"HOW ROBUST R U?": EVALUATING TASK-ORIENTED DIALOGUE SYSTEMS ON SPOKEN CONVERSATIONS

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis,
Karthik Gopalakrishnan, Behnam Hedayatnia, Dilek Hakkani-Tür

Amazon Alexa AI, Sunnyvale, CA, USA

ABSTRACT

Most prior work in dialogue modeling has been on written conversations mostly because of existing data sets. However, written dialogues are not sufficient to fully capture the nature of spoken conversations as well as the potential speech recognition errors in practical spoken dialogue systems. This work presents a new benchmark on spoken task-oriented conversations, which is intended to study multi-domain dialogue state tracking and knowledge-grounded dialogue modeling. We report that the existing state-of-the-art models trained on written conversations are not performing well on our spoken data, as expected. Furthermore, we observe improvements in task performances when leveraging n-best speech recognition hypotheses such as by combining predictions based on individual hypotheses. Our data set enables speech-based benchmarking of task-oriented dialogue systems.

Index Terms— spoken dialogue systems, dialogue state tracking, knowledge-grounded dialogue generation

1. INTRODUCTION

Recently, more public data sets and benchmarks have become available for dialogue research on task-oriented conversations in various domains [1, 2, 3, 4]. However, most data sets include only written conversations collected by crowdsourcing via web interfaces, which differ from spoken conversations for the following reasons. First, there are differences between the style of spoken and written conversations, even for the same context, intention, and semantics. Second, spoken conversations tend to have extra noise from grammatical errors, disfluencies or barge-ins, which are rarely encountered when processing written texts. In addition, speech recognition errors bring about even more challenges for developing spoken dialogue systems in practice.

Figure 1 compares a written and a spoken conversation, and shows many differences in terms of wording and expressions between the two examples even for the same content. The spoken example includes disfluencies and speech recognition errors highlighted as underlined text. Moreover, no punctuation, capitalization, or sentence segmentation is available in raw speech recognizer outputs.

Written Conversation

User	I need a hotel in Fisherman's Wharf
Agent	Is there a particular price range you are looking for?
User	I'm looking in the expensive price range
Agent	The Suite at Fisherman's Wharf may work for you
User	Do you know how much the parking is?
Agent	It would cost 25 dollars per day.

Spoken Conversation

User	hi ummm i'm looking for a place at uhhh to stay at fisherman's wharf at a hotel in the expensive pressure engine
Agent	sure let me see ok so there is one called the suites at fisherman's wharf is that something that would be interesting to you
User	can you tell me how much parking coast
Agent	sure okay this hotel charges twenty five dollars per day

Fig. 1. Examples of written and spoken conversations

There have been extensive studies towards robust language understanding against spoken input, mostly for single-turn intent classification and slot filling tasks [5, 6, 7, 8, 9, 10, 11, 12, 13]. Nonetheless, the research communities have rarely addressed these issues on more contextual dialogue tasks including dialogue state tracking, dialogue policy learning, or end-to-end dialogue generation, which are as important as the single-turn understanding tasks in fully working dialogue systems. This is mainly due to the lack of rich, annotated spoken data for such multi-turn dialogue tasks.

To benchmark the robustness of conversational models on spoken conversations, this work introduces a new data set with spoken task-oriented dialogues. We release this data as the official validation set of our public benchmark challenge under DSTC10 [1] for the following two subtasks: 1) multi-domain dialogue state tracking [3] and 2) knowledge-grounded dialogue modeling [14]. The remainder of this paper presents the data and task details as well as the analysis result showing how state-of-the-art methods from existing written data perform on these benchmark tasks.

2. RELATED WORK

There has been a lot of studies towards improving the single-turn spoken language understanding (SLU) robustness to au-

---

[1]https://dstc10.dstc.community/

tomatic speech recognition (ASR) errors. The most common line of work has focused on utilizing multiple recognition hypotheses given from ASR systems in the form of word confusion networks [5, 6, 7, 8, 9] or word lattices [10, 11, 12, 13]. Recently, more proactive approaches have been explored for ASR error correction [15, 16] and data augmentation [17] as well.

The earlier dialog state tracking challenges (DSTCs) aimed to address the multi-turn dialogue problems in spoken conversations. DSTC2 [18] and DSTC3 [19] data sets included n-best ASR hypotheses as well as word confusions, which was intended for speech-oriented studies. However, the dialogue research community hasn't paid much attention to this aspect, due to the lack of critical ASR issues on these single-domain human-machine conversations that were restricted only to a small domain ontology. DSTC4 [20] and DSTC5 [21] targeted to extend it with multi-domain human-human conversations, but only manual transcriptions were included in the challenge data sets without ASR outputs.

On the other hand, many recent task-oriented dialogue data sets [1, 2, 3, 4] include written conversations collected by crowdsourcing with no consideration of speech-specific aspects in spoken dialogue systems. There have been studies focusing on the speech robustness issues on task-oriented [22] and open-domain conversations [23], but they were restricted to simulated ASR errors on top of written conversations.

3. DATA

To study speech-based task-oriented dialogue modeling, we collected spoken human-human dialogues about touristic information for San Francisco. Each session was collected by pairing two participants: one as a user and the other as an agent. We provided a set of specific goals to the user-side participant before each session. The agent-side participant had access to the domain database including both structured information and unstructured text snippets. We recorded 107 sessions, which are around 4 hours in total, and manually transcribed all the utterances. This data is released as the official validation set of our public benchmark challenge under DSTC10 Track 2.

Table 1 summarizes the data details in comparison with two other data sets. MultiWOZ 2.0 [3] and its variants [24, 25] include crowd-sourced written conversations about seven different domains including hotel, restaurant, attraction, train, taxi, hospital, and police station in Cambridge, UK. Following the MultiWOZ data collection set-up, we recently released new written dialogues as a part of the official test set for the DSTC9 Track 1 [26]. This data was collected for a new locale, San Francisco, for three target domains: hotel, restaurant, and attraction, but with almost three times more entities than the MultiWOZ ontology entries. In addition, this data includes the turns grounded on the knowledge snippets from the FAQ list for the entities. All these infrastructure and

domain ontology for this written data were re-used for our spoken data collection. The difference is that the DSTC10 data came from the recorded conversations instead of the written texts from crowd-sourcing.

To benchmark the robustness of models in practical spoken dialogues systems, we test on the ASR output instead of manual transcript for each user turn. In terms of the ASR model, we took the wav2vec 2.0 [27] model pre-trained on 960 hours of Librispeech [28] and fine-tuned it with 10% of our data. Then, we run 10-best decoding with an external language model built with KenLM [29] on all the written texts from MultiWOZ and DSTC9 data sets. This ASR pipeline finally achieved a 24.09% WER at 1-best and 21.89% oracle WER at 10-best hypotheses on the user utterances of the remaining 90% of our data.

4. TASKS

This section describes two benchmark tasks that we propose in this work. As shown in Figure 3, we decouple between turns that could be handled by conventional task-oriented conversational models with no extra knowledge and turns that require external knowledge resources, following the architecture in [14]. In the first API/DB-based pipeline, we focus only on dialogue state tracking as the first target task. For the other knowledge access branch, our task 2 includes all three subtasks: 1) Knowledge-seeking Turn Detection, 2) Knowledge Selection, and 3) Knowledge-grounded Response Generation introduced in [14].

4.1. Task 1: Multi-domain Dialogue State Tracking

Dialogue state tracking (DST) aims to estimate the system's belief states after each interaction with the user, which is a key problem in task-oriented conversational modeling. The belief states are defined to represent the latest user goals in a dialogue context from the beginning to the target user turn of a given conversation.

In this benchmark, we address the multi-domain DST task on human-human conversations, which have been actively explored by the dialogue research community mainly with MultiWOZ and its variants [3, 24, 25]. Following previous work, we also represent the user goals as a set of slot-value pairs defined for each domain and take the slot-level value prediction performances and the joint goal accuracy [18] as the evaluation metrics. Figure 2 presents an example conversation with the ground-truth DST annotations for the first two user turns.

Our task differs from most previous DST benchmarks in the following two aspects. First, we focus on the DST performances on spoken conversations rather than written ones. The latter has been widely used by previous DST studies because of its cost-efficiency in large scale data collection. We believe that those written data sets are not enough to fully reflect the actual human behaviors for spoken conversations.

Table 1. Comparisons of the data sets.

Data	Split	Locale	Modality	Dialogues		Domain Ontology		
				# sessions	# turns	# domains	# entities	# snippets
MultiWOZ 2.0	all	Cambridge	written	8,438	113,556	7	289	-
DSTC9: Track 1	test	San Francisco	written	903	8,501	3	855	15,086
DSTC10: Track 2	val	San Francisco	spoken	107	2,292	3	855	15,086

Speaker	Utterance	Task	Ground-truth Annotations
User	hi ummm i'm looking for a place at uhhh to stay at fisherman's wharf at a hotel in the moderate pressure engine	Task 1	hotel-area: fisherman's wharf hotel-type: hotel hotel-pricerange: moderate
Agent	sure let me see what can i find for you ok unfortunately we're not showing any result is there any specification that i can change to possibly find you something		
User	is there wholesale in the expensive pressure engine student	Task 1	hotel-area: fisherman's wharf hotel-type: hotel hotel-pricerange: expensive
Agent	sure let me see ok so there is one called the suites at fisherman's wharf is that something that would be interesting to you		
User	can you tell me how much parking coast	Task 2	Q: What is the parking cost at the Suites at Fisherman's Wharf? - A: Parking costs $25 per day.
Agent	sure okay this hotel charges twenty five dollars per day		

Fig. 2. An example conversation with the ground-truth labels for both tasks.

So we propose to take DST models trained on existing written data, evaluate on our new collection of spoken conversations, and eventually try to improve the model performance in face of a mismatch between training and test data sets.

In addition, our new data uses the ASR output instead of manual transcripts for the user turns, as described in Section 3. The goal is to evaluate how robust each DST model is against ASR errors. Although ASR errors are expected to be a very critical issue in developing spoken dialogue systems in practice, it hasn't been studied for the DST task.

4.2. Task 2: Knowledge-grounded Dialogue Modeling

Recently, we introduced a new benchmark on task-oriented conversational modeling with unstructured knowledge access [14], which aims to incorporate external unstructured knowledge into the end-to-end dialogue response generation problem. As shown in Figure 3, this task focuses on the knowledge access branch, including the following three sub-tasks:

- **Knowledge-seeking Turn Detection** decides whether to trigger the knowledge access branch for a given utterance and dialogue history

- **Knowledge Selection** selects proper knowledge snippets from the domain knowledge base for the knowledge seeking turn

- **Knowledge-grounded Response Generation** generates a system response given a triple of input utterance,

dialog context, and the selected knowledge snippet

Figure 2 shows an example knowledge-seeking turn along with its ground-truth knowledge snippet and a reference response. We organized a challenge track on this task under DSTC9 [26, 30], which had more than 100 submissions from 24 teams in total.

In this work, we propose to extend this DSTC9 track by replacing written conversations with spoken ones, as in the first task (Section 4.1). The issue between written and spoken conversations has been partially discussed in our DSTC9 track with a spoken subset in the test data [26]. But that was only on manual transcripts, while the new data for this work includes the ASR outputs for the user turns.

5. BASELINE MODELS

To investigate the existing model behaviors on our spoken data, we took state-of-the-art models on written data for both tasks as baselines.

5.1. Task 1 Baseline

We use TriPy [2] [31] as a baseline for the DST task. It is based on a fine-tuned BERT [32] on the DST objectives with copy mechanisms from three different sources: user utterance, system utterance, and previous dialogue states. This model achieved 55.30% in joint goal accuracy on MultiWOZ 2.1, which is also used as the DST baseline for DialoGLUE [33].
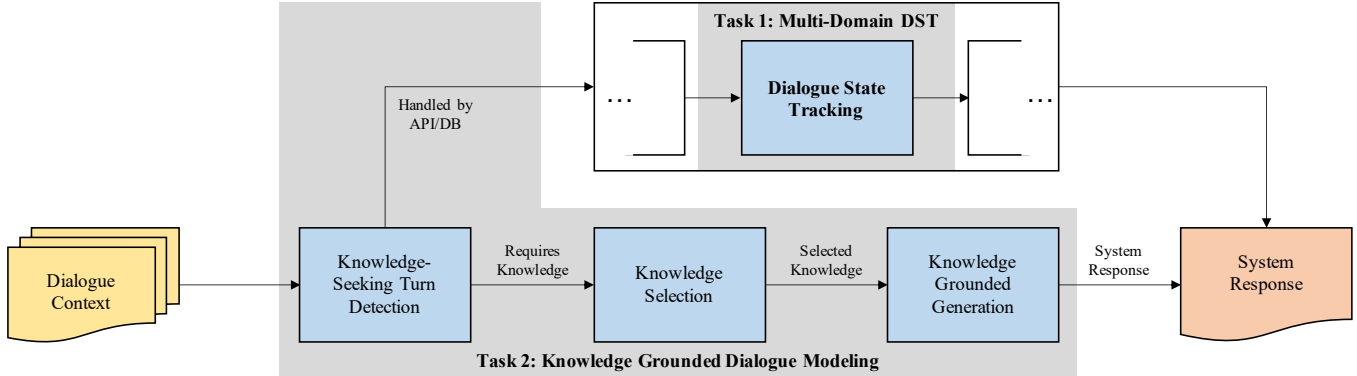
---

[2]https://gitlab.cs.uni-duesseldorf.de/general/dsml/trippy-public

Fig. 3. An overview of the benchmark tasks: multi-domain dialogue state tracking and knowledge-grounded dialogue modeling.

5.2. Task 2 Baselines

For task 2, we take two baselines: one is the official baseline of DSTC9 [3] [26] and the other is Knover [4] [34] from the DSTC9 track winner. The official DSTC9 baseline uses the fine-tuned GPT-2 [35] for each sub-task. For Knover, we use the model for their submission #0. It has the following three updates from the baseline: 1) replacing GPT-2 with PLATO-2 [36]; 2) multi-scale negative sampling for knowledge selection; and 3) response generation with beam search instead of nucleus sampling. Note that the DSTC9 winning entry was a further enhanced version from this model by schema guided turn detection and model ensemble, however, that model is not publicly available.

6. EVALUATION

6.1. Task 1 Results

Table 2 compares the TripPy baseline performances on the DSTC10 and MultiWOZ 2.1 data sets. Note that all the evaluations on the DSTC10 data were done on the official task 1 validation set including 936 target turns. It achieves extremely low performance on the spoken data in joint goal accuracy and also a significantly worse score in slot-level accuracy compared to that on MultiWOZ. Considering that there are too many false negatives from the model predictions, we report more detailed results in precision, recall and F1 scores for *none* predictions and other values from the system separately. The results show that the baseline model is not working well on our test data, especially in value predictions.

To utilize n-best ASR results, we first took every combination of the ASR hypotheses of the user turns and calculated the sum of the turn-level language model scores to select the top-n context-level hypotheses among them. Then, we run the TripPy model for each context separately and finally aggregated all the predicted values by taking the union of the
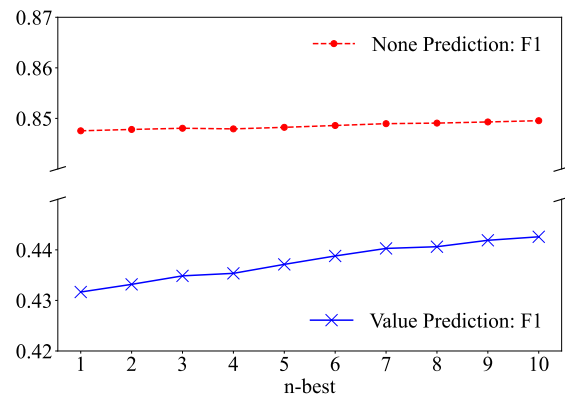

Fig. 4. Value and none prediction performances with n-best ASR hypotheses

n results for each slot. Figure 4 presents the value and none prediction performances with different n values. The use of multiple ASR hypotheses helped to achieve slightly higher scores than using only the top-1 results, but the differences were not significant.

To further investigate how much ASR errors affect DST performance, we run the model on two additional configurations: one with manual transcripts with no ASR errors, and the other with the 1-best ASR results not only for the user turns, but also for the agent turns. Figure 5 compares the model performance on those three different settings and shows the impact of the ASR errors on all three metrics. In particular, the value prediction score in F1 drops 5% and 6%, when the ASR errors are introduced in the user turns and both turns, respectively. But even with no ASR error, the value prediction score at an F1 of 48.61% is much lower than 90.80% on the original MultiWOZ. This indicates that the other aspects including different modalities (spoken vs written) and locales (Cambridge vs San Francisco) between training and test data are as critical as the ASR errors, and

Table 2. Multi-domain dialogue state tracking results by the TripPy baseline on the DSTC10 validation and MultiWOZ 2.1 test data sets.

Data	Joint Goal Accuracy	Slot Accuracy	Value Prediction			None Prediction		
			Precision	Recall	F1	Precision	Recall	F1
DSTC10 (1-best)	0.0043	0.7208	0.5977	0.3378	0.4317	0.7473	0.9788	0.8475
MultiWOZ	0.5427	0.9738	0.9056	0.9103	0.9080	0.9802	0.9934	0.9868

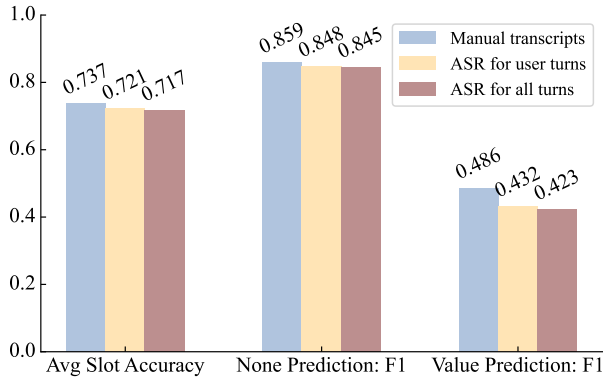**Fig. 5**. Comparisons of the task 1 baseline model performances on manual transcripts and ASR results

Fig. 6. Comparisons of the task 2 baseline performances on manual transcripts and ASR results

Table 3. Statistics of the task 2 data sets. † indicates the spoken data with manual transcriptions only.

Dataset	Split	Locale	Modality	# dialogs	total # instances
DSTC9	Test	CAM	written	977	2,084
	Test	SF	written	900	1,834
	Test	SF	spoken†	107	263
DSTC10	Valid	SF	spoken	107	263

they all result in performance degradation.

6.2. Task 2 Results

This section reports the task 2 baseline performances on both the DSTC9 test set and DSTC10 validation set (data statistics are shown in Table 3). We treat each instance including a target turn and its dialogue context independently from the others in the data and evaluated the prediction results at the point of the final target turn only. Table 4 shows the knowledge-seeking turn detection and knowledge selection performance of the baseline models. For both tasks, the models achieve significantly worse performance on the spoken DSTC10 data compared to the whole DSTC9 test set. We notice there are many false negatives in the turn detection task, which was the key factor of the performance degradation due to the low recalls. Comparing these two baseline models on the spoken data, the DSTC9 baseline method achieves slightly better scores for the detection task, while Knover is better for two
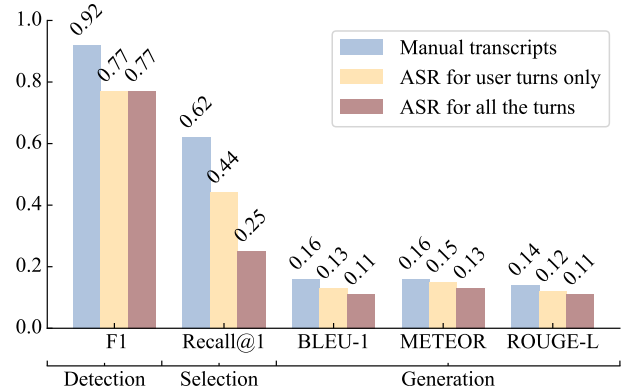
of the three selection metrics (MRR@5 and Recall@1). The response generation results in Table 5 also show the same pattern, with lower scores than the ones on the DSTC9 test set. While Knover consistently outperformed the baseline on the DSTC9 test set for all the metrics, there was no dominant result by either system on the DSTC10 validation data.

Figure 6 shows the baseline model performance in different configurations: with manual transcripts, and 1-best ASR outputs only for user turns or for all the turns. We can see that the ASR errors have a much bigger impact on these tasks compared to the same experimental results on the DST task. There is a larger degradation in the detection and selection performances in face of ASR errors.

To improve the robustness of the baseline model against the ASR errors, especially for the low recall on the detection task, we developed a simple heuristic to utilize the n-best ASR hypotheses. First, we run the model on each of the n-best hypotheses of the target turn. We still kept the 1-best results for the other previous turns in the dialogue context. Then, if there is at least one positive prediction from the n runs, we treat the instance as a knowledge-seeking turn. Figure 7 shows that using n-best hypotheses with this simple ensemble method helps to improve the detection performance.

Finally, we conducted a human evaluation with the following crowd-sourcing tasks introduced in [26]:

- Accuracy: This task asks crowd workers to score the accuracy of a system output based on the provided ref-

Table 4. Knowledge-seeking turn detection and knowledge selection results by the DSTC9 and Knover baselines on the DSTC10 validation and DSTC9 test data sets.

| | | Detection | | | Selection | | |
Data	Model	Precision	Recall	F1	MRR@5	Recall@1	Recall@5
DSTC10	Baseline	**0.9851**	**0.6346**	**0.7719**	0.5405	0.4444	0.6901
	Knover	0.9701	0.6250	0.7602	**0.5782**	**0.5263**	**0.6667**
DSTC9	Baseline	0.9933	0.9021	0.9455	0.7263	0.6201	0.8772
	Knover	0.9941	0.9430	0.9679	0.9181	0.8870	0.9554

Table 5. Response generation results in automated metrics on the DSTC10 validation and DSTC9 test data sets.

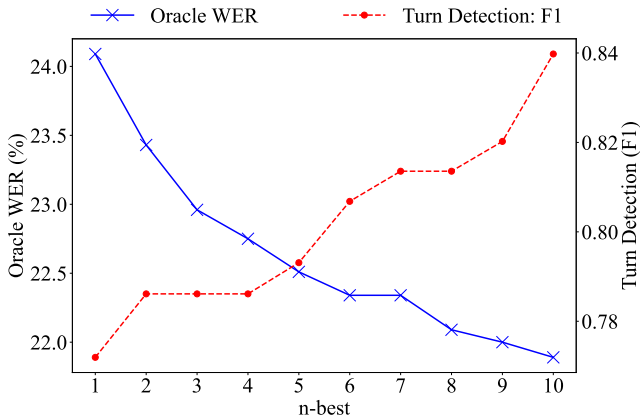| Data	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
DSTC10	Baseline	**0.1338**	0.0539	0.0235	**0.0128**	0.1457	**0.1650**	0.0443	0.1181
	Knover	0.1301	**0.0635**	**0.0330**	0.0111	**0.1592**	0.1578	**0.0536**	**0.1245**
DSTC9	Baseline	0.3031	0.1732	0.1005	0.0655	0.2983	0.3386	0.1364	0.3039
	Knover	0.3726	0.2402	0.1556	0.1064	0.3802	0.4103	0.1936	0.3665

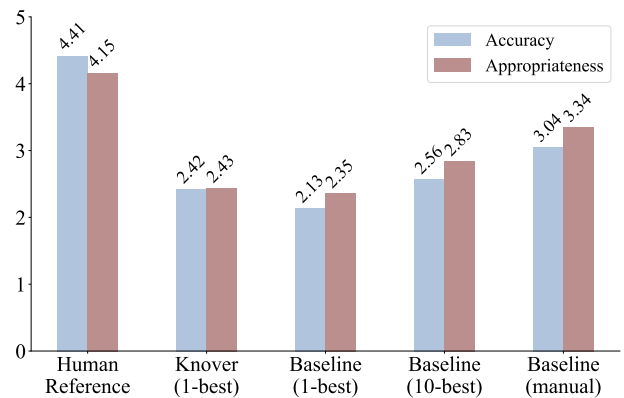**Fig. 7**. Speech recognition and knowledge-seeking turn detection performances with n-best ASR hypotheses

Fig. 8. Human evaluation results

erence knowledge on a scale of 1-5.

- Appropriateness: This task asks crowd workers to score how well a system output is naturally connected to a given conversation on a scale of 1-5.

Figure 8 compares the human evaluation results of five different configurations in Accuracy and Appropriateness. As expected, there were still significant differences between all the model outputs and the ground-truth human responses. Comparing between two base models, Knover was better in both metrics than the DSTC9 baseline, which is consistent with the official DSTC9 results on the written test set. Again our ensemble heuristic with n-best ASR hypotheses helped to boost the baseline performance on the human evaluation results. Furthermore, the gap between the results on ASR outputs and manual transcripts implies more room for potential improvements with better methods.

7. CONCLUSIONS

This paper introduced a new data set on multi-domain dialogue state tracking and knowledge-grounded dialogue modeling tasks, aiming towards more robust multi-turn dialogue processing on spoken conversations. From the evaluations, we observed that the baseline models built on the existing written data were not performing well on the new spoken data for both tasks. These results show the importance of more speech-oriented studies to improve the robustness of spoken dialogue systems.

To support research in this direction, we organize a public benchmark challenge under DSTC10 and release the data introduced in this work as the official validation set for the challenge participants. Furthermore, we plan to collect more spoken dialogues and release them as the main challenge test set. We believe such community efforts will make advancement in the state-of-the-art in spoken dialogue processing.

8. REFERENCES

[1] Layla El Asri, Hannes Schulz, Shikhar Kr Sarma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman, "Frames: a corpus for adding memory to goal-oriented dialogue systems," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 207–219.

[2] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young, "A network-based end-to-end trainable task-oriented dialogue system," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 438–449.

[3] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic, "Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 5016–5026.

[4] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan, "Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 8689–8696.

[5] Gokhan Tur, Jerry Wright, Allen Gorin, Giuseppe Riccardi, and Dilek Hakkani-Tür, "Improving spoken language understanding using word confusion networks," in *Seventh International Conference on Spoken Language Processing*, 2002.

[6] Dilek Hakkani-Tür, Frédéric Béchet, Giuseppe Riccardi, and Gokhan Tur, "Beyond asr 1-best: Using word confusion networks in spoken language understanding," *Computer Speech & Language*, vol. 20, no. 4, pp. 495–514, 2006.

[7] Matthew Henderson, Milica Gašić, Blaise Thomson, Pirros Tsiakoulis, Kai Yu, and Steve Young, "Discriminative spoken language understanding using word confusion networks," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 176–181.

[8] Gokhan Tur, Anoop Deoras, and Dilek Hakkani-Tür, "Semantic parsing using word confusion networks with conditional random fields.," in *Interspeech 2013*. Citeseer, 2013, pp. 2579–2583.

[9] Ryo Masumura, Yusuke Ijima, Taichi Asami, Hirokazu Masataki, and Ryuichiro Higashinaka, "Neural confnet classification: Fully neural network based spoken utterance classification using word confusion networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6039–6043.

[10] Faisal Ladhak, Ankur Gandhe, Markus Dreyer, Lambert Mathias, Ariya Rastrow, and Björn Hoffmeister, "Latticernn: Recurrent neural networks over lattices," in *Interspeech 2016*, 2016, pp. 695–699.

[11] Leonid Velikovich, "Semantic model for fast tagging of word lattices," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 398–405.

[12] Chao-Wei Huang and Yun-Nung Chen, "Adapting pre-trained transformer to lattices for spoken language understanding," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 845–852.

[13] Chao-Wei Huang and Yun-Nung Chen, "Learning spoken language representations with neural lattice language modeling," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020, pp. 3764–3769.

[14] Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur, "Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2020, pp. 278–289.

[15] Yue Weng, Sai Sumanth Miryala, Chandra Khatri, Runze Wang, Huaixiu Zheng, Piero Molino, Mahdi Namazifar, Alexandros Papangelis, Hugh Williams, Franziska Bell, and Gokhan Tur, "Joint contextual modeling for asr correction and language understanding," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6349–6353.

[16] Mahdi Namazifar, John Malik, Li Erran Li, Gokhan Tur, and Dilek Hakkani Tür, "Correcting automated and manual speech transcription errors using warped language models," *arXiv preprint arXiv:2103.14580*, 2021.

[17] Longshaokan Wang, Maryam Fazel-Zarandi, Aditya Tiwari, Spyros Matsoukas, and Lazaros Polymenakos, "Data augmentation for training dialog models robust to speech recognition errors," in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 2020, pp. 63–70.

[18] Matthew Henderson, Blaise Thomson, and Jason 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)*, Philadelphia, PA, U.S.A., June 2014, pp. 263–272, Association for Computational Linguistics.

[19] Matthew Henderson, Blaise Thomson, and Jason D. Williams, "The third dialog state tracking challenge," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 324–329.

[20] Seokhwan Kim, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, and Matthew Henderson, "The fourth dialog state tracking challenge," in *Dialogues with Social Robots*, pp. 435–449. Springer, 2017.

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

[22] Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao, "Raddle: An evaluation benchmark and analysis platform for robust task-oriented dialog systems," *arXiv preprint arXiv:2012.14666*, 2020.

[23] Karthik Gopalakrishnan, Behnam Hedayatnia, Longshaokan Wang, Yang Liu, and Dilek Hakkani-Tur, "Are neural open-domain dialog systems robust to speech recognition errors in the dialog history? an empirical study," *arXiv preprint arXiv:2008.07683*, 2020.

[24] Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur, "Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines," *arXiv preprint arXiv:1907.01669*, 2019.

[25] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen, "Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines," in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*, 2020, pp. 109–117.

[26] Seokhwan Kim, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, and Dilek Hakkani-Tur, "Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access track in dstc9," 2021.

[27] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[29] Kenneth Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[30] Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, Maxine Eskenazi, Ahmad Beirami, Eunjoon, Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba, "Overview of the ninth dialog system technology challenge: Dstc9," 2020.

[31] Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic, "Trippy: A triple copy strategy for value independent neural dialog state tracking," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2020, pp. 35–44.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[33] Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur, "Dialoglue: A natural language understanding benchmark for task-oriented dialogue," 2020.

[34] Huang He, Hua Lu, Siqi Bao, Fan Wang, Hua Wu, Zhengyu Niu, and Haifeng Wang, "Learning to select external knowledge with multi-scale negative sampling," 2021.

[35] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language models are unsupervised multitask learners," 2019.

[36] Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu, "Plato-2: Towards building an open-domain chatbot via curriculum learning," 2021.