INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

ConvKT: Conversation-Level Knowledge Transfer for Context Aware End-to-End Spoken Language Understanding

Vishal Sunder[1], *Eric Fosler-Lussier*[1], *Samuel Thomas*[2], *Hong-Kwang J Kuo*[2], *Brian Kingsbury*[2]

[1]The Ohio State University, Columbus, OH, USA
[2]IBM Research AI, Yorktown Heights, NY, USA

sunder.9@osu.edu, fosler@cse.ohio-state.edu, {sthomas, hkuo, bedk}@us.ibm.com

Abstract

Dialog history enhances downstream classification performance in both speech and text based dialog systems. However, there still exists a gap in dialog history integration in a fully end-to-end (E2E) spoken dialog system (SDS) versus a textual dialog system. Text-based dialog systems use large language models (LLMs) to encode long-range dependencies by attending to the entire conversation as a contiguous token sequence. This is not possible in an E2E SDS, as speech sequences can be intractably long. We propose a convolution subsampling approach to make the speech sequence of a conversation tractable and use a conformer to attend to the speech-based conversation in a fine-grained manner. This model is further enhanced via a conversation-level knowledge transfer from a LLM using a token-level alignment strategy. Finetuning the E2E model pretrained this way gives significant gains, of up to 8%, over strong non-contextual baselines in the E2E dialog act classification task on two datasets.

Index Terms: speech understanding, spoken dialog systems, knowledge transfer

1. Introduction

In recent years, there has been a surge in the popularity of end-to-end (E2E) spoken language understanding (SLU), due to advancements made in building robust speech processing models, speech representation learning [1, 2, 3] and techniques for knowledge transfer from large language models (LLMs) [4, 5, 6], like BERT [7], into speech encoders. Unlike the traditional approach of using automatic speech recognition (ASR) in combination with natural language understanding (NLU) models, an E2E model is more compact and robust to ASR errors, making it preferable for many applications [8, 9, 10].

SLU finds diverse applications in spoken dialog systems (SDS) [11, 12], where a user and agent interact with or without task objectives. In such a system, the integration of dialog history in the model is crucial for downstream language understanding tasks [13, 14, 15, 16, 17] as it provides critical contextual information. It has been shown recently that E2E systems can integrate dialog history directly into the SLU model as speech, thereby maintaining the model's resilience to ASR errors and leveraging context from dialog history efficiently [17].

In the literature, there are broadly two ways of integrating dialog history into a language understanding model, either coarse or fine-grained level (see Figure 1).

Coarse-grained integration (CG): This follows a hierarchical setup where each utterance in the dialog is encoded separately into a representation and a sequence of such utterance representations is encoded by a high level sequence encoder. Such a setup was a popular choice to model conversations in text-based dialog systems before the introduction of LLMs [13, 14]. Recently, this approach was used in an E2E SDS to incorporate dialog history in speech form directly [17].

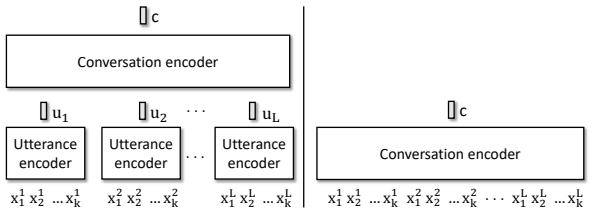

Figure 1: ***Left**: Coarse-grained integration of dialog history. Here, x_i^j represents the i^{th} token in the j^{th} utterance in the dialog. Each utterance is encoded separately into a representation u_j and the sequence $\{u_j\}_{j=1}^{L}$ is passed to another sequence encoder to get the context representation c.* ***Right**: Fine-grained integration of dialog history. The entire sequence of tokens in the conversation $\{\{x_i^j\}_{i=1}^{k}\}_{j=1}^{L}$ is passed as a single contiguous sequence to the conversation encoder.*

Fine-grained integration (FG): This approach is used primarily in text-based dialog systems using LLMs for language understanding [18, 19]. Here, an entire conversation is passed as a single contiguous sequence into a LLM which acts as a conversation encoder. Since LLMs are pretrained on large amounts of text data using long sequences as input to transformer-based models, they are good at learning attention weights which span over long sequences. This helps them capture long-range dependencies in a dialog context which in turn is useful for downstream understanding tasks.

One of the advantages of the FG approach over CG is the ability to capture long-range dependencies at the token level. However, as mentioned above, this requires some form of pretraining of transformers [18]. For speech based E2E models, it is not clear how to pretrain these models in a similar fashion using speech based tokens. Using pretrained transformer-based speech encoders like wav2vec2.0 [1] or HuBERT [2] is impractical as speech sequences are much longer than their text counterparts and concatenating speech based utterances as a single sequence will make the sequence too long to process.

In this paper, we propose to use a FG approach to integrate dialog history in E2E SDS and handle the long sequence problem by using a convolution sub-sampling approach. Instead of mapping a sequence of speech frames in an utterance into a single representation, as done in CG, we first map this sequence into a shorter sequence of speech representations where each representation spans a longer duration than individual frames. These shorter sequences can now be concatenated across the

10.21437/Interspeech.2023-2018

conversation and fed to a conversation encoder.

This approach has the advantage that the output of the conversation encoder in the speech modality can be compared, on a token by token basis, with the corresponding output of a LLM based conversation encoder operating on text. We leverage this to our benefit by using a recently proposed tokenwise contrastive pretraining criterion [4, 6] to perform a knowledge transfer from a LLM to the speech-based conversation encoder. This step also takes care of the necessary pretraining required to capture long-range token-level dependencies by using the knowledge present in LLMs gained through large-scale text-only data.

We show that our proposed model when fine-tuned, outperforms previous state-of-the-art (SOTA) results for E2E dialog act classification task on the Switchboard and HarperValley-Bank datasets. Furthermore, our model performs consistently better as the amount of context information is varied and is also robust to ASR errors compared to the traditional cascaded setup.

2. Method

2.1. Speech-based conversation model

The proposed E2E conversation model consists of three parts.

Utterance encoder: Every utterance in a dialog is encoded using a speech encoder. We used a medium-sized conformer-based speech encoder [20], of a pretrained RNN-T based ASR model. This encoder was trained using 80-dimensional, global mean and variance normalized log-mel filterbank features, extracted every 10 ms using a 25 ms window. We concatenated four consecutive frames such that each output frame from the encoder represents 40 ms of speech.

Convolution subsampling: Simply concatenating all the output sequences from the utterance encoder can lead to a sequence which can be computationally expensive to encode using an attention-based encoder. Thus, to make the dialog sequence more tractable, we first pass the concatenated sequence through 256 convolution channels with a kernel size and stride of 3. This effectively reduces the dialog sequence length by a factor of 3, thus giving us output frames which are 120 ms in duration.

Conversation encoder: The output of the convolution layer is now treated as a sequence of tokens that corresponds to the entire dialog. An attention-based conversation encoder can now be trained on this sequence such that long-term, token-level dependencies are captured just like an LLM does on text-based conversations. We use a 16-layered conformer to encode the conversation. This conformer follows the same design as the medium-sized conformer in Gulati et al. [20] and is trained using the knowledge transfer approach that we describe next.

2.2. Conversation-level Knowledge Transfer (ConvKT)

As mentioned previously, pretraining the conversation encoder is crucial to capturing long-range dependencies. We use tokenwise contrastive learning to transfer BERT's[1] token-level knowledge to the speech-based encoder. We extend this technique to transfer knowledge through the last 6 layers of the 16-layer conformer instead of just the last layer. We describe the process below, borrowing some notations from [4].

Here, we are looking to align the output of layer l_c of the conformer-based conversation encoder with the output of layer l_b of BERT. For this paper, we use all tuples (l_c, l_b) in the set $\mathbb{L} = \{(11,2),(12,4),(13,6),(14,8),(15,10),(16,12)\}$,

[1] https://huggingface.co/bert-base-uncased

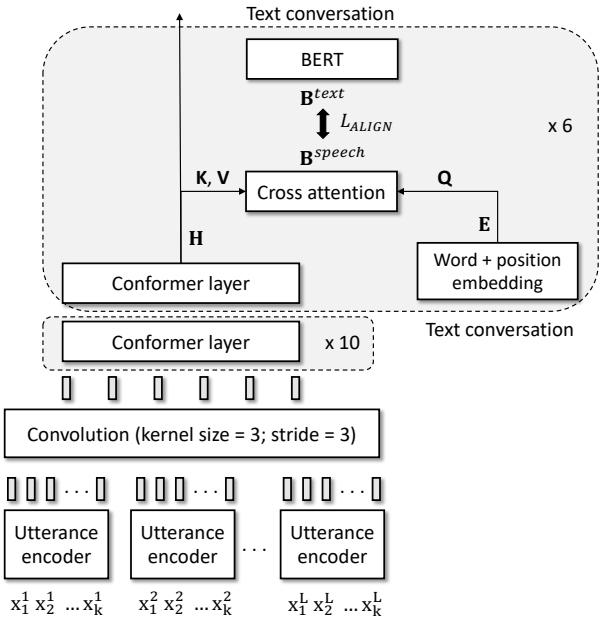

Figure 2: *Each utterance in a dialog is encoded using a speech encoder. All sequences in the dialog are concatenated and passed through a convolution layer which cuts down the sequence length by a third. This sequence is passed through a series of conformer layers. Knowledge transfer from BERT takes place at the last six layers of the conformer model.*

such that we use equally spaced layers from BERT following Shleifer et al. [21]. Let the output of any layer l_c of the conformer[2] be $\mathbf{H} \in \mathbb{R}^{T\times 768}$ and the output of layer l_b of BERT be $\mathbf{B}^{text} \in \mathbb{R}^{n\times 768}$. $\mathbf{H}$ is converted to $\mathbf{B}^{speech} \in \mathbb{R}^{n\times 768}$ through a cross-attention between non-contextual (NC) word embeddings, $\mathbf{E} \in \mathbb{R}^{n\times 768}$ of the conversation text and $\mathbf{H}$ as shown in Figure 2. Note that for each of the 6 layers where the knowledge transfer happens, we have separate NC embeddings and cross-attention weights. However, the NC embeddings are all initialized with the WordPiece embedding layer of BERT.

The cross-attention follows a dot-product attention mechanism with weights $\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v \in \mathbb{R}^{768\times 768}$. The query, key and value iare now computed as,

$$\mathbf{Q} = \mathbf{E}\mathbf{W}_q$$
$$\mathbf{K} = \mathbf{H}\mathbf{W}_k$$
$$\mathbf{V} = \mathbf{H}\mathbf{W}_v$$

Now, the contextual embeddings, $\mathbf{B}^{speech}$ are computed as,

$$\mathbf{B}^{speech} = \text{softmax}(\mathbf{Q}\mathbf{K}^{\mathbb{T}})\mathbf{V}$$

Each row in $\mathbf{B}^{speech}$ now has a one-to-one correspondence with each row in $\mathbf{B}^{text}$. Next, we concatenate the rows of $\mathbf{B}^{speech}$ and $\mathbf{B}^{text}$ obtained for all $(l_c, l_b) \in \mathbb{L}$ across the batch, thus giving us $\mathbf{B}^{speech}_{full}$ and $\mathbf{B}^{text}_{full} \in b \times 768$.

To align the speech and text tokens, a contrastive loss is computed between $\mathbf{B}^{speech}_{full}$ and $\mathbf{B}^{text}_{full}$ as,

$$L_{ALIGN} = -\frac{\tau}{2b}\sum_{i=1}^{b}(\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{b}\exp(s_{ij}/\tau)} + \log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{b}\exp(s_{ji}/\tau)})$$

[2] The number of hidden units in the conformer is 256 but we convert it to 768 using a linear layer for knowledge transfer.

Here, s_{ij} is the cosine similarity between the i^{th} row of $\mathbf{B}_{full}^{text}$ and the j^{th} row of $\mathbf{B}_{full}^{speech}$ and τ is the temperature set to 0.07 following previous work [4]. Minimizing this loss can align speech embeddings with BERT embeddings at the token level. It is important that the embeddings **E** are non-contextual, otherwise the cross-attention does not use the speech embeddings **H** in a meaningful way, ignoring the required context from speech.

3. Experiments

3.1. Training procedure

Pretraining: First, we pretrain a conformer-based ASR model following Saon et al.'s [22] training configuration on 2000 hours of the Fisher dataset. This does not include the Switchboard data as we use it for evaluating the downstream task. The transcription network from this model is used as the utterance encoder. Once trained, the utterance encoder is kept frozen and the rest of the conversation model in Figure 2 is trained using the ConvKT approach above. This is also done on the Fisher dataset as its format is open-ended dyadic conversations. Each training instance is a sequence of 8 utterances in the Fisher dataset. The BERT model was kept frozen during this pretraining.

ConvKT was performed on 8 V100 GPUs for 50 epochs using a batch size of 512. We used the AdamW optimizer and a OneCycleLR policy with a peak learning rate of 5e-4.

Finetuning: After pretraining the conversation encoder, we finetune it on the downstream classification task. We pool the final layer output of the speech-based conversation encoder by using the learnt NC [CLS] embedding to attend over the sequence using the learned cross-attention module. Thus, we get a BERT-like [CLS] token embedding from speech-only data for the entire conversation which is then passed to a classifier. *We do not adapt our model using in-domain transcripts of the classification data* using ConvKT as we assume a realistic scenario where no such transcriptions are available. We used a conversation length of 8 for finetuning where 7 utterances form the dialog context and the 8th utterance is to be labelled. For finetuning, we adapt the model E2E including the utterance encoders.

3.2. Dialog act (DA) classification datasets

For downstream evaluation, we focus on the DA classification task and use two conversation datasets.

Switchboard (SWB)[23]: We use the NXT format Switchboard corpus which is a version of the Switchboard telephonic speech corpus annotated with 42 dialog acts. This dataset contains 193k utterances in the training set, 23k in the validation set and 5k in the test set. We finetune our model on this dataset using multiclass cross-entropy loss.

HarperValleyBank (HVB)[11]: The HVB dataset consists of 1446 simulated spoken conversations between bank employees and customers. There are 16 dialog acts in total and the data is annotated with multiple dialog acts per utterance. The data contains 1174 conversations in the training set and 199 conversations in the test set. For this data, we finetune our model using multilabel binary cross-entropy loss.

4. Results

The results of our experiments are shown in Table 1. We included the oracle performance of the BERT model when trained using human annotated transcriptions of the two datasets. BERT-utt refers to BERT finetuned with only the utterance text for the classification task, while in BERT-conv, the entire conversation up to 7 preceding utterances is also given as input. Row (3) refers to the traditional cascaded baseline where we use an off-the-shelf ASR model to decode the test set and run it through the BERT-conv model. For a fair comparison with our E2E model which does not assume access to in-domain transcripts, we do not adapt the ASR model to the SWB or HVB datasets. The ASR model we used is the RNN-T based model that we pretrained and mentioned in section 3.1. As observed in previous work [27, 17], an unadapted RNN-T gives errorful transcriptions and in some cases just outputs a blank token due to acoustic mismatch. This is catastrophic for the SLU model as evidenced by its poor performance. Rows (4) and (5) show results for other models from previous work on the HVB dataset although these do not use dialog history.

Table 1: *Results on DA classification on SWB and HVB. We use accuracy for SWB and Macro-F1 score for HVB.*

Model	SWB	HVB
Oracle		
(1) BERT-utt	75.36	56.90
(2) BERT-conv	78.12	63.50
Baselines		
(3) ASR $\rightarrow$ BERT-conv	56.29	51.81
(4) Wu et al. [11]	-	45.50
(5) Thomas et al. [24]	-	55.33
(6) Ortega et al. [25]	67.40	-
(7) ESPnet-SLU [26]	68.70	47.10
(8) HIER-S [17]	72.85	57.73
Our E2E models		
(9) Conformer-utt	70.72	53.28
(10) ConvRAND	71.35	57.70
(11) ConvKT-SL	73.21	59.66
(12) ConvKT-ML	**73.98**	**59.78**

Rows (6), (7) and (8) are models that use dialog history, however, the HIER-S model is the only one that is E2E. This is similar to the CG setup in the introduction. Finally, the last 4 rows show our implementation of various E2E models. Row (9) is an E2E model that does not use dialog history but rather finetunes just the conformer-based utterance encoder for the SLU task. ConvRAND refers to our E2E conversation model in figure 2, however, it is not pretrained using the ConvKT criteria. Thus, it underperforms compared to the CG based HIER-S model possibly due to lack of pretraining which makes it unable to capture long-range dependencies in the FG setup. ConvKT-SL (single-layer) in row (11) refers to the E2E conversation model pretrained with ConvKT but the knowledge transfer from BERT only takes place at the last layer, i.e. $\mathbb{L} = \{(16, 12)\}$. Whereas, ConvKT-ML (multi-layer) in row (12) is our full model where knowledge transfer happens in the last six layers.

Our full E2E model, ConvKT-ML outperforms the best non-contextual models by 7.7% and 8.0% on SWB and HVB respectively. The conversation model is able to utilize dialog history effectively as shown by these gains. Furthermore, compared to the previous best HIER-S model, both ConvKT-SL and ConvKT-ML perform better. Also, note that utilizing multiple layers for knowledge transfer (ConvKT-ML) is beneficial over a

single-layer variant (ConvKT-SL). Next, we show that this benefit is consistent even when we vary the context length.

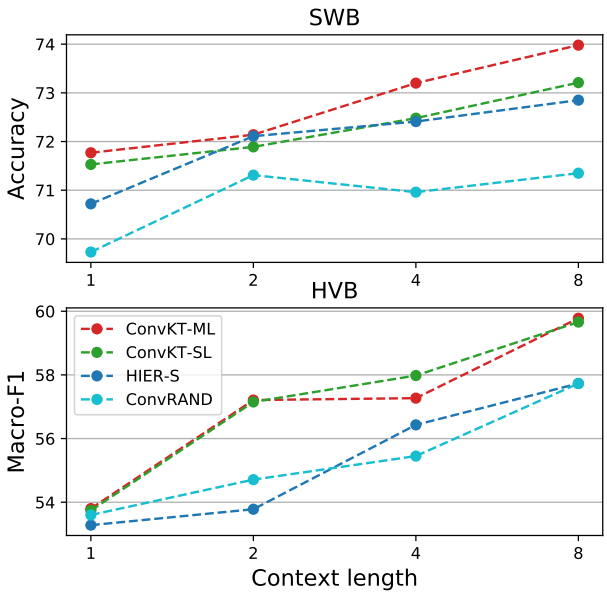

Figure 3: *Performance variation of dialog history based E2E models as the amount of context information is varied.*

Effect of context length: We further evaluate how varying the number of utterances in the dialog history affects the performance of different models. For this, we use only dialog history based E2E models from Table 1; the trend is shown in Figure 3. The performance of all models improves as the context length is increased. The only exception seems to be the ConvRAND model on the SWB dataset, whose performance drops slightly as the context length is increased from 2 to 4. This may be because this model is not pretrained, and hence faces difficulty in capturing long-term dependencies effectively.

On the SWB dataset, we see that ConvKT-ML performs better for all context lengths whereas on HVB, ConvKT-SL and ConvKT-ML perform almost equally well for most context lengths. Overall, our FG based ConvKT models outperform the CG based HIER-S model even with different context lengths.

Cross-modal embedding alignment: One of the motivations for using the ConvKT pretraining is to achieve an alignment between WordPieces and the corresponding locations in the speech. It has been shown previously that such an alignment is indeed achieved at the utterance level [4]. Here, we wish to see if this is consistent even at the conversation level where it might be harder to learn this alignment on much longer sequences. For this, we plot the heatmap of the cross-attention at the last layer of our model and show this in Figure 4.

The top and the bottom parts show this alignment in the ConvKT-SL and ConvKT-ML models respectively. We note that there is a monotonic alignment between the WordPiece tokens and the speech embeddings from our E2E conversation model for both ConvKT-SL and ConvKT-ML. Previously, this type of monotonic alignment has been seen in attention-based ASR and knowledge transfer models [4, 6, 28] at the utterance level. In this paper, we see for the first time that this alignment can be achieved at the conversation level and can improve context aware spoken language understanding.

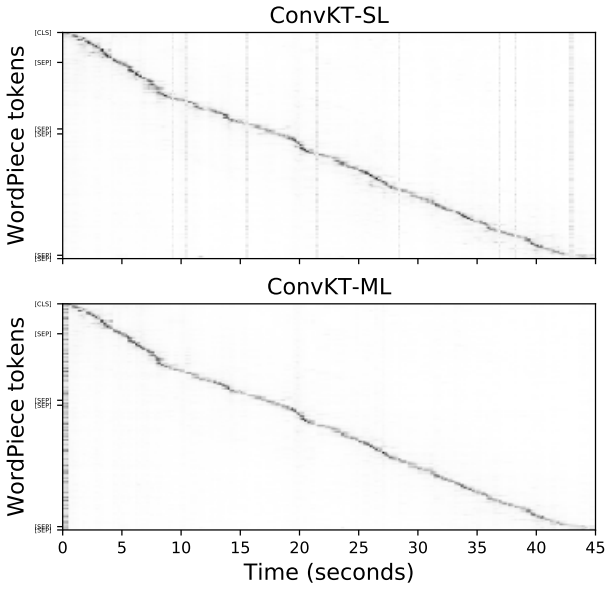

Figure 4: *Attention heatmap of the cross-attention layer which shows how the speech embeddings align with the WordPiece BERT embeddings in a 45 seconds long conversation from the Fisher data. [SEP] on the y-axis represent points where the speaker changes. Each pixel denotes 120 ms of audio.*

Note that in the ConvKT-SL alignment, there are a few intermediate regions in the speech that have high weights which means that these regions are always part of the contextual representation of any other speech token. These might be regions where the models embeds useful contextual information. However, with the ConvKT-ML model, we do not see such regions prominently scattered around. Rather, the only region where we see consistently high attention weights are the first few tokens. This shows that context information is more localized in the ConvKT-ML model than in the ConvKT-SL model. We hypothesize that as the knowledge transfer to ConvKT-ML is deeper, the model relies confidently on a localized region in the speech to embed context rather than scattering it across the signal. Future work can investigate this phenomenon in more depth.

5. Conclusion

In this paper, we show how an E2E SLU model can integrate dialog history in speech form in a token level manner for improved performance. This is analogous to how LLMs encode the entire text conversation for dialog tasks. Thus, knowledge transfer from LLMs to our E2E conversation model can be done in a fine-grained manner at the token level. To do this, we propose the ConvKT mechanism which is a tokenwise contrastive learning based knowledge transfer technique. Extensive experiments on two datasets show that our proposed model can improve downstream E2E SLU and consistently improves performance even with varied context lengths. Future work should look at how an E2E conversation model can be built in a self-supervised way without needing parallel transcripts.

6. Acknowledgement

This work was supported by the National Science Foundation under Grant No. 2008043.

7. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[3] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec2.0/HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[4] V. Sunder, E. Fosler-Lussier, S. Thomas, H.-K. J. Kuo, and B. Kingsbury, "Tokenwise contrastive pretraining for finer speech-to-BERT alignment in end-to-end speech-to-intent systems," *Interspeech*, 2022.

[5] Y. Higuchi, B. Yan, S. Arora, T. Ogawa, T. Kobayashi, and S. Watanabe, "BERT meets CTC: New formulation of end-to-end speech recognition with pre-trained masked language model," *arXiv preprint arXiv:2210.16663*, 2022.

[6] V. Sunder, S. Thomas, H.-K. J. Kuo, B. Kingsbury, and E. Fosler-Lussier, "Fine-grained textual knowledge transfer to improve rnn transducers for speech recognition and understanding," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[8] M. Rao, A. Raju, P. Dheram, B. Bui, and A. Rastrow, "Speech to semantics: Improve ASR and NLU jointly via all-neural interfaces," *Interspeech*, 2020.

[9] A. Raju, M. Rao, G. Tiwari, P. Dheram, B. Anderson, Z. Zhang, C. Lee, B. Bui, and A. Rastrow, "On joint training with interfaces for spoken language understanding," *arXiv preprint arXiv:2106.15919*, 2021.

[10] S. Seo, D. Kwak, and B. Lee, "Integration of pre-trained networks with continuous token interface for end-to-end spoken language understanding," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7152–7156.

[11] M. Wu, J. Nafziger, A. Scodary, and A. Maas, "HarperValleyBank: A domain-specific spoken dialog corpus," *arXiv preprint arXiv:2010.13929*, 2020.

[12] V. Sunder, P. Serai, and E. Fosler-Lussier, "Building an ASR error robust spoken virtual patient system in a highly class-imbalanced scenario without speech data," *arXiv preprint arXiv:2204.05183*, 2022.

[13] C. Bothe, C. Weber, S. Magg, and S. Wermter, "A context-based approach for dialogue act recognition using simple recurrent neural networks," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[14] V. Raheja and J. Tetreault, "Dialogue act classification with context-aware self-attention," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 3727–3733.

[15] N. Tomashenko, C. Raymond, A. Caubrière, R. De Mori, and Y. Estève, "Dialogue history integration into end-to-end signal-to-concept spoken language understanding systems," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8509–8513.

[16] J. Ganhotra, S. Thomas, H.-K. J. Kuo, S. Joshi, G. Saon, Z. Tüske, and B. Kingsbury, "Integrating dialog history into end-to-end spoken language understanding systems," *Interspeech*, 2021.

[17] V. Sunder, S. Thomas, H.-K. J. Kuo, J. Ganhotra, B. Kingsbury, and E. Fosler-Lussier, "Towards end-to-end integration of dialog history for improved spoken language understanding," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7497–7501.

[18] C.-S. Wu, S. C. Hoi, R. Socher, and C. Xiong, "TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 917–929.

[19] S. Mehri, M. Eric, and D. Hakkani-Tur, "Dialoglue: A natural language understanding benchmark for task-oriented dialogue," *arXiv preprint arXiv:2009.13570*, 2020.

[20] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *Interspeech*, 2020.

[21] S. Shleifer and A. M. Rush, "Pre-trained summarization distillation," *arXiv preprint arXiv:2010.13002*, 2020.

[22] G. Saon, Z. Tüske, D. Bolanos, and B. Kingsbury, "Advancing RNN transducer technology for speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5654–5658.

[23] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The NXT-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language resources and evaluation*, vol. 44, pp. 387–419, 2010.

[24] S. Thomas, H.-K. J. Kuo, B. Kingsbury, and G. Saon, "Towards reducing the need for speech training data to build spoken language understanding systems," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7932–7936.

[25] D. Ortega, C.-Y. Li, G. Vallejo, P. Denisov, and N. T. Vu, "Context-aware neural-based dialog act classification on automatically generated transcriptions," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7265–7269.

[26] S. Arora, S. Dalmia, P. Denisov, X. Chang, Y. Ueda, Y. Peng, Y. Zhang, S. Kumar, K. Ganesan, B. Yan *et al.*, "ESPnet-SLU: Advancing spoken language understanding through espnet," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7167–7171.

[27] A. Mani, S. Palaskar, N. V. Meripo, S. Konam, and F. Metze, "ASR error correction and domain adaptation using machine translation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6344–6348.

[28] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.