ASR Adaptation for E-commerce Chatbots using Cross-Utterance Context and Multi-Task Language Modeling

Ashish Shenoy Sravan Bodapati Katrin Kirchhoff
Amazon AWS AI, USA
{ashenoy, sravanb, katrinki}@amazon.com

Abstract

Automatic Speech Recognition (ASR) robustness toward slot entities are critical in e-commerce voice assistants that involve monetary transactions and purchases. Along with effective domain adaptation, it is intuitive that cross utterance contextual cues play an important role in disambiguating domain specific content words from speech. In this paper, we investigate various techniques to improve contextualization, content word robustness and domain adaptation of a Transformer-XL neural language model (NLM) to rescore ASR N-best hypotheses. To improve contextualization, we utilize turn level dialogue acts along with cross utterance context carry over. Additionally, to adapt our domain-general NLM towards e-commerce on-the-fly, we use embeddings derived from a finetuned masked LM on in-domain data. Finally, to improve robustness towards in-domain content words, we propose a multi-task model that can jointly perform content word detection and language modeling tasks. Compared to a non-contextual LSTM LM baseline, our best performing NLM rescorer results in a content WER reduction of 19.2% on e-commerce audio test set and a slot labeling F1 improvement of 6.4%.

1 Introduction

Task-oriented conversations in voice chatbots deployed for e-commerce usecases such as shopping (Maarek, 2018), browsing catalog, scheduling deliveries or ordering food are predominantly short-form audios. Moreover, these dialogues are restricted to a narrow range of multi-turn interactions that involve accomplishing a specific task (Mari et al., 2020). The back and forth between a user and the chatbots are key to reliably capture the user intent and slot entities referenced in the spoken utterances. As shown in previous works (Irie et al., 2019; Parthasarathy et al., 2019; Sun et al., 2021), rather than decoding each utterance independently, there can be benefit in decoding these utterances based on context from previous turns. In the case of grocery shopping for example, knowing that the context is "what kind of laundry detergent?" should help in disambiguating "pods" from "pause". Another common aspect in e-commerce chatbots is that the speech patterns differ among sub-categories of usecases (Eg. shopping clothes vs ordering fast food). Hence, some chatbot systems allow users to provide pre-defined grammars or sample utterances that are specific for their usecase (Gandhe et al., 2018). These user provided grammars are then predominantly used to perform domain adaptation on an n-gram language model. Recently (Shenoy et al., 2021) showed that these can be leveraged to bias a Transformer-XL (TXL) LM rescorer on-the-fly.

While there has been extensive previous work on improving contextualization of TXL LM using historical context, none of the approaches utilize signals from a natural language understanding (NLU) component such as turn level dialogue acts. This paper investigates how to utilize dialogue acts along with user provided speech patterns to adapt a domain-general TXL LM towards different e-commerce usecases on-the-fly. We also propose a novel multi-task architecture for TXL, where the model jointly learns to perform domain specific slot detection and LM tasks. We use perplexity (PPL) and word error rate (WER) as our evaluation metrics. We also evaluate on downstream NLU metrics such as intent classification (IC) F1 and slot labeling (SL) F1 to capture the success of these conversations. The overall contributions of this work can be summarized as follows :

- We show that a TXL model that utilizes turn level dialogue act information along with long

Proceedings of the 4th Workshop on e-Commerce and NLP (ECNLP 4), pages 18–25

span context helps with contextualiziation and improves WER and IC F1 in e-commerce chatbots.

- To improve robustness towards e-commerce domain specifc slot entities, we propose a novel TXL architecture that is jointly trained on slot detection and LM tasks which significantly improves content WERR and SL F1.

- We show that adapting the NLM towards user provided speech patterns by using BERT on domain specific text is an efficient and effective method to perform on-the-fly adaptation of a domain-general NLM towards e-commerce utterances.

2 Related Work

Incorporating cross utterance context has been well explored with both recurrent and non-recurrent NLMs. With LSTM NLMs, long span context is usually propogated without resetting hidden states across sentences or using longer sequence lengths (Xiong et al., 2018a; Irie et al., 2019; Khandelwal et al., 2018; Parthasarathy et al., 2019). In (Xiong et al., 2018b), along with longer history, information about turn taking and speaker overlap is used to improve contextualization in human to human conversations. With transformer architecture based on self attention (Vaswani et al., 2017) (Dai et al., 2019) showed that by utilizing segment wise recurrence Transformer-XL (TXL) (Dai et al., 2019) is able to effectively leverage long span context while decoding. More recently, improving contextualization of the TXL models included adding a LSTM fusion layer to complement the advantages of recurrent with non-recurrent models (Sun et al., 2021). (Shenoy et al., 2021) incorporated a non-finetuned masked LM fusion in order to make the domain adaptation of TXL models quick and on-the-fly using embeddings derived from customer provided data and incorporated dialogue acts but only with an LSTM based LM. While (Sunkara et al., 2020) tried to fuse multi-model features into a seq-to-seq LSTM based network. In (Sharma, 2020) cross utterance context was effectively used to perform better intent classification with e-commerce voice assistants.

For domain adaptation, previous techniques explored include using an explicit topic vector as classified by a separate domain classifier and incorporating a neural cache (Mikolov and Zweig, 2019; Li et al., 2018; Raju et al., 2018; Chen et al., 2015). (Irie et al., 2018) used a mixture of domain experts which are dynamically interpolated. It is also shown in (Liu et al., 2020), that using a hybrid pointer network over contextual metadata can also help in transcribing long form social media audio. Joint learning NLU tasks such as intent detection and slot filling have been explored with RNN based LMs in (Liu and Lane, 2016) and more recently in (Rao et al., 2020), where they show that a jointly trained model consisting of both ASR and NLU tasks interfaced with a neural network based interface helps incorporate semantic information from NLU and improves ASR that comprises a LSTM based NLM. In (Yang et al., 2020) tried to incorporate joint slot and intent detection into a LSTM based rescorer with a goal of improving accuracy on rare words in an end-to-end ASR system.

However, none of the previous work utilize dialogue acts with a non-recurrent based LM such as Transformer-XL nor optimize towards improving robustness of in-domain slot entities. In this paper we experiment and study the impact of utilizing dialogue acts along with a masked language model fusion to improve contextualization and domain adaptation. Additionally, we also propose a novel multi-task architecture with TXL LM that improves the robustness towards in-domain slot entity detection.

3 Approach

A standard language model in an ASR system computes a probability distribution over a sequence of words $W = w_0, ..., w_N$ auto-regressively as:

$$p(W) = \prod_{i=1}^{N} p(w_i|w_1, w_2, ..., w_{i-1}) \quad (1)$$

In our experiments, along with historical context, we condition the LM on additional contextual metadata such as dialogue acts :

$$p(W) = \prod_{i=1}^{N} p(w_i|w_1, w_2, ..., w_{i-1}, c_1, c_2, ..., c_k) \quad (2)$$

Where $c_1, c_2, ...c_k$ are the turn based lexical representation of the contextual metadata. For baseline, we use a standard LSTM LM as summarized below :

$$\begin{aligned} embed_i &= E_{ke}^T w_{i-1} \\ c_i, h_i &= LSTM(h_{i-1}, c_{i-1}, embed_i) \\ p(w_i|w_{<i}) &= Softmax(W_{ho}^T h_i) \end{aligned} \quad (3)$$

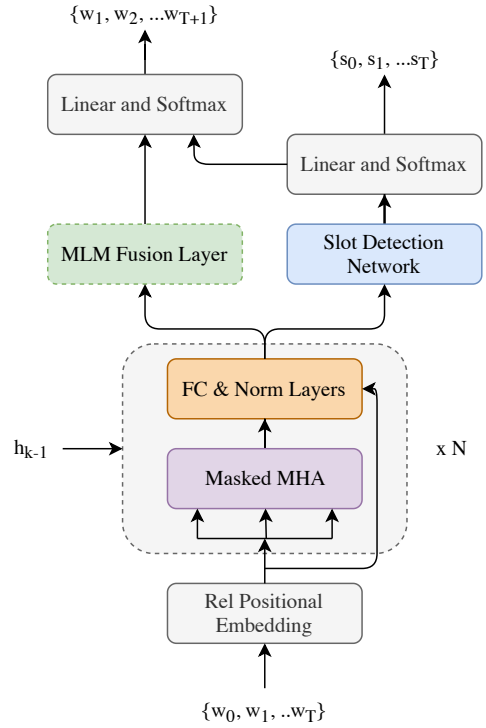

Figure 1: Transformer-XL language model architecture jointly trained with slot detection task with an optional MLM fusion layer

Utterance	i	want	my	shoes	delivered	to	seattle	next	thursday
Annotation	O	O	O	SLOT	O	O	SLOT	SLOT	SLOT

Figure 2: Example utterance with slots annotated

where $embed_i$ is a fixed size lower dimensional word embedding and the LSTM outputs are projected to word level outputs using W_{ho}^T. A $Softmax$ layer converts the word level outputs into final word level probabilities.

3.1 Transformer-XL based NLM

Although recurrent language models help in modeling long range dependencies to certain extent, they still suffer from the fuzzy far away problem (Khandelwal et al., 2018). Vanilla transformer LMs on the other hand use fixed segment lengths which leads to context fragmentation. To address these limitations and model long range dependencies, TXL models add segment-level recurrence and use a relative positional encoding scheme (Dai et al., 2019). Hence we choose to use a TXL LM directly. The cached hidden representations from previous segments helps contextual information flow across segment boundaries. If $s_k = [w_{k,1}, ..., w_{k,T}]$ and $s_{k+1} = [x_{k+1,1}, ..., x_{k+1,T}]$ are two consecutive segments of length T and h_k^n is the n-th layer hidden state produced for the k-th segment s_k, then, the n-th layer hidden state for segment s_{k+1} is produced as follows:

$$\begin{aligned} \tilde{h}_{k+1}^{n-1} &= [SG(h_k^{n-1}) \circ h_{k+1}^{n-1}] \\ q_{k+1}^n, k_{k+1}^n, v_{k+1}^n &= \tilde{h}_{k+1}^{n-1} W_q^\intercal \\ h_{k+1}^n &= TL(q_{k+1}^n, k_{k+1}^n, v_{k+1}^n) \end{aligned} \tag{4}$$

where $SG(.)$ stands for stop gradient and TL stands for Transformer Layer. To carry over context from previous turns, we train and evaluate the model by concatenating all the turns, including the bot responses, in a single conversation session. The model is trained with a cross entropy objective as defined below :

$$\mathbf{L_{LM}} = -\frac{1}{T}\left[\sum_{i=1}^{T} log(P(w_i \mid w_{<i}, s_{<i}))\right] \tag{5}$$

During inference time, we cache a fixed length hidden representation from previous segments. We also use the generated bot responses to perform a forward pass and carry over the context to the next user turn.

3.2 Slot detection and language modeling multi-task learning

To make the our domain-general model robust to e-commerce specific slot entities, we propose a multi-task learning approach to training the TXL LM. We train our models on both LM and slot detection tasks. Similar to slot filling, slot detection is a sequence classification task that involves predicting if a word, w_i at time step i is a domain specific slot entity. We use a separate slot detection network, consisting of a simple multi-layer perceptron, and use the final layer hidden representation from the TXL network as inputs to the network. Figure 2 shows an example utterance with the slot annotations. Formally, let $s = (s_0, s_1, ..., s_T)$ be the slot label sequence, corresponding to a word sequence $w = w_0, w_1, ..., w_T$ in the k-th segment. We model the slot label output s_t as a conditional distribution over input word sequence up to time step t, $w_{\leq t}$ similar to (Liu and Lane, 2016) :

$$\begin{aligned} h_k^n &= TL(q_k^n, k_k^n, v_k^n) \\ p(s_t|w_{\leq t}) &= SlotLabelDist(h_t^n) \end{aligned} \tag{6}$$

We use a cross-entropy training objective for the

slot detection task as below :

$$\mathbf{L_{SD}} = -\frac{1}{T}\left[\sum_{i=1}^{T} log(P(s_i \mid w_{\leq i}))\right] \quad (7)$$

To incorporate this semantic information about the word from previous time step into the NLM, we use the logits from the slot detection network to condition the probability distribution of the next word in the sequence as shown in Figure 1.

The total loss is then computed using a linear combination of LM and slot detection losses:

$$\mathbf{L_{total}} = \mathbf{L_{LM}} + \alpha_{SD}\mathbf{L_{SD}} \quad (8)$$

where α_{SD} is the weight for the slot detection loss.

3.3 Transformer-XL LM conditioning on dialogue acts

Dialogue acts (DA) in a conversation represent the intention of an utterance and is intended towards capturing the action that an agent is trying to accomplish (Austin, 1975). An example conversation snippet with DA is shown in Table 1. DA classification is typically performed in a separate component that is part of a downstream NLU system and consumes the outputs generated by ASR. The classified DA is an important contextual signal that provides hints about the type of speech pattern that can be expected in the next turn. We utilize these signals to train our TXL models. Specifically, we augment the training data with the dialogue act information prefixed to the user turns and surround them with explicit <dialogue_act> tags. The expectation is that the TXL LM learns the usage patterns associated with different dialogue acts and this information should help narrow down the search space for the model to content words relevant to the current dialogue context.

3.4 Domain adaptation using contextual semantic embeddings

In production chatbots, it is common for bot developers to provide example speech patterns, in the form of sample sentences or explicit grammars, which can then be used to bias the n-gram language models in a ASR system (Gandhe et al., 2018). This pre-defined set of speech patterns is a useful source of contextual information that can be also used to bias NLMs as well. As demonstrated in (Shenoy et al., 2021), pretrained masked language models (MLM) such as BERT, can be used to derive a fixed size semantic representation from this lexical information. Large pretrained MLMs are gaining widespread popularity and are considered as powerful language learners (Radford et al., 2016; Brown et al., 2020). However, the sentence or document embeddings derived from such an MLM without finetuning on in-domain data is shown to be inferior in terms of the ability to capture semantic information that can be used in similarity related tasks (Reimers and Gurevych, 2019). Instead of using the [CLS] vector to obtain sentence embeddings, in this paper we take the average of context embeddings from last two layers as these are shown to be consistently better than using [CLS] vector (Reimers and Gurevych, 2019; Li et al., 2020).

Actor	Utterance	Dialogue Act
Bot	how can i help you today	general-welcome
User	hi i want to track my online shopping order	inform-intent
Bot	sure! what is the order number?	request
User	my order number is abcdef	inform
Bot	your order is scheduled to be delivered tomorrow	inform
User	thanks	thank-you
Bot	do you need help with anything else?	req-more

Table 1: A sample user bot conversation snippet showing example dialogue acts.

We use a simple fusion method as experimented in (Shenoy et al., 2021) where the hidden state from the last layer of the TXL decoder is concatenated with the BERT derived embedding. This is then followed by a single projection layer with a non-linear activation function σ, such as $sigmoid$.

$$g_t = \sigma(W[h_t^{TXL}; e^{MLM}] + b) \quad (9)$$

Where h_t^{TXL} is the hidden state from the last transformer decoder and e^{MLM} is the BERT derived embedding from in domain sample utterances. The intuition here is that the model learns to associate the domain specific BERT derived embedding with the occurrences of jargon specific to that domain. Thus providing different BERT vectors derived from different domain texts should allow the model to adapt towards such domains on-the-fly.

	Model	Retail				Fastfood			
		CWERR	IC F1	SL F1	p-value	CWERR	IC F1	SL F1	p-value
1	Non-contextual LSTM	–	–	–	–	–	–		
2	TXL	1.0%	0.5%	0.4%	0.083	12.3%	0.4%	0.9%	**0.048**
3	+ Dialogue Acts (DA)	1.2%	1.2%	1.3%	0.057	14.4%	1.0%	1.2%	**0.041**
4	+ Joint Slot Detection (SD)	4.3%	2.0%	3.3%	**0.046**	16.3%	0.9%	2.1%	**0.015**
5	+ Joint SD + DA	8.6%	2.1%	3.3%	**0.048**	17.3%	1.3%	2.7%	**0.009**
6	+ BERT Fusion	6.4%	2.8%	2.3%	**0.030**	18.2%	1.8%	4.8%	**0.004**
7	+ BERT Fusion + DA	9.6%	2.9%	2.7%	**0.023**	19.2%	1.7%	4.8%	**0.004**
8	+ Joint SD + DA + BERT Fusion	11.8%	3.8%	4.3%	**0.037**	19.2%	2.1%	6.4%	**0.002**

Table 2: ASR and NLU improvements on two e-commerce sub-domains : Retail and Fastfood. CWERR - Content Word Error Reduction, IC F1 - Relative Intent Classification F1 Improvement, SL F1 - Relative Slot Labeling F1 Improvement, MPSSWE p-value test on WERR where significant improvements are in bold

Model	PPLR_{gen}	PPLR_{ecom}
TXL	–	
+ Dialogue Acts (DA)	3.4%	9.9%
+ Joint Slot Detection (SD)	8.5%	11.4%
+ BERT Fusion (BF)	16.3%	23.3%
+ Joint SD + DA	9.8%	13.0%
+ BF + DA	21.5%	25.3%
+ BF + DA + Joint SD	21.0%	25.8%

Table 3: Relative perplexity reduction (PPLR) from the various TXL models on a general domain eval set (PPL_{gen}) and on e-commerce domain eval set (PPL_{ecom}).

4 Experimental Setup

4.1 Dataset

We required task-oriented dialogue datasets with actor, dialogue acts and the slot entities annotated. Since no single dataset was large enough to train a reliable language model, we used a combination of Schema-Guided Dialogue Dataset (Rastogi et al., 2019), MultiWOZ 2.1 (Eric et al., 2019; Budzianowski et al., 2018), MultiDoGo (Peskov et al., 2019) along with anonymized in-house datasets that belong to two e-commerce usecases : retail and fastfood delivery. The final LM training data consisted of 260k training samples, 56k validation and evaluation samples and around 9.9 million running words. We used a vocabulary of size 25k. We evaluated our models on anonymized in-house 8kHz close-talk audio. These audio comprised of task-oriented conversations with multiple speakers and acoustic conditions representative of real world usage and belonged to the same two usecases mentioned above. The average number of turns in the audio dataset was 5.

4.2 ASR setup and NLM setup

We used a hybrid ASR model comprising of a regular-frame-rate (RFR) model trained on cross-entropy loss, followed by sMBR (Ghoshal and Povey, 2013). The first pass LM we used was a domain-general Kneser-Ney (KN) (Kneser and Ney, 1995) smoothed 4-gram model estimated on a weighted mix of datasets spanning multiple domains. The final vocabulary size of the n-gram LM was 500k words. All our NLM rescorers used a 4-layer Tranformer-XL[1] decoder, each of size 512 with 4 attention heads. The input word embedding size was 512. We used a segment and memory length of 25. During model training we applied a dropout rate of 0.3 to both the slot detection network and TXL. For the slot detection layer we used a 3 layer MLP and used the final layer hidden representation from the TXL as the output. To obtain the BERT embedding from in-domain speech patterns, we finetune huggingface[2] pretrained BERT mode on the retail and fastfood text corpus. The derived BERT embedding size used was 768. During inference, we extract n-best hypothesis with `n<=50` from the lattice generated by the first pass ASR model. We rescored the n-best hypothesis by multiplying the acoustic score with the acoustic scale and adding it to the scores obtained from the TXL rescorer. We used a fixed α_{SD} of 0.8 for the slot detection loss.

5 Results and Discussion

Table 3 summarizes the relative perplexity reductions (PPLR). Since we are optimizing our models to improve on the e-commerce domain specific con-

[1]https://github.com/kimiyoung/transformer-xl
[2]https://github.com/huggingface/transformers

tent words we directly report the relative content word error rate reductions (CWERR) in Table 2 along with the relative impact on the downstream NLU tasks of IC and SL. For computing CWERR, we remove all the stop words comprising of commonly used function words, such as conjunctions and prepositions from the transcriptions and evaluate only on content words. We also report statistical significance of our CWER improvements using matched pairs sentence segment word error test (MPSSWE). All the WER numbers are relative to a non-contextual LSTM baseline. The gap in the performance between the two domains we tested on is reflective of the underlying training corpus distribution, which has more text belonging to the fastfood domain.

Perplexity gains indicate effective domain adaptation We report both general domain and e-commerce domain PPLR. Overall, the contextualization and domain adaptation techniques help with the PPL dropping in both cases. The jointly trained model on in-domain slot detection however clearly helps more in the e-commerce case. Moreover, since we used BERT that was finetuned on e-commerce text we again see larger gains in the domain specific testset when compared to the general domain testset (23.3% vs 16.3%).

Using system dialogue acts improves intent detection: From our experiments that train the TXL LMs with dialogue act information, it is clear that dialogue acts helps with relatively marginal gains in PPL (3.4% on generic and 9.9% on e-commerce) and WER (1.2% Retail, 14.4% Fastfood). When compared to other techniques we explored, we see that the impact on intent classification was higher in proportion to the gain in WER, which indicates that dialogue acts are valuable contextual signals to help with intent conveying phrases.

Slot detection loss yields improvements on domain specific content words: Rows 4 and 5 of Table 2 report the content WERR, IC and SL F1s that we obtain by incorporating the joint LM and slot detection (SD) loss. As expected, the multi-task model improves on the content words significantly (1.2% to 4.3% on Retail, 12.3% to 16.3% on Fastfood). This WER improvement also carries over to a higher SL F1 improvement, but a relatively small IC F1 improvement. This is again indicative that the improvements are mainly on recognition of in-domain slot entities and the auxiliary function words that are important to recognize intents do not benefit as much.

Domain adaptation using BERT fusion provides maximum gains: Rows 6 and 7 in Table 2 illustrate the performance of the TXL LM that incorporates the BERT embedding fusion layer. Compared to the model trained with joint slot detection loss, BERT fusion model performs better on all ASR and the NLU metrics. It is evident from the results that the BERT embeddings that are derived from different user provided text helps the model effectively adapt to the domain that the embedding was derived from. The gains are amplified when complemented with the dialogue acts ability to improve on intent carrying words and the joint slot detection model leading to a WERR improving from 12.3% to 19.2% on the fastfood domain and 1% to 11.8% on the retail domain. This also carries over to an improvement on IC and SL F1 of 3.8%, 4.3% on retail and 2.1%, 6.4% on fastfood.

6 Conclusion

In this paper we explored different ways to robustly adapt a domain-general Transformer-XL NLM to rescore N-best hypotheses from a hybrid ASR system for task-oriented e-commerce speech conversations. We demonstrated that Transformer-XL LM trained with turn level dialogue acts benefits intent classification by improving the recognition of content words. Additionally, we show that using semantic embeddings derived from a masked language model finetuned on e-commerce domain can be effectively used to adapt a domain-general TXL LM for e-commerce domain utterance rescoring task. Finally, we introduced a new TXL training loss function to jointly predict content words along with language modeling task, this when combined with BERT fusion and dialogue acts, amplifies the WER, IC F1 and SL F1 gains. We have also shown these improvements to be statistically significant. Future work can look at integrating these methods into an end-to-end ASR system for both rescoring task and first pass LM fusion.

References

John Langshaw Austin. 1975. How to do things with words.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,

Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *CoRR*, abs/1810.00278.

X. Chen, T. Tan, Xunying Liu, Pierre Lanchantin, M. Wan, Mark J. F. Gales, and Philip C. Woodland. 2015. Recurrent neural network language model adaptation for multi-genre broadcast speech recognition. In *Interspeech 2015, Dresden, Germany, September 6-10, 2015*, pages 3511–3515. ISCA.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tür. 2019. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *CoRR*, abs/1907.01669.

Ankur Gandhe, Ariya Rastrow, and Björn Hoffmeister. 2018. Scalable language model adaptation for spoken dialogue systems. In *SLT Workshop 2018, Athens, Greece*, pages 907–912. IEEE.

Arnab Ghoshal and Daniel Povey. 2013. Sequence-discriminative training of deep neural networks. In *in Proc. INTERSPEECH*.

Kazuki Irie, Shankar Kumar, Michael Nirschl, and Hank Liao. 2018. Radmm: Recurrent adaptive mixture model with applications to domain robust language modeling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 6079–6083, Calgary, Canada.

Kazuki Irie, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. Training language models for long-span cross-sentence evaluation. In *ASRU Singapore*, pages 419–426. IEEE.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *ICASSP*, pages 181–184. IEEE Computer Society.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Ke Li, Hainan Xu, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur. 2018. Recurrent neural network language model adaptation for conversational speech recognition. In *Interspeech, Hyderabad, India, 2-6 September 2018*, pages 3373–3377.

Bing Liu and Ian Lane. 2016. Joint online spoken language understanding and language modeling with recurrent neural networks. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 22–30, Los Angeles. Association for Computational Linguistics.

Da-Rong Liu, Chunxi Liu, Frank Zhang, Gabriel Synnaeve, Yatharth Saraf, and Geoffrey Zweig. 2020. Contextualizing ASR lattice rescoring with hybrid pointer network language model. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 3650–3654. ISCA.

Yoelle Maarek. 2018. Alexa and Her Shopping Journey. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, page 1. ACM.

Alex Mari, Andreina Mandelli, and René Algesheimer. 2020. The evolution of marketing in the context of voice commerce: A managerial perspective. In *HCI in Business, Government and Organizations - 7th International Conference, HCIBGO 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings*, volume 12204 of *Lecture Notes in Computer Science*, pages 405–425. Springer.

Tomas Mikolov and Geoffrey Zweig. 2019. Context dependent recurrent neural network language model. In *SLT*, pages 234–239. IEEE.

Sarangarajan Parthasarathy, William Gale, Xie Chen, George Polovets, and Shuangyu Chang. 2019. Long-span language modeling for speech recognition. *CoRR*, abs/1911.04571.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (MultiDoGO): Strategies toward curating and annotating large scale dialogue data. In *Proc EMNLP-IJCNLP*, pages 4526–4536.

Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks.

Anirudh Raju, Behnam Hedayatnia, Linda Liu, Ankur Gandhe, Chandra Khatri, Angeliki Metallinou, Anu Venkatesh, and Ariya Rastrow. 2018. Contextual language model adaptation for conversational agents. In *Interspeech, Hyderabad, India*, pages 3333–3337. ISCA.

Milind Rao, Anirudh Raju, Pranav Dheram, Bach Bui, and Ariya Rastrow. 2020. Speech to semantics: Improve ASR and NLU jointly via all-neural interfaces. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 876–880. ISCA.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset. *arXiv e-prints*, page arXiv:1909.05855.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Arpit Sharma. 2020. Improving intent classification in an E-commerce voice assistant by using inter-utterance context. In *Proceedings of The 3rd Workshop on e-Commerce and NLP*, pages 40–45, Seattle, WA, USA. Association for Computational Linguistics.

Ashish Shenoy, Sravan Bodapati, Monica Sunkara, Srikanth Ronanki, and Katrin Kirchhoff. 2021. Adapting long context nlm for asr rescoring in conversational agents.

G. Sun, C. Zhang, and P. C. Woodland. 2021. Transformer language models with lstm-based cross-utterance information representation.

Monica Sunkara, Srikanth Ronanki, Dhanush Bekal, Sravan Bodapati, and Katrin Kirchhoff. 2020. Multimodal semi-supervised learning framework for punctuation prediction in conversational speech. In *Interspeech 2020, Shanghai, China, 25-29 October 2020*, pages 4911–4915. ISCA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

W. Xiong, L. Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke. 2018a. The microsoft 2017 conversational speech recognition system. In *ICASSP, Calgary, Canada*, pages 5934–5938. IEEE.

Wayne Xiong, Lingfeng Wu, Jun Zhang, and Andreas Stolcke. 2018b. Session-level language modeling for conversational speech. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2764–2768, Brussels, Belgium. Association for Computational Linguistics.

Chao-Han Huck Yang, Linda Liu, Ankur Gandhe, Yile Gu, Anirudh Raju, Denis Filimonov, and Ivan Bulyko. 2020. Multi-task language modeling for improving speech recognition of rare words. *CoRR*, abs/2011.11715.