Towards Universal Dialogue Act Tagging for Task-Oriented Dialogues

Shachi Paul, Rahul Goel, Dilek Hakkani-Tür

Amazon

{shachp, goerahul, hakkanit}@amazon.com

Abstract

Machine learning approaches for building task-oriented dialogue systems require large conversational datasets with labels to train on. We are interested in building task-oriented dialogue systems from human-human conversations, which may be available in ample amounts in existing customer care center logs or can be collected from crowd workers. Annotating these datasets can be prohibitively expensive. Recently multiple annotated task-oriented human-machine dialogue datasets have been released, however their annotation schema varies across different collections, even for well-defined categories such as dialogue acts (DAs). We propose a Universal DA schema for task-oriented dialogues and align existing annotated datasets with our schema. Our aim is to train a Universal DA tagger (U-DAT) for task-oriented dialogues and use it for tagging human-human conversations. We investigate multiple datasets, propose manual and automated approaches for aligning the different schema, and present results on a target corpus of human-human dialogues. In unsupervised learning experiments we achieve an F1 score of 54.1% on system turns in human-human dialogues. In a semi-supervised setup, the F1 score increases to 57.7% which would otherwise require at least 1.7K manually annotated turns. For new domains, we show further improvements when unlabeled or labeled target domain data is available.

Index Terms: dialogue act tagging, spoken dialogue systems, human-human conversations

1. Introduction

Dialogue acts aim to portray the meaning of utterances at the level of illocutionary force, capturing a speaker's intention in producing that utterance [1]. Recent work in task-oriented dialogue systems proposed a set of core DAs that describe interactions at the level of intentions [2, 3, 4]. With these, the system actions as output by a dialogue system policy are commonly represented as the system DAs and associated entities [5]. Previous work on dialogue policy learning and end-to-end training of dialogue systems rely on supervised learning approaches to estimate system actions at each turn, given the dialogue state or the previous conversation. These models can then be fine-tuned with reinforcement learning [6, 7].

In this work, we build an RNN-based DA tagger for tagging human-human task-oriented conversations with DAs from a Universal DA schema that is representative of the commonly-used acts in task-oriented dialogue systems. Our long term goal is to use these annotated human-human dialogues to train end-to-end dialogue systems to predict system actions for new dialogue-task domains. Here, we focus on automatically annotating system-side DAs on human-human dialogues. Such human-human dialogues for the new domain can be found in existing customer care center logs or collected via crowdsourcing by pairing two crowdworkers [8] or asking a single crowd worker to write self dialogues [9]. Previous work on DA tagging mainly focused on human-human social interactions, such as the Switchboard corpus [10], with little or no attention to task-oriented dialogues.

Recently, multiple annotated task-oriented human-machine dialogue datasets have been released [11, 4], fostering research in this area. Hence, we focus on learning to tag DAs from these human-machine dialogues, and applying the learned models to human-human dialogues for task-oriented systems. However, the annotation schema varies across different corpora, even for well-defined categories, such as DAs. Towards this goal we experiment with various alignment schemes, propose a Universal schema of DAs across multiple existing corpora, align the corpora accordingly to train a Universal DA tagger (U-DAT).

We use U-DAT to tag human-human multi-domain dialogues (MultiWOZ-2.0 [8]). In our semi-supervised learning experiments we achieve an F1 score of 57.7% on system-turns on human-human data, which requires at least 1.7K manually annotated turns. We examine the potential of domain adaptation of the U-DAT by leave-one-domain-out experiments. In presence of a new domain we compare the performance of DA tagging using unsupervised, semi-supervised(self-training) and supervised approaches. For these domains, we show further improvements when unlabeled or labeled target domain data is available, providing guidelines on bootstrapping a new domain without any DA annotations.

Our work has multiple novel contributions including a new hierarchical recurrent neural network based approach for tagging DAs, a Universal DA schema for task-oriented dialogues, alignment of multiple datasets to the universal schema, using the aligned corpus for training of U-DAT for human-human dialogue annotation and showcasing alternatives when bootstrapping a new domain.

2. Related Work

Since the publication of the seminal work on a machine learning approach for DA tagging [12], multiple learning approaches have been proposed for this task, including maximum entropy taggers [13], conditional random fields [14], and dynamic Bayesian networks [15]. Recent studies investigated recurrent and convolutional neural networks with a pooling layer for short-text classification tasks, such as DA tagging [16]. But these works don't take into account the dialogue context. However, in a task-oriented conversation, there is a strong correlation between system and user acts. For example, a user usually *inform*s when a system *request*s information. Our work represents short user utterances using recurrent neural networks, and additionally models dialogue context using a hierarchical recurrent neural network. Such dialogue-level models have also been proposed in [17] for dialogue act tagging of human-human social phone conversations. Previous studies mainly considered DA tagging of multi-human conversations, such as the Switchboard [10] corpus and meetings, such as the ICSI meeting corpus [18] whereas, our focus lies on modeling system-side DAs.

In dialogue systems, the system utterances are also generated from system actions and are hence, observable. Thus, in our context representation we include past system DAs in addition to system utterances. For user utterances as well as system-side DAs in human-human conversations, we use the predicted DAs. Domain adaptation of DA tagging with unlabeled data was also investigated by [19] for two human-human conversation genre, telephone speech and face-to-face meetings. However, that work did not have annotation mismatch issues across different datasets.

3. DA Tagging for Dialogue Systems

Let a dialogue D with N turns be denoted as a series of user and system utterances, u_i, i.e. $D = u_1, u_2, ..., u_N$ and A be the predefined set of M DAs i.e. $A = a_1, a_2..., a_M$. Given an utterance u_i and its conversation history, DA tagging aims to predict the set of DAs $A_i \subset A$ of u_i.

We use a deep neural network based model for DA tagging. The input to the model is the utterance u_i and the conversation context C_i which is a function of the past utterances and their corresponding DAs i.e. $C_i = f((u_1, A_1), ..., (u_{i-1}, A_{i-1}))$. Since the utterance u_i can be classified into one or more DAs in A, the model makes a binary-decision $y_j \in [0, 1]$ for every DA $a_j \in A_{1..M}$ for candidature.

For a dialogue D_k and every DA class $a_j \in A$, we minimize the following cross-entropy loss:

$$L = -\sum_{j=1}^{M}\sum_{i=1}^{N} \log\left(P\left(y_j | C_i^k, u_i^k\right)\right). \qquad (1)$$

We use the following encoders to represent context:

1. A bi-directional LSTM to encode each utterance, $u_i = w_1^i, ..., w_{n_i}^i$, where n_i denotes the number of tokens in u_i and the final utterance representation for utterance z_i is obtained by concatenating the last hidden layer of the forward LSTM, $\overrightarrow{LSTM}$, and the first hidden layer of the backward LSTM, $\overleftarrow{LSTM}$:
$$z_i = \overleftarrow{LSTM}(u_i) \oplus \overrightarrow{LSTM}(u_i)$$
2. A hierarchical, uni-directional LSTM to encode the dialogue level information, e_i:
$$e_i = LSTM(z_1, ...z_{i-1})$$
3. An indicator number, g_i, representing whether the agent is user or the system, i.e., $g_i = 0$, if $u_i^{agent} = user$, $g_i = 1$ otherwise.
4. Encoding over past DA(s) p_i, where the final representation is obtained by concatenating the many-hot representations of past-DAs. A DA vector is represented as a many-hot vector d_i of dimension M, where we mark the true DAs as 1.
$$p_i = d_1 \oplus d_2 \oplus ... \oplus d_{i-1}$$

The final encoded context C_i is given by:

$$C_i = e_i \oplus g_i \oplus p_i \qquad (2)$$

C_i is then fed into a feed forward network FF_j, along with u_i, for each DA. The context encoders are shared for all acts.

$$y_j = sigmoid(FF_j(z_i \oplus C_i)) \qquad (3)$$

4. Datasets and Experiments

Our aim is to train a Universal DA tagger using public datasets, but the label spaces across these datasets are not aligned. Therefore, we need a unified representation of all the acts present across the datasets. We obtain this representation by manually going through the datasets and aligning semantically similar sentences to the same DA. We chose the Google Simulated Dialogue (GSim) dataset [4] and the DSTC2 dataset [20] for our experimentation as they are both inspired by the CUED schema [2] for DAs. The GSim data has two parts and was collected by generating dialogue flows for movie (GSim-M) and restaurant (GSim-R) booking domains, where the individual turns from simulation in terms of DAs and associated arguments were then converted to natural language by crowd workers. DSTC2 contains human-machine interactions collected for the second dialogue state tracking challenge [11]. To experiment with DA tagging on human-human conversational interactions, we use the MultiWOZ-2.0 [8], which was collected by assigning tasks (such as, booking a restaurant table and a cab to get there) and roles (such as, user and agent) to two crowd workers, paired to accomplish the task. The three datasets are summarized and compared across various metrics in Table 1. The last two rows of the table show the vocabulary size of the system turns and unique number and percentage of system turns after delexicalization, which replaces the entity values with an entity type. The percentage of unique turns is obtained by dividing the unique number of system turns with the total number of system turns. Since the GSim and MultiWOZ-2.0 dialogues were written by crowdworkers, they include lots of variation in the output system turns, whereas DSTC2 system turns were generated by the participating systems, and have much less richness for building DA taggers for system acts, but provides consistent annotations.

Table 1: *Data statistics of various datasets*

Data Sets:	GSim-R	GSim-M	DSTC2	MultiWOZ-2.0
# Dialogues Train	1,116	384	1,612	8,438
# Dialogues Dev	349	120	506	1,000
# Dialogues Test	775	264	1,117	1,000
Avg # Turns/Dialogue	5.5	5.1	7.2	6.7
# Sys Dialogue Acts	7	7	12	14
SysTurn Vocab Size	577	349	229	15,408
#Uniq SysTurns	3,878	1,247	306	49,460
%Uniq SysTurns	76.6%	78.4%	2.6%	87.2%

Experimental Setup: Our model architecture consists of four encoders: the utterance encoder, hierarchical dialogue encoder, past DAs encoder and an agent encoder. Our utterance encoder is a bi-directional LSTM with hidden layer size of 128. The utterance representation is the final state of the biLSTM. The hierarchical dialogue encoder is an LSTM which takes the utterance representation as input and its hidden size is 256. The past DAs vector is a concatenation of the many-hot representations of past DAs wherein each DA is many-hot over a set of 20 DAs. The agent encoding is an indicator number representing the agent of the turn - 0 for the user, 1 for the system. We concatenate these representations and pass it through a feed-forward network to make a binary decision per DA. For training, we use ADAM for optimization with a learning rate of 0.001 and default parameters. Our batch size is 100 for training. We initialize our word embeddings with pretrained fastText [21] embeddings and fine-tune during training.

5. Universal DA Schema

5.1. Union of acts based on namespace

In order to align the respective acts in the datasets (GSim and DSTC2), we first took a union of all the acts based on their names to create a unified representation. Figure 1 represents the distribution of DAs used for the system side in these datasets. Since our final aim is to tag human-human conversations (MultiWOZ-2.0 [8]) with our unified set of acts, we have also included the distribution of acts in MultiWOZ-2.0 for com-

Figure 1: *Distribution of system acts across datasets*

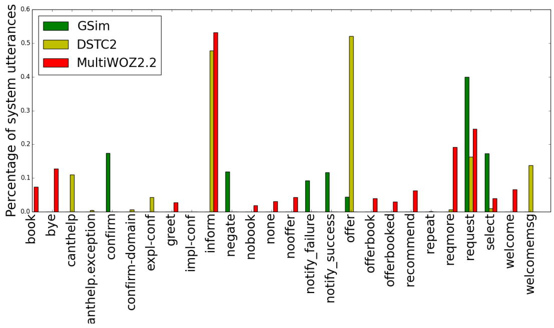

Table 2: *Examples of manual alignment of acts in all datasets.*

GSim	DSTC2	MultiWOZ-2.0	Univ DA Schema
notify_failure	canthelp.exception	NoBook	sys_notify_failure
confirm	confirm-domain expl-conf		sys_expl_confirm
offer	offer	Recommend	offer

pleteness, after stripping off the domain-name from the acts. It can be observed from the distribution that apart from a few common acts like *inform*, *request* etc., these datasets do not share the same namespace for DAs, and even when the names are the same, there may be differences in their semantics, as the distributions of the acts are very different. For example, MultiWOZ-2.0 does not include the *offer* act, whereas it appears in about 7% of the system turns for GSim and 50% of the system turns for DSTC2. Similarly, GSim and MultiWOZ-2.0 both have $select$, about 20% and 5% of the turns respectively, whereas it is observed rarely in DSTC2.

5.2. Tackling Annotation Mismatch: Manual alignments

Due to the lack of a shared namespace of acts, we manually assessed the semantics of the acts in the datasets and found some obvious alignments. Table 2 includes example alignments.

Post-alignment, many acts in these datasets were shared between the user and system such as *inform* and *negate*. However, we observed that these acts do not share the same semantics and wording and hence the flow of the conversation varies based on which agent the turn belongs to. Thus, to curate our unified schema of acts - we made a finer distinction between user/system acts i.e. *negate* from the user is a *user-negate* whereas from system, is *system-negate*. Finally, we train our DA tagger with the manually-aligned DSTC2 and GSim data.

To gauge the effectiveness of manual-alignments, we trained our DA tagger on one dataset and tested it on the other to see the inter-dataset and intra-dataset confusion. The results of these experiments are listed in Table 3 as *Baseline* numbers. The best result for each test-set is highlighted in the table. As expected, for each test-set, we obtain the best F1 scores when we use the matching training-set. On the combined test-set, the model trained after combining all the datasets performs the best.

5.3. Machine-aided alignments

After manually aligning the acts across datasets, we still observed poor performance on the task. Looking at the various training and validation set DAs in the manually curated unified representation, we noticed some semantically similar acts which were confusing our tagger. Some examples are:

1. **Mod1: offer/select**- *I found a show for 7.30 pm/I found shows for 5 pm and 7 pm*. We merge these acts.
2. **Mod2: user-request/sys-request**- *What is the phone number?/What kind of food would you like?* We merge these acts.
3. **Mod3: affirm(x=y)/affirm + inform(x=y)**- *affirm* with slots is equivalent to separate *affirm* and *inform* DAs, for eg. 'yes, 7pm' can become *affirm, inform(time=7pm)* from *affirm(time=7pm)*. We split them.
4. **Mod4: reqalts/reqmore**- *Is there anything else?/Can i help you with anything else?* We merge these acts.

Table 3: *Comparison of models trained on manually-aligned(baseline) vs the final machine-aligned universal schema (univ) based on F1 scores. Inter-dataset numbers represent the setting where the train/test data belong to different datasets. In intra-dataset, they belong to the same one. 'All' is the combination of respective partitions of all datasets*

		Training Set			
Test Sets		Gsim-R	Gsim-M	DSTC2	All
Gsim-R	Baseline	**0.867**	0.706	0.324	**0.897**
	Univ	**0.892**	0.751	0.453	**0.916** (U-DAT)
Gsim-M	Baseline	0.801	**0.904**	0.382	**0.908**
	Univ	0.850	**0.914**	0.474	**0.921** (U-DAT)
DSTC2	Baseline	0.434	0.365	**0.909**	0.899
	Univ	0.564	0.496	**0.920**	0.917 (U-DAT)
All	Baseline	0.560	0.477	0.742	**0.900**
	Univ	0.659	0.583	0.786	**0.921** (U-DAT)
Avg of inter-dataset scores (Baseline/Univ)				0.439/0.555	
Avg of intra-dataset scores (Baseline/Univ)				0.898/0.912	

Table 4: *Effect of squashing/splitting different acts on inter and intra-dataset average F1 score. Each column displays the effect of addition of modification to the one on its left.*

	manually-aligned	+Mod1	+Mod2	+Mod3	+Mod4
Inter-dataset	0.439	0.436	0.512	0.527	0.555
Intra-dataset	0.898	0.903	0.896	0.890	0.912

We merged/split DAs like the aforementioned ones, as they can easily be restored using other information. For example, if multiple results are offered, we could convert an *offer* act to a *select* act, or depending on the agent, we can convert a *request* act to a *user-request* or a *sys-request*. The effect of these transformations on inter and intra-dataset F1 scores is shown in Table 4.

After performing all these transformations, we curated a Universal DA schema of 20 acts which capture the entirety of all the acts present in these datasets. We present these in Table 5[1]. We compare the F1 scores of DA tagging models trained using this schema with our original baseline (models trained on manually-aligned acts) in Table 3. The best-performing model is obtained by combining all the GSim and DSTC2 datasets using the Universal DA schema. We refer to this model as U-DAT.

6. DA Tagging of Human-Human Datasets

For experimenting with DA annotation of human-human(HH) dialogues, we used MultiWOZ-2.0[8] as our dataset. This version of the dataset only has DAs for the system turns.

To do an evaluation on MultiWOZ-2.0, we first need to map the dataset to our Universal DA schema. The distribution of acts in MultiWOZ-2.0 can be seen in Figure 1. However,

[1] We will release the alignment of acts in all datasets with the Universal DA schema as supplementary material with the final version of the paper.

Table 5: *Universal DA schema*

ack, affirm, bye, deny, inform, repeat, reqalts, request, restart, thank-you, user-confirm, sys-impl-confirm, sys-expl-confirm, sys-hi, user-hi, sys-negate, user-negate, sys-notify-failure, sys-notify-success, sys-offer

during manual assessment, we found that while most of the acts in MultiWOZ-2.0 dataset aligned well with our Universal DA schema, the annotations in *inform/select/recommend* and *general-domain* space of acts were inconsistent in MultiWOZ-2.0. For example, *select* was often confused with *inform*. Additionally, the MultiWOZ-2.0 act annotation space lacks granularity for expressing intent. Thus, to maximally align the MultiWOZ-2.0 with the Universal DA schema, we used heuristics. For example, we check for presence of keywords like 'bye', 'thank you' etc. to label the *bye* and *thank_you* class of system acts. [2] To evaluate the effectiveness of our heuristics, we manually annotated a smaller subset (524 turns) of the MultiWOZ-2.0 test-set with DAs in our Universal DA schema, we call this as the *univ-testset*. We then trained 2 DA tagging models on MultiWOZ-2.0 - one with the DA labels mapped to the Universal DA schema using heuristics (say *heuristics-model*) and one without (say *no-heuristics-model*). On *univ-testset*, we got an F1 score of 0.609 with *no-heuristics-model* and 0.716 with *heuristics-model*, which validates the effectiveness of our heuristics.

Due to labeling inconsistency in MultiWOZ-2.0 as described above, to do an accurate evaluation on HH datasets, we use *univ-testset* as our test-set henceforth. For training and validation, we use the standard dataset partitions.

6.1. Adaptation to Human-Human dialogues

In addition to the unsupervised (w.r.t. HH data) U-DAT model, we train 2 other models[3] in semi-supervised and supervised settings.

- **Semi-supervised HH U-DAT:** Our aim is to see the quality of DA annotations without any labeled HH dialogues. For this, we labeled the MultiWOZ-2.0 corpus with U-DAT. Then, we trained another model with the estimated DA labels. This model is semi-supervised, as it doesn't use any manually labeled data, but uses the labels generated by U-DAT.
- **Supervised HH U-DAT:** We trained a supervised DA tagging model with the manually annotated DA labels mapped to our Universal DA schema.

We plot learning curves by varying the amount of data used in each model in Figure 2. In this plot, the red line corresponds to the performance of **U-DAT**, a system-side F1 score of 0.541. The blue line is the performance of **Semi-supervised HH U-DAT** with all the MultiWOZ-2.0 training dialogues, a system-side F1 score of 0.577. These two lines show that, if unannotated data is available from the HH conversations, we can improve the DA tagging F1 score by 3.6% absolute. As can be seen from the green curve obtained with **Supervised HH U-DAT**, we would need over 1700 manually annotated examples to reach the best semi-supervised learning F1 score. This provides useful guidelines on the amount of data required for accurate DA tagging.

6.2. Analysis of domain adaptation via self-training

To gauge the extent of domain adaptation by self-training (semi-supervised) over HH data, we also performed leave-one-domain-out experiments. For each domain X, we train 3 models on HH data (mapped to Universal DA schema) with the following settings:

[2] We will release the heuristics as a supplementary material with the final version of the paper.

[3] Due to the absence of user-acts in MultiWOZ-2.0, we removed the past DA encoder from the context encoder of both the models.

Figure 2: *DA tagging: System-side F1 score learning curves using U-DAT and models trained using HH data on univ-testset.*

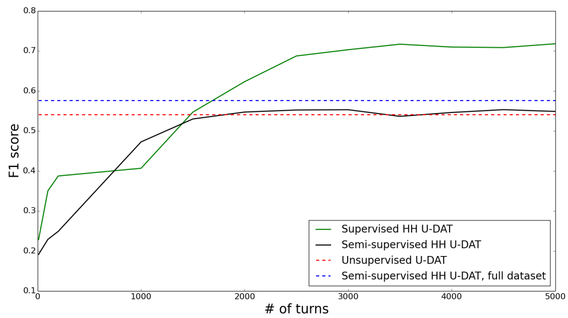

Table 6: *Per-domain F1 score results for domain adaptation via self-training (semi-supervised training)*

Domain	U-DAT	HH-UDAT	HH-UDAT	HH-UDAT
Domain data	N/A	no	yes	yes
Supervision	no	yes	semi	yes
Restaurant	0.613	0.703	0.721	0.735
Hotel	0.571	0.685	0.697	0.701
Train	0.523	0.723	0.666	0.724
Taxi	0.709	0.727	0.787	0.784
Attraction	0.444	0.672	0.689	0.728
Average	0.572	0.702	0.712	0.734

- *HH-UDAT, no domain data, supervised*: We use 3000 turns of manually-labeled out-of-domain(OOD) data i.e. we exclude turns from domain X. Since the data is manually-labeled, this model is supervised.
- *HH-UDAT, w-domain data, semi-supervised*: In addition to OOD data used above, we use 300 turns of data from X labeled using U-DAT. This model is semi-supervised w.r.t X.
- *HH-UDAT, w-domain data, supervised*: In addition to OOD data, we use the manual-labels of the 300 turns of X data used above. This model is supervised w.r.t X.

The results of these experiments are listed in table 6. From the results, we can observe improvements when unlabeled (0.712 vs 0.702) or labeled (0.734 vs 0.702) target domain data is available.

7. Conclusions

We are interested in DA tagging of human-human conversations with the final goal of end-to-end training of task-oriented dialogue systems, so that we can generate system actions for a given dialogue context. In this work, we investigated multiple annotated human-machine conversation datasets, with differences in DA schema. We discussed manual and automatic approaches for aligning these different schema, and presented results on a target corpus of human-human dialogues. We demonstrated that without manually annotating any new human-human conversations, we achieve an F1 score of 57.7%, which requires at least 1.7K turns of manually annotated human-human dialogue data. We provided learning curves to present performance improvement with different amounts of manually and automatically labeled data which provides useful guidelines on the amount of data required for accurate DA tagging. In the presence of a new domain, we compared the performance of DA tagging using unsupervised, semi-supervised and supervised approaches. For these domains, we showed further improvements when unlabeled or labeled target domain data is available. As future work, we intend to further explore domain adaptation and use these annotated human-human conversations to train end-to-end task-oriented dialogue systems.

8. References

[1] J. L. Austin, *How to do things with words*. Oxford university press, 1975.

[2] S. Young, "Cued standard dialogue acts," *Report, Cambridge University Engineering Department, 14th October*, vol. 2007, 2007.

[3] H. Bunt, "The semantics of dialogue acts," in *Proceedings of the Ninth International Conference on Computational Semantics*. Association for Computational Linguistics, 2011, pp. 1–13.

[4] P. Shah, D. Hakkani-Tür, B. Liu, and G. Tur, "Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT)*, vol. 3, 2018, pp. 41–51.

[5] M. Gašić, F. Lefevre, F. Jurčíček, S. Keizer, F. Mairesse, B. Thomson, K. Yu, and S. Young, "Back-off action selection in summary space-based pomdp dialogue systems," in *IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 456–461.

[6] J. D. Williams, K. Asadi, and G. Zweig, "Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning," in *55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[7] B. Liu, G. Tur, D. Hakkani-Tür, P. Shah, and L. Heck, "Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT)*, 2018.

[8] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[9] B. Krause, M. Damonte, M. Dobre, D. Duma, J. Fainberg, F. Fancellu, E. Kahembwe, J. Cheng, and B. Webber, "Edina: Building an open domain socialbot with self-dialogues," *Alexa Prize Proceedings*, 2017.

[10] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *icassp*. IEEE, 1992, pp. 517–520.

[11] M. Henderson, B. Thomson, and J. D. Williams, "The second dialog state tracking challenge." in *SIGDIAL Conference*, 2014, pp. 263–272.

[12] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[13] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–1061.

[14] M. Zimmermann, "Joint segmentation and classification of dialog acts using conditional random fields," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[15] G. Ji and J. Bilmes, "Dialog act tagging using graphical models," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–33.

[16] J. Y. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," *arXiv preprint arXiv:1603.03827*, 2016.

[17] Y. Liu, K. Han, Z. Tan, and Y. Lei, "Using context information for dialog act classification in dnn framework," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2170–2178.

[18] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The icsi meeting recorder dialog act (mrda) corpus," in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, 2004.

[19] A. Margolis, K. Livescu, and M. Ostendorf, "Domain adaptation with unlabeled data for dialog act tagging," in *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 45–52.

[20] M. Henderson, B. Thomson, and J. D. Williams, "The second dialog state tracking challenge," in *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2014, pp. 263–272.

[21] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.