Chapter 1: See Other File

Chapter 2: See Other File

Chapter 3: See Other File

Chapter 4: See Other File

Chapter 5: See Other File

Chapter 6: See Other File

Chapter 8: Quantity and (Mostly) Quality Through Hybridization

As a dovetail between crowd-driven and expert-driven data sources, we propose a hybrid solution that pairs a **crowd-worker** *with* an **expert**. This creates a verisimilitude of a customer, simulated by a worker from the crowd, interacting with a customer service agent, simulated by an actual professional customer service agent. The resulting dataset illustrates the stark contrast in the language generated by anonymous crowd workers and experts. Furthermore, it demonstrates how NLP **generation** and **annotation** can be scaled through the crowd, while being quality controlled by an expert.[1]

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 8.1: 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".

8.1 The Goal of Creating Goal-Oriented Dialogues

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 refer to a flight-booking intent. Hence, entire *conversations*, rather than independent utterances, must be generated.

NLU would benefit from large, varied, and ideally human-generated datasets. Joint-training and transfer learning techniques (Dong et al., 2015; Devlin et al., 2019) benefit natural language processing tasks; however, these approaches have yet to become widely used in dialogue tasks due to a lack of large-scale datasets.

[1]Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang,Adel Youssef, and Mona Diab. Multi-domain goal-oriented dialogues(multidogo): Strategies toward curating and annotating large scale dialogue data. 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 4518–4528, 2019.

Peskov planned and implemented some of the crowd-sourcing tasks, supervised the data collection thereof, wrote some of the task instructions, performed data analysis, and wrote most of the paper.

Furthermore, end-to-end neural approaches benefit from such training data more than past work on goal-oriented dialogues structured around slot filling (Lemon et al., 2006; Wang and Lemon, 2013).

Conveniently, the training data for NLU occurs organically. 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 fulfill 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).

We curate, annotate, and evaluate a large scale multi-domain set of goal oriented dialogues to address the prior limitations. 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 conversation fulfill a role (Kelley, 1984). Popular goal-oriented datasets, DSTC (Williams et al., 2016) and `MultiWOZ` (Budzianowski et al., 2018) use this approach. Hence, our dataset is gathered from workers in the crowd paired with professional annotators using Wizard-of-Oz. The dataset generated, `MultiDoGO`, comprises over 86K raw conversations of which 54,818 conversations are annotated at the turn level; this is a geometric increase over the number of utterances generated in Chapter 8. We investigate multiple levels of annotation granularity: annotating a subset of the data

on both turn and sentence levels. Generating and annotating such data given its contextual setting is nontrivial. We furthermore illustrate the efficacy of our devised approaches and annotation decisions against intrinsic metrics and via extrinsic evaluation by applying neural baselines for **Dialogue Acts**, **Intent Classification**, and **Slot Labeling**.

8.2 Existing Dialogue Datasets

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 (Chapter **??**) rely on dialogue to answer questions, but the task differs from our use case of modeling goal oriented conversational AI, hence leading to different evaluation considerations (Choi et al., 2018; Reddy et al., 2019).

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). Some dialogue datasets specialize in a single domain. In addition to the datasets mentioned in Background Section **??**,

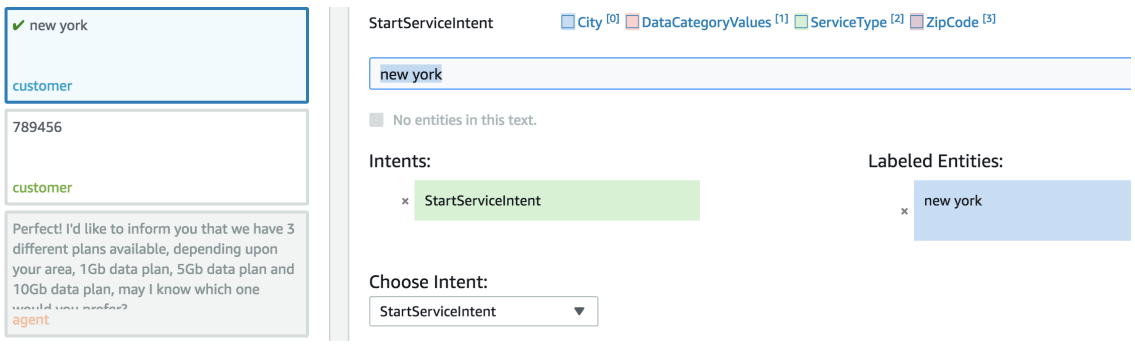

Figure 8.1: Crowd sourced annotators select an intent and choose a slot in our custom-built Mechanical Turk interface. Entire conversations are provided for reference. Detailed instructions are provided to users, but are not included in this figure. Options are unique per domain.

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 et al., 2018).[2]

8.3 `MultiDoGO` Dataset Generation

Generating and **annotating** a dataset of this scale requires task design, data collection, and post-task quality control.

8.3.1 Defining Dialogues

NLU uses specific terminology. A turn is a sequence of 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 multiple sentences. We do use the term

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

utterance to refer to a unit (turn or sentence, spoken or written by a participant).[3]

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.

8.3.2 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 HITs in the past. We give each agent a prompt listing the supported request types (dialog acts) and pieces of information (slots) needed to complete each request to structure goal-oriented conversations between the customer and agent,. We also specify criteria such as minimal conversation length, number of goals, and number of complex requests to increase conversation diversity (Figure 8.2). We explicitly request that neither agents nor customers use any personally identifiable information. At an

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

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 instructions without stopping the conversation. `MultiDoGO` follows a familiar Wizard-of-Oz elicitation procedure and curates data for multiple domains akin to previous data collection efforts such as `MultiWOZ`. However, `MultiDoGO` comprises more varied domains, is a magnitude larger, and is curated with prompts to ensure diverse conversations.

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 request multiple intents such as "I'd like to find my routing number and check my balance."[4] To that end, our collection procedure deliberately attempts to guide the dialogue flow to ensure diversity in dialogue policies.

8.4 Data Annotation

Annotation classifies the thousands of conversations in our dataset. Of particular interest, a direct comparison of using experts versus the crowd is made in

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

Section 8.4.2. Our annotators use a web interface (Figure 8.1) to select the appropriate intent class for an utterance out of a list of provided options. They use their cursors to highlight slot value character spans within an utterance and then select the corresponding slot label from a list of options to annotate slot labels. The output of this slot labeling process is a list of ⟨slot-label, slot-value, span⟩ triplets for each utterance.

8.4.1 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 to produce more fine-grained speech act tags than past iterations of dialog datasets. Intuitively, agent DAs are consistent across domains and more general 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 8.1.

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. 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 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).

8.4.2 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 generating 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

Dialogue Act	Intent Classes	Slot Labels
0.701	0.728	0.695

Table 8.2: Inter Source Annotation Agreement (ISAA) scores quantifying the agreement of crowd sourced and professional annotations.

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. A 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 conversations per domain for 15,000 conversations in total—at the sentence as well as turn level to assess the design choice on downstream accuracy. The remainder of our dataset is annotated only at the turn level.

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 8.7. 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 8.3. 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 measure the agreement of these crowd sourced and professionally sourced annotations. ISAA is a relaxation of Cohen κ, 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). We report ISAA for the DA, IC, and SL tasks in Table 8.2. Crowd sourced and professional annotations have a substantial degree of shared annotations. Therefore, the crowd can be used for **annotation** for NLP tasks, if the annotations are verified to be comparable to experts.

8.4.3 Quality Control

We institute three processes to enforce data quality, possible due to the use of **experts**. 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 follow instructions around coherence, whether the dialogue achieved the purported goal, etc., to decide

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 8.3: 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.

on the chosen rating. We keep conversations with “Good" or “Excellent" ratings in subsequent annotation to maximize the quality of our dataset.

Secondly, each conversation is annotated at least twice. We resolve inconsistent annotations by selecting the annotation given by the majority of annotators for an item. We calculate inter-annotator agreement with Fleiss’ κ and find “substantial agreement”, according to the metric.[5] Our annotators must pass a qualification test as well as maintain an on-going level of accuracy in randomly distributed test

[5]We use Fleiss’ κ unlike in the earlier profession/crowd worker comparison as we have more than two annotators for this task.

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 8.4: 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.

questions throughout their annotation. Third, we pre-process our data to remove issues, such as duplicate conversations and improperly entered slot value spans. Further pre-processing details are in Section 8.5.

8.4.4 Dataset Characterization and Statistics

The `MultiDoGO` dataset is the most diverse dialog dataset due to covering more domains and being generated, rather than scraped from existing and dubiously reliable data sources (e.g., Ubuntu forums). Table 8.3 shows the statistics for `MultiDoGO` raw conversations generated, rated as Excellent or Good, and annotated for DA, IC and SL. Table 8.4 shows the number of conversations per domain reflecting the specific biases used.

Metric	DSTC 2	WOZ2.0	M2M	MULTIWOZ	MULTIDOGO
Number of Dialogues	1,612	600	1,500	8,438	40,576
Total Number of Turns	23,354	4,472	14,796	115,424	813,834
Total Number of Tokens	199,431	50,264	121,977	1,520,970	9,901,235
Avg. Turns per Dialog	14.49	7.45	9.86	15.91	20.06
Avg. Tokens Per Turn	8.54	11.24	8.24	13.18	12.16
Total Unique Tokens	986	2,142	1,008	24,071	70,003
Number of Unique Slots	8	4	14	25	73
Number of Slot Values	212	99	138	4,510	55,816
Number of Domains	1	1	1	7	6
Number of Tasks	1	1	2	2	3

Table 8.5: `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.

`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 8.5 provides comparative statistics.

We provide summary statistics for the subset of our data annotated at both turn and sentence granularity in Table 8.7. 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

Domain	**#Conv**	**#Turn**	**#Turn/Conv**	**#Sentence**	**#Intent**	**#Slot**
Airline	2,500	39,616	15.8 (15)	66,368	11	15
Fast Food	2,500	46,246	18.5 (18)	73,305	14	10
Finance	2,500	46,001	18.4 (18)	70,828	18	15
Insurance	2,500	41,220	16.5 (16)	67,657	10	9
Media	2,500	35,291	14.1 (14)	65,029	16	16
Software	2,500	40,093	16.0 (15)	70,268	16	15

Table 8.6: Data statistics by domain. Conversation length is in *average (median)* number of turns per conversation. Inter-annotator agreement (IAA) is measured with Fleiss' κ for the three annotation tasks: Agent DA (DA), Customer IC (IC), and Slot Labeling (SL).

Domain	Turn-level IAA	Sentence-level IAA
Airline	0.514/0.808/0.802	0.670/0.788/0.771
Fast Food	0.314/0.700/0.624	0.598/0.725/0.607
Finance	0.521/0.827/0.772	0.700/0.735/0.714
Insurance	0.521/0.862/0.848	0.703/0.821/0.826
Media	0.499/0.812/0.725	0.678/0.802/0.758
Software	0.508/0.748/0.745	0.709/0.764/0.698

Table 8.7: Inter-annotator agreement (IAA) is measured with Fleiss' κ for the three annotation tasks: Agent DA (DA), Customer IC (IC), and Slot Labeling (SL).

annotations. DA annotations have 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.

8.5 Dialogue Classification Baselines

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

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

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 8.2: 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.

		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 8.8: Dialogue act (DA), Intent class (IC), and slot labeling (SL) F1 scores by domain for the majority class, LSTM, and ELMobaselines 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.

word embeddings, a hidden state of size 512, and two fully connected output layers for slot labels and intent classes. The second model, ELMo, resembles LSTM archi-

	Airline		Fast Food		Finance		Insurance		Media		Software	
A	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 8.9: Joint training of ELMo on all agent DA data leads to a slight increase in test performance. However, we expect stronger joint models that use 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.

tecture but it additionally 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 use 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. In addition, we use 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. Similarly, we use a span based F1 score, implemented in the seqeval[6] library, to evaluate SL performance.

8.5.1 Results

DA/IC/SL Results. Table 8.8 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. LSTM, and ELMo outperform MFC across all domains at the turn and sentence level. 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 over the LSTMbaseline. 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 has a 13.17 point F1 gain and turn level ELMo has a 12.67 point F1 gain. ELMo increases sentence and turn level SL F1 scores by 12.38 and 9.86 F1 points for the airlines domain. Both LSTM and ELMo yield similar F1 scores on DA classification for which the difference in performance of these models is within one F1 point across all domains. The Fast Food domain yields the overall lowest absolute F1 scores. Recall that Fast Food had the most diverse dialogues (biases) as per Table 8.4 and the lowest IAA as per Table 8.7.

Sentence vs. Turn Level Annotation Units. Turn level annotations

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

increase the difficulty of the DA classification task in our LSTM and ELMo results. This finding is evidenced by DA accuracy 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 is driven by a corresponding increase in the ambiguity of turn level dialogue acts. This assertion of greater turn level DA ambiguity is supported by the lower inter annotator agreement (IAA) scores on turn level DA, which range from 0.314 to 0.521, than the 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 have similar IC F1 and SL F1 scores on turn and sentence level annotations. We posit 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 8.9. 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, there is room for 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.

8.6 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.

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.

By pairing **crowd**-sourced labor (Chapter **??**) with **experts** (Chapters **??**), we ensure quality and diversity in **generated** conversations while scaling to multiple domains and tasks. We show that by adopting a modular annotation strategy, the crowds can reliably **annotate** dialogues at a level commensurate with trained professional annotators. Without the expert, our data would be just as large, but it could not be trusted.

There is a stark difference in quality of the **generated** language between the crowd-sourced workers and the experts, in this case Amazon Customer Service agents. The crowd-sourced workers have a financial incentive to complete the task as quickly as possible and contribute sentences that are often prosaic, ungrammatical, or repeated. In our case, these incentives mimic those of the usual customer and does not undermine the realism of the conversation. But, should datasets be *large* or should they be *accurate* in future work where these incentives are not desirable?

Bibliography

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. In *Proceedings of Empirical Methods in Natural Language Processing*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

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)*, 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 Association for Computational Linguistics*.

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 *Demonstrations*.

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.

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*, 7:249–266.

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.