-of-Oz approach. In each conversation, the data associates assumes the role of an agent while the MTurkers act as customers. In an effort to source competent MTurkers, we require that each MTurker have a Human Intelligence Task (HIT) accuracy minimum of 90%, a location in the United States, and have completed a significant number of HITs in the past. To facilitate goal-oriented conversations between the customer and agent, we give each agent a prompt listing the supported request types (dialog acts)and pieces of information (slots) needed to complete each request. We also specify criteria such as minimal conversation length, number of goals,number of complex requests, etc, to increase conversation diversity. See Figure 2 for an example prompt. In addition, we explicitly request that neither agents nor customers use any personally identifiable information. At an implementation level,we create a custom, web interface for the MTurkers and data associates that displays our instructions next to the current dialogue. This allows each participant to quickly refer to our guidelines without stopping the conversation. Despite following a familiar wizard-of-oz elicitation procedure, and curating data for multiple domains in a fashion similar to previous data collection efforts such as `MultiWOZ`, `MultiDoGO` comprises more varied domains, it is collected at an unprecedented scale, and it is curated with control over generating explicit biases in the conversations to allow for diverse conversation representation. To our knowledge this is a novel collection strategy as we explicitly guide/prod the participants in a dialogue to engage in conversations with specific biases such as intent change, slot change, multi-intent, multiple slot values, slot overfilling and slot deletion. For example, in the Fast Food domain, participants were instructed to pretend that they were ordering fast food from a drive-thru. After making their initial order, they were instructed to change their mind about what they were ordering ("I'd like a burger. No wait, can you make that a chicken sandwich?"). In the Financial domain, we asked participants to make sure that they requested multiple intents such as "I'd like to find my routing number and check my balance."[3] To that end, our collection procedure deliberately attempts to guide the dialogue flow to ensure diversity in dialogue policies.

3.2 Domain Selection

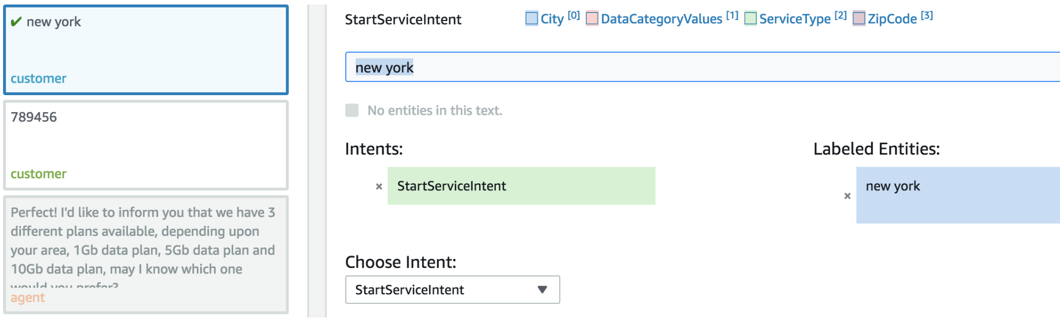
Our primary criteria for domain selection are two-fold: covering a broad sweep of industries that use goal-oriented dialogues and selecting domains where conversational interfaces are already in use or likely to be implemented in the future. This set of criteria is especially well matched with domains that frequently involve customer support. Furthermore, there is a shortage of publicly available data in the domains we provide, such as Fast Food and Finance. To fulfill both of these needs, we include multiple domains in the `MultiDoGO` dataset. Ultimately, we curate conversations for six domains: Airline, Fast Food, Finance, Insurance, Media, and Software Support. When considered independently, the corpus of dialogues for each of these domains is the largest collection of human-elicited dialogues available for financial advice and help, media support, enterprise software support (non-technical level support, unlike the Ubuntu forum dataset (Lowe et al., 2015)), fast food, and insurance.

Domains in our Data:[4] **Airline** domain dialogues focus on booking airline flights, select-

[2] The Google Airlines dataset has not been released to date.

[3] For a full list of conversational biases with examples, please see the appendix.

[4] Detailed data collection and annotation schemata for each domain can be found in the appendix.



✔ new york

customer

789456

customer

Perfect! I'd like to inform you that we have 3 different plans available, depending upon your area, 1Gb data plan, 5Gb data plan and 10Gb data plan, may I know which one would you prefer?

agent

StartServiceIntent

new york

□ No entities in this text.

Intents: StartServiceIntent

Labeled Entities: new york

Choose Intent: StartServiceIntent

Figure 1: Crowd sourced annotators select an intent and choose a slot in our custom-built Mechanical Turk interface. Entire conversations are provided for reference. Detailed instructions are provided to users, but are not included in this figure. Options are unique per domain.

ing or changing seat assignments, and requesting boarding passes; **Fast Food** domain is the least similar to the others, as the intents primarily involve ordering food and the slots quantify their order. For example, the OrderBurgerIntent contains slots for size, quantity, and ingredients; **Finance** domain simulates dialogues a customer may have with a bank. These include opening a bank account, checking their balance, and reporting a lost credit card; **Insurance** domain simulates users calling about their insurance policy or requesting the fulfillment of a policy on their car or phone; **Media** domain simulates dialogues a customer may have ordering a service or paying bills related to telecommunications. This is our largest domain; **Software** domain involves customers inquiring about software services: products, outages, promotions, and bills. The majority of intents are domain specific.

3.3 Domain Schemata and Guidelines

Prior to dialogue collection, we develop schemata for each domain. These schemata are the set of slot labels, slot value types, and intents that pertain to the domain. To determine which slots, values, intents, and dialog acts to include, we rely on real word reference points. For instance, we populate slots for the Fast Food domain by identifying menu items, such as sodas, that are shared among popular fast food menus. Using this schema, we then write two sets of instructions. One set of instructions is for the “agents” and the other is for the “customers”. The agents’ instructions are meticulously detailed as we expect them to “structure” the conversation and appropriately respond to out of domain requests. Since our agents are trained for their role, we have high confidence in their ability to follow complex guidelines. In contrast, taking into consideration that the customer role is to be carried out by crowd-sourced workers, i.e. lay people, we create simplified instructions that are less detailed and shorter in length. For each task, we provide an annotated conversation, explain each answer option, and (most importantly) provide examples. Before scaling up our data collection, we run a pilot for each task and identify commonly missed questions. We use this pilot process to revise the instructions and add relevant examples iteratively.

4 Data Annotation

4.1 Annotated Dialogue Tasks

Our dataset has three types of annotation: Agent dialogue acts [DA], customer intent classes [IC], and slot labels [SL]. We intentionally decouple Agent and customer speech act tags into the categories DA and IC, respectively, to produce more fine-grained speech act tags than past iterations of dialog datasets. Intuitively, agent DAs are consistent across domains and more abstract in nature, since agents have a standard form of response. On the other hand, customer ICs are domain-specific and can entail reserving a hotel room or ordering a burger, depending on the domain. A conversation example with annotations is provided in Table 1.

Agent Dialogue Acts (DA) Agent dialogue acts are the most straightforward of our annotation tasks. There are eight possible DAs in all domains: *ElicitGoal, ElicitSlot, ConfirmGoal, ConfirmSlot, EndGoal, Pleasantries, Other*. The names are self-explanatory. *Elicit Goal/Slot* indicates that the agent is gathering information. *Confirm Goal/Slot* indicates that the agent is confirming previously provided information. The *EndGoal* and *Pleasantries* tags, identify non-task related actions. *Other* indicates that the selected utterance was not one of the other possible tags. Agent dialogue acts are consistent across domains and are often abstract (e.g. ElicitIntent, ConfirmSlot).

Customer Intent Classes (IC): Unlike Agent DA, customer IC vary for each domain and are more concrete. For example, the Airline domain has a "BookFlight" IC, Fast Food has an "OrderMeal" IC, and Insurance has an "OrderPolicy" IC in our annotation schema. Customer intents can overlap across domains (e.g. OpeningGreeting, ClosingGreeting) and other times be domain specific (e.g. RequestCreditLimitIncrease, OrderBurger, BookFlight).

Slot Labels (SL): Slot Labeling (SL) is a task contingent on Customer Intent Classes. Certain intents require that additional information, namely slot values, be captured. For instance, to open a bank account, one must solicit the customer's social security number. Slots can overlap across intents (e.g. Name, SSN Number) or they can be unique to a domain-specific intent (e.g. CarPolicy).

4.2 Data Annotation Procedure

Our annotators utilize a web interface, depicted in Figure 1, to select the appropriate intent class for an utterance out of a list of provided options. To annotate slot labels, our annotators use their cursors to highlight slot value character spans within an utterance and then select the corresponding slot label from a list of options. The output of this slot labeling process is a list of ⟨slot-label, slot-value, span⟩ triplets for each utterance.

4.3 Annotation Design Decisions

Decoupled Agents and Customers Label sets: Agents and customers have notably different goals and styles of communication. However, past dialogue datasets do not make this distinction at speech act schema level. Specificity is important for handling unique customer requests, but a relatively formulaic approach is required of agents across different industries. Our distinction between the customer and agent roles creates training data for a bot that explicitly simulates agents.

Annotation Unit Granularity: Sentence vs. Turn Level An important decision, which is often under discussed, is the proper semantic unit of text to annotate in a dialogue. Commonly, datasets provide annotations at the turn level (Budzianowski et al., 2018; Asri et al., 2017; Mihail et al., 2017). However, turn level annotations can introduce confusion for IC datasets, given multiple intents may be present in different sentences of a single turn. For instance, consider the turn "I would like to book a flight to San Francisco. Also, I want to cancel a flight to Austin." Here, the first sentence has the BookFlight intent and the second sentence has the CancelFlight intent. An turn level annotation of this utterance would yield the multi-class intent (BookFlight, CancelFlight). In contrast, a sentence level annotation of this utterance identifies that the first sentence corresponds to BookFlight while the second corresponds to CancelFlight. We annotate a subset our data, 2,500 conversation per domain for 15,000 conversations in total, at the sentence as well as turn level to access the impact of this design choice on downstream performance. The remainder of our dataset is annotated only at the turn level.

ISAA		
DA	IC	SL
0.701	0.728	0.695

Table 2: Dialogue act (DA), Intent class (IC), and slot labeling (SL) Inter Source Annotation Agreement (ISAA) scores quantifying the agreement of crowd sourced and professional annotations.

Professional vs. Crowd-Sourced Workers for annotation For annotation, we compare and contrast professional annotators to crowd sourced annotators on a subset of data. Professional annotators assign DA, IC, and SL tags to the 15,000 conversations annotated at both the turn and sentence level; statistics for these conversations are given in table 6. In an effort to decrease annotation cost, we employ crowd source annotators via Mechanical Turk to label an additional 54,818 conversations rated as Good or Excellent quality during data collection. We provide statistics for this set of crowd annotated data in Table 3. To compare the quality of crowd sourced annotations against professional annotations, we use both strategies to annotate a shared subset of 8,450 conversations. We devise an Inter Source Annotation Agreement (ISAA) metric to quantify the agreement of these crowd sourced and professionally sourced annotations. ISAA is a relaxation of Cohen Kappa, intended to count partial agreement of multi-tag labels. ISAA defines two sets of tags, A and B, to be in agreement if there is at least one "shared" tag in both A and B. A and B reflect the majority labels agreed upon per source (professionals or crowd workers). Using ISAA we find that crowd sourced and professional annotations have a substantial degree of shared annotations. We report ISAA for the DA, IC, and SL tasks in Table 2.

4.4 Quality Control

We institute three processes to enforce data quality. During data collection, our data associates report on the quality of each conversation. Specifically, the data associates grade the conversation on a scale from "Unusable", "Poor", "Good", to "Excellent". They were provided with guidelines to help decide on the chosen rating such as coherence, whether the dialogue achieved the purported goal, etc. To ensure high data quality we only utilize conversations with "Good" or "Excellent" ratings in subsequent annotation.

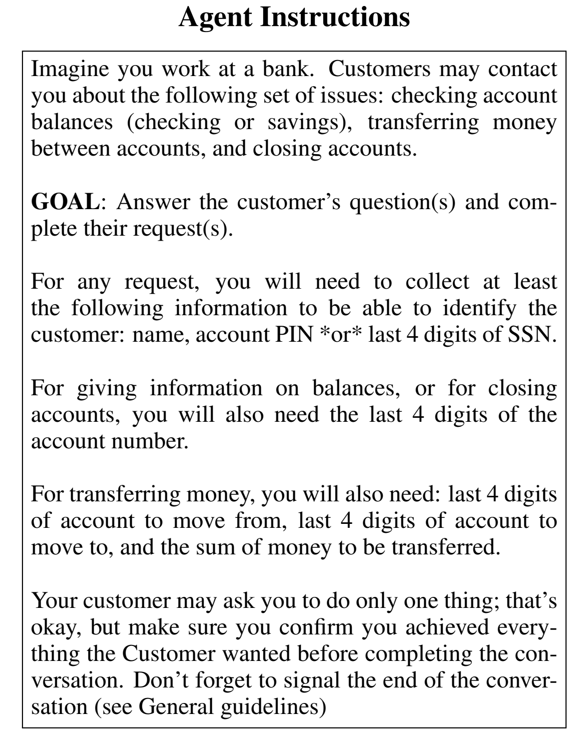
Secondly, for data annotation, each conversation is annotated at least twice. We remove inconsistent annotations by selecting the annotation given by the majority of annotators per item. We calculate inter-annotator agreement with Fleiss Kappa and find "substantial agreement", according to the metric. Our annotators must pass a qualification test as well as maintain an on-going level of accuracy in randomly distributed test questions throughout their annotation. Third, we preprocess our data to remove issues such as duplicate conversations and improperly entered slot value spans. We refer readers to our discussion of preprocessing in Section 5 for further detail.

4.5 Dataset Characterization and Statistics

`MultiDoGO` dataset is more diverse by virtue of covering more domains, but more importantly, it is more controlled since it was curated rather than being scraped from existing data sources that are not necessarily synchronous (Ubuntu). Table 3 shows the statistics for `MultiDoGO` raw conversations harvested, rated as Excellent or Good, and annotated for DA, IC and SL.

Table 4 shows the number of conversations per domain reflecting the specific biases used.

`MultiDoGO` is several orders of magnitude larger than comparable datasets as reflected in nearly every dimension: the number of conversations, the length of the conversation, the number of domains, and the diversity of the utterances used. Table 5 illustrates a comparative statistics to existing data sets.



Agent Instructions

Imagine you work at a bank. Customers may contact you about the following set of issues: checking account balances (checking or savings), transferring money between accounts, and closing accounts.

GOAL: Answer the customer's question(s) and complete their request(s).

For any request, you will need to collect at least the following information to be able to identify the customer: name, account PIN *or* last 4 digits of SSN.

For giving information on balances, or for closing accounts, you will also need the last 4 digits of the account number.

For transferring money, you will also need: last 4 digits of account to move from, last 4 digits of account to move to, and the sum of money to be transferred.

Your customer may ask you to do only one thing; that's okay, but make sure you confirm you achieved everything the Customer wanted before completing the conversation. Don't forget to signal the end of the conversation (see General guidelines)

Figure 2: Agents are provided with explicit fulfillment instructions. These are quick-reference instructions for the Finance domain. Agents serve as one level of quality control by evaluating a conversation between Excellent and Unusable.

We provide summary statistics for the subset of our data annotated at both turn and sentence granularity in Table 6. This describes the total size of the data per domain in number of conversations, turns, the unique number of intents and slots, and inter-annotator agreement (IAA) for both turn and sentence level annotations. It is worth observing that the DA annotations achieve a much higher IAA in Sentence level annotations compared to Turn level annotation, most notably in the Fast Food domain. IC and SL annotations reflect a slightly higher IAA in Turn level annotation granularity compared to Sentence level.

5 Dialogue Classification Baselines

To establish baseline performance for the `MultiDoGO` dataset we pre-process, create dataset splits, and evaluate the performance of three baseline models for each domain.

Pre-processing: We pre-process the corpus of dialogues for each domain to remove duplicate

Domain	**Elicited**	**Good/Excellent**	**IC/SL**	**DA/IC/SL**
Airline	15100	14205	7598	6287
Fast Food	9639	8674	7712	4507
Finance	8814	8160	8002	6704
Insurance	14262	13400	7799	7434
Media	33321	32231	19877	12891
Software	5562	4924	3830	2753
Total	**86698**	**81594**	**54818**	**40576**

Table 3: Total number of conversations per domain: raw conversations Elicited; Good/Excellent is the total number of conversations rated as such by the agent annotators; (IC/SL) is the number of conversations annotated for Intent Classes and Slot Labels only; (DA/IC/SL) is the total number of conversations annotated for Dialogue Acts, Intent Classes, and Slot Labels.

Bias	**Airlines**	**Fast Food**	**Finance**	**Insurance**	**Media**	**Software**
IntentChange		1443				
MultiIntent	2200	1913	1799	1061	607	2295
MultiValue		354				
Overfill			1486	2763		
SlotChange	4207	2011	2506	3321	570	2085
SlotDeletion		333				
Total	6407	6054	5791	7145	1177	4380

Table 4: Number of conversations per domain collected with specific biases. Fast Food had the maximum number of biases. MultiIntent and SlotChange are the most used biases.

Metric	DSTC 2	WOZ2.0	FRAMES	KVRET	**M2M**	**MultiWOZ**	**MultiDoGO**
Number of Dialogues	1,612	600	1,396	2,425	1,500	8,438	40,576
Total Number of Turns	23,354	4,472	19,986	12,732	14,796	115,424	813,834
Total Number of Tokens	199,431	50,264	251,867	102,077	121,977	1,520,970	9,901,235
Avg. Turns per Dialog	14.49	7.45	14.60	5.25	9.86	15.91	20.06
Avg. Tokens Per Turn	8.54	11.24	12.60	8.02	8.24	13.18	12.16
Total Unique Tokens	986	2,142	12,043	2,842	1,008	24,071	70,003
Number of Unique Slots	8	4	61	13	14	25	73
Number of Slot Values	212	99	3,871	1,363	138	4,510	55,816
Number of Domains	1	1	1	3	1	7	6
Number of Tasks	1	1	1	2	2	2	3

Table 5: MultiDoGO is several times larger in nearly every dimension to the pertinent datasets as selected by Budzianowski et al. (2018). We provide counts for the training data, except for FRAMES, which does not have splits. Our number of unique tokens and slots can be attributed to us not relying on carrier phrases.

Domain	#Conv	#Turn	#Turn/Conv	#Sentence	#Intent	#Slot	Turn-level IAA	Sentence-level IAA
Airline	2,500	39,616	15.8 (15)	66,368	11	15	0.514/0.808/0.802	0.670/0.788/0.771
Fast Food	2,500	46,246	18.5 (18)	73,305	14	10	0.314/0.700/0.624	0.598/0.725/0.607
Finance	2,500	46,001	18.4 (18)	70,828	18	15	0.521/0.827/0.772	0.700/0.735/0.714
Insurance	2,500	41,220	16.5 (16)	67,657	10	9	0.521/0.862/0.848	0.703/0.821/0.826
Media	2,500	35,291	14.1 (14)	65,029	16	16	0.499/0.812/0.725	0.678/0.802/0.758
Software	2,500	40,093	16.0 (15)	70,268	16	15	0.508/0.748/0.745	0.709/0.764/0.698

Table 6: Data statistics by domain. Conversation length is shown in *average (median)* number of turns per conversation. Inter-annotator agreement (IAA) is measured with Fleiss' Kappa for the three annotation tasks: Agent DA (DA), Customer IC (IC), and Slot Labeling (SL).

conversations and utterances with inconsistent annotations. The most common source of inconsistent annotations in our dataset is imprecise selection of slot label spans by annotators, which results in sub-token slot labels. While much of this inconsistent data could likely be recovered by mapping each character span to the nearest token span, we drop these utterances to ensure these errors have no effect on our experimental results. Our post-processed data is pruned to approximately 90% of the original size. We form splits for each domain at the conversation level by randomly assigning 70% of conversations to train, 10% to development, and 20% to test. Conversation level splits enable the application of contextual models to our dataset, as each conversation is assigned to a single

		Airline			Fast Food			Finance		
Model	Annot	DA	IC	SL	DA	IC	SL	DA	IC	SL
MFC	S	60.57	33.69	38.71	57.14	25.42	61.92	51.73	37.37	34.07
LSTM	S	97.20	90.84	74.16	90.40	86.09	72.93	93.90	90.06	69.09
ELMo	S	**97.32**	**91.88**	**86.55**	**91.03**	**87.95**	**77.51**	**94.07**	**91.15**	**77.36**
MFC	T	33.04	32.79	37.73	33.07	25.33	61.84	36.52	38.16	34.31
LSTM	T	**84.25**	89.15	75.78	**66.41**	87.35	73.57	76.19	92.30	70.92
ELMo	T	84.04	**89.99**	**85.64**	65.69	**88.96**	**79.63**	**76.29**	**94.50**	**79.47**
		Insurance			Media			Software		
Model	Annot	DA	IC	SL	DA	IC	SL	DA	IC	SL
MFC	S	56.87	38.37	53.75	57.02	30.42	82.06	58.14	33.32	53.96
LSTM	S	**94.73**	93.30	75.27	**94.27**	92.35	90.84	93.22	90.95	69.48
ELMo	S	94.63	**94.27**	**88.45**	**94.27**	**93.32**	**93.99**	**93.66**	**92.25**	**76.04**
MFC	T	36.39	39.42	54.66	29.90	31.82	78.83	36.79	33.78	54.84
LSTM	T	**75.37**	94.75	76.84	**77.94**	94.35	87.33	**83.32**	89.78	72.34
ELMo	T	75.34	**95.39**	**89.51**	77.81	**94.76**	**91.48**	82.97	**90.85**	**76.48**

Table 7: Dialogue act (DA), Intent class (IC), and slot labeling (SL) F1 scores by domain for the majority class, LSTM, and ELMo baselines on data annotated at the sentence (S) and turn (T) level. Bold text denotes the model architecture with the best performance for a given annotation granularity, i.e. sentence or turn level. Red highlight denotes the model with the best performance on a given task across annotation granularities.

	Airline		Fast Food		Finance		Insurance		Media		Software	
Annot	Single	Joint	Single	Joint	Single	Joint	Single	Joint	Single	Joint	Single	Joint
S	97.32	**97.44**	91.03	**91.26**	94.07	**94.27**	94.63	**94.99**	94.27	**94.47**	93.66	**94.00**
T	84.04	**84.64**	**65.69**	65.35	**76.29**	75.68	75.34	**75.89**	77.81	**78.56**	82.97	**83.76**

Table 8: Joint training of ELMo on all agent DA data leads to a slight increase in test performance. However, we expect stronger joint models that leverage transfer learning should see a larger improvement. Bold text denotes the training strategy, i.e. single domain (Base) or multi-domain (Joint), with the best performance for a given annotation granularity. Red highlight denotes the strategy with the highest DA F1 score across annotation granularities.

split. However, our conversation level splits result in imbalanced intent and slot label distributions.

Models: We evaluate the performance of two neural models on each domain. The first is a bi-directional LSTM (Hochreiter and Schmidhuber, 1997) with GloVe word embeddings, a hidden state of size 512, and two fully connected output layers for slot labels and intent classes respectively. The second model, ELMo, is similar to the LSTM architecture but it addiitonally uses pre-trained ELMo (Peters et al., 2018) embeddings in addition to GloVe word embeddings, which are kept frozen during training. We combine these ELMo and GloVe embeddings via concatenation. As a sanity check, we also include a most frequent class (MFC) baseline. The MFC baseline assigns the most frequent class label in the training split to every utterance u' in the test split for both DA and IC tasks. To adapt the MFC baseline to SL, we compute the most frequent slot label MFC(w) for each word type w in the training set. Then given a test utterance u', we assign the pre-computed, most frequent slot MFC(w') to each word $w' \in u'$ if w' is present in the training set. If a given word $w' \in u'$ is not present the training set, we assign the *other* slot label, which denotes the absence of a slot, to w'. We utilize the AllenNLP (Gardner et al., 2017) library to implement these models and evaluate our performance. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001 to train the LSTM and ELMo models for 50 epochs, using batch sizes 256 and 128, respectively. In addition, we employ early stopping on the validation loss with a tolerance of 10 epochs to prevent over fitting.

Evaluation Metrics: We report micro F1 score to evaluate DA and IC performance of our models. Similarly, we use a span based F1 score, implemented in the seqeval library [5], to evaluate SL performance.

5.1 Results

DA/IC/SL Results. Table 7 presents the MFC, LSTM, and ELMo results for each domain, on the subset of 15,000 conversations annotated at both the turn and sentence levels. In general for both granularities Turn and Sentence, both LSTM,

[5] https://github.com/chakki-works/seqeval

and ELMo outperform MFC significantly across all domains. Relative to the LSTM, we find that ELMo obtains a modest increase in IC accuracy of 0.41 to 2.20 F1 points and a significant increase in SL F1 score on all domains. Concretely, ELMo boosts SL F1 performance by 3.16 to 13.17 F1 points. We see the biggest SL gains on the Insurance domain, where sentence level ELMo achieves the 13.17 point F1 gain and turn level ELMo achieves a 12.67 point F1 gain. Performance gains on the Airline domain are also large; here, ELMo increases sentence and turn level SL F1 score by 12.38 and 9.86 F1 points, respectively. Both LSTM and ELMo yield similar performance in terms of F1 score on DA classification for which the difference in performance of these models is within one F1 point across all domains. In general, the FastFood domain yields the overall lowest absolute F1 scores. Recall that Fastfood had the most diverse dialogues (biases) as per Table 4 and the lowest IAA as per Table 6.

Sentence vs. Turn Level Annotation Units. Regarding the performance of the LSTM and ELMo models on sentence vs. turn level annotation units, our results suggest that turn level annotations increase the difficulty of the DA classification task. This finding is evidenced by DA performance of our models on the Fast Food domain, for which F1 score is up to 25 F1 points lower for turn level annotations than sentence level annotations. We believe the increased difficulty of turn level DA relative to sentence level DA is driven by a corresponding increase in the confusability of turn level dialogue acts. This assertion of greater turn level DA confusability is supported by the lower inter annotator agreement (IAA) scores on turn level DA, which range from 0.314 to 0.521, relative to IAA scores for sentence level DA, which range from 0.598 to 0.709. This experimental result highlights the importance of collecting sentence level annotations for conversational DA datasets. Somewhat surprisingly, our models achieve similar IC F1 and SL F1 scores on turn and sentence level annotations. We hypothesize that the choice of annotation unit has a lesser impact on the IC and SL tasks because customer utterances are more likely to focus on a single speech act, whereas Agent utterances may be more complex in comparison and include a greater number of speech acts.

Joint Training on Agent DA. Agent DA classification naturally lends itself to joint training, given agent DAs are shared among all domains. To explore the benefits of multi-domain training, we jointly train an agent DA classification model on all domains and report test results for each domain separately. These results are provided in Table 8. This straightforward technique leads to a consistent but less than one point improvement in F1 scores. We expect that more sophisticated transfer learning methods (Liu et al., 2017; Howard and Ruder, 2018) could generate larger improvements for these domains.

Overall, our results demonstrate that there is still headroom for performance improvement, especially for the SL task, across all domains. Consequently, `MultiDoGO` should be a relevant benchmark for developing new state-of-the-art NLU models for the foreseeable future.

6 Future Directions

The data collection and annotation methodology that we use to gather `MultiDoGO` can efficiently scale across languages. Several pilot experiments aimed at collecting Spanish dialogues in the same domains have shown preliminary success in quality assessment. The production of a NLU dataset with parallel data in multiple languages would be a boon to the cross-lingual research community. To date, cross-lingual NLU research (Upadhyay et al., 2018; Schuster et al., 2018) has relied on much smaller parallel corpora.

7 Conclusion

We present `MultiDoGO`, a new Wizard-of-Oz dialogue dataset that is the largest human-generated, multi-domain corpora of conversations to date. The scale and range of this data provides a test-bed for future work in joint training and transfer learning. Moreover, our comparison of sentence and turn level annotations provides insight into the effect of annotation granularity on downstream model performance. By pairing crowd-source labor with professional data annotators, we balance the cost, diversity, and quality of these conversations in a scalable manner. We show that by adopting a modular annotation strategy, the crowds can reliably annotate dialogues at a level commensurate with trained professional annotators.

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. *arXiv preprint arXiv:1704.00057*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1723–1732.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

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, June 24-27, 1990*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Oliver Lemon, Kallirroi Georgila, James Henderson, and Matthew Stuttle. 2006. An isu dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the talk in-car system. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 119–122. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Eric Mihail, Krishnan Lakshmi, Charette Francois, and Manning Christopher. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the Special Interest Group on Discourse and Dialogue*.

Silvia Pareti and Tatiana Lando. 2018. Dialog intent structure: A hierarchical schema of linked dialog acts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), pages 6034–6038. IEEE.

Zhuoran Wang and Oliver Lemon. 2013. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In *Proceedings of the SIGDIAL 2013 Conference*, pages 423–432.

Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. Airdialogue: An environment for goal-oriented dialogue research. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3844–3854.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Jason Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.