INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Improving Conversation-Context Language Models with Multiple Spoken Language Understanding Models

Ryo Masumura, Tomohiro Tanaka, Atsushi Ando, Hosana Kamiyama,
Takanobu Oba, Satoshi Kobashikawa, Yushi Aono

NTT Media Intelligence Laboratories, NTT Corporation, Japan
ryou.masumura.ba@hco.ntt.co.jp

Abstract

In this paper, we integrate fully neural network based conversation-context language models (CCLMs) that are suitable for handling multi-turn conversational automatic speech recognition (ASR) tasks, with multiple neural spoken language understanding (SLU) models. A main strength of CCLMs is their capacity to take long-range interactive contexts beyond utterance boundaries into consideration. However, it is hard to optimize the CCLMs so as to fully exploit the long-range interactive contexts because conversation-level training datasets are often limited. In order to mitigate this problem, our key idea is to introduce various SLU models that are developed for spoken dialogue systems into the CCLMs. In our proposed method (which we call "SLU-assisted CCLM"), hierarchical recurrent encoder-decoder based language modeling is extended so as to handle various utterance-level SLU results of preceding utterances in a continuous space. We expect that the SLU models will help the CCLMs to properly understand semantic meanings of long-range interactive contexts and to fully leverage them for estimating a next utterance. Our experiments on contact center dialogue ASR tasks demonstrate that SLU-assisted CCLMs combined with three types of SLU models can yield ASR performance improvements.

Index Terms: multi-turn conversational automatic speech recognition, conversation-context language models, spoken language understanding, hierarchical recurrent encoder-decoder

1. Introduction

Language models (LMs), which compute generative probabilities of word sequences, are assuming a key role in automatic speech recognition (ASR). In traditional ASR systems, LMs are essential as well as acoustic models and pronunciation models [1]. In addition, in end-to-end ASR systems, LMs are effective in improving ASR performance even though the end-to-end ASR systems partly possess the ability of the LMs [2, 3].

It is known that LMs play especially an important role in transcribing long-duration speech tasks rather than short-duration speech tasks. So far, a lot of studies have tried to capture long-range linguistic context information [4–9]. While traditional n-gram LMs consider only a few words as context information [10], LMs based on recurrent neural networks (RNNs) including long short-term memory RNNs (LSTM-RNNs) can take account of longer linguistic contexts within an utterance [4, 5]. In addition, large-context LMs including document-context LMs and conversation-context LMs (CCLMs), which can capture long-range linguistic contexts beyond utterances boundaries, are suitable for handing multi-turn conversational ASR tasks (e.g., contact center dialogues, service center dialogues, meetings, etc.) [6–9].

For building the document-context LMs or the CCLMs, adequate document-level text datasets or conversation-level text datasets are essential. However, it is hard to collect sufficient conversation-level text datasets for the CCLMs since they need to manually transcribe conversational speech datasets, while document-level text datasets can be collected from public text resources such as the Web. Therefore, it is difficult to optimize the CCLMs from the limited datasets so as to fully leverage the long-range interactive contexts.

In order to help to efficiently capture the long-range interactive contexts, our key idea is to leverage the spoken language understanding (SLU) models [11–13] used in spoken dialogue systems for improving the CCLMs. In the spoken dialogue systems, the SLU models help to generate response texts by properly understanding the semantic meaning of the input texts over multiple turns. In fact, roles of the CCLMs in ASR are similar to those in generating the response texts in the spoken dialogue systems. Therefore, we can expect that the SLU models will also help to improve the CCLMs. Furthermore, recent SLU models are composed of neural networks, so it can be expected that CCLMs based on neural networks will be compatible with neural SLU models [14–18].

In this paper, we propose SLU-assisted CCLMs that integrate CCLMs with multiple SLU models. The SLU-assisted CCLMs are fully formed from neural networks in which hierarchical recurrent encoder-decoder based language modeling [19–22] is combined with multiple neural SLU models in a continuous space (see Section 3). In the work reported in this paper, we leveraged multiple spoken utterance classification models [16–18], i.e., dialogue act, topic type, and question type classification models, for improving the CCLMs. This makes it possible to naturally utilize the utterance-level semantic meaning of preceding interactive contexts in CCLMs.

SLU-assisted CCLMs are related to the joint modeling of a slot filling and RNN-based LM, in which word-by-word semantic parsing results are utilized for estimating the next word within an utterance [23]. On the other hand, SLU-assisted CCLMs utilize utterance-level SLU results of all preceding utterances for estimating the next utterance. In addition, the proposed method is closely related to neural conversation models using SLU results, which were developed for generating the response text in the spoken dialogue systems [24, 25]. To the best of our knowledge, this paper is the first study that leverages SLU models for improving multi-turn conversational ASR.

In our experiments on Japanese contact center dialogue datasets, we introduce three types of spoken utterance classification models: dialogue act classification, topic type classification, and question type classification, each of which was developed for spoken dialogue systems [26–28]. We demonstrate that combining multiple SLU models with CCLMs is effective in improving multi-turn conversational ASR tasks.

http://dx.doi.org/10.21437/Interspeech.2019-1534

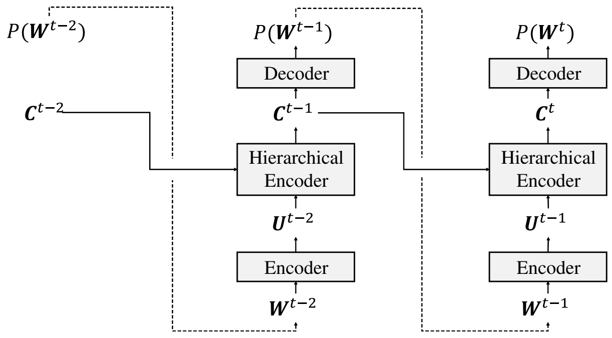

Figure 1: *Network structure of CCLM.*

2. Conversation-Context LMs

This section describes conversation-context LMs (CCLMs) based on a hierarchical recurrent encoder-decoder [9, 19–21]. The CCLM estimates the generative probability of a word in a target utterance from all preceding utterances and all preceding words in the target utterance. A conversation composed of a sequence of utterances is defined as $\mathcal{W} = \{\boldsymbol{W}^1, \cdots, \boldsymbol{W}^T\}$ where T represents the number of utterances in the conversation. The t-th utterance is represented as the word sequence $\boldsymbol{W}^t = \{w_1^t, \cdots, w_{N^t}^t\}$ where N^t represents the number of words in the t-th utterance. The CCLMs define the generative probability of a conversation $\mathcal{W}$ as

$$\begin{aligned} P(\mathcal{W};\boldsymbol{\Theta}) &= \prod_{t=1}^{T} P(\boldsymbol{W}^t|\boldsymbol{W}^1,\cdots,\boldsymbol{W}^{t-1};\boldsymbol{\Theta}) \\ &= \prod_{t=1}^{T}\prod_{n=1}^{N_t} P(w_n^t|w_1^t,\cdots,w_{n-1}^t, \\ &\qquad\qquad \boldsymbol{W}^1,\cdots,\boldsymbol{W}^{t-1};\boldsymbol{\Theta}) \end{aligned} \tag{1}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^{\text{enc}}, \boldsymbol{\theta}^{\text{henc}}, \boldsymbol{\theta}^{\text{dec}}\}$ represents the model parameters. A CCLM based on a hierarchical recurrent encoder-decoder is composed of an encoder, a hierarchical encoder, and a decoder. Figure 1 shows the network structure of the CCLM. In the encoder, utterance-level information in preceding contexts is converted into a continuous representation. The continuous representation that embeds $\boldsymbol{W}^{t-1}$ is defined as

$$\boldsymbol{U}^{t-1} = \texttt{Encoder}(\boldsymbol{W}^{t-1};\boldsymbol{\theta}^{\text{enc}}), \tag{2}$$

where $\texttt{Encoder}()$ is the function of the encoder that is modeled using word-level LSTM-RNNs. Additionally, in the hierarchical encoder, continuous representations of all preceding utterances are converted into a continuous representation. The continuous representation that summarizes all preceding utterances from $\boldsymbol{W}^1$ to $\boldsymbol{W}^{t-1}$ is defined as

$$\begin{aligned} \boldsymbol{C}^t &= \texttt{HierarchicalEncoder}(\boldsymbol{U}^1,\cdots,\boldsymbol{U}^{t-1};\boldsymbol{\theta}^{\text{henc}}) \\ &= \texttt{HierarchicalEncoder}(\boldsymbol{C}^{t-1},\boldsymbol{U}^{t-1};\boldsymbol{\theta}^{\text{henc}}), \end{aligned} \tag{3}$$

where $\texttt{HierarchicalEncoder}()$ is the function of the hierarchical encoder that is modeled using utterance-level LSTM-RNNs. The decoder estimates the generative probability of a word using the continuous representation that embeds all preceding utterances and preceding words in the target utterance. The generative probability of w_n^t is defined as

$$\begin{aligned} &P(w_n^t|w_1^t,\cdots,w_{n-1}^t,\boldsymbol{W}^1,\cdots,\boldsymbol{W}^{t-1};\boldsymbol{\Theta}) \\ &\qquad = \texttt{Decoder}(w_1^t,\cdots,w_{n-1}^t,\boldsymbol{C}^t;\boldsymbol{\theta}^{\text{dec}}), \end{aligned} \tag{4}$$

where $\texttt{Decoder}()$ is the function of the decoder that is represented as an auto-regressive generative model.

3. Proposed Method

This section details spoken language understanding (SLU)-assisted CCLMs. The proposed method involves integrating multiple neural SLU models into CCLMs. In the SLU-assisted CCLMs using M types of SLU models, the generative probability of a conversation $\mathcal{W}$ is defined by

$$\begin{aligned} &P(\mathcal{W};\boldsymbol{\Theta},\boldsymbol{\Phi}_{(1)},\cdots,\boldsymbol{\Phi}_{(M)}) \\ &= \prod_{t=1}^{T}\prod_{n=1}^{N_t} P(w_n^t|w_1^t,\cdots,w_{n-1}^t, \\ &\qquad\qquad \boldsymbol{W}^1,\cdots,\boldsymbol{W}^{t-1};\boldsymbol{\Theta},\boldsymbol{\Phi}_{(1)},\cdots,\boldsymbol{\Phi}_{(M)}), \end{aligned} \tag{5}$$

where $\boldsymbol{\Phi}_{(m)}$ is the model parameter set for the m-th SLU model.

3.1. Spoken Language Understanding Models

In this paper, we use spoken utterance classification models as the SLU models. The spoken utterance classification assumes the role of determining utterance-level semantic label $l \in \mathcal{L}$ from given utterance $\boldsymbol{W} = \{w_1, \cdots, w_N\}$ where $\mathcal{L}$ represents the label set. The spoken utterance classification model defines a conditional probability for each label given utterance, denoted as $P(l|\boldsymbol{W};\boldsymbol{\Phi})$, where $\boldsymbol{\Phi} = \{\boldsymbol{\phi}^{\text{enc}}, \boldsymbol{\phi}^{\text{cls}}\}$ is the model parameter set. In neural spoken utterance classification models, the conditional probabilities are modeled by neural networks in an end-to-end manner. The models comprise an SLU encoder and an SLU classifier. The SLU encoder converts utterance-level semantic information into a continuous representation as is done in the hierarchical recurrent encoder-decoder models. A continuous representation that embeds $\boldsymbol{W}$ is defined as

$$\boldsymbol{Z} = \texttt{SluEncoder}(\boldsymbol{W};\boldsymbol{\phi}^{\text{enc}}), \tag{6}$$

where $\texttt{SluEncoder}()$ is the function of the SLU encoder. In the SLU classifier, conditional probabilities of the utterance-level semantic labels are computed by

$$P(l|\boldsymbol{W};\boldsymbol{\Phi}) = \texttt{SluClassifier}(\boldsymbol{Z};\boldsymbol{\phi}^{\text{cls}}), \tag{7}$$

where $\texttt{SluClassifier}()$ is the function of the SLU classifier that includes a softmax layer.

3.2. Integrated Network Structure

In SLU-assisted CCLMs, we utilize M types of SLU models for extracting utterance-level semantic vector representations from each preceding utterance. The parameter set of the m-th SLU model is defined as $\boldsymbol{\Phi}_{(m)} = \{\boldsymbol{\phi}_{(m)}^{\text{enc}}, \boldsymbol{\phi}_{(m)}^{\text{cls}}\}$. In order to effectively capture semantic meaning of preceding utterances, we examined two semantic vector representations: a hidden vector representation and a posteriorgram representation. Figure 2 shows the network structure of SLU-assisted CCLMs with the hidden vector representations, and Figure 3 shows that with the posteriorgram representations.

The hidden vector representation utilizes a hidden vector that can be extracted from individual SLU encoders since they embed semantic information effective in addressing SLU problems. The hidden vector representation for the $t-1$-th utterance, denoted $\boldsymbol{V}^{t-1}$, is defined as

$$\boldsymbol{V}^{t-1} = [\boldsymbol{U}^{t-1}, \boldsymbol{Z}_{(1)}^{t-1}, \cdots, \boldsymbol{Z}_{(M)}^{t-1}]^\top, \tag{8}$$

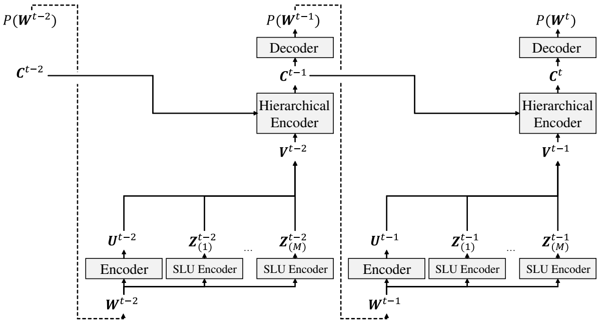

Figure 2: *Network structure of SLU-assisted CCLM with hidden vector representations.*

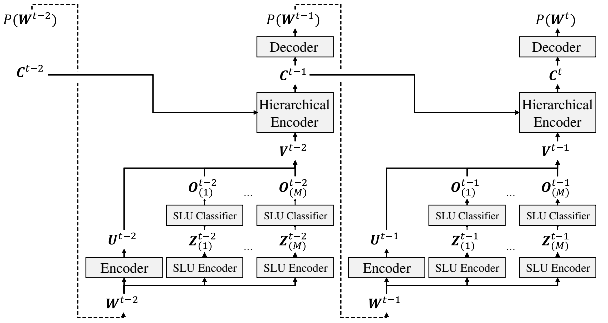

Figure 3: *Network structure SLU-assisted CCLM with posteriorgram representations.*

$$\boldsymbol{U}^{t-1} = \texttt{Encoder}(\boldsymbol{W}^{t-1}; \boldsymbol{\theta}^{\text{enc}}), \tag{9}$$

$$\boldsymbol{Z}^{t-1}_{(m)} = \texttt{SluEncoder}(\boldsymbol{W}^{t-1}; \boldsymbol{\phi}^{\text{enc}}_{(m)}), \tag{10}$$

where $\boldsymbol{Z}^{t-1}_{(m)}$ is the hidden vector in the m-th SLU model.

On the other hand, the posteriorgram representation utilizes posterior probabilities extracted from individual SLU classifiers since the posterior probabilities explicitly include estimated semantic label information. The posteriorgram representation of the $t-1$-th utterance is defined as

$$\boldsymbol{V}^{t-1} = [\boldsymbol{U}^{t-1}, \boldsymbol{O}^{t-1}_{(1)}, \cdots, \boldsymbol{O}^{t-1}_{(M)}]^{\top}, \tag{11}$$

$$\boldsymbol{O}^{t-1}_{(m)} = \texttt{SluClassifier}(\boldsymbol{Z}^{t-1}_{(m)}; \boldsymbol{\phi}^{\text{cls}}_{(m)}), \tag{12}$$

where $\boldsymbol{O}^{t-1}_{(m)}$ is the posterior probabilities in the m-th SLU model.

The hierarchical encoder handles either the hidden vector representations or the posteriorgram representations of all preceding utterances. The continuous representation that summarizes all preceding utterances from $\boldsymbol{W}^1$ to $\boldsymbol{W}^{t-1}$ is defined as

$$\begin{aligned}\boldsymbol{C}^{t} &= \texttt{HierarchicalEncoder}(\boldsymbol{V}^{1}, \cdots, \boldsymbol{V}^{t-1}; \boldsymbol{\theta}^{\text{henc}}) \\ &= \texttt{HierarchicalEncoder}(\boldsymbol{C}^{t-1}, \boldsymbol{V}^{t-1}; \boldsymbol{\theta}^{\text{henc}}).\end{aligned} \tag{13}$$

The decoder computes the generative probability of the n-th word in the t-th utterance as

$$\begin{aligned}P(w^{t}_{n} | w^{t}_{1}, \cdots, w^{t}_{n-1}, \\ \boldsymbol{W}^{1}, \cdots, \boldsymbol{W}^{t-1}; \boldsymbol{\Theta}, \boldsymbol{\Phi}_{(1)}, \cdots, \boldsymbol{\Phi}_{(M)}) \\ = \texttt{Decoder}(w^{t}_{1}, \cdots, w^{t}_{n-1}, \boldsymbol{C}^{t}; \boldsymbol{\theta}^{\text{dec}}),\end{aligned} \tag{14}$$

where $\boldsymbol{C}^{t}$ involves multiple SLU results of all preceding utterances. We can therefore expect that the SLU-assisted CCLMs efficiently leverage long-range interactive contexts.

Table 1: *Contact center dialogue datasets.*

Data	# of dialogues	# of utterances	# of words
Train	2,545	309,401	3,318,235
Valid	45	6,492	48,511
Test	46	6,796	50,116

Table 2: *Datasets for training SLU models.*

Type	Data	# of utterance	# of words
Dialogue act (DA)	Train	201,092	1,665,476
	Valid	4,190	34,345
Topic type (TT)	Train	40,350	627,418
	Valid	4,036	63,673
Question type (QT)	Train	55,328	556,755
	Valid	4,257	42,698

3.3. Training

The model parameter sets of SLU-assisted LMs are optimized in two steps. In the first step, M types of utterance-level training datasets $\{\mathcal{S}_{(1)}, \cdots, \mathcal{S}_{(M)}\}$ are used for optimizing parameter sets $\{\boldsymbol{\Phi}_{(1)}, \cdots, \boldsymbol{\Phi}_{(M)}\}$, where the m-th utterance-level dataset is composed as $\mathcal{S}_{(m)} = \{(\boldsymbol{W}_1, l_1), \cdots, (\boldsymbol{W}_{I_{(m)}}, l_{I_{(m)}})\}$. In the second step, conversation-level training datasets $\mathcal{D} = \{\mathcal{W}_1, \cdots, \mathcal{W}_J\}$ are used for training parameter sets $\boldsymbol{\Theta}$. In the first step, the m-th parameter set $\boldsymbol{\Phi}_{(m)}$ is optimized by

$$\hat{\boldsymbol{\Phi}}_{(m)} = \underset{\boldsymbol{\Phi}_{(m)}}{\text{argmin}} - \sum_{i=1}^{I_{(m)}} \log P(l_i | \boldsymbol{W}_i; \boldsymbol{\Phi}_{(m)}). \tag{15}$$

In the second step, $\boldsymbol{\Theta}$ is optimized so as to fully leverage pretrained parameters $\{\hat{\boldsymbol{\Phi}}_{(1)}, \cdots, \hat{\boldsymbol{\Phi}}_{(M)}\}$ for estimating the generative probability of words. The optimization is followed by

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\text{argmin}} - \sum_{j=1}^{J} \log P(\mathcal{W}_j; \boldsymbol{\Theta}, \hat{\boldsymbol{\Phi}}_{(1)}, \cdots, \hat{\boldsymbol{\Phi}}_{(M)}). \tag{16}$$

In this way, we can freely add various SLU models to improve the SLU-assisted CCLMs.

4. Experiments

In our experiments, home-made Japanese contact center dialogue datasets [9], which include several topics ("finance", "internet provider", "local government", etc.), were prepared for constructing LMs and evaluating ASR performance. One dialogue means one telephone call between one operator and one customer. We manually transcribed conversational speech exchanges and divided them into a training set, a validation set and a test set. Note that any SLU labels were not annotated to the manual transcriptions. In the training set, on average one dialogue included about 121 utterances and one utterance included about 10 words. The vocabulary size of the training set was 25K words. Table 1 shows the details. In addition, home-made Japanese datasets of three different spoken utterance classification tasks, i.e., dialogue act (DA), topic type (TT), and question type (QT) classification, were prepared for constructing the SLU models. Each dataset was constructed for developing spoken dialogue systems [27, 28]. Semantic labels in each dataset were manually annotated. The number of labels was set respectively to 28 for DA ("greeting", "proposal", etc.), 168 for TT ("person", "movie", etc.), and 15 for QT ("true/false", "explanation: method", etc.). Each dataset was individually divided into training (Train) and validation (Valid) sets. Table 2 shows details of the datasets.

Table 3: *Experimental results in terms of PPL and WER (%).*

	Models	SLU models	Semantic vector representations for preceding utterances	Valid		Test	
				PPL	WER (%)	PPL	WER (%)
(1).	HPYLM	-	-	26.97	23.58	25.54	23.37
(2).	LSTMLM	-	-	20.37	21.68	19.45	21.65
(3).	CCLM	-	-	14.78	20.80	14.17	20.60
(4).	SLU-assisted CCLM	DA	Hidden vector representation	14.38	20.45	13.79	20.36
(5).	SLU-assisted CCLM	TT	Hidden vector representation	13.90	20.42	13.36	20.32
(6).	SLU-assisted CCLM	QT	Hidden vector representation	14.42	20.49	13.74	20.36
(7).	SLU-assisted CCLM	DA, TT, QT	Hidden vector representation	**13.80**	**20.20**	**13.16**	**20.12**
(8).	SLU-assisted CCLM	DA	Posteriorgram representation	14.80	20.52	14.21	20.42
(9).	SLU-assisted CCLM	TT	Posteriorgram representation	14.08	20.62	13.47	20.40
(10).	SLU-assisted CCLM	QT	Posteriorgram representation	14.44	20.59	13.78	20.50
(11).	SLU-assisted CCLM	DA, TT, QT	Posteriorgram representation	13.98	20.36	13.30	20.25

For ASR evaluation, we used a senone-based LSTM-RNN acoustic model. The acoustic features were 40 dimensional log mel-filterbank coefficients appended with delta and acceleration coefficients; the frame size was 20 ms and the frame shift was 10 ms. For acoustic modeling, we stacked 2-layer LSTM-RNNs with 512 cells, two fully connected layers that had 1,024 hidden units with rectified linear units, and a softmax layer with 3,072 outputs. The speech recognizer was a weighted finite state transducer (WFST) based decoder [29]. We formed SLU models for constructing SLU-assisted CCLMs. In our neural spoken utterance classification models, the encoder was composed from a word embedding layer with 256 units and a 2-layer LSTM-RNN with 256 units with a self-attention mechanism [30]. The classifier was formed by combining a sigmoid layer with 256 units and a softmax layer. In these setups, words that appeared once or less in the training set were treated as unknown words. The optimization algorithm we used was Adam. Classification accuracy values for the validation set were respectively 65.7 % for DA, 78.7 % for TT, and 87.2 % for QT.

We constructed the following LMs: HPYLM, LSTMLM, CCLM, and multiple SLU-assisted CCLMs. HPYLM is a 3-gram hierarchical Pitman-Yor LM, which is the baseline n-gram LM [31]. HPYLM was introduced in the ASR decoder by converting the WFST. In addition, we constructed the following neural LMs. LSTMLM is an LSTM-RNN based LM composed of a word embedding layer with 650 units, a 1-layer LSTM-RNN with 650 units, and a softmax layer. The CCLM and the SLU-assisted CCLMs are models detailed in Sections 2 and 3. In these models, the encoders were composed of a 1-layer LSTM-RNN with 650 units, the hierarchical encoders were composed of a 1-layer LSTM-RNN with 650 units, and the decoders were composed of a word embedding layer with 650 units, a 1-layer LSTM-RNN with 650 units, and a softmax layer. To optimize each neural LM, we used mini-batch stochastic gradient descent where the learning rate was altered following the validation loss. In each LSTM-RNN layer, variational dropout was applied [32]; the dropout rate was set to 0.6. When using the neural LMs in ASR experiments, 100-best rescoring was conducted by interpolating HPYLM with each neural LM. The interpolation weights were optimized using the validation set.

4.1. Results

Experimental results in terms of perplexity (PPL) and word error rate (WER) are shown in Table 3. In the table, "SLU models" represents which SLU models were used for the SLU-assisted CCLMs and "Semantic vector representations for preceding utterances" means which semantic vector representations were extracted using the SLU models. Note that ground truth SLU labels cannot be utilized for the SLU-assisted CCLMs since they were not annotated to the dialogue datasets.

Line (1) shows baseline results, i.e., no neural LMs were used. Results obtained with LSTMLM are shown by line (2) and those obtained with CCLM are shown by line (3). They show that the neural network LM, which can take longer contexts than HPYLM, offers better ASR performance than HPYLM. They also show that CCLM outperforms LSTMLM. This indicates that long-range interactive contexts beyond utterance boundaries are effective in improving multi-turn conversational ASR performance. Lines (4)-(7) show the results obtained by SLU-assisted CCLM using the hidden vector representation and lines (8)-(11) show those obtained by SLU-assisted CCLM using the posteriorgram representation. Lines (4)-(6) and (8)-(10) used a single SLU model, while lines (7) and (11) simultaneously introduced three SLU models. The results show that each SLU-assisted CCLM improved perplexity and ASR performance compared to ordinary CCLM. These results confirm that the outputs of SLU models can be improved to capture the contexts of preceding utterances. Additionally, the hidden vector representation was superior to the posteriorgram representation, indicating that the former includes more effective semantic information than the latter. We assume that the latter explicitly involves classification errors of individual spoken utterance classification tasks. The best ASR performance was obtained by SLU-assisted CCLM with three types of SLU models. This confirms that it is an effective way to utilize multiple SLU models for efficiently capturing the contexts of preceding utterances.

5. Conclusions

This paper described a method that integrates conversation-context language models (CCLMs) with multiple spoken language understanding (SLU) models for multi-turn conversational automatic speech recognition (ASR) tasks. One strength of the method is to efficiently capture semantic meaning of previous interactive contexts for estimating the next utterance as it can explicitly utilize multiple utterance-level SLU results of the preceding utterances. Experiments in a contact center dialogue ASR task showed that a CCLM integrated with spoken utterance classification models yielded ASR performance improvements. In addition, we verified the effectiveness of simultaneously utilizing multiple SLU models. In future work, we will integrate the CCLMs with not only the spoken utterance classification models but also slot filling models.

6. References

[1] R. Rosenfeld, “Two decades of statistical language modeling: Where do we go from here?” *Proceedings of the IEEE*, vol. 88, pp. 1270–1278, 2000.

[2] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4945–4949, 2015.

[3] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, “Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM,” *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 949–953, 2017.

[4] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, “Recurrent neural network based language model,” *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1045–1048, 2010.

[5] M. Sundermeyer, H. Ney, and R. Schluter, “From feedforward to recurrent LSTM neural networks for language models,” *IEEE/ACM Transactions of Audio, Speech and Language processing*, vol. 23, no. 3, pp. 517–529, 2015.

[6] R. Lin, S. Liu, M. Yang, M. Li, M. Zhou, and S. Li, “Hierarchical recurrent neural network for document modeling,” *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 899–907, 2015.

[7] T. Wang and K. Cho, “Larger-context language modelling with recurrent neural network,” *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1319–1329, 2016.

[8] B. Liu and I. Lane, “Dialogue context language modeling with recurrent neural networks,” *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5715–5719, 2017.

[9] R. Masumura, T. Tanaka, A. Ando, H. Masataki, and Y. Aono, “Role play dialogue aware language models based on conditional hierarchical recurrent encoder-decoder,” *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1259–1263, 2018.

[10] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” *Computer Speech and Language*, vol. 13, pp. 359–393, 1999.

[11] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, “The ATIS spoken language systems pilot corpus,” *In Proc. DARPA speech and natural language workshop*, pp. 96–101, 1990.

[12] Y. He and S. Young, “A data-driven spoken language understanding system,” *In Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 583–588, 2003.

[13] C. Raymond and G. Riccardi, “Generative and discriminative algorithms for spoken language understanding,” *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1605–1608, 2007.

[14] K. Yao, G. Zweig, M.-Y. Hwang, Y. Shi, and D. Yu, “Recurrent neural networks for language understanding,” *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2524–2528, 2013.

[15] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Yur, X. He, L. Heck, G. Tur, and D. Yu, “Using recurrent neural networks for slot filling in spoken language understanding,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 530–539, 2015.

[16] S. Ravuri and A. Stolcke, “Recurrent neural network and LSTM models for lexical utterance classification,” *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 135–139, 2015.

[17] S. Ravuri and A. Stolcke, “A comparative study of neural network models for lexical intent classification,” *In Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 368–374, 2015.

[18] S. Ravuri and A. Stolcke, “A comparative study of recurrent neural network models for lexical domain classification,” *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6075–6079, 2016.

[19] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. G. Simonsen, and J.-Y. Nie, “A hierarchical recurrent encoder-decoder for generative context-aware query suggestion,” *In Proc. ACM International on Conference on Information and Knowledge Management (CIKM)*, pp. 553–562, 2015.

[20] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, “Building end-to-end dialogue systems using generative hierarchical neural network models,” *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3776–3783, 2016.

[21] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, “A hierarchical latent variable encoder-decoder model for generating dialogues,” *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3295–3301, 2017.

[22] R. Masumura, T. Tanaka, T. Moriya, Y. Shinohara, T. Oba, and Y. Aono, “Large context end-to-end automatic speech recognition via extension of hierarchical recurrent encoder-decoder models,” *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5661–5665, 2019.

[23] B. Liu and I. Lane, “Joint online spoken language understanding and language modeling with recurrent neural networks,” *In Proc. Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pp. 22–30, 2016.

[24] T.-H. Wen, M. Gasic, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young, “Semantically conditioned LSTM-based natural language generation for spoken dialogue systems,” *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1711–1721, 2015.

[25] H. Kumar, A. Agarwal, and S. Joshi, “Dialogue-act-driven conversation model: An experimental study,” *In Proc. International Conference on Computational Linguistics (COLING)*, pp. 1246–1256, 2018.

[26] R. Masumura, Y. Ijima, T. Asami, H. Masataki, and R. Higashinaka, “Neural ConfNet classification: Fully neural network based spoken utterance classification using word confusion networks,” *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6039–6043, 2018.

[27] R. Masumura, T. Tanaka, R. Higashinaka, H. Masataki, and Y. Aono, “Multi-task and multi-lingual joint learning of neural lexical utterance classification based on partially-shared modeling,” *In Proc. International Conference on Computational Linguistics (COLING)*, pp. 3586–3596, 2018.

[28] R. Masumura, Y. Shinohara, R. Higashinaka, and Y. Aono, “Adversarial training for multi-task and multi-lingual joint modeling of utterance intent classification,” *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 633–639, 2018.

[29] T. Hori, C. Hori, Y. Minami, and A. Nakamura, “Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition,” *IEEE transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1352–1365, 2007.

[30] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, “Hierarchical attention networks for document classification,” *In Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1480–1489, 2016.

[31] Y. W. Teh, “A hierarchical Bayesian language model based on Pitman-Yor processes,” *In Proc. Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL)*, pp. 985–992, 2006.

[32] Y. Gal and Z. Ghahramani, “A theoretically grounded application of dropout in recurrent neural networks,” *In Proc. International Conference on Neural Information Processing System (NIPS)*, pp. 1027–1035, 2016.