INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Contrastive Learning Based ASR Robust Knowledge Selection For Spoken Dialogue System

Zhiyuan Zhu[1], *Yusheng Liao*[1], *Yu Wang*[1,2,*], *Yunfeng Guan*[1,*]

[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
[2]Shanghai AI Laboratory

{zzysjtu_iwct, liao20160907, yuwangsjtu, yfguan69}@sjtu.edu.cn

Abstract

The construction of knowledge-based, task-oriented systems for spoken conversations is a challenging task. Given the spoken dialogue history information, a knowledge selection model selects the appropriate knowledge snippet from an unstructured knowledge base. However, the performance of this model is sensitive to automatic speech recognizer (ASR) recognition errors. To address this problem, we propose a method called CLKS, which develops a knowledge selection model that is robust to ASR recognition errors. This approach involves: 1) To leverage a wide range of information from various ASR outputs, we employ the self-attention mechanism to aggregate the representation of the N-best hypotheses of the dialogue history. 2) We use the written dialogue representation to guide the aggregated spoken dialogue representation to select the correct knowledge candidate through contrastive learning. Experimental results on the DSTC10 dataset demonstrate the effectiveness of our method.

Index Terms: Contrastive Learning, N-best Aggregation, Spoken Dialogue System, Knowledge Retrieval.

1. Introduction

The speech-based task-oriented dialogue systems are popular for assisting users with specific tasks through natural and information-rich conversations. Unlike written language-based dialogue systems [1, 2, 3, 4], these systems rely heavily on the performance of automatic speech recognition (ASR) systems and an external knowledge base to build a meaningful conversation [5]. Knowledge-based task-oriented spoken dialogue systems first use an ASR system to transcribe speech into text, and then a dialogue system to generate an appropriate response by combining external knowledge and dialogue information.

Most publicly available task-oriented systems mainly focus on written conversations [6, 7, 8]. However, the characteristics of written conversations are different from that of practical spoken ones, which contain speaker disfluency and interruptions phenomena. A dialogue system trained on a written conversation corpus but deployed in a spoken conversation scenario suffers from enormous performance degradation due to the mismatch between the written and ASR output transcripts. One of the main reasons is that recognition errors are inevitable in the ASR output. Although some previous works have attempted to develop ASR robust knowledge-based spoken dialogue systems [5, 9] to mitigate the effects of ASR errors, recognition errors are still high for the knowledge entities which often contain rare words [10, 11] and there is not much training data for these words, thereby affecting the accuracy of knowledge selection models. Thus, building an ASR error robust knowledge selection system is an essential but challenging task.

One of the most common lines of works to reduce the impact of recognition errors focuses on utilizing the N-best hypotheses of ASR instead of the one-best hypothesis to recover from ASR recognition errors [12, 13, 14, 15], or using ASR error detection and correction [16, 17, 18] methods based on a pre-trained language model (LM) to modify the recognition output. However, these works ignore the link between the written and corresponding spoken transcripts, which is crucial for improving the robustness of the model to ASR errors. Some works use knowledge from the written context to create a robust representation of the spoken context [19, 20, 21, 22], but rare works have been investigated in spoken conversation tasks where user intent is more complicated. Other works focus on using sophisticated sampling methods to improve knowledge selection [23, 24] performance. However, these methods are sensitive to the expansion of the knowledge base.

Recently, the Track 2 [25] proposed in the Tenth Dialog System Technology Challenge (DSTC10) aims to incorporate unstructured external knowledge into a spoken task-oriented dialogue system and attracted wide attention. This paper focuses mainly on the second subtask of DSTC10 Track 2. A contrastive learning-based robust knowledge selection method (CLKS) using the N-best hypotheses from ASR output is proposed to narrow the gap between the written and spoken representations in the semantic space and improve the robustness of the model to ASR errors. Experimental results and ablation studies conducted on the DSTC10 Track 2 dataset validate the effectiveness and robustness of our proposed method, which outperforms the DSTC10 Track 2 baseline system by about 9.7% compared to R@1. In this paper, we contribute to the knowledge selection task of the spoken dialogue system from two aspects:

- To leverage the information from the various N-best hypotheses output by ASR, we aggregate the representation of the N-best hypotheses of the spoken dialogue history using a sentence-level self-attention mechanism.
- We use the sentence-level representation of the written dialogue history to guide the aggregated representation of the spoken dialogue history to select the correct knowledge from the unstructured knowledge base through contrastive learning, which further improves the robustness of the model.

2. Methodology

To formalize the task of knowledge selection in spoken task-oriented conversational systems, we define the written dialogue

* Corresponding authors

This work was supported by National Key R&D Program of China (No.2022ZD0162101), Shanghai Science and Technology Committee (No.21511101100) and Shanghai Key Laboratory of Digital Media Processing and Transmissions (STCSM 22DZ2229005)

10.21437/Interspeech.2023-644

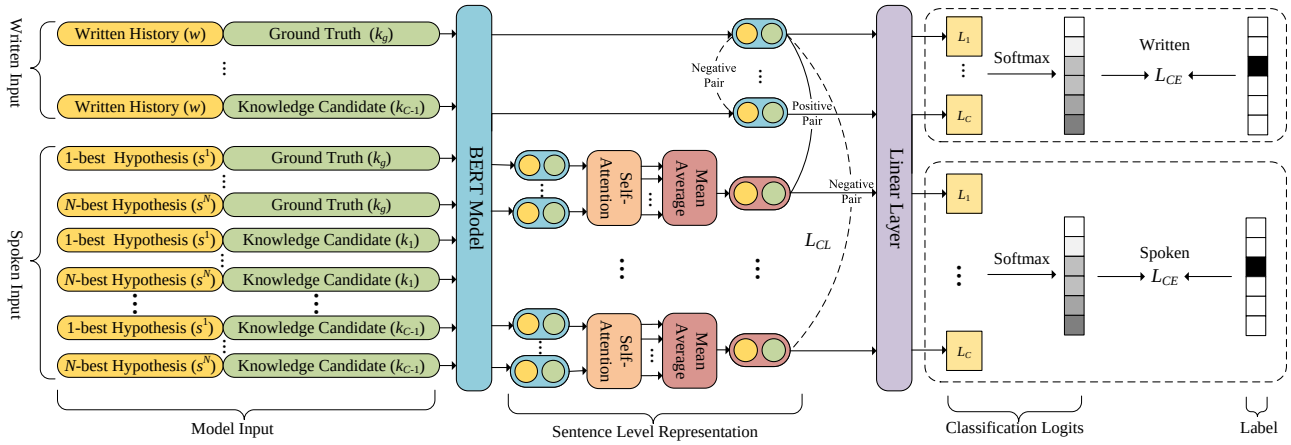

Figure 1: *We aggregate ASR N-best hypotheses by a self-attention mechanism to obtain a spoken domain representation. We use cross-entropy and contrastive loss to fine-tune the model with the sentence-level written and spoken dialogue representation.*

context of the t-th utterance as $W_t = \{w_{t-u+1}, \cdots, w_{t-1}, w_t\}$ and the knowledge snippets in the external knowledge base as $K = \{k_1, \cdots, k_M\}$, where u is the size of the dialogue history window and M is the size of the knowledge base. Following [5], we randomly sample $C-1$ knowledge snippets from the knowledge base as negative candidates. In conjunction with the ground truth knowledge k_g, we use a total of C knowledge candidates for further training of the knowledge selection model. We use the TTS-ASR pipeline to generate the dialogue context hypotheses $S_t^i = \{s_{t-u+1}^i, \cdots, s_{t-1}^i, s_t^i\}$ corresponding to the spoken version transcripts of W_t, where $i = 1, 2, ..., N$ represents the i-th best hypothesis of the ASR outputs. As shown in Figure 1, the written and ASR transcripts of dialogue history are concatenated with each knowledge snippet as the written and spoken input of the model, respectively.

In summary, by augmenting the dialogue data with the TTS-ASR pipeline and concatenating dialogue history with knowledge candidates, each written domain training sample in the original dataset is augmented from C to $(N+1)\times C$ training samples for model training.

2.1. Dialogue Data Augmentation

To bridge the gap between the manual transcripts and the spoken conversation contexts, we utilize a TTS-ASR pipeline to obtain a spoken version of the conversation transcripts $\{S_t^i\}, i = 1, \cdots, N$ when given the corresponding written one W_t.

First, We synthesize the original DSTC9 Track1 [11] training set into speech using a commercial text-to-speech system[1]. Then the synthesized speech is transcribed into text by a wave2vec 2.0-based Connectionist Temporal Classification (CTC) acoustic model [26] trained on 960 hours [27] of Librispeech dataset and an external language model built using KenLM [28]. We train this external language model using the DSTC9 Track1 training set and MultiWoz2.1 [29] dataset. This ASR pipeline finally achieved an 18.89% WER on the user utterances in the training data. Some examples of spoken conversation utterances generated by our ASR pipeline are shown in Table 1. The ASR recognition errors are marked with bold italics. It can be seen that the rare words that appear in the hotel names can easily be misidentified. On the other hand, the punctuation in the original utterance is omitted. These ASR recognition noises can confuse the downstream task and affect the overall system performance.

[1]https://cloud.google.com/text-to-speech

Table 1: *Comparison between written and spoken transcripts.*

Written Transcripts	Spoken Transcripts
Can I get the address of the **Gonville**, please?	can i get the address of the ***gonvil*** please
I need to be in **Broxbourne** by 08:00.	i need to be in ***brokborn*** by eight o'clock
Can you give me the **postcode** and star rating for **Lovell** Lodge and tell me if they have wifi?	can you give me the ***postcoat*** and star rating for ***love*** lodge and tell me if they have wifi

2.2. ASR N-best Hypotheses Aggregation

As shown in Figure 1, to improve the robustness of the model to ASR errors, we aggregate the information of the N-best hypotheses based on a self-attention mechanism.

Specifically, we first concatenate the written and N-best ASR hypotheses transcripts of the same dialogue history with C different knowledge candidates as model input, which can be written as written history-knowledge pairs $\{[W_t, k_1], \cdots, [W_t, k_C]\}$ and spoken history-knowledge pairs $\{[S_t^i, k_1], \cdots, [S_t^i, k_C]\}_{i\in 1,\cdots,N}$. Then, these two types of history-knowledge pairs were passed through BERT [30] to obtain the last hidden states of each input token. Finally, feature extraction is performed by applying average pooling on the last hidden states to obtain the sentence-level representation of the history-knowledge pair, which can be formulated as $H_w = [\mathbf{h}_{w,1}, \mathbf{h}_{w,2}, \cdots, \mathbf{h}_{w,C}] \in \mathbb{R}^{d\times C}$ and $H_s^i = [\mathbf{h}_{s,1}^i, \mathbf{h}_{s,2}^i, \cdots, \mathbf{h}_{s,C}^i] \in \mathbb{R}^{d\times C}$, $i \in 1, \cdots, N$, where d is the dimension of sentence level representation vector.

After feature extraction, the spoken sentence representation matrix $\tilde{H}_s^j = [\mathbf{h}_{s,j}^1, \mathbf{h}_{s,j}^2, \cdots, \mathbf{h}_{s,j}^N]_{j=1,\cdots,C} \in \mathbb{R}^{d\times N}$ that collected according to the same knowledge candidate k_j is aggregated by averaging the representation output using a self-attention mechanism, which can be formulated as:

$$\hat{\mathbf{h}}_{s,j} = \frac{1}{N}\,\mathrm{softmax}\left(\frac{(Q\tilde{H}_s^j)(K\tilde{H}_s^j)^T}{\sqrt{d}}\right)V\tilde{H}_s^j \times \overrightarrow{\mathbf{1}}, \quad (1)$$

where $\overrightarrow{\mathbf{1}} \in \mathbb{R}^{N\times 1}$ is all-ones vector and $Q, K, V \in \mathbb{R}^{d\times d}$ are the learnable parameters of the model. The aggregated representation of the N-best hypotheses of C history-knowledge pairs at the level of spoken transcripts can be written as $\hat{H}_s = [\hat{\mathbf{h}}_{s,1}, \hat{\mathbf{h}}_{s,2}, \cdots, \hat{\mathbf{h}}_{s,C}]$. We call this step N-best aggregation.

The written and spoken sentence level representation H_w and $\hat{H}_s$ are passed through linear layer $W \in \mathbb{R}^{1\times d}$ to obtain classification distributions $p_w = [p_{w,1}, p_{w,2}, \cdots, p_{w,C}] = \mathrm{softmax}(WH_w)$ and $p_s = [p_{s,1}, p_{s,2}, \cdots, p_{s,C}] = \mathrm{softmax}(W\hat{H}_s)$. We use a cross-entropy loss on the classification logits to guide the network to choose the ground truth knowledge in C knowledge candidates, which can be written as:

$$\mathcal{L}^W_{CE} + \mathcal{L}^S_{CE} = -\log(p_{w,g}) - \log(p_{s,g}), \tag{2}$$

where g is the index of ground truth knowledge.

2.3. Supervised Contrastive Learning

To improve the knowledge selection capacity of our model, we propose a contrastive learning-based method to fine-tune the BERT model to learn the joint representation of the spoken dialogue history and the knowledge snippets. In contrastive learning formulation, we consider the pair of written and spoken sentence representation $(\mathbf{h}_{w,i}, \hat{\mathbf{h}}_{s,j})$ as a positive pair only if i and j are both equal to the index of ground truth knowledge g. We construct negative samples of representation $\mathbf{h}_{w,g}$ from two parts, including in-domain (i.e. written) negative samples $B^{in}_w = \{\mathbf{h}_{w,i}\}_{i\neq g}$ and out-domain (i.e. spoken) negative samples $B^{out}_w = \{\hat{\mathbf{h}}_{s,j}\}_{j\neq g}$. The final negative samples for $\mathbf{h}_{w,g}$ is $B_w = B^{in}_w \bigcup B^{out}_w$. We can also construct negative samples $B_s = B^{in}_s \bigcup B^{out}_s$ for $\hat{\mathbf{h}}_{s,g}$. The supervised contrastive learning loss can thus be written as Equation 3, where $s(\cdot,\cdot)$ is a cosine similarity function and τ is a temperature hyper-parameter, the whole process is illustrated in Figure 1.

$$\begin{aligned}\mathcal{L}_{CL} = -\frac{1}{2}\Big(&\log\frac{e^{s(\mathbf{h}_{w,g},\hat{\mathbf{h}}_{s,g})/\tau}}{\sum_{\mathbf{h}\in B_w} e^{s(\mathbf{h}_{w,g},\mathbf{h})/\tau}} \\ &+ \log\frac{e^{s(\hat{\mathbf{h}}_{s,g},\mathbf{h}_{w,g})/\tau}}{\sum_{\mathbf{h}\in B_s} e^{s(\hat{\mathbf{h}}_{s,g},\mathbf{h})/\tau}}\Big),\end{aligned} \tag{3}$$

In this way, we leverage the 'clean' sentence-level written dialogue history representation to guide the aggregated spoken domain dialogue history representation to choose the correct knowledge candidate. As we need to fine-tune the model shown in Figure 1, the final training loss can be written as follows:

$$\mathcal{L}_{ft} = \mathcal{L}^W_{CE} + \mathcal{L}^S_{CE} + \mathcal{L}_{CL}. \tag{4}$$

During inference, we use only spoken dialogue history as model input to select knowledge from the knowledge base using the fine-tuned model's representation extraction capacity.

3. Experiments

3.1. Datasets

We use the DSTC10 Track2 dataset and an augmented spoken version of the DSTC9 Track1 dataset. The statistics of the dataset are shown in Table 3. The training set is a spoken version of the DSTC9 training set, which is generated by our data augmentation method described in subsection 2.1. The DSTC10 dataset differs from the spoken DSTC9 dataset in that the DSTC10 challenge only provides validation and test sets, and the texts are from the output of an ASR system, the details of this ASR system can be found in [5]. The knowledge base consists of question-answer pairs, which were collected from frequently asked questions (FAQ) pages and cover four different domains: Hotel, Restaurant, Train, and Taxi. The document of hotels and restaurants are further divided into entities. The number of snippets in the external knowledge base of the DSTC10 dataset is about four times that of the original DSTC9 dataset, totaling 10586 snippets.

3.2. Experimental Setting

In this paper, we focus on the knowledge selection sub-task of DSTC10 Track2. To verify the effectiveness of our approach, we compare CLKS with the following three baseline models. For a fair comparison, we use our augmented spoken dialogue data as training input for each baseline method[2].

- DSTC10: DSTC10 Track2 baseline[3] is an official baseline of the DSTC10 challenge. We replace the original loss function with Equation 2 to utilize both written and spoken transcripts.
- Knover: Knover [24] utilizes a multi-scale negatives sample to strengthen the ability of fine-grained relevance estimation. To make a fair comparison with our method, we replace the pre-trained model used in the original paper with BERT and use the loss function of Equation 2.
- RADGE: RADGE [31] constructs a multi-task learning architrave in combination with an entity-recognition model that enables the model to select knowledge based not only on knowledge text information but also on domain and entity.

All methods are evaluated using several standard IR metrics including Recall (R@1 and R@5) and Mean Reciprocal Rank (MRR@5) [11]. Models are trained with batch size 16 and with 10 epochs on a single RTX 3090 GPU. We use an AdamW optimizer with a learning rate of 6.25e-5 and an ϵ of 1e-8. All results are measured on the DSTC10 Track2 test set. The number of knowledge candidates C used to train our model is 6, the number of ASR hypotheses N used for aggregation is 5, and the temperature τ in contrastive learning is 0.06. The window size u is set to 128 tokens.

3.3. Evaluation Results

The performance of the proposed CLKS method shown in Table 2 consistently achieves good performance gains on all metrics compared to the DSTC10, Knover, and RADGE methods. CLKS yields an R@1 of 0.6794 which corresponds to a relative improvement of 9.7% over the DSTC10 baseline system.

Table 2: *The main results of the proposed CLKS method. CLKS outperforms other baselines on all metrics and achieves an R@1 improvement of 9.7% over the DSTC10 baseline.*

SYSTEM	MRR@5	R@1	R@5
Knover [24]	0.6522	0.5652	0.7862
DSTC10 [11]	0.7040	0.6193	0.8346
RADGE [31]	0.7137	0.6589	0.7906
CLKS (ours)	**0.7461**	**0.6794**	**0.8419**
w/o CL	0.7135	0.6442	0.8214
w/o (CL and N-best Agg.)	0.7040	0.6193	0.8346

The sampling-based method Knover does not perform well on the test set of DSTC10 Track 2 when we train Knover using

[2]Since Knover and RADGE didn't release their source code, we re-implement their method based on our augmented dialogue data.

[3]https://github.com/alexa/alexa-with-dstc10-track2-dataset

Table 3: *Dataset statistics: Spoken DSTC9 is the data augmentation version of the original dataset using the data augmentation method described in subsection 2.1. The DSTC10 dataset is provided by the DSTC10 challenge.*

Datasets	Split	Modality	Dialogs	Knowledge Domains	Knowledge Entities	Knowledge Snippets
Spoken DSTC9	Train	Written	71348	4	143	2900
DSTC10 dataset	Val	Spoken	263	3	855	10586
	Test		1988	3	855	10586

our spoken DSTC9 dataset (as mentioned above, the DSTC10 dataset does not provide a training set). The main reason may come from the large discrepancy between the knowledge base of DSTC9 and DSTC10. The sampling method proposed by Knover may bias the model towards the knowledge base of DSTC9 and affect the generalization ability of the model. The comparison between Knover and CLKS shows the robustness of CLKS to the scale variations of the knowledge base.

3.4. Ablation Study

The effect of the various components of our proposed method on the final result is shown in the last three rows of Table 2. The result of CLKS without CL (i.e. w/o CL) shows that performance in R@1 deteriorates by about 5.2% when no contrastive learning is used for robust representation learning. The result of CLKS without CL and N-best Agg. (i.e. w/o CL and N-best Agg.) shows that performance in R@1 deteriorates by about 8.8% compared to CLKS when we use only one-best hypotheses and no contrastive learning.

Table 4: *Ablation study on the different number of aggregated hypotheses. We set the number of knowledge candidates to 6. The performance increase when we aggregate more hypotheses.*

Aggregated N-best	MRR@5	R@1	R@5
1	0.7025	0.6325	0.8097
3	0.7268	0.6545	0.8316
5	**0.7461**	**0.6794**	**0.8419**

In the second experiment, we vary the number of N-best hypotheses to investigate the performance of the CLKS method. As can be seen in Table 4, the more N-best hypotheses aggregated, the more robust the knowledge selection model is to ASR errors, leading to better performance.

Table 5: *Ablation study on the different numbers of knowledge candidates (one ground truth + the number of negative samples). We fix the number of N-best hypotheses to aggregate at 5.*

Knowledge candidates	MRR@5	R@1	R@5
2 (1+1)	0.7069	0.6193	0.8258
4 (1+3)	0.7260	0.6559	0.8360
6 (1+5)	**0.7461**	**0.6794**	**0.8419**

The results in Table 5 show that the performance of the CLKS method gradually improves as the number of input knowledge candidates increases when the number of N-best hypotheses to aggregate is fixed at 5.

To demonstrate the effectiveness of our proposed method with various pre-trained models, we combine our representation learning method with DeBERTa [32] and ELECTRA [33]. The results are shown in Table 6. We use the 'base' size of each pre-trained model. 'Baseline' means that we do not use contrastive learning and N-best aggregation for model training, and 'CLKS' means that we use CLKS method for model fine-tuning. As can be seen in Table 6, the performance of 'CLKS' consistently outperforms that of 'Baseline' on all metrics, which demonstrates the effectiveness and generalization of CLKS for different pre-trained models.

Table 6: *Ablation study of compatibility with the different pre-trained model of CLKS.*

Pre-trained model		MRR@5	R@1	R@5
BERT [30]	Baseline	0.7040	0.6193	0.8346
	CLKS	**0.7461**	**0.6794**	**0.8419**
DeBERTa [32]	Baseline	0.6896	0.6208	0.7950
	CLKS	**0.7411**	**0.6647**	**0.8492**
ELECTRA [33]	Baseline	0.7347	0.6662	0.8345
	CLKS	**0.7601**	**0.6911**	**0.8565**

For the case study, the knowledge selected by the different methods for a given dialogue history is shown in Table 7. It can be seen that the baseline approach is susceptible to similar FAQs and it selects the wrong restaurant. On the other hand, the CLKS method can choose the correct knowledge.

Table 7: *Case study: knowledge selected by different methods.*

Dialogue History	
...	
U: uh can i give the address and zip code for dolus in stone son please	
S: sure address is seven five two jackson street zip code nine four one three three	
U: do you know if they offer take-out	
Selected Knowledge	
DSTC10 baseline	Q: Dos Restaurant ***One Seven*** offer takout? A: No, takeout is not offered at One Seven
CLKS	Q: Does ***Delicious Dim Sum*** offer take-out?", A: This restaurant offers take-out.
ORACLE	Q: Does ***Delicious Dim Sum*** offer take-out?", A: This restaurant offers take-out.

4. Conclusions

This paper presents a robust unstructured knowledge selection method, named CLKS, which introduces an ASR N-best aggregation method and a contrastive learning-based method for better knowledge selection of spoken dialogue. Experiments on the DSTC10 dataset demonstrate the effectiveness of the proposed framework and show the potential of the proposed method in bridging the knowledge selection gap between written and spoken inputs.

5. References

[1] M. Ghazvininejad, C. Brockett, M.-W. Chang, and *et al.*, "A knowledge-grounded neural conversation model," in *Proc. AAAI*, 2018.

[2] A. Madotto, C.-S. Wu, and P. Fung, "Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems," in *Proc. ACL*, vol. 1, 2018, p. 1468.

[3] X. Zhao, W. Wu, C. Xu, and *et al.*, "Knowledge-grounded dialogue generation with pre-trained language models," in *Proc. EMNLP*, 2020.

[4] M. Rony, R. Usbeck, and J. Lehmann, "Dialokg: Knowledge-structure aware task-oriented dialogue generation," *arXiv preprint arXiv:2204.09149*, 2022.

[5] S. Kim, Y. Liu, D. Jin, and *et al.*, ""how robust ru?": Evaluating task-oriented dialogue systems on spoken conversations," in *Proc. ASRU*, 2021.

[6] J. Liu, R. Takanobu, J. Wen, and *et al.*, "Robustness testing of language understanding in task-oriented dialog," in *Proc. ACL*, 2021.

[7] G. Karthik, H. Behnam, M. Longshaokan, and *et al.*, "Are neural open-domain dialog systems robust to speech recognition errors in the dialog history? an empirical study," in *Proc. Interspeech*, 2020.

[8] L. Wang, M. Fazel-Zarandi, A. Tiwari, and *et al.*, "Data augmentation for training dialog models robust to speech recognition errors," in *Proc. NLPC AI*, 2020.

[9] X. Tian, X. Huang, D. He, and *et al.*, "Tod-da: Towards boosting the robustness of task-oriented dialogue modeling on spoken conversations," *arXiv e-prints*, pp. arXiv–2112, 2021.

[10] G. Sun, C. Zhang, and P. C. Woodland, "Minimising biasing word errors for contextual asr with the tree-constrained pointer generator," *ArXiv*, vol. abs/2205.09058, 2022.

[11] S. Kim, E. Mihail, G. Karthik, and *et al.*, "Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access," in *Proc. SIGDIAL*, 2020.

[12] W. Haoyu, C. John, and *et al.*, "Leveraging asr n-best in deep entity retrieval," in *Proc. Interspeech*, 2021.

[13] K. Ganesan, P. Bamdev, B. Jaivarsan, and *et al.*, "N-best ASR transformer: Enhancing SLU performance using multiple ASR hypotheses," in *Proc. ACL*, 2021.

[14] Z. Wang, Y. Le, Y. Zhu, and *et al.*, "Building robust spoken language understanding by cross attention between phoneme sequence and asr hypothesis," in *Proc. ICASSP*, 2022.

[15] C.-W. Huang and T.-N. Chen, "Learning asr-robust contextualized embeddings for spoken language understanding," in *Proc. ICASSP*, 2020.

[16] Y. Weng, S.-S. Miryala, C. Khatri, and *et al.*, "Joint contextual modeling for asr correction and language understanding," in *Proc. ICASSP*, 2020.

[17] N. Mahdi, M. John, L. Erran, and *et al.*, "Correcting automated and manual speech transcription errors using warped language models," in *Proc. Interspeech*, 2021.

[18] J. Shin, Y. Lee, and K. Jung, "Effective sentence scoring method using bert for speech recognition," in *Asian Conference on Machine Learning*, 2019.

[19] Y.-H. Chang and Y.-N. Chen, "Contrastive learning for improving asr robustness in spoken language understanding," *arXiv preprint arXiv:2205.00693*, 2022.

[20] C.Jaejin, R. Pappagari, Z.Piotr, and *et al.*, "Non-contrastive self-supervised learning of utterance-level speech representations," in *Proc. INTERSPEECH*, 2022.

[21] Y. Jiang, S. Bidisha, and H. Li, "Knowledge distillation from bert transformer to speech transformer for intent classification," in *Proc. INTERSPEECH*, 2021.

[22] C. You, N. Chen, and Y. Zou, "Contextualized attention-based knowledge transfer for spoken conversational question answering," in *Proc. INTERSPEECH*, 2021.

[23] J. Han, J. Shin, H. Song, and *et al.*, "External knowledge selection with weighted negative sampling in knowledge-grounded task-oriented dialogue systems," *ArXiv*, vol. abs/2209.02251, 2022.

[24] H. He, H. Lu, S. Bao, and *et al.*, "Learning to select external knowledge with multi-scale negative sampling," *ArXiv*, vol. abs/2102.02096, 2021.

[25] S. Kim, Y. Liu, D. Jin, and *et al.*, "Knowledge-grounded task-oriented dialogue modeling on spoken conversations track at dstc10," in *AAAI 2022 Workshop on Dialog System Technology Challenge*, 2022.

[26] A. Baevski, Y. Zhou, A. Mohamed, and *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NIPS*, 2020.

[27] V. Panayotov, G. Chen, D. Povey, and *et al.*, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[28] M. Henderson, B. Thomson, and J. Williams, "The third dialog state tracking challenge," in *Proc. SLT workshop*, 2014.

[29] M. Eric, R. Goel, S. Paul, and *et al.*, "MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines," in *Proc. LREC*, 2020.

[30] J. Devlin, M.-W. Chang, K. Lee, and *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NACL*, 2019.

[31] L. Tang, Q. Shang, K. Lv, and *et al.*, "Radge: Relevance learning and generation evaluating method for task-oriented conversational systems," in *AAAI 2021, Workshop on DSTC9*, vol. 7, 2021.

[32] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations*, 2021.

[33] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *ICLR*, 2020.