Multi-task and Multi-lingual Joint Learning of Neural Lexical Utterance Classification based on Partially-shared Modeling

Ryo Masumura, Tomohiro Tanaka, Ryuichiro Higashinaka,
Hirokazu Masataki and Yushi Aono
NTT Media Intelligence Laboratories, NTT Corporation, Japan
ryou.masumura.ba@hco.ntt.co.jp

Abstract

This paper is an initial study on multi-task and multi-lingual joint learning for lexical utterance classification. A major problem in constructing lexical utterance classification modules for spoken dialogue systems is that individual data resources are often limited or unbalanced among tasks and/or languages. Various studies have examined joint learning using neural-network-based shared modeling; however, previous joint learning studies focused on either cross-task or cross-lingual knowledge transfer. In order to simultaneously support both multi-task and multi-lingual joint learning, our idea is to explicitly divide state-of-the-art neural lexical utterance classification into language-specific components that can be shared between different tasks and task-specific components that can be shared between different languages. In addition, in order to effectively transfer knowledge between different task data sets and different language data sets, this paper proposes a partially-shared modeling method that possesses both shared components and components specific to individual data sets. We demonstrate the effectiveness of the proposed adversarial training using Japanese and English data sets with three different lexical utterance classification tasks.

1 Introduction

Modern spoken dialogue systems use multiple lexical utterance classification modules that can detect dialogue act (Stolcke et al., 2000; Khanpour et al., 2016), intent (Tur et al., 2011), domain (Xu and Sarikaya, 2014), question type (Wu et al., 2005), etc. to properly understand natural languages. The modules are typically trained using the machine learning technologies developed for individual language-specific systems (Higashinaka et al., 2014). A common issue is that the data resources for such individual training are often limited or unbalanced among different tasks and/or different languages.

For modeling lexical utterance classification, various modeling methods have been examined. Recently, neural lexical utterance classification, which is a fully neural-network-based modeling method, has demonstrated substantial performance without the use of manual feature engineering. The networks include long short-term memory recurrent neural networks (LSTM-RNNs) (Ravuri and Stolcke, 2015b; Ravuri and Stolcke, 2015a; Ravuri and Stolcke, 2016), convolution neural networks (Kim, 2014), and more advanced networks (Zhou et al., 2016a; Yang et al., 2016; Sawada et al., 2017).

In addition, neural networks are suitable for performing joint learning; the paucity of data is tackled by transferring knowledge between different tasks or different languages. Various joint learning methods have been examined for leveraging different tasks or different language data sets in the natural language processing field. Multi-task joint learning can transfer knowledge between tasks by sharing task-invariant layers (Collobert and Weston, 2008; Liu et al., 2015; Liu et al., 2016c; Zhang and Weng, 2016). In lexical utterance classification, multi-task joint learning has been shown to effectively improve individual tasks (Liu et al., 2016b; Liu et al., 2016a; Liu et al., 2017). In addition, multi-lingual joint learning can transfer knowledge between languages, mainly from the resource-rich language to the resource-poor language. The knowledge transfer is achieved by learning common semantic representations for different

Proceedings of the 27th International Conference on Computational Linguistics, pages 3586–3596
Santa Fe, New Mexico, USA, August 20-26, 2018.

languages. Usually, word-aligned or sentence-aligned parallel data sets are employed for joint learning (Guo et al., 2016; Duong et al., 2016)

However, most existing joint learning approaches focus only on cross-tasks or cross-lingual knowledge transfer. In fact, task-aligned multi-lingual data sets have been rarely utilized for joint learning (Mogadala and Rettinger, 2016; Pappas and Popescu-Belis, 2017). We can expect to enhance lexical utterance classification performance by achieving effective knowledge transfer among both different tasks and different languages.

In this paper, we propose multi-task and multi-lingual joint learning; it can enhance neural lexical utterance classification by flexibly transferring knowledge among both different tasks and different languages. The proposed method is closely related to multi-task sequence-to-sequence learning (Luong et al., 2016) including many-to-many neural machine translation (Firat et al., 2016; Firat et al., 2017; Schwenk and Douze, 2017). While input and output components are easily distinguished in sequence-to-sequence models, neural lexical utterance classification methods are not explicitly divided into input and output components. Our idea is to divide the neural lexical utterance classification into two components. The language-specific components converts words to hidden representations while task-specific components convert the hidden representations into prediction probabilities. The former can be shared between different tasks and the latter can be shared between different languages.

In addition, in order to perform effective joint learning by simultaneously using multi-task and multi-lingual data sets, this paper examines two joint modeling strategies. The fully-shared modeling strategy is often used in various joint learning methods. Fully-shared modeling can share knowledge between tasks or languages based on task-invariant components and language-invariant components, however, classification performance in some data sets is deteriorated. Therefore, this paper proposes the partially-shared modeling strategy; it introduces not only shared components between tasks or languages but also exclusive components that handle task-language combinations. It can be expected that the latter are suitable for multi-task and multi-lingual joint learning since they allow us to accumulate just shareable knowledge.

Main contributions are summarized as follows.

- This paper proposes multi-task and multi-lingual joint learning of neural lexical utterance classification. For neural lexical utterance classification, we introduce a state-of-the-art model structure based on bidirectional LSTM-RNNs with a self-attention mechanism. We demonstrate the superiority of multi-task and multi-lingual joint learning over multi-task joint learning and multi-lingual joint learning.

- This paper proposes partially-shared modeling for multi-task and multi-lingual joint learning. Partially-shared modeling can be utilized for various neural network based joint learning schemes. We demonstrate the superiority of partially-shared modeling over fully-shared modeling. In addition, we reveal the properties of partially-shared modeling.

- This paper introduces a new corpus for evaluating multi-task and multi-lingual lexical utterance classification methods. The corpus includes Japanese and English data sets with three different lexical utterance classification tasks. The tasks are dialogue act classification, extended named entity classification (Sekine and Nobata, 2004; Higashinaka et al., 2012), and question type classification.

2 Neural Lexical Utterance Classification

This section details neural lexical utterance classification. Lexical utterance classification is the problem of determining the correct label $l \in \{l_1, \cdots, l_K\}$ of given utterance $\mathcal{W} = \{w_1, \cdots, w_T\}$. In neural lexical utterance classification, conditional probabilities for each label given utterance, $P(l|\mathcal{W}, \mathbf{\Theta})$, can be modeled by neural networks in an end-to-end manner where $\mathbf{\Theta}$ is the model parameter. Various model structures can be used for neural lexical utterance classification. In this work, we use bidirectional LSTM-RNNs (BLSTM-RNNs) with a self-attention mechanism (Yang et al., 2016; Zhou et al., 2016b).

2.1 Modeling

In our neural lexical utterance classification, each word in input utterance $\mathcal{W}$ is first converted into a continuous representation. The continuous representation of the t-th word is defined as:

$$\boldsymbol{w}_t = \texttt{EMBED}(w_t; \boldsymbol{\theta}_{\texttt{w}}), \tag{1}$$

where $\texttt{EMBED}()$ is a linear transformational function to embed a word into a continuous vector and $\boldsymbol{\theta}_{\texttt{w}}$ is the trainable parameter. Next, each word representation is converted into a hidden representation that takes neighboring word context information into consideration. The hidden representation for the t-th word is calculated as:

$$\boldsymbol{h}_t = \texttt{BLSTM}(\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_T\}, t; \boldsymbol{\theta}_{\texttt{h}}), \tag{2}$$

where $\texttt{BLSTM}()$ is a function of the BLSTM-RNN layer and $\boldsymbol{\theta}_{\texttt{h}}$ is the trainable parameter.

The hidden representations are summarized as a sentence representation by using a self-attention mechanism that can consider the importance of individual hidden representations. The sentence continuous representation $\boldsymbol{s}$ is calculated as:

$$\boldsymbol{z}_t = \texttt{tanh}(\boldsymbol{h}_t; \boldsymbol{\theta}_{\texttt{z}}), \tag{3}$$

$$\boldsymbol{s} = \sum_{t=1}^{T} \frac{\exp(\boldsymbol{z}_t^\top \bar{\boldsymbol{z}})}{\Sigma_{j=1}^{T} \exp(\boldsymbol{z}_j^\top \bar{\boldsymbol{z}})} \boldsymbol{h}_t, \tag{4}$$

where $\texttt{tanh}()$ is a non-linear transformational function with tanh activation, $\boldsymbol{\theta}_{\texttt{z}}$ is the trainable parameter, and $\bar{\boldsymbol{z}}$ is a trainable context vector, which is used for measuring the importance of individual hidden representations. The output layer produces predicted probabilities $\boldsymbol{O}$ by:

$$\boldsymbol{o} = \texttt{LINEAR}(\boldsymbol{s}; \boldsymbol{\theta}_{\texttt{o}}), \tag{5}$$

$$\boldsymbol{O} = \texttt{SOFTMAX}(\boldsymbol{o}), \tag{6}$$

where $\texttt{LINEAR}()$ is a linear transformational function and $\boldsymbol{\theta}_{\texttt{o}}$ is the trainable parameter. $\texttt{SOFTMAX}()$ is a softmax activation to convert $\boldsymbol{o}$ into predicted probabilities. The k-th dimension in $\boldsymbol{O}$ corresponds to $P(l_k|\mathcal{W}, \boldsymbol{\Theta})$, and $\boldsymbol{\Theta}$ corresponds to $\{\boldsymbol{\theta}_{\texttt{w}}, \boldsymbol{\theta}_{\texttt{h}}, \boldsymbol{\theta}_{\texttt{z}}, \bar{\boldsymbol{z}}, \boldsymbol{\theta}_{\texttt{o}}\}$.

2.2 Optimization

The parameter can be optimized by minimizing the cross entropy between reference and estimated probabilities:

$$\hat{\boldsymbol{\Theta}} = \operatorname*{argmin}_{\boldsymbol{\Theta}} - \sum_{\mathcal{W} \in \mathcal{D}} \sum_{k=1}^{K} \hat{\boldsymbol{O}}_{\mathcal{W}}^{l_k} \log \boldsymbol{O}_{\mathcal{W}}^{l_k}, \tag{7}$$

where $\hat{\boldsymbol{O}}_{\mathcal{W}}^{l_k}$ and $\boldsymbol{O}_{\mathcal{W}}^{l_k}$ are, respectively, the reference probability and the estimated probability of label l_k for $\mathcal{W}$. $\mathcal{D}$ denotes the training data set.

3 Multi-task and Multi-lingual Neural Lexical Utterance Classification

This section presents multi-task and multi-lingual joint learning of neural lexical utterance classification. We split neural lexical utterance classification by considering two types of components: language-specific components and task-specific components.

Language-specific components can be shared between tasks, where words in an utterance are converted into hidden representations. The language-specific components can be simplified as:

$$\boldsymbol{h}_t = \texttt{W2H}(\mathcal{W}, t; \boldsymbol{\Theta}_{\texttt{W2H}}), \tag{8}$$

where $\texttt{W2H}()$ is a function that compiles Eqs. (1) and (2). $\boldsymbol{\Theta}_{\texttt{W2H}}$ corresponds to $\{\boldsymbol{\theta}_{\texttt{w}}, \boldsymbol{\theta}_{\texttt{h}}\}$.

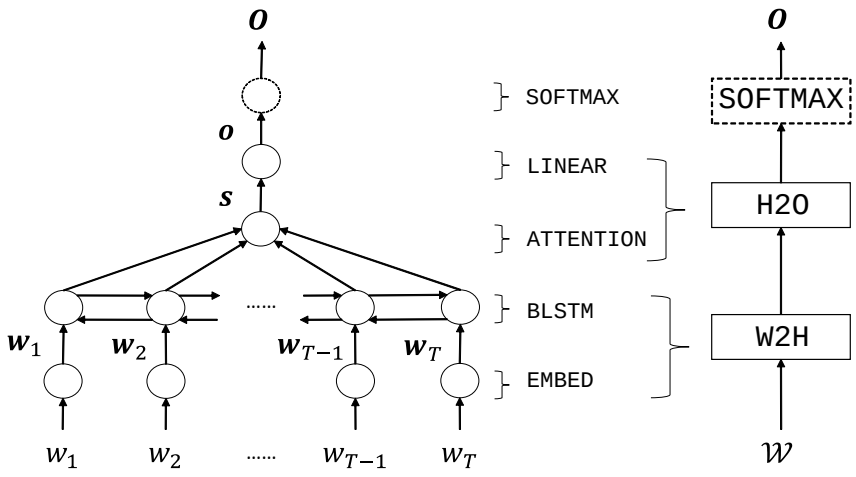

Figure 1: Neural lexical utterance classification and its simplified model structure.

Task-specific components can be shared between languages, where hidden representations are converted into predicted probabilities. The task-specific components can be simplified as:

$$\boldsymbol{o} = \texttt{H2O}(\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T\}; \boldsymbol{\Theta}_{\texttt{H2O}}), \tag{9}$$

where $\texttt{H2O}()$ is a function that compiles Eqs. (3) to (5) and $\boldsymbol{\Theta}_{\texttt{H2O}}$ represents $\{\boldsymbol{\theta}_{\texttt{z}}, \bar{\boldsymbol{z}}, \boldsymbol{\theta}_{\texttt{o}}\}$.

Figure 1 shows a detailed model structure and a simplified model structure of BLSTM-RNN with the attention mechanism. Both the dotted line square and a dotted line circle represent softmax activation. White squares are simplified components in the neural lexical utterance classification.

3.1 Fully-shared Modeling

Fully-shared modeling forms universal hidden representations that are completely invariant to differences in tasks or languages. In this case, language-specific components are fully-shared between the same language data sets and task-specific components are fully-shared between the same task data sets. The t-th universal hidden representation is calculated as:

$$\boldsymbol{h}_t = \texttt{W2H}(\mathcal{W}^{(i)}, t; \boldsymbol{\Theta}^{(i)}_{\texttt{W2H}}), \tag{10}$$

where $\boldsymbol{\Theta}^{(i)}_{\texttt{W2H}}$ is the shared parameter that handles the i-th language and $\mathcal{W}^{(i)}$ denotes the input utterance in the i-th language. The universal hidden representation can be input to any task-specific component. The predicted probabilities for the j-th task, denoted as $\boldsymbol{O}^{(j)}$, are calculated as:

$$\boldsymbol{o}^{(j)} = \texttt{H2O}(\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T\}; \boldsymbol{\Theta}^{(j)}_{\texttt{H2O}}), \tag{11}$$

$$\boldsymbol{O}^{(j)} = \texttt{SOFTMAX}(\boldsymbol{o}^{(j)}), \tag{12}$$

where $\boldsymbol{\Theta}^{(j)}_{\texttt{H2O}}$ is the task-specific shared parameter that handles the j-th task.

Figure 2 shows the model structure of multi-task fully-shared modeling for a language and two tasks. Figure 3 shows the model structure of multi-lingual fully-shared modeling for two languages and a task. Figure 4 shows the model structure of multi-task and multi-lingual fully-shared modeling for two tasks and two languages. Gray squares are shared components between languages or between tasks, and white squares are non-shared components.

3.2 Partially-shared Modeling

Partially-shared modeling introduces not only shared components between tasks or languages but also exclusive components that handle task-language combinations. Therefore, our hidden representations

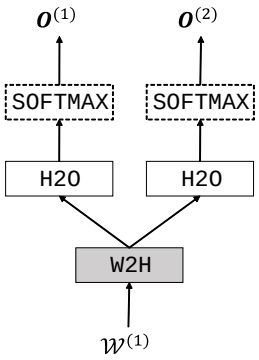

Figure 2: Multi-task fully-shared modeling.

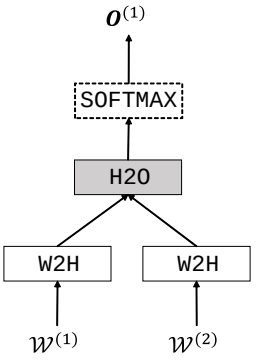

Figure 3: Multi-lingual fully-shared modeling.

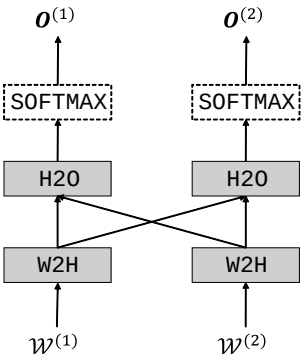

Figure 4: Multi-task and multi-lingual fully-shared modeling.

are designed to support such combinations. The t-th hidden representation for the combination of the i-th language and the j-th task, denoted as $\boldsymbol{h}_t^{(i,j)}$, is calculated as:

$$\bar{\boldsymbol{h}}_t^{(i,j)} = \texttt{W2H}(\mathcal{W}^{(i,j)}, t; \boldsymbol{\Theta}_{\texttt{W2H}}^{(i,j)}), \tag{13}$$

$$\bar{\boldsymbol{h}}_t^{(i)} = \texttt{W2H}(\mathcal{W}^{(i,j)}, t; \boldsymbol{\Theta}_{\texttt{W2H}}^{(i)}), \tag{14}$$

$$\boldsymbol{h}_t^{(i,j)} = \bar{\boldsymbol{h}}_t^{(i,j)} + \bar{\boldsymbol{h}}_t^{(i)}, \tag{15}$$

where $\boldsymbol{\Theta}_{\texttt{W2H}}^{(i,j)}$ is the exclusive parameter specific to the combination of the i-th language and the j-th task, and $\boldsymbol{\Theta}_{\texttt{W2H}}^{(i)}$ is the shared parameter that handles the i-th language. The predicted probabilities for the combination of the i-th language and j-th task, denoted as $\boldsymbol{O}^{(i,j)}$, are calculated as:

$$\bar{\boldsymbol{o}}^{(i,j)} = \texttt{H2O}(\{\boldsymbol{h}_1^{(i,j)}, \cdots, \boldsymbol{h}_T^{(i,j)}\}; \boldsymbol{\Theta}_{\texttt{H2O}}^{(i,j)}), \tag{16}$$

$$\bar{\boldsymbol{o}}^{(j)} = \texttt{H2O}(\{\boldsymbol{h}_1^{(i,j)}, \cdots, \boldsymbol{h}_T^{(i,j)}\}; \boldsymbol{\Theta}_{\texttt{H2O}}^{(j)}), \tag{17}$$

$$\boldsymbol{O}^{(i,j)} = \texttt{SOFTMAX}(\bar{\boldsymbol{o}}^{(i,j)} + \bar{\boldsymbol{o}}^{(j)}), \tag{18}$$

where $\boldsymbol{\Theta}_{\texttt{H2O}}^{(i,j)}$ is the exclusive parameter specific to the combination of the i-th language and the j-th task, and $\boldsymbol{\Theta}_{\texttt{H2O}}^{(j)}$ is the shared parameter that handles the j-th task.

Figure 5 shows the model structure of multi-task partially-shared modeling for a language and two tasks. Figure 6 shows the model structure of multi-lingual partially-shared modeling for two tasks and a language. Figure 7 shows the model structure of multi-task and multi-lingual partially-shared modeling for two tasks and two languages. Note that Figures 5-7 correspond to Figures 2-4.

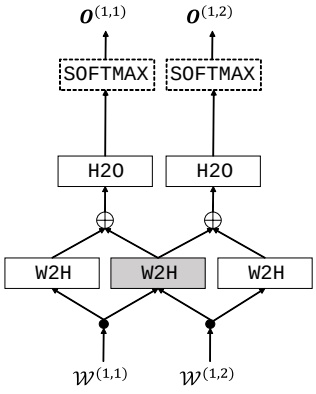

Figure 5: Multi-task partially-shared modeling.

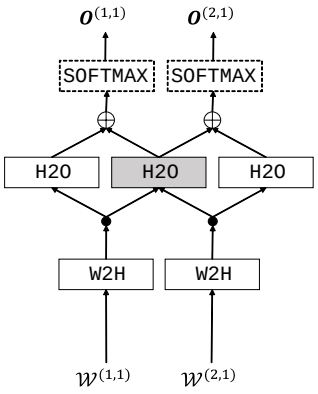

Figure 6: Multi-lingual partially-shared modeling.

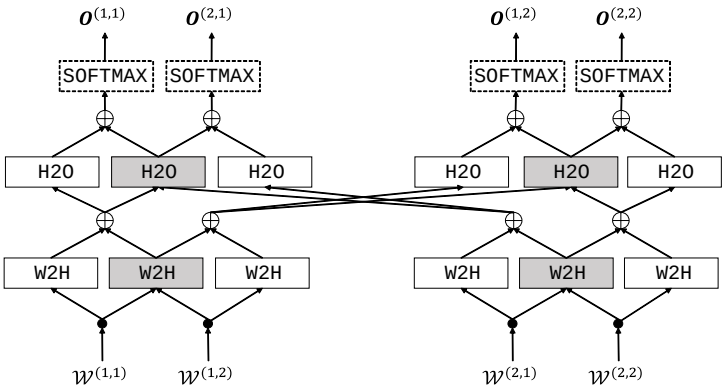

Figure 7: Multi-task and multi-lingual partially-shared modeling.

3.3 Joint Optimization

In multi-task and multi-lingual joint leaning, all parameters, denoted as $\mathbf{\Theta}$, can be jointly optimized by using all data sets. Given I languages and J tasks, joint optimization of the model parameter $\mathbf{\Theta}$ follows:

$$\hat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}}{\operatorname{argmin}} - \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{\mathcal{W} \in \mathcal{D}^{(i,j)}} \sum_{k=1}^{K^{(j)}} \hat{\boldsymbol{O}}_{\mathcal{W}}^{l_k} \log \boldsymbol{O}_{\mathcal{W}}^{l_k}, \tag{19}$$

where $\mathcal{D}^{(i,j)}$ denotes the training data set for the combination of the i-th language and the j-th task, and $|\mathcal{D}^{(i,j)}|$ means the number of utterances. $K^{(j)}$ represents the number of labels in the j-th task.

Basically, $\mathbf{\Theta}$ can be gradually updated by repeating mini-batch training using individual data sets. In this case, an optimizer with a learning rate is prepared for individual data sets and individual learning rates fall when the cross entropy loss for a target validation data set increases. The training epoch is stopped when the averaged loss for all validation data sets is not improved. Details of the joint optimization procedure are shown in **Algorithm 1**.

4 Experiments

4.1 Data

Our experiments employed Japanese (Ja) and English (En) data sets created for three different lexical utterance classification tasks. The tasks were dialogue act (DA) classification, extended named entity (ENE) classification (Sekine and Nobata, 2004; Higashinaka et al., 2012), and question type (QT) classification; natural language texts were used for the lexical utterances and individual label sets were unified

Algorithm 1 :Joint Optimization procedure of multi-task and multi-lingual joint leaning.

Input: Training data sets $\mathcal{D}^{(1,1)}, \cdots, \mathcal{D}^{(1,J)}, \cdots, \mathcal{D}^{(I,1)}, \cdots, \mathcal{D}^{(I,J)}$
Validation data sets $\bar{\mathcal{D}}^{(1,1)}, \cdots, \bar{\mathcal{D}}^{(1,J)}, \cdots, \bar{\mathcal{D}}^{(I,1)}, \cdots, \bar{\mathcal{D}}^{(I,J)}$
Output: Model parameter $\mathbf{\Theta}$

```
Initialize Θ randomly
while true do
  for t = 1 to number of mini-batches in training data sets do
    Select language i randomly
    Select task j randomly
    Pick mini-batch D_t from D^(i,j) randomly
    Update Θ by computing gradient of D_t
  end for
  for i = 1 to I do
    for j = 1 to J do
      Compute current validation loss for D̄^(i,j)
      if previous validation loss for D̄^(i,j) < current validation loss for D̄^(i,j) then
        Decrease learning rate for D^(i,j)
      end if
    end for
  end for
  if previous averaged validation loss < current averaged validation loss then
    break
  end if
end while
return Θ
```

Table 1: Number of utterances in individual data sets.

Language	Task	#labels	Train	Valid	Test
Japanese	DA	28	201,092	4,190	4,190
	ENE	168	40,350	4,036	4,036
	QT	15	55,328	4,257	4,257
English	DA	28	25,171	3,147	3,147
	ENE	168	25,005	3,230	3,230
	QT	15	22,213	2,777	2,777

between Japanese and English. For example, the task of English ENE classification is to obtain the requested ENE type for a question. Each of the data sets were divided into training (Train), validation (Valid), and test (Test) sets. Table 1 shows the number of utterances in individual data sets where #labels represents the number of labels. Table 2 shows English utterances and label examples for individual tasks.

4.2 Setups

We evaluated non-shared modeling, fully-shared modeling, and partially-shared modeling. For the shared modeling methods, multi-task joint learning, multi-lingual joint learning, multi-task and multi-lingual joint learning were examined. The multi-task joint learning used three classification tasks for optimizing each language. The multi-lingual joint learning used both Japanese and English data sets for optimizing each task. The multi-task and multi-lingual joint learning used all data sets. Several modeling parameters were unified. Word representation size was set to 128, LSTM-RNN unit size was set to 200, and context vector size in the attention mechanism was set to 200. Dropout was used for `EMBED`() and `BLSTM`(), and the dropout rate was set to 0.5. In these setups, words that appeared once or less in

Table 2: English utterances and label examples in individual tasks.

Task	Utterance	Label
DA	Hello, how are you today?	GREETING
	I am so sorry to hear of your son's accident.	SYMPATHY/AGREE
	Lets go to school an hour early today.	PROPOSAL
ENE	What is the highest mountain in the world?	MOUNTAIN
	Who is president of the united states?	PERSON
	What is the name of the most recent Star Wars movie?	MOVIE
QT	Do you like egg salad?	TRUE/FALSE
	How do you correct a hook in a golf swing?	EXPLANATION:METHOD
	Why is blood red?	EXPLANATION:CAUSE

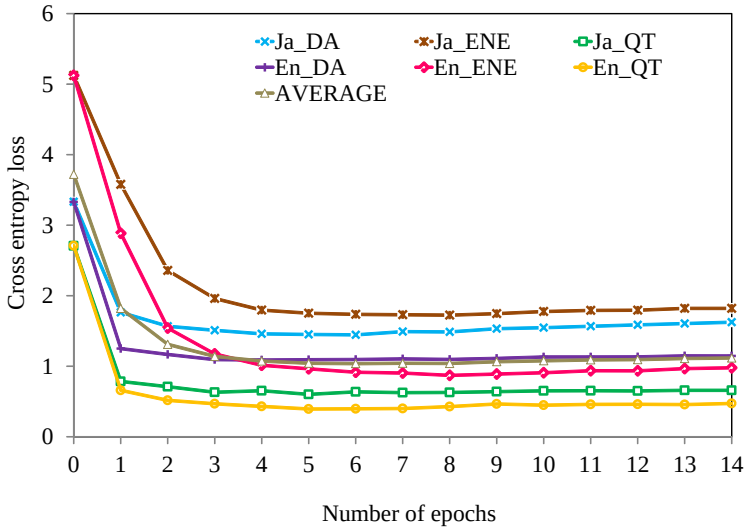

Figure 8: Cross entropy loss for validation sets.

the training data sets were treated as unknown words. For joint learning, mini-batch stochastic gradient descent was used for the individual optimizers. Initial learning rate for individual data sets was set to 0.1. The cutoff threshold for gradient clipping was set to 1.0. Training was stopped when the averaged validation loss was not improved in 5 consecutive iterations.

4.3 Results

Figure 8 demonstrates the change in cross entropy validation loss for individual validation sets when performing multi-task and multi-lingual joint learning based on partially-shared modeling. As epoch number increased, both cross entropy validation losses for individual data sets and averaged validation loss for all data sets (AVERAGE) decreased. This indicates that joint optimization procedure used in algorithm 1 worked well.

Table 2 shows the experimental results in terms of utterance classification accuracy for test sets. For each setup, we constructed five models by varying the initial parameters and evaluated the average accuracy.

First, (a) demonstrates the results of non-shared modeling trained using only an individual data set. These results are the baseline of this evaluation. In fully-shared modeling, (b) to (d), the classification performance deteriorated in some cases, while performance improvements were achieved in other cases. In English QT, multi-task and multi-lingual joint learning was inferior to multi-task joint learning or multi-lingual joint learning. This indicates that fully-shared modeling, which learns universal hidden representations, are not suitable for supporting both cross-lingual and cross-task knowledge transfer.

Table 2: Experimental results: utterance classification accuracy (%) for individual test sets.

		Joint learning		Japanese			English		
		task	language	DA	ENE	QT	DA	ENE	QT
(a).	Non-shared modeling	-	-	66.5	79.1	87.7	61.8	64.7	83.5
(b).	Fully-shared modeling	√	-	66.5	79.6	89.3	60.6	64.4	83.7
(c).		-	√	66.7	78.7	87.2	61.4	64.3	83.0
(d).		√	√	66.5	79.7	89.3	60.5	65.4	82.6
(e).	Partially-shared modeling	√	-	66.6	80.9	89.4	62.0	64.8	83.7
(f).		-	√	**66.9**	79.7	88.0	61.9	65.0	83.8
(g).		√	√	**66.9**	**81.8**	**89.7**	**62.3**	**65.8**	**84.0**

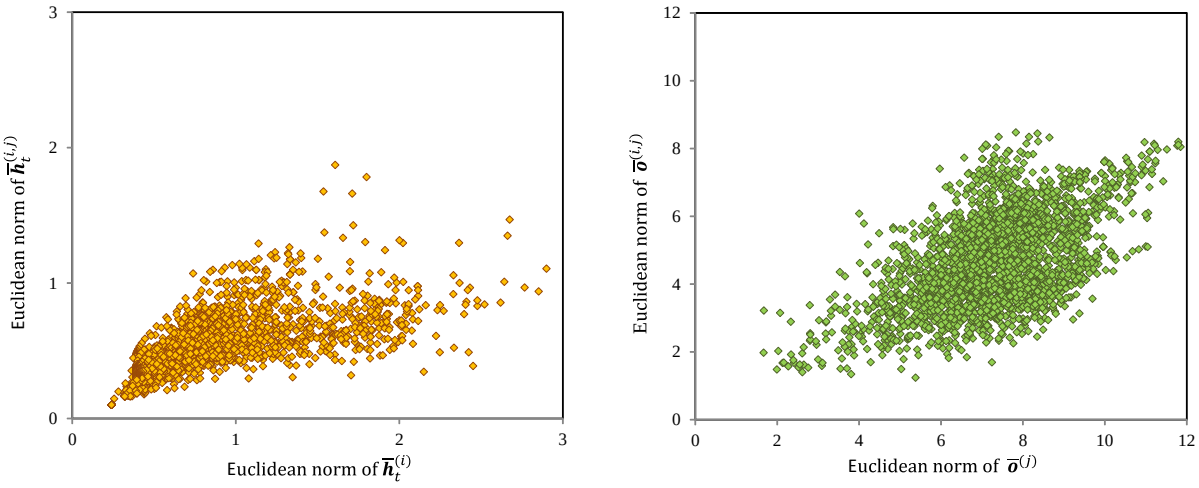

Figure 9: Euclidean norm results in partially-shared modeling.

On the other hand, partially-shared modeling, (e) to (g), improved the classification performance in all data sets compared to (a). In addition, multi-task and multi-lingual joint learning outperformed multi-task joint learning and multi-lingual joint learning. These results confirm that partially-shared modeling can effectively transfer knowledge between different data sets. We conducted sign test for verifying the effectiveness of multi-task and multi-lingual joint learning based on partially-shared modeling. In Japanese ENE classification, Japanese QT classification, and English ENE classification, statistically significant performance improvements ($p < 0.05$) were achieved by (g) compared to (a). Furthermore, in Japanese ENE classification, English DA classification, and English QT classification, significant performance improvements ($p < 0.05$) were also achieved by (g) compared to (d).

We investigated how shared components and exclusive components worked in partially-shared modeling. The left side of Figure 9 presents the Euclidean norm of both $\bar{\boldsymbol{h}}_t^{(i,j)}$ and $\bar{\boldsymbol{h}}_t^{(i)}$, while the right side presents the Euclidean norm of both $\bar{\boldsymbol{o}}^{(i,j)}$ and $\bar{\boldsymbol{o}}^{(j)}$ when classifying English question type validation data sets. These results show that shared components are more influential than exclusive components. This indicates that the shared components accumulate shareable knowledge, while exclusive components were utilized to offset the small differences between tasks or between languages.

5 Conclusions

This paper proposed multi-task and multi-lingual joint learning of neural lexical utterance classification for effectively leveraging data sets of different tasks and different language. For neural lexical utterance classification, we proposed BLSTM-RNN; it uses a self-attention mechanism and introduces language-specific and task-specific components. Each component can be effectively trained by partially-shared

modeling. Experiments on Japanese and English data sets created for three different tasks showed that the proposed multi-task and multi-lingual joint learning based on partially-shared modeling can transfer knowledge more effectively than multi-task or multi-lingual joint learning based on fully-shared modeling.

References

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. *In Proc. International Conference on Machine Learning (ICML)*.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1285–1295.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *In Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 866–875.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T. Yarman Vural, and Yoshua Bengio. 2017. Multi-way, multilingual neural machine translation. *Computer Speech & Language*, pages 236–252.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pages 2734–2740.

Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, Toshiro Makino, and Yoshihiro Matsuo. 2012. Creating an extended named entity dictionary from wikipedia. *In Proc. International Conference on Computational Linguistics (COLING)*, pages 1163–1178.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. *In Proc. International Conference on Computational Linguistics (COLING)*, pages 928–9239.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. *In Proc. International Conference on Computational Linguistics (COLING)*, pages 2012–2021.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. *In Proc. Annual Conference of the North American Chapter of the ACL (NAACL)*, pages 912–921.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016a. Deep multi-task learning with shared memory. *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 118–127.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016b. Recurrent neural network for text classification with multi-task learning. *In Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2873–2879.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016c. Implicit discourse relation classification via multi-task neural networks. *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pages 2750–2756.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–10.

Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. *In Proc. International Conference on Learning Representations (ICLR)*.

Aditya Mogadala and Achim Rettinger. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. *In Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 692–702.

Nikolaos Pappas and Andrei Popescu-Belis. 2017. Multilingual hierarchical attention networks for document classification. *In Proc. International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1015–1025.

Suman Ravuri and Andreas Stolcke. 2015a. A comparative study of neural network models for lexical intent classification. *In Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 368–374.

Suman Ravuri and Andreas Stolcke. 2015b. Recurrent neural network and LSTM models for lexical utterance classification. *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 135–139.

Suman Ravuri and Andreas Stolcke. 2016. A comparative study of recurrent neural network models for lexical domain classification. *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6075–6079.

Naoki Sawada, Ryo Masumura, and Hiromitsu Nishizaki. 2017. Parallel hierarchical attention networks with shared memory reader for multi-stream conversational document classification. *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3311–3315.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *In Proc. Workshop on Representation Learning for NLP*, pages 157–167.

Satoshi Sekine and Chikashi Nobata. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. *In Proc. Language Resources and Evaluation Conference (LREC)*, pages 1977–1980.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martion, Carol Van Ess-Dykema, and Marie Metter. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Gokhan Tur, Dilek Hakkani-Tur, Larry Heck, and Suresh Parthasarathy. 2011. Sentence simplification for spoken language understanding. *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5628–5631.

Chung-Hsien Wu, Jui-Feng Yeh, and Ming-Jun Chen. 2005. Domain-specific FAQ retrieval using independent aspects. *ACM Transactions on Asian Language Information Processing*, 4(1):1–17.

Puyang Xu and Ruhi Sarikaya. 2014. Contextual domain classification in spoken language understanding systems using recurrent neural network. *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 136–140.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. 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)*, pages 1480–1489.

Xiaodong Zhang and Houfeng Weng. 2016. A joint model of intent determination and slot filling for spoken language understanding. *In Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2993–2999.

Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016a. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *In Proc. International Conference on Computational Linguistics (COLING)*, pages 3485–3496.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016b. Attention-based bidirectional long short-term memory networks for relation classification. *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 207–212.