

HIERARCHICAL KNOWLEDGE DISTILLATION FOR DIALOGUE SEQUENCE LABELING

Shota Orihashi, Yoshihiro Yamazaki, Naoki Makishima, Mana Ihori,
Akihiko Takashima, Tomohiro Tanaka, Ryo Masumura

NTT Media Intelligence Laboratories, NTT Corporation, Japan

ABSTRACT

This paper presents a novel knowledge distillation method for dialogue sequence labeling. Dialogue sequence labeling is a supervised learning task that estimates labels for each utterance in the target dialogue document, and is useful for many applications such as dialogue act estimation. Accurate labeling is often realized by a hierarchically-structured large model consisting of utterance-level and dialogue-level networks that capture the contexts within an utterance and between utterances, respectively. However, due to its large model size, such a model cannot be deployed on resource-constrained devices. To overcome this difficulty, we focus on knowledge distillation which trains a small model by distilling the knowledge of a large and high performance teacher model. **Our key idea is to distill the knowledge while keeping the complex contexts captured by the teacher model.** To this end, **the proposed method, hierarchical knowledge distillation, trains the small model by distilling not only the probability distribution of the label classification, but also the knowledge of utterance-level and dialogue-level contexts trained in the teacher model by training the model to mimic the teacher model's output in each level.** Experiments on dialogue act estimation and call scene segmentation demonstrate the effectiveness of the proposed method.

Index Terms— Knowledge distillation, dialogue sequence labeling, dialogue act estimation, call scene segmentation

1. INTRODUCTION

With the progress of automatic speech recognition technologies, expectations for the understanding and utilization of linguistic information present in human-to-human dialogs are increasing. For example, by understanding contact center telephone dialogue documents, a service for discovering customer needs and issues with the center has been developed [1–6].

In this paper, we focus on utterance-level dialogue sequence labeling, a key component in dialogue document understanding. Dialogue sequence labeling is often modeled as a supervised learning task that estimates labels for each utterance when given a dialogue document; it is useful in many

applications such as topic segmentation [7–9], dialogue act estimation [10–15], and call scene segmentation [16–18]. To understand dialogue documents, it is necessary to consider who spoke what and in what order. Therefore, these techniques often adopt a hierarchically-structured model consisting of an utterance-level network and a dialogue-level network to capture contexts not only within an utterance but also between utterances [16]. In addition, an effective self-supervised pretraining method using only unlabeled data has been proposed [18].

Capturing dialogue documents precisely demands a large number of parameters for both the utterance-level and the dialogue-level networks. However, label inference using such large models requires a rich computation environment. Unfortunately, it is difficult to prepare such an environment, especially when we need to process multiple inferences in parallel or process inference on a device with low computing power such as a mobile device. Therefore, using a large model hinders the adoption of various services.

To overcome the difficulties created by using large models, we focus on knowledge distillation; a small student model with just a few parameters is trained by distilling the knowledge of a large and high performance teacher model so replicate the teacher's performance [19–21]. In recent years, knowledge distillation techniques have been successful in the natural language processing field; examples include neural machine translation [22–24] and compression of bidirectional encoder representations from Transformers (BERT) [25–30]. The strength of knowledge distillation is that the student model can be trained to mimic the behavior of the teacher model. For dialogue sequence labeling, it is especially important to mimic the behavior of the teacher model faithfully to keep full performance while reducing the model size because dialogue sequence labeling is a task in which complex contexts at the utterance-level and the dialogue-level must be precisely captured. Knowledge distillation is seen as potentially able to overcome the difficulty of using a large model for dialogue sequence labeling, but no truly effective knowledge distillation technique for dialogue sequence labeling has been described so far.

In this paper, we propose a novel knowledge distillation method for dialogue sequence labeling. Our key idea is to

train a small student model by distilling the knowledge of utterance-level and dialogue-level contexts while retaining the complex contexts captured by the large teacher model. To this end, our method, hierarchical knowledge distillation, not only trains the student model so that the probability distribution of its output labels approaches that of the teacher model, but also trains the student model so that the outputs of the utterance-level and the dialogue-level networks of the student model approach those of the teacher model. By distilling the knowledge of complex contexts from the large teacher model via hierarchical knowledge distillation, our method enables us to train the small student model without losing the ability to capture contexts within an utterance and between utterances as captured by the teacher model. Our experiments on dialogue act estimation using the switchboard dialogue act (SwDA) corpus [31, 32] and call scene segmentation using a simulated Japanese contact center dialogue dataset demonstrate the effectiveness of the proposed method.

Our contributions are summarized as follows:

- We provide an effective knowledge distillation method for dialogue sequence labeling that distills not only the probability distribution of the label classification [20], but also the knowledge of utterance-level and dialogue-level contexts. To the best of our knowledge, this is the first method to achieve knowledge distillation for dialogue sequence labeling.
- We conduct ablation experiments on dialogue act estimation and call scene segmentation tasks that analyze the effectiveness of the proposed method. We also provide the results achieved by combining self-supervised pretraining [18] and the proposed method.

2. RELATED WORK

2.1. Utterance-level dialogue sequence labeling

Utterance-level dialogue sequence labeling is being used for topic segmentation [7-9], dialogue act estimation [10-15], and call scene segmentation [16-18]. Hierarchically structured models consisting of utterance-level and dialogue-level neural networks are often used to efficiently capture contexts within an utterance and between utterances, and an effective self-supervised pretraining method has been proposed [18]. If a hierarchical model is used for dialogue sequence labeling, a large number of parameters are needed to train a model that offers high accuracy. In this paper, to train a highly accurate model that has just a few parameters, we introduce a knowledge distillation technique to utterance-level dialogue sequence labeling.

2.2. Knowledge distillation

Knowledge distillation is a technique to train a small student model efficiently by utilizing the knowledge of a large and

high performance teacher model without significant performance loss [19]. One of the early methods trains the student model so that the probability distribution of the output label of the student model approaches that of the teacher model by utilizing soft target loss [20]. Another method trains the student model so that the hidden layers' outputs of the student model approach those of the teacher model [21]. Successful knowledge distillation techniques have recently been reported in the natural language processing field [22-24, 27-30]. In this paper, we propose a knowledge distillation method for dialogue sequence labeling. To retain the ability to capture contexts within an utterance and between utterances, we train the student model so that the outputs of the utterance-level and the dialogue-level networks of the student model approach those of the teacher model.

3. UTTERANCE-LEVEL DIALOGUE SEQUENCE LABELING

This section describes utterance-level dialogue sequence labeling in dialogue documents. This task estimates utterance-level label sequence $\mathbf{Y} = \{y_1, \dots, y_T\}$ from input utterance sequence $\mathbf{X} = \{x_1, \dots, x_T\}$ using neural networks, where the t -th utterance, x_t , consists of token sequence $\{w_{t,1}, \dots, w_{t,K_t}\}$; K_t is number of tokens in the t -th utterance. The t -th label, y_t , is an element of \mathcal{Y} , where \mathcal{Y} is the set of labels. Label types are task dependent, for example dialogue act labels for dialogue act estimation and call scene labels for call scene segmentation.

In our dialogue sequence labeling, y_t is estimated from $\{x_1, \dots, x_t\}$ in an online manner. For this, conditional probability $P(y_t | x_1, \dots, x_t, \Theta)$ is modeled, where Θ represents a model parameter. The t -th label can be categorized by:

$$\hat{y}_t = \arg \max_{y_t \in \mathcal{Y}} P(y_t | x_1, \dots, x_t, \Theta). \quad (1)$$

In this paper, we assume that $P(y_t | x_1, \dots, x_t, \Theta)$ is modeled by the Transformer encoder [25] and hierarchical long short-term memory recurrent neural networks (LSTM-RNNs). Figure 1 shows the structure of the labeling model.

In the utterance-level network, each token is first converted into a continuous vector representation. The continuous vector representation of the k -th token in the t -th utterance is given by:

$$\mathbf{w}_{t,k} = \text{Embedding}(w_{t,k}; \theta^w), \quad (2)$$

where $\text{Embedding}()$ is a linear transformational function that embeds a symbol into a continuous vector, and θ^w is the trainable parameter. Continuous vectors $\mathbf{w}_{t,k}$ are then converted into $\mathbf{q}_{t,k}$ for input to the Transformer encoder block as:

$$\mathbf{q}_{t,k} = \text{AddPosEnc}(\mathbf{w}_{t,k}), \quad (3)$$

where $\text{AddPosEnc}()$ is a function that adds a continuous vector in which position information is embedded. The

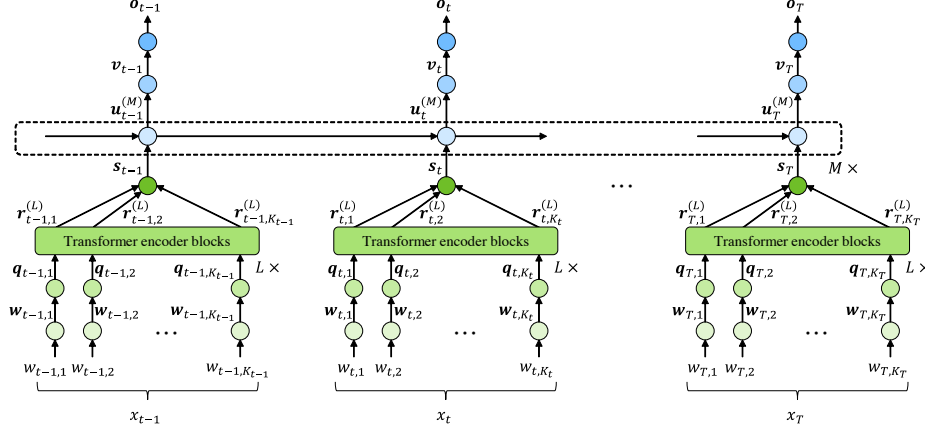


Fig. 1. Structure of dialogue sequence labeling model.

Transformer encoder forms hidden representations $\mathbf{R}_t^{(L)} = \{\mathbf{r}_{t,1}^{(L)}, \dots, \mathbf{r}_{t,K_t}^{(L)}\}$ from $\mathbf{Q}_t = \{\mathbf{q}_{t,1}, \dots, \mathbf{q}_{t,K_t}\}$ by using L Transformer encoder blocks. The l -th Transformer encoder block forms the l -th hidden representations $\mathbf{R}_t^{(l)}$ from the lower layer outputs $\mathbf{R}_t^{(l-1)}$ as:

$$\mathbf{R}_t^{(l)} = \text{TransformerEnc}(\mathbf{R}_t^{(l-1)}; \theta^r), \quad (4)$$

where $\mathbf{R}_t^{(0)} = \mathbf{Q}_t$, and $\text{TransformerEnc}()$ is a Transformer encoder block that consists of a scaled dot product multi-head self-attention layer and a position-wise feed-forward network [25]. θ^r represents the trainable parameter. The hidden representations are then summarized as an utterance representation by a self-attention mechanism [33]. The t -th utterance continuous representation is calculated as:

$$\mathbf{s}_t = \text{SelfAttention}(\mathbf{r}_{t,1}^{(L)}, \dots, \mathbf{r}_{t,K_t}^{(L)}; \theta^s), \quad (5)$$

where $\text{SelfAttention}()$ is a transformational function that converts into a fixed-length vector by the self-attention mechanism; θ^s is the trainable parameter.

In the dialogue-level network, interaction information from start-of-dialogue to the t -th utterance is incrementally embedded into a continuous vector representation. The t -th continuous vector representation that embeds all dialogue context sequential information up to the t -th utterance $\mathbf{u}_t^{(M)}$ is calculated from $\{\mathbf{s}_1, \dots, \mathbf{s}_t\}$ by using M LSTM-RNN layers. The m -th LSTM-RNN layer forms the m -th hidden representations $\mathbf{u}_t^{(m)}$ from the lower layer outputs $\{\mathbf{u}_1^{(m-1)}, \dots, \mathbf{u}_t^{(m-1)}\}$ as:

$$\mathbf{u}_t^{(m)} = \text{LSTM}(\mathbf{u}_1^{(m-1)}, \dots, \mathbf{u}_t^{(m-1)}; \theta^u), \quad (6)$$

where $\mathbf{u}_t^{(0)} = \mathbf{s}_t$, $\text{LSTM}()$ is a function of the unidirectional LSTM-RNN layer, and θ^u represents the trainable parameter.

In the output layer, predictive probabilities of the labels for the t -th utterance \mathbf{o}_t are defined using logits $\mathbf{v}_t =$

$[v_{t,1}, \dots, v_{t,|\mathcal{Y}|}]$ as:

$$\mathbf{v}_t = \text{FeedForward}(\mathbf{u}_t^{(M)}; \theta^v), \quad (7)$$

$$\mathbf{o}_t = \text{Softmax}(\mathbf{v}_t), \quad (8)$$

where $\text{FeedForward}()$ is a function of a fully-connected feed forward neural network, θ^v is a trainable parameter, and $\text{Softmax}()$ is a softmax function. \mathbf{o}_t corresponds to $P(y_t | x_1, \dots, x_t, \Theta)$.

The model parameters $\Theta = \{\theta^w, \theta^r, \theta^s, \theta^u, \theta^v\}$ can be optimized by preparing training dataset $\mathcal{D} = \{(\mathbf{X}^1, \bar{\mathbf{Y}}^1), \dots, (\mathbf{X}^N, \bar{\mathbf{Y}}^N)\}$, where \mathbf{X}^n and $\bar{\mathbf{Y}}^n$ are input utterance sequence and reference utterance-level label sequence in the n -th dialogue, respectively. In this case, cross-entropy loss, named hard target loss, is computed by:

$$\mathcal{L}_{\text{HT}} = -\frac{1}{N} \sum_{n=1}^N \left(\frac{1}{T_n} \sum_{t=1}^{T_n} \sum_{y \in \mathcal{Y}} \bar{o}_{t,y}^n \log o_{t,y}^n \right), \quad (9)$$

where $\bar{\mathbf{o}}_t^n = [\bar{o}_{t,1}^n, \dots, \bar{o}_{t,|\mathcal{Y}|}^n]$ and $\mathbf{o}_t^n = [o_{t,1}^n, \dots, o_{t,|\mathcal{Y}|}^n]$ are the reference and estimated probabilities of label y for the t -th utterance in the n -th dialogue, respectively. T_n is the number of utterances in the n -th dialogue. Note that $\bar{\mathbf{o}}_t^n$ is a one-hot vector.

When self-supervised pretraining [18] is utilized, parameters $\{\theta^w, \theta^r, \theta^s, \theta^u\}$ are initialized by pretraining using unlabeled data, and then parameters Θ are optimized with \mathcal{L}_{HT} in the same way as above.

4. PROPOSED METHOD

This section details our proposed knowledge distillation method for utterance-level dialogue sequence labeling. The main idea of the proposed method, hierarchical knowledge distillation, is to train a small student model by distilling

the knowledge of utterance-level and dialogue-level complex contexts captured by a large teacher model. To this end, our method trains the student model so that the probability distribution of the output labels approaches that of the large teacher model by utilizing soft target loss [20]. Not only that, our method trains the student model so that the outputs of the utterance-level and the dialogue-level networks of the student model approach those of the teacher model. By distilling the knowledge to capture complex contexts trained in the teacher model, our method efficiently trains the small student model so that it offers high accuracy.

Figure 2 outlines the proposed method. Our hierarchical knowledge distillation proposal trains the student model by distilling the knowledge of the teacher model by optimizing the student model using a loss function that is a combination of four components: hard target loss (9), soft target loss, utterance-level context loss, and dialogue-level context loss.

4.1. Soft target loss

Soft target loss aims to bring the student model’s probability distribution of the output labels closer to that of the teacher model [20]. To calculate soft target loss, the probability distribution is computed from logits \mathbf{v}_t by:

$$\mathbf{z}_t = \text{SoftmaxWithTemperature}(\mathbf{v}_t; \tau), \quad (10)$$

where $\text{SoftmaxWithTemperature}()$ is a softmax function with temperature, and τ is a hyper-parameter that represents temperature [20]. Soft target loss is thus defined as:

$$\mathcal{L}_{\text{ST}} = -\frac{\tau^2}{N} \sum_{n=1}^N \left(\frac{1}{T_n} \sum_{t=1}^{T_n} \sum_{y \in \mathcal{Y}} \tilde{z}_{t,y}^n \log z_{t,y}^n \right), \quad (11)$$

where $\tilde{\mathbf{z}}_t^n = [\tilde{z}_{t,1}^n, \dots, \tilde{z}_{t,|\mathcal{Y}|}^n]$ and $\mathbf{z}_t^n = [z_{t,1}^n, \dots, z_{t,|\mathcal{Y}|}^n]$ are the probabilities of label y for the t -th utterance in the n -th dialogue estimated by the teacher model and the student model, respectively.

4.2. Utterance-level context loss

The proposed method aims to train the student model so that the utterance-level network of the student model mimics that of the teacher model. For this, utterance-level context loss is defined as the difference between the outputs of the utterance-level networks of the student and the teacher models. Utterance-level context loss is defined using mean squared error (MSE) as:

$$\mathcal{L}_{\text{UC}} = \frac{1}{N} \sum_{n=1}^N \left(\frac{1}{T_n} \sum_{t=1}^{T_n} \|\tilde{\mathbf{s}}_t^n - \mathbf{s}_t^n\|_2^2 \right), \quad (12)$$

where $\tilde{\mathbf{s}}_t^n$ and \mathbf{s}_t^n are the t -th utterance continuous vector representations of the teacher model and the student model, respectively. Note that the proposed method assumes that $\tilde{\mathbf{s}}_t^n$ and \mathbf{s}_t^n have equal size.

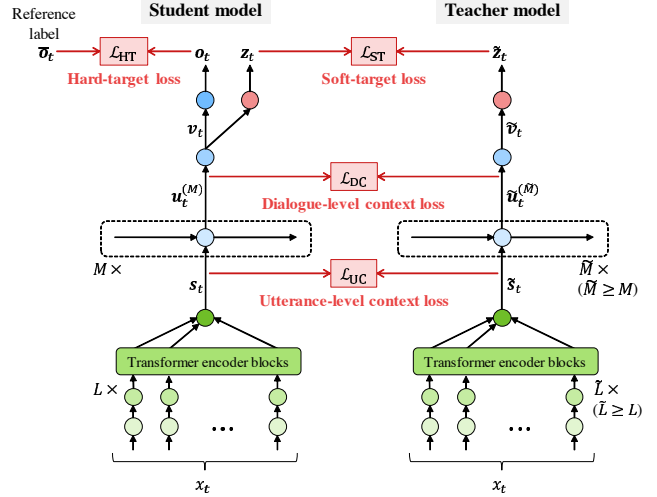


Fig. 2. Outline of the proposed method.

4.3. Dialogue-level context loss

The proposed method also aims to train the student model so that the dialogue-level network of the student model mimics that of the teacher model. For this, dialogue-level context loss is defined as the difference between the outputs of the dialogue-level networks of the student and the teacher models. Dialogue-level context loss is defined using MSE as:

$$\mathcal{L}_{\text{DC}} = \frac{1}{N} \sum_{n=1}^N \left(\frac{1}{T_n} \sum_{t=1}^{T_n} \|\tilde{\mathbf{u}}_t^{(\tilde{M}),n} - \mathbf{u}_t^{(M),n}\|_2^2 \right), \quad (13)$$

where $\tilde{\mathbf{u}}_t^{(\tilde{M}),n}$ and $\mathbf{u}_t^{(M),n}$ are the t -th continuous vector representations that embed all dialogue context sequential information up to the t -th utterance, of the teacher model and the student model, respectively. \tilde{M} is the number of layers for the teacher model’s dialogue-level network ($\tilde{M} \geq M$). Note that the proposed method assumes that $\tilde{\mathbf{u}}_t^{(\tilde{M}),n}$ and $\mathbf{u}_t^{(M),n}$ have equal size.

4.4. Training

For training, the parameters of the student model Θ are optimized by using training dataset \mathcal{D} with the application of combined loss. Combined loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{HT}} + \lambda \mathcal{L}_{\text{ST}} + \alpha \mathcal{L}_{\text{UC}} + \beta \mathcal{L}_{\text{DC}}, \quad (14)$$

where λ , α , and β are hyper-parameters.

When self-supervised pretraining [18] is utilized, parameters $\{\theta^w, \theta^r, \theta^s, \theta^u\}$ are initialized in the pretraining using unlabeled data, and then parameters Θ are optimized with \mathcal{L} in the same way as above.

Table 1. Details of the simulated Japanese contact center dialogue dataset.

Business type	#calls	#utterances	#tokens
Finance	59	6,081	55,933
Internet provider	57	3,815	47,668
Government unit	73	5,617	48,998
Mail-order	56	4,938	46,574
PC repair	55	6,263	55,101
Mobile phone	61	5,738	51,061
All	361	32,452	305,351

5. EXPERIMENT

5.1. Datasets

We evaluated the proposed knowledge distillation method on two dialogue sequence labeling tasks: dialogue act estimation and call scene segmentation.

For dialogue act estimation, we used SwDA corpus [31, 32]. SwDA corpus consists of 1,155 telephone calls between two people with no specific topic; it holds 205K utterances and 1.4M tokens. Each utterance is tagged with one dialogue act label, and each dialogue act label summarizes syntactic, semantic and pragmatic information about the corresponding utterance. SwDA corpus originally used over 200 kinds of dialogue act labels, but labels are usually clustered into 43 label-types such as *statement-non-opinion*, *acknowledge (backchannel)*, *statement-opinion*, and *agree/accept* [31]. Following this, we used the clustered 43 dialogue act label-types. We split the SwDA corpus into 1,115 training dialogues and 19 test dialogues following the conventional approach [32].

For call scene segmentation, we used a simulated Japanese contact center dialogue dataset consisting of 361 dialogues in six business fields. Details of the dataset are shown in Table 1. One dialogue means one telephone call between one operator and one customer; all utterances were manually transcribed. Each dialogue was divided into speech units using LSTM-RNN-based speech activity detection [34] trained from various Japanese speech samples. We manually set five labels to define call scenes: *opening*, *requirement confirmation*, *response*, *customer confirmation*, and *closing* [16]. We split the dataset into 324 training dialogues and 37 test dialogues. Only for call scene segmentation, we additionally prepared 4,000 unlabeled dialogues collected from various Japanese contact centers, and an additional 500 million unlabeled Japanese sentences collected from the Web to utilize self-supervised pretraining [18].

5.2. Setups

We first trained the teacher model from the labeled dataset. For dialogue act estimation, we trained the teacher model

Table 2. Details of the models and sizes.

	L	M	#units	#parameters
Teacher	8	2	2,048	13.11M
S1	1	1	256	2.47M
S2	2	2	512	3.65M

Table 3. Results in terms of classification accuracy for dialogue act estimation (%).

	S1	S2
Teacher (common to S1 and S2)	72.79	72.79
Student		
Baseline	71.43	71.75
Knowledge distillation w/o $\mathcal{L}_{UC}, \mathcal{L}_{DC}$	72.40	72.44
Knowledge distillation w/o \mathcal{L}_{UC}	72.54	72.66
Knowledge distillation w/o \mathcal{L}_{DC}	72.60	72.67
Proposed knowledge distillation	72.69	72.81

using only labeled dataset from scratch. For call scene segmentation, we trained the teacher model utilizing the self-supervised pretraining [18] using unlabeled dialogues and unlabeled sentences before training by using a labeled dataset. To evaluate the proposed knowledge distillation, we constructed student models by the following two training procedures. In **Baseline**, training used only the labeled dataset from scratch. In **Knowledge distillation**, training used only the labeled dataset and acquired the knowledge of the teacher model by utilizing the knowledge distillation proposal. Only for call scene segmentation, we constructed additional student models by the following two training procedures. In **Pretraining**, we utilized the self-supervised pretraining as used by the teacher model, and then trained using the labeled dataset. In **Pretraining + Knowledge distillation**, we utilized the self-supervised pretraining as used by the teacher model, and then trained utilizing the knowledge distillation proposal using the labeled dataset.

Our experiments examined student models of two sizes: S1 and S2. Details of the models and their size, together with the teacher model, are shown in Table 2. In Table 2, L and M are the number of layers for utterance-level network and the dialogue-level network, respectively. Also, “#units” represents the inner outputs in the position-wise feed forward networks for Transformer encoder blocks. For all models, we defined the token vector representation as a 256-dimensional vector, and unit size of LSTM-RNN was set to 256. For the Transformer encoder blocks, the dimensions of the output continuous representations were set to 256, and the number of heads in the multi-head attentions was set to 4. Note that the teacher model is common to S1 and S2.

For training, the mini-batch size was set to five dialogues. The optimizer was RAdam [35] with the default setting. For knowledge distillation, parameters τ , λ , α , and β were set to 5.0, 0.1, 0.05, and 0.05, respectively. We constructed five

Table 4. Results in terms of classification accuracy for call scene segmentation (%).

		S1	S2
Teacher (common to S1 and S2)		91.28	91.28
Student	Baseline	87.22	87.94
	Pretraining	89.05	89.46
	Knowledge distillation w/o $\mathcal{L}_{UC}, \mathcal{L}_{DC}$	87.54	88.23
	Knowledge distillation w/o \mathcal{L}_{UC}	88.83	88.86
	Knowledge distillation w/o \mathcal{L}_{DC}	88.99	89.20
	Proposed knowledge distillation	89.81	91.10
	Pretraining + Knowledge distillation w/o $\mathcal{L}_{UC}, \mathcal{L}_{DC}$	88.65	88.82
	Pretraining + Knowledge distillation w/o \mathcal{L}_{UC}	89.42	89.62
	Pretraining + Knowledge distillation w/o \mathcal{L}_{DC}	89.49	89.86
	Pretraining + Proposed knowledge distillation	89.98	91.26

models by varying the initial parameters, and evaluated their average classification accuracy. Note that a part of the training dialogues was used for early stopping.

5.3. Results

The resulting classification accuracy values for dialogue act estimation are shown in Table 3. In the table, line 1 shows ideal accuracy achieved by the teacher model. Line 2 shows results yielded by training the student models from scratch. The results show that there is a performance gap between line 1 and line 2; this is due to a reduction in the number of parameters, see Table 2. Lines 3–6 show the results of knowledge distillation. Line 3 shows the results yielded by using only hard and soft target losses without utterance-level and dialogue-level context losses, which follows a previous method [20]. The results on line 3 show performance improvements over the baseline, but the improvements were limited. Lines 4 and 5 show the results yielded by applying the knowledge distillation proposal without utterance-level context loss or dialogue-level context loss, respectively. The results on lines 4 and 5 show that the utilization of utterance-level or dialogue-level context loss improved performance compared with line 3. Line 6 shows the results achieved by the knowledge distillation proposal; the proposed method attained the best performance. Especially for S2, the accuracy of the proposed method exceeds that of the teacher model.

The resulting classification accuracy values for call scene segmentation are shown in Table 4. In the table, line 1 shows ideal accuracy values achieved by the teacher model. Line 2 shows results yielded by training the student models from scratch, and line 3 shows results yielded by utilizing the self-supervised pretraining. The results show that there is a performance gap between line 1 and lines 2 and 3 due to parameter reduction. Lines 4–7 show the results of knowledge distillation without pretraining the student models. Line 4 shows that using only hard and soft target losses yielded poor knowledge distillation performance. Lines 5 and 6 show that applying the knowledge distillation proposal without utterance-level con-

text loss or dialogue-level context loss yielded limited performance improvements. Line 7 shows that the knowledge distillation proposal exceeds the performance of the results on lines 4–6. In addition, lines 8–11 show the results yielded by applying knowledge distillation with pretraining. The results on lines 8–10 show that full knowledge distillation performance was not attained when only a part of the loss was used. Note that lines 8–10 demonstrate improved performance compared to lines 4–6 due to pretraining. Line 11 shows that applying the proposed knowledge distillation with pretraining yielded the best performance of all other methods examined. S2 allowed the proposed method to most closely approach the accuracy of the teacher model.

The performance improvements attained by the knowledge distillation proposal are considered to be due to the fact that the proposed method could train the student model without losing the ability of the teacher model to capture contexts within an utterance and between utterances. Our results show that the proposed knowledge distillation method is an effective way of improving performance in small student models.

6. CONCLUSIONS

This paper has proposed a novel knowledge distillation method, hierarchical knowledge distillation, for dialogue sequence labeling. The key advance of our method is to distill the knowledge of the utterance-level and dialogue-level contexts captured by a large teacher model. To this end, our method utilizes utterance-level and dialogue-level context losses so that the outputs of the utterance-level and the dialogue-level networks of the student model approach those of the teacher model. Experiments on dialogue act estimation and call scene segmentation tasks showed that our method allows small student models to achieve better performance and that combining utterance-level and dialogue-level context losses is an effective approach to knowledge distillation for dialogue sequence labeling.

7. REFERENCES

- [1] J. Mamou, D. Carmel, and R. Hoory, “Spoken document retrieval from call-center conversations,” in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2006, pp. 51–58.
- [2] R. J. Byrd, M. S. Neff, W. Teiken, Y. Park, K.-S. F. Cheng, S. C. Gates, and K. Visweswariah, “Semi-automated logging of contact center telephone calls,” in *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, 2008, pp. 133–142.
- [3] R. Higashinaka, Y. Minami, H. Nishikawa, K. Dohsaka, T. Meguro, S. Takahashi, and G. Kikui, “Learning to model domain-specific utterance sequences for extractive summarization of contact center dialogues,” in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2010, pp. 400–408.
- [4] A. Tamura, K. Ishikawa, M. Saikou, and M. Tsuchida, “Extractive summarization method for contact center dialogues based on call logs,” in *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2011, pp. 500–508.
- [5] C. Chastagnol and L. Devillers, “Analysis of anger across several agent-customer interactions in French call centers,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4960–4963.
- [6] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, “Hierarchical LSTMs with joint learning for estimating customer satisfaction from contact center calls,” in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 1716–1720.
- [7] J. Yu, X. Xiao, L. Xie, E. S. Chng, and H. Li, “A DNN-HMM approach to story segmentation,” in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, pp. 1527–1531.
- [8] E. Tsunoo, P. Bell, and S. Renals, “Hierarchical recurrent neural network for story segmentation,” in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2919–2923.
- [9] E. Tsunoo, O. Klejch, P. Bell, and S. Renals, “Hierarchical recurrent neural network for story segmentation using fusion of lexical and acoustic features,” in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 525–532.
- [10] Q. H. Tran, I. Zukerman, and G. Haffari, “A hierarchical neural model for learning sequences of dialogue acts,” in *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2017, vol. 1, pp. 428–437.
- [11] H. Kumar, A. Agarwal, R. Dasgupta, and S. Joshi, “Dialogue act sequence labeling using hierarchical encoder with CRF,” in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018, pp. 3440–3447.
- [12] Z. Chen, R. Yang, Z. Zhao, D. Cai, and X. He, “Dialogue act recognition via CRF-attentive structured network,” in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2018, pp. 225–234.
- [13] W. Jiao, H. Yang, I. King, and M. R. Lyu, “HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 397–406.
- [14] V. Raheja and J. Tetreault, “Dialogue act classification with context-aware self-attention,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 3727–3733.
- [15] Y. Yu, S. Peng, and G. H. Yang, “Modeling long-range context for concurrent dialogue acts recognition,” in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 2019, pp. 2277–2280.
- [16] R. Masumura, S. Yamada, T. Tanaka, A. Ando, H. Kamiyama, and Y. Aono, “Online call scene segmentation of contact center dialogues based on role aware hierarchical LSTM-RNNs,” in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 811–815.
- [17] S. Orihashi, M. Ihori, T. Tanaka, and R. Masumura, “Unsupervised domain adaptation for dialogue sequence labeling based on hierarchical adversarial training,” in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 1575–1579.
- [18] R. Masumura, N. Makishima, M. Ihori, A. Takashima, T. Tanaka, and S. Orihashi, “Large-context conversational representation learning: Self-supervised learning for conversational documents,” in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 1012–1019.

- [19] C. Bucilua, R. Caruana, and A. Niculescu-Mizil, “Model compressions,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006, pp. 535–541.
- [20] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in *Proceedings of the Deep Learning and Representation Learning Workshop, NIPS*, 2014.
- [21] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “FitNets: Hints for thin deep nets,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
- [22] Y. Kim and A. M. Rush, “Sequence-level knowledge distillation,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016, pp. 1317–1327.
- [23] M. Freitag, Y. Al-Onaizan, and B. Sankaran, “Ensemble distillation for neural machine translation,” *arXiv preprint arXiv:1702.01802*, 2017.
- [24] X. Tan, Y. Ren, D. He, T. Qin, Z. Zhao, and T.-Y. Liu, “Multilingual neural machine translation with knowledge distillation,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.
- [26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional Transformers for language understanding,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186.
- [27] S. Sun, Y. Cheng, Z. Gan, and J. Liu, “Patient knowledge distillation for BERT model compression,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 4323–4332.
- [28] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” in *Proceedings of the Workshop on Energy Efficient Machine Learning and Cognitive Computing, NeurIPS*, 2019.
- [29] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, “TinyBERT: Distilling BERT for natural language understanding,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4163–4174.
- [30] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, “MobileBERT: A compact task-agnostic BERT for resource-limited devices,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 2158–2170.
- [31] D. Jurafsky, E. Shriberg, and D. Biasca, “Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13,” Tech. Rep., University of Colorado at Boulder, 1997.
- [32] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, “Dialogue act modeling for automatic tagging and recognition of conversational speech,” *Computational linguistics*, vol. 26, no. 3, pp. 339–371, 2000.
- [33] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2016, pp. 1480–1489.
- [34] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, “Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 483–487.
- [35] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.