Adversarial Training for Multi-task and Multi-lingual Joint Modeling of Utterance Intent Classification

Ryo Masumura and **Yusuke Shinohara** and **Ryuichiro Higashinaka** and **Yushi Aono**
NTT Media Intelligence Laboratories, NTT Corporation
1-1, Hikarinooka, Yokosuka-shi, Kanagawa, 239-0847, Japan
ryou.masumura.ba@hco.ntt.co.jp

Abstract

This paper proposes an adversarial training method for the multi-task and multi-lingual joint modeling needed for utterance intent classification. In joint modeling, common knowledge can be efficiently utilized among multiple tasks or multiple languages. This is achieved by introducing both language-specific networks shared among different tasks and task-specific networks shared among different languages. However, the shared networks are often specialized in majority tasks or languages, so performance degradation must be expected for some minor data sets. In order to improve the invariance of shared networks, the proposed method introduces both language-specific task adversarial networks and task-specific language adversarial networks; both are leveraged for purging the task or language dependencies of the shared networks. The effectiveness of the adversarial training proposal is demonstrated using Japanese and English data sets for three different utterance intent classification tasks.

1 Introduction

In natural language processing fields, full neural network based methods are suitable for joint modeling as they can simultaneously utilize multiple task data sets or multiple language data sets to improve the performance achieved for individual tasks or languages (Collobert and Weston, 2008). It is known that joint modeling can address the data scarcity problem.

Key natural language processing technologies for spoken dialogue systems include utterance intent classification, which is needed to detect intent labels such as dialogue act (Stolcke et al., 2000; Khanpour et al., 2016), domain (Xu and Sarikaya, 2014), and question type (Wu et al., 2005) from input utterances (Ravuri and Stolcke, 2015a,b, 2016). One problem is that the training data are often limited or unbalanced among different tasks or different languages. Therefore, our motivation is to leverage both multi-task joint modeling and multi-lingual joint modeling to enhance utterance intent classification.

The multi-task and multi-lingual joint modeling can be composed by introducing both task-specific networks, which are shared among different languages, and language-specific networks, which are shared among different tasks (Masumura et al., 2018; Lin et al., 2018). Although joint modeling is mainly intended to improve classification performance in resource-poor tasks or languages, its classification performance is degraded in some minor data sets. This is because the language-specific networks often depend on majority tasks, while the task-specific networks often depend on majority languages. What are needed are task-specific networks that are invariant to languages, and language-specific networks that are invariant to tasks.

In order to explicitly improve the invariance of language and task-specific networks, this paper introduces adversarial training (Goodfellow et al., 2014). Our idea is to train language-specific networks so as to be insensitive to the target task, while training task-specific networks to be insensitive to language. To this end, we introduce multiple domain adversarial networks (Ganin et al., 2016), language-specific task adversarial networks, and task-specific language adversarial networks, into a state-of-the-art fully neural network based joint modeling; we adopt the bidirectional long short-term memory recurrent neural networks (BLSTM-RNNs) with attention mechanism (Yang et al., 2016; Zhou et al., 2016). To the best of our knowledge, this paper is the first study to employ adversarial training for multi-input and multi-output joint modeling.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 633–639
Brussels, Belgium, October 31 - November 4, 2018.

Experiments on Japanese and English data sets demonstrate the effectiveness of the adversarial training proposal. To support spoken dialogue systems, three different utterance intent classification tasks are examined: dialogue act, topic type, and question type classification.

2 Related Work

Joint Modeling: In natural language processing research, joint modeling is usually split into multi-task joint modeling and multi-lingual joint modeling. Multi-task joint modeling has been shown to effectively improve individual tasks (Collobert and Weston, 2008; Liu et al., 2016a,b; Zhang and Weng, 2016; Liu et al., 2016c). In addition, multi-lingual joint modeling is achieved by learning common semantic representations among different languages (Guo et al., 2016; Duong et al., 2016; Zhang et al., 2016, 2017b). In addition, a few work have examined multi-task and multi-lingual joint modeling (Masumura et al., 2018; Lin et al., 2018). Different from the previous work, our novelty is to introduce adversarial training for multi-task and multi-lingual joint modeling.

Adversarial Training: The concept of adversarial training was first proposed by Goodfellow et al. (2014), and many studies in the machine learning field have focused on adversarial training. Adversarial training has been well utilized in text classification (Ganin et al., 2016; Chen et al., 2016; Liu et al., 2017; Miyato et al., 2017; Chen and Cardie, 2018). Most natural language processing papers adopted either the language invariant approach (Chen et al., 2016; Zhang et al., 2017a) or the task invariant approach (Ganin et al., 2016; Liu et al., 2017; Chen and Cardie, 2018). This paper aims to fully utilize both task adversarial training and language adversarial training. To this end, we simultaneously introduce language-specific task adversarial networks and task-specific language adversarial networks.

3 Proposed Method

This section details our adversarial training method for multi-task and multi-lingual joint modeling of utterance intent classification.

In the j-th task utterance intent classification for the i-th language input utterance, intent label $l^{(j)} \in \{1, \cdots, K^{(j)}\}$ is estimated from input utterance $\mathcal{W}^{(i)} = \{w_1^{(i)}, \cdots, w_T^{(i)}\}$ where $i \in \{1, \cdots, I\}$ and $j \in \{1, \cdots, J\}$. Utterance intent classification is followed by estimation of the probabilities of each intent label given input utterance, $P(l^{(j)}|\mathcal{W}^{(i)}, \boldsymbol{\Theta}^{(i,j)})$ where $\boldsymbol{\Theta}^{(i,j)}$ is the trainable model parameter for the combination of the i-th language and the j-th task. In multi-task and multi-lingual joint modeling, $\{\boldsymbol{\Theta}^{(1,1)}, \cdots, \boldsymbol{\Theta}^{(I,J)}\}$ are jointly trained from I language and J task data sets.

3.1 Main Joint Network

The proposed method is founded on a fully neural network that employs I language-specific networks, J task-specific networks, and J classification networks as well as Masumura et al. (2018).

The language-specific network can be shared between multiple tasks, where words in the input utterance are converted into language-specific hidden representations. Each word in the i-th language input utterance $\mathcal{W}^{(i)}$ is first converted into a continuous representation. Next, each word representation is converted into a hidden representation that uses BLSTM-RNNs to take neighboring word context information into account. The t-th language-specific hidden representation for the i-th language is given by:

$$\boldsymbol{w}_t^{(i)} = \texttt{EMBED}(w_t^{(i)}; \boldsymbol{\theta}_{\texttt{h}}^{(i)}), \quad (1)$$

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

where $\texttt{EMBED}()$ is a linear transformational function for word embedding, $\texttt{BLSTM}()$ is a function of the BLSTM-RNN layer, and $\boldsymbol{\theta}_{\texttt{h}}^{(i)}$ is the trainable parameter for the i-th language-specific network.

In addition, task-specific networks can be shared between multiple languages, where the language-specific hidden representations are converted into task-specific hidden representations. The t-th language-specific hidden representation for the j-th task is given by:

$$\boldsymbol{u}_t^{(j)} = \texttt{BLSTM}(\{\boldsymbol{h}_1^{(i)}, \cdots, \boldsymbol{h}_T^{(i)}\}, t; \boldsymbol{\theta}_{\texttt{u}}^{(j)}), \quad (3)$$

where $\boldsymbol{\theta}_{\texttt{u}}^{(j)}$ is the trainable parameter for the j-th task-specific network.

In classification networks for each task, the task-specific hidden representations are summarized as sentence representation $\boldsymbol{s}^{(j)}$ by using a self-attention mechanism that can consider the importance of individual hidden representations (Yang et al., 2016; Zhou et al., 2016; Sawada et al., 2017). Next, predicted probabilities of the j-th

task intent labels, $\boldsymbol{o}^{(j)} \in \mathbb{R}^{K^{(j)}}$, are given by:

$$\boldsymbol{s}^{(j)} = \texttt{ATTENSUM}(\{\boldsymbol{h}_1^{(i)}, \cdots, \boldsymbol{h}_T^{(i)}\}; \boldsymbol{\theta}_{\texttt{o}}^{(j)}), \quad (4)$$

$$\boldsymbol{o}^{(j)} = \texttt{SOFTMAX}(\boldsymbol{s}^{(j)}; \boldsymbol{\theta}_{\texttt{o}}^{(j)}), \quad (5)$$

where $\texttt{ATTENSUM}()$ is a weighted sum function with self-attention, $\texttt{SOFTMAX}()$ is a transformational function with softmax activation, and $\boldsymbol{\theta}_{\texttt{o}}^{(j)}$ is the trainable parameter for the j-th classification network. In the main joint networks of the proposal, $\boldsymbol{\Theta}^{(i,j)}$ corresponds to $\{\boldsymbol{\theta}_{\texttt{h}}^{(i)}, \boldsymbol{\theta}_{\texttt{u}}^{(j)}, \boldsymbol{\theta}_{\texttt{o}}^{(j)}\}$.

3.2 Adversarial Networks

The proposed method combines a language-specific task adversarial network with a task-specific language adversarial network. The task adversarial network is used for training the language-specific networks to be insensitive to target task labels, and the language adversarial network is used for training the task-specific networks to be insensitive to target language labels. In order to efficiently use stochastic gradient descent based training for optimizing the adversarial networks, we use gradient reversal layers, which allow the input vectors during forward propagation, and sign inversion of the gradients during back propagation, to be utilized (Ganin et al., 2016).

The i-th language-specific task adversarial network estimates task labels from the i-th language-specific hidden representations. The predicted probabilities of task labels, $\boldsymbol{x}^{(i)} \in \mathbb{R}^{J}$, are given by:

$$\tilde{\boldsymbol{h}}_t^{(i)} = \texttt{GRL}(\boldsymbol{h}_t^{(i)}), \quad (6)$$

$$\tilde{\boldsymbol{h}}^{(i)} = \texttt{ATTENSUM}(\{\tilde{\boldsymbol{h}}_1^{(i)}, \cdots, \tilde{\boldsymbol{h}}_T^{(i)}\}; \boldsymbol{\theta}_{\texttt{x}}^{(i)}), \quad (7)$$

$$\boldsymbol{x}^{(i)} = \texttt{SOFTMAX}(\tilde{\boldsymbol{h}}^{(i)}, \boldsymbol{\theta}_{\texttt{x}}^{(i)}), \quad (8)$$

where $\texttt{GRL}()$ represents the gradient reversal layer, and $\boldsymbol{\theta}_{\texttt{x}}^{(i)}$ is the trainable parameter. The j-th task-specific language adversarial network estimates language labels from the j-th task-specific hidden representations. The predicted probabilities of language labels, $\boldsymbol{y}^{(j)} \in \mathbb{R}^{I}$, are given by:

$$\tilde{\boldsymbol{u}}_t^{(j)} = \texttt{GRL}(\boldsymbol{u}_t^{(j)}), \quad (9)$$

$$\tilde{\boldsymbol{u}}^{(j)} = \texttt{ATTENSUM}(\{\tilde{\boldsymbol{u}}_1^{(j)}, \cdots, \tilde{\boldsymbol{u}}_T^{(j)}\}; \boldsymbol{\theta}_{\texttt{y}}^{(j)}), \quad (10)$$

$$\boldsymbol{y}^{(j)} = \texttt{SOFTMAX}(\tilde{\boldsymbol{u}}^{(j)}, \boldsymbol{\theta}_{\texttt{y}}^{(j)}), \quad (11)$$

where $\boldsymbol{\theta}_{\texttt{y}}$ is the trainable parameter.

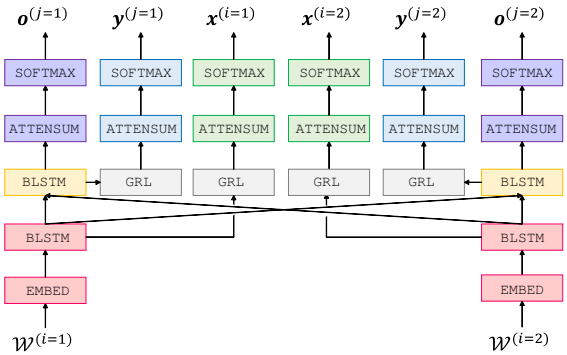

Figure 1: Proposed network structure for two tasks and two languages.

The proposed network structure shown in Figure 1 includes both joint networks and adversarial networks for two tasks and two languages. The red components are language-specific networks, the orange components are task-specific networks, and the purple components are classification networks. In addition, green components are language-specific task adversarial networks, and blue components are task-specific language adversarial networks.

3.3 Training

Our adversarial training proposal jointly optimizes all parameters in both the main joint networks and the adversarial networks by using all training data sets $\{\mathcal{D}^{(1,1)}, \cdots, \mathcal{D}^{(I,J)}\}$ where $\mathcal{D}^{(i,j)}$ represents the sets of the input utterances and the reference. The cross-entropy loss functions of each network are defined as:

$$L_{\boldsymbol{o}} = -\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{|\mathcal{D}^{(i,j)}|}\sum_{k=1}^{K^{(j)}} \hat{\boldsymbol{o}}_{n,k}^{(j)} \log \boldsymbol{o}_{n,k}^{(j)}, \quad (12)$$

$$L_{\boldsymbol{x}} = -\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{|\mathcal{D}^{(i,j)}|}\sum_{j'=1}^{J} \hat{\boldsymbol{x}}_{n,j'}^{(i)} \log \boldsymbol{x}_{n,j'}^{(i)}, \quad (13)$$

$$L_{\boldsymbol{y}} = -\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{n=1}^{|\mathcal{D}^{(i,j)}|}\sum_{i'=1}^{I} \hat{\boldsymbol{y}}_{n,i'}^{(j)} \log \boldsymbol{y}_{n,i'}^{(j)}, \quad (14)$$

where $L_{\boldsymbol{o}}$, $L_{\boldsymbol{x}}$, and $L_{\boldsymbol{y}}$ are the cross entropy loss terms for the classification networks, the task adversarial networks, and the language adversarial networks. $\hat{\boldsymbol{o}}_{n,k}^{(j)}$, $\hat{\boldsymbol{x}}_{n,j'}^{(i)}$, and $\hat{\boldsymbol{y}}_{n,i'}^{(j)}$ are the reference probabilities, and $\boldsymbol{o}_{n,k}$, $\boldsymbol{x}_{n,j'}$, and $\boldsymbol{y}_{n,i'}$ are the estimated probabilities of the k-th label in the j-th task classification network, the j'-th task in the i-th language-specific task adversarial network, and

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
TT	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

Table 2: Examples of English data sets.

the i'-th language in the j-th task-specific language adversarial network for $\mathcal{W}_n$, respectively.

Due to use of gradient reversal layers, individual parameters are gradually updated as follows:

$$\boldsymbol{\theta}_{\mathtt{o}}^{(j)} \leftarrow \boldsymbol{\theta}_{\mathtt{o}}^{(j)} - \epsilon \frac{\partial L_{\boldsymbol{o}}}{\partial \boldsymbol{\theta}_{\mathtt{o}}^{(j)}}, \tag{15}$$

$$\boldsymbol{\theta}_{\mathtt{y}}^{(j)} \leftarrow \boldsymbol{\theta}_{\mathtt{y}}^{(j)} - \epsilon\beta \frac{\partial L_{\boldsymbol{y}}}{\partial \boldsymbol{\theta}_{\mathtt{y}}^{(j)}}, \tag{16}$$

$$\boldsymbol{\theta}_{\mathtt{u}}^{(j)} \leftarrow \boldsymbol{\theta}_{\mathtt{u}}^{(j)} - \epsilon\left(\frac{\partial L_{\boldsymbol{o}}}{\partial \boldsymbol{\theta}_{\mathtt{u}}^{(j)}} - \beta \frac{\partial L_{\boldsymbol{y}}}{\partial \boldsymbol{\theta}_{\mathtt{u}}^{(j)}}\right), \tag{17}$$

$$\boldsymbol{\theta}_{\mathtt{x}}^{(i)} \leftarrow \boldsymbol{\theta}_{\mathtt{x}}^{(i)} - \epsilon\alpha \frac{\partial L_{\boldsymbol{x}}}{\partial \boldsymbol{\theta}_{\mathtt{x}}^{(i)}}, \tag{18}$$

$$\boldsymbol{\theta}_{\mathtt{h}}^{(i)} \leftarrow \boldsymbol{\theta}_{\mathtt{h}}^{(i)} - \epsilon\left(\frac{\partial L_{\boldsymbol{o}}}{\partial \boldsymbol{\theta}_{\mathtt{h}}^{(i)}} - \alpha \frac{\partial L_{\boldsymbol{x}}}{\partial \boldsymbol{\theta}_{\mathtt{h}}^{(i)}} - \beta \frac{\partial L_{\boldsymbol{y}}}{\partial \boldsymbol{\theta}_{\mathtt{h}}^{(i)}}\right), \tag{19}$$

where α and β are hyper parameters of the parameter update, and ϵ is the learning rate. Note that adversarial training is suppressed by setting α and β to 0.0. In training, we prepared optimizers for individual data sets. The individual learning rates fall when the validation loss of the target classification network increases.

4 Experiments

Our experiments employed Japanese and English data sets created for three different utterance intent classification tasks. The tasks, dialogue act (DA) classification, topic type (TT) classification, and question type (QT) classification, are intended to support spoken dialogue systems. For example, the task of English DA classification is to obtain a DA label from an input utterance. We used natural language texts as the input utterances and individual label sets were unified between Japanese and English. Data sets employed in experiments were corpora that were made for constructing spoken dialogue systems (Masumura et al., 2018). 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.

Language	Task	#labels	Train	Valid	Test
Japanese	DA	28	201 K	4 K	4 K
	TT	168	40 K	4 K	4 K
	QT	15	55 K	4 K	4 K
English	DA	28	25 K	3 K	3 K
	TT	168	25 K	3 K	3 K
	QT	15	22 K	2 K	2 K

Table 1: Number of utterances in individual data sets.

4.1 Setups

We examined single-task and mono-lingual modeling, multi-task joint modeling, multi-lingual join modeling, and multi-task and multi-lingual joint modeling with or without adversarial training.

We unified network configurations as follows. Word representation size was set to 128, BLSTM-RNN unit size was set to 400, and sentence representation was set to 400. Dropout was used for $\mathtt{EMBED}()$ and $\mathtt{BLSTM}()$, and the dropout rate was set to 0.5. Words that appeared only once in the training data sets were treated as unknown words. We used mini-batch stochastic gradient descent, in which initial learning rate was set to 0.1. We optimized hyper-parameters of adversarial training (α and β) for the validation sets by varying them from 0.001 to 1.0. Other hyper parameters were also optimized for the validation sets.

4.2 Results

Table 3 shows the results in terms of utterance classification accuracy. For each setup, we constructed five models by varying the initial parameters and evaluated the average accuracy. Line (1) shows baseline results: single-task and mono-lingual modeling. Lines (2) and (3) show results

	Joint modeling		Adversarial Training		Japanese			English		
	Multi-task	Multi-lingual	Task-invariant	Language-invariant	DA	TT	QT	DA	TT	QT
(1).	-	-	-	-	66.6	79.1	87.7	61.8	64.5	83.4
(2)	√	-	-	-	66.5	79.6	89.3	60.6	64.4	**83.7**
(3)	√	-	√	-	66.5	80.6	89.5	61.6	65.7	**83.7**
(4)	-	√	-	-	66.7	78.7	87.2	61.4	64.3	83.0
(5)	-	√	-	√	66.9	79.8	88.2	61.8	64.8	83.3
(6).	√	√	-	-	66.6	79.7	89.3	60.5	65.4	82.6
(7).	√	√	√	-	67.3	81.1	89.6	61.5	66.1	83.5
(8).	√	√	-	√	66.7	80.7	89.5	60.9	**66.7**	83.0
(9).	√	√	√	√	**67.6**	**81.3**	**90.0**	**61.9**	**66.7**	**83.7**

Table 3: Experimental results: utterance classification accuracy (%) for individual test sets.

with only performing multi-task joint modeling, and lines (4) and (5) show results with only performing multi-lingual joint modeling. Note that lines (3) and (5) show the results achieved with adversarial training. Line (6) shows multi-task and multi-lingual joint modeling results: adversarial training was suppressed by setting both α and β to 0.0. Lines (7)–(9) shows the results achieved with adversarial training. Note that setting with bold values achieved the highest performance in our evaluation.

First, in lines (2) and (4), the classification performance deteriorated in some cases, while performance improvements were achieved in other cases. On the other hand, in lines (3) and (5), classification performance in each data sets was improved by introducing adversarial training. This indicates that adversarial training was effective in improving the performance of joint modeling.

Next, line (6) shows that, relative to line 1, multi-task and multi-lingual joint modeling can improve the classification performance for Japanese TT, Japanese QT, and English TT, but classification performance was degraded for English DA and English QT. This indicates that it is difficult to simultaneously improve the classification performance for all data sets because joint modeling often depends on majority tasks or majority languages. In addition, lines (7) and (8) show the introduction of either task adversarial networks or language adversarial networks yielded better performance than line (6) for all data sets. This indicates that adversarial training was effective in improving the performance of multi-task and multi-lingual joint modeling. The best results were achieved by using both language-specific task adversarial networks and task-specific language adversarial networks, line (9). These results confirm that task adversarial networks and language adversarial networks well complement each other. Of particular benefit, the proposed method demonstrated greater classification performance improvements when the number of training utterances per label was small.

5 Conclusions

We have proposed an adversarial training method for the multi-task and multi-lingual joint modeling needed to enhance utterance intent classification. Our adversarial training proposal utilizes both task adversarial networks and language adversarial networks for improving task-invariance in language-specific networks and language-invariance in task-specific networks. Experiments showed that the adversarial training proposal could well realize the benefits of joint modeling in all data sets.

References

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. *arXiv preprint arXiv:1802.05694*.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv preprint arXiv:1606.01614*.

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.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky.

2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:1–35.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *In Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680.

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.

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.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A multi-lingual multi-task architecture for low-resource sequence labeling. *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages pp.799–809.

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.

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.

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.

Ryo Masumura, Tomohiro Tanaka, Ryuichiro Higashinaka, Hirokazu Masataki, and Yushi Aono. 2018. 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)*, pages pp.3586–3596.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. *In Proc. International Conference on Learning Representation (ICLR)*.

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.

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.

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.

Meng Zhang, Yang Liu, Huanbo Luan, Yiqun Liu, and Maosong Sun. 2016. Inducing bilingual lexica from non-parallel data with earth mover's distance regularization. *In Proc. International Conference on Computational Linguistics (COLING)*, pages 3188–3198.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1959–1970.

Meng Zhang, Haoruo Peng, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Bilingual lexicon induction from non-parallel data with minimum supervision. *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pages 3379–3384.

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, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. 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.