Multi-Source Domain Generalization Using Domain Attributes for Recurrent Neural Network Language Models

Naohiro TAWARA†‡, Atsunori OGAWA†, Members, Tomoharu IWATA†, Senior Member, Hiroto ASHIKAWA††*, Nonmember, Tetsunori KOBYASHI††, and Tetsuji OGAWA††††, Members

SUMMARY Most conventional multi-source domain adaptation techniques for recurrent neural network language models (RNNLMs) are domain-centric. In these approaches, each domain is considered independently and this makes it difficult to apply the models to completely unseen target domains that are unobservable during training. Instead, our study exploits domain attributes, which represent common knowledge among such different domains as dialects, types of wordings, styles, and topics, to achieve domain generalization that can robustly represent unseen target domains by combining the domain attributes. To achieve attribute-based domain generalization system in language modeling, we introduce domain attribute-based experts to a multi-stream RNNLM called recurrent adaptive mixture model (RADMM) instead of domain-based experts. In the proposed system, a long short-term memory is independently trained on each domain attribute as an expert model. Then by integrating the outputs from all the experts in response to the context-dependent weight of the domain attributes of the current input, we predict the subsequent words in the unseen target domain and exploit the specific knowledge of each domain attribute. To demonstrate the effectiveness of our proposed domain attributes-centric language model, we experimentally compared the proposed model with conventional domain-centric language model by using texts taken from multiple domains including different writing styles, topics, dialects, and types of wordings. The experimental results demonstrated that lower perplexity can be achieved using domain attributes.

key words: domain attribute, mixture-of-experts, recurrent adaptive mixture model, domain generalization, language model

1. Introduction

Language modeling is an important module for many applications, such as speech recognition, machine translation, and so on. Linguistic characteristics, such as frequently appearing words and contextual information (wordings), can significantly vary according to domains including topics [1], styles [2], and dialects [3]. This implies that language models should be constructed for respective target domains. However, it is often impractical to prepare large-scale training data on individual target domains. The simplest approach to address this issue is using external resources as augmented training data [4], [5]. Another approach is domain adaptation [1], [2], [6]–[14] in which the model trained in the source domain is adapted to the target domain by re-training the whole network [6], or adaptation layers [7] that use the target domain data. More recently, multiple domains have been leveraged as source domains to represent the target domains that do not match any individual source domains [8], [11], [12], [15], [16]. A typical example of multi-source domain adaptation is based on mixture-of-experts (MoEs) [17], which trains expert models for individual domains and combines them with equal [15] or weighted contributions that are calculated with metric learning [16], or neural network-based gating functions [13], [14]. By introducing state-dependent gating functions such as recurrent neural networks, the weight of each expert can be dynamically changed contingent on the context of the words [11], [12]. These multi-domain adaptation techniques can leverage the knowledge of multiple source domains to capture the target domain. However, the effectiveness of adaptation can deteriorate in the completely unseen target domains.

In this paper, we assume that these problems, which are caused by the shortages and lacks in the source and target domains data, will be reduced by dividing the domains into sub-domains based on their attributes and sharing them among domains with the identical attributes. For example, consider two journals, one on robotics and another on social science as two domain articles. Both articles are commonly described in an academic style, and the topic of each article is robotics and social science, respectively. On the other hand, science essays on robotics and social science are written in a popular style, although their topics are identical as those of the journals. This example indicates that some domains share similar linguistic characteristics when they have the same attribute, such as topics and styles. In this paper, we refer to these individual linguistic characteristics as domain attributes and exploit them by introducing a domain attribute-centric adaptation.

Figure 1 depicts a conceptual image of our domain attribute-centric and the conventional domain-centric approaches for a case that contains eight source domains, $D_1$ – $D_8$, and one unknown target domain $D_9$. In the conventional domain-based MoEs model, the model for the target domain $D_9$ is obtained by combining the expert models trained on the source domains $D_1$ – $D_8$. However, these experts on the source domains can be often over-fitted to each domain when the amount of data is limited, yielding a degradation
of performance in the target domain. On the other hand, in our domain attribute-based MoEs model, each domain is firstly decomposed into two types of domain attributes, including three styles ($S_1$, $S_2$, and $S_3$) and three topics ($T_1$, $T_2$, and $T_3$). Then, the model for the target domain $D_0$ is obtained by combining the expert models trained on each style and topic. These domain attributes-centric experts are more reliable than domain-centric ones because each expert is trained using larger data by sharing the data among the domains with the same attributes (e.g. $T_1$ is trained on $D_1$, $D_2$, and $D_3$).

To achieve this concept, we propose introducing the domain attributes to train each expert of domain adaptive mixture model (D-RADMM)[12] instead of the conventional domains. The proposed method trains a long short-term memory (LSTM) as an expert on each domain attribute and dynamically combines them depending on the context of the input using a state-dependent gating function. Compared with the conventional domain-centric approach [12], [15], [16], the proposed domain attribute-centric approach achieves a more generalized model since each expert is trained on a specific attribute that can be shared among different domains. This enables a domain generalization [18] of language models to the data on unknown domains. Here, the attributes of the target domain are known or accurately estimated in some applications such as automatic speech recognition. For example, in a case where we adapt a language model trained on publicly available lecture speeches to conversation speeches where only a limited amount of data is available, both the speaking styles and subjects of lecture and conversation can be used as known domain attributes. In this case, we can introduce a domain attribute indicator that represents the attribute composing the target domains, and utilize it as an input of the model to explicitly provide the information of the target domain. In the conventional domain-centric approach, on the other hand, we cannot utilize such indicators because there is no expert identical with the target domain.

The remainder of this paper is organized as follows. Section 2 briefly reviews relevant works. The domain generalization of RNNLM using domain attributes is proposed in Sect. 3, and its effectiveness is examined in Sect. 4. A summary and future works are presented in Sect. 5.

2. Relevant Works of Domain Generalization

Domain generalization [18] is a technique that improves the generalization performance of the target domain without data on the target domain during training. Several attempts have been made to extract domain-invariant representations from data on multiple domains using kernel methods [19], [20] or neural networks [21], [22]. Recent adversarial methods on multi-source domain adaptation [23] also seek domain-invariant representations. In addition, approaches that explicitly address the differences among domains have proven effective [22], [24].

Particularly, a combination of domain-specific and domain-invariant representations was successfully applied to domain adaptation in language modeling [2], [8], [12]. For example, in a work by Irie et al. [12], which is the most related to ours, LSTM experts are trained on multiple linguistic domains and their outputs are integrated in response to the weight of the current input to the respective domains. This scheme is called a recurrent adaptive mixture model (RADMM). In our study, RADMM is extended such that the expert LSTMs would be trained not on domains as in RADMM but on domain attributes, such as description styles and topics. By forming experts for domain attributes, prediction reliability can be improved if the training and testing data are on different domains but on the same domain attributes (e.g., same style or same topic). In contrast, without the domain attributes, we can only use the knowledge from the domains exactly the same as the target one.

3. Domain Generalization Using Domain Attributes

3.1 Domain Attributes

This study assumes that linguistic domains can be decomposed into a combination of such attributes as description styles (e.g., academic and popular writing), topics (e.g., humanities and technology), types of wordings (e.g., spoken and written languages), and dialects (e.g., American and British English). If these domain attributes of the target domain are known, the word distribution and context structure that are characteristic of the domain can be estimated and exploited for word prediction without data on the target domain while training the language models. In this study, we assume that the attributes of the target domain are known both in the training and testing phases. We provide a discrete vector that represents the domain attributes that comprise the input document as an auxiliary input to the language model.

3.2 Domain Attribute Expert-Based Recurrent Adaptive Mixture Model (DA-RADMM)

To extract linguistic knowledge useful for word prediction of individual domain attributes, RADMM [12] is extended
We assume that the vocabulary size is $V$, and that a word vector at time $t$, denoted by $w(t) \in \mathbb{R}^V$, is represented by a 1-of-$K$ encoding. Then using $L \in \mathbb{R}^{E \times V}$ of the projection layer, $w(t)$ is mapped to the lower-dimensional continuous vector $p(t) \in \mathbb{R}^E$ as

$$p(t) = Lw(t).$$

$p(t)$ is taken as an input to each expert LSTM $LSTM_k$ ($k \in \{0, \ldots, N-1\}$, where $N$ denotes the number of defined domain attributes (e.g., $N = 6$ in the example of Fig. 1). LSTM $c$ denotes a common LSTM and other LSTMs express expert LSTMs for the relevant domain attributes. Each expert’s output, $h_k(t) \in \mathbb{R}^H$, is calculated as

$$h_k(t) = LSTM_k(p(t), h_k(t-1); U_k, R_k),$$

where $H$ denotes the number of units in each LSTM, $U_k \in \mathbb{R}^{H \times E}$ denotes the weights between the Input layer and Experts, and $R_k \in \mathbb{R}^{H \times H}$ denotes the weights of a recurrent path of an expert LSTM. In addition, the Mixer LSTM’s output $h_m(t) \in \mathbb{R}^{H_m}$ is calculated as

$$h_m(t) = LSTM_m(p(t), h_m(t-1); U_m, R_m).$$

$U_m \in \mathbb{R}^{H_m \times E}$ and $R_m \in \mathbb{R}^{H_m \times H_m}$ denote the weight matrix between the Input layer and the Mixer and that of the recurrent path of the Mixer LSTM, respectively. $h_m(t)$ is then concatenated with indicator vector $a = [a_1, \ldots, a_N]^T \in \{0, 1\}^V$, where $a_j$ assumes a value of one for the domain attribute corresponding to the input data, and zero otherwise:

$$g'(t) = \text{concat}(h_m(t), a).$$

Resulting mixture weights $g(t) = [g_0(t), g_1(t), \ldots, g_N(t)] \in \mathbb{R}^{N+1}$ are computed as

$$g(t) = \text{softmax}(W_g g'(t)),$$

where $W_g \in \mathbb{R}^{(N+1) \times (H_m+H)}$ denotes the weight parameters and $g_k(t) \in g(t)$ denotes the mixture weights to LSTM $k$. Resulting feature $s(t) \in \mathbb{R}^H$ is then obtained as

$$s(t) = \sum_{k=0}^{N} g_k(t) h_k(t).$$

The probabilistic distribution of subsequently appearing words $y(t) \in \mathbb{R}^V$ is emitted using $s(t)$ as

$$y(t) = \text{softmax}(W_y s(t)),$$

where $W_y \in \mathbb{R}^{V \times H}$ denotes the weights between the Experts and the Output layer. The original RADMM [12] uses two LSTM layers. However, our preliminary experiments showed a consistent degradation of performance in conditions described in Sect. 4 because of a relatively limited amount of training data compared with the original paper. We, therefore, used only one LSTM layer for all experiments in Sect. 4.
DA-RADMM follows a three-stage training procedure as same as the original RADMM. Each component is trained, and then the DA-RADMM is trained as follows:

1. Train the Input layer, the common LSTM, and the Output layer parameters using all of the training data. During the training, the mixture weights of other experts are fixed to 0.
2. Copy the parameters of the common LSTM into each expert LSTM to initialize the experts. Fix the Input layer and the Output layer parameters, and update each expert LSTM only using data on the corresponding domain attributes.
3. Take the common LSTM, all the expert LSTMs, the Input layer, and the Output layer from the previous stages to initialize the final DA-RADMM. Then train the Mixer and the Output layer parameters fixing all the LSTMs and the Input layer parameters.

### 4. Experiments

To demonstrate the effectiveness of exploiting domain attributes, we experimentally compared our four models using written texts on multiple domains with different styles and topics. They were compared in terms of perplexity as follows:

- **Baseline LSTMLM (Conventional):** an LSTMLM trained using data on domains other than the evaluation target domain. No domain attribute information was used.
- **D-RADMM (Conventional):** RADMM-based language models with domain-independent common LSTM and domain-specific expert LSTMs, consistent with a previously presented model [12]. No domain attribute information was used.
- **DA-RADMM-C (Proposed):** DA-RADMM-based language models with domain attribute-independent common LSTM and domain attribute-specific expert LSTMs. Attribute indicator a is not given to the Mixer, i.e., it does not process Eq. (4), and the mixture weights are determined only using contextual information.
- **DA-RADMM-I (Proposed):** DA-RADMM-based language models, which do not include the Mixer LSTM, determine the mixture weights only from attribute indicator a.
- **DA-RADMM-CI (Proposed):** DA-RADMM-based language models illustrated in Fig. 3. The mixture weights are determined using both the contextual information and the domain attribute information of the current input by integrating the output of the mixer LSTM and attribute indicator a.

Note that the amount of data used to train the baseline and common LSTMs is identical in all methods, although the amount used to train each expert differed among the conventional D-RADMM and the proposed DA-RADMMs. Moreover, the number of experts and the size of the whole model

| Style     | Topic       | # of words (train) | # of words (dev) | # of words (test) | Vocab size |
|-----------|-------------|--------------------|------------------|-------------------|------------|
| Academic  | Humanities  | 289k               | 12.7k            | 13.0k             | 10.1k      |
| Popular   | Technology  | 282k               | 11.7k            | 10.6k             | 10.1k      |

is different among the baseline LSTMLM, the conventional D-RADMM, and the proposed DA-RADMMs. We therefore discuss the relationship between the model size and the performance as well as the performance itself. To demonstrate the general effectiveness of our domain attribute-centric approach, we explored two cases using different domain attributes.

### 4.1 Case Study 1: Domain Attributes for Writing Styles and Topics

The first experiment explored the effectiveness of domain attributes composed of writing styles and topics.

#### 4.1.1 Experimental Materials

To investigate the validity of using writing styles and topics as domain attributes, we used the International Corpus of English (ICE) [25] as a text material. The ICE includes spoken and written English documents collected in each country around the world. We selected the written English subsets of six countries, Canada (ICE-CAN), Hong Kong (ICE-HK), India (ICE-IND), Jamaica (ICE-JA), The Philippines (ICE-PHI), and Singapore (ICE-SIN). We examined four domains with a combination of two styles and two topics (Academic-writing + Humanities), (Academic-writing + Technology), (Popular-writing + Humanities), and (Popular-writing + Technology). The data obtained from three domains other than the evaluation target domain were used for training. Note that in the training set, texts from the target domain were excluded from the training set, but texts were included from the different attributes of the target domain. 90% of the data were used for training, and the remainder were used as test data. 10% of the training data were used as development data for the hyper-parameter tuning and early stopping. Words that appeared in the training data less than twice were mapped to unknown words. Table 1 lists the number of words and vocabulary sizes of the training, development, and testing data in each domain. Tables 2 and 3 detail the data used to train each expert in D-RADMM and DA-RADMMs, respectively. Each column lists the names of domains, including Academic_Humanities (AH), Academic_Technology (AT), Popular_Humanities (PH), and Popular_Technology (PT), used to train each expert in each evaluation setting. Note that each target domain was excluded from the training data of each expert in our domain generalization setting. As Table 2 show, all experts in D-RADMM were trained on a single domain. In contrast, all experts in DA-RADMM
would be trained on a combination of domains when more than two attributes in styles or topics are used as the example shown in Fig. 1. Note that, however, the expert that corresponds to the target domain are trained on a single domain in our experiment because we could use only one attribute excluding the attribute of the target domain.

### 4.1.2 Experimental Conditions

All the neural networks were trained on an NVIDIA 1080Ti GPU using a batch size of 8. All the LSTMs have tied input, forget gate, and recurrent projection, as in Irie et al.’s work [12]. All our implementations were based on Chainer [26]. Table 4 lists the RNNLM parameters investigated for the ICE experiment. We selected the numbers of projection layer units $E$, hidden layers units $H$, and the length of the back-propagation through time (BPTT) with which the lowest perplexities were obtained on the development set in each domain. We used $l_2$-regularization of $1.0 \times 10^{-6}$ to prevent over-fitting. We used an Adam optimizer [27] with an initial learning rate of 0.1, $\beta_1$ of 0.9, and $\beta_2$ of 0.999 to optimize the models. The learning rate was initialized to 0.1 and halved when the logarithmic likelihood ratio of the current epoch and the previous epoch on the development set was less than 1.003. The computational time to train the common LSTM, each expert, and the Mixer model was about 67, 10, and 52 minutes, respectively.

### 4.1.3 Results

Table 5 lists the test set perplexities included in the five models that we compared. The results demonstrate that the proposed DA-RADMM-CI outperformed the conventional models in all domains. The conventional domain expert-based RADMM (D-RADMM) did not yield explicit improvements in terms of the perplexity over the Baseline LSTLM. In this case, useful knowledge in the target domain might not be acquired by D-RADMM because a small number of domains and a small amount of data were used in the present experiment. The proposed DA-RADMM, on the other hand, yielded improvements over the conventional D-RADMM. In addition, the DA-RADMM without the attribute indicator, i.e., D-RADMM-C, was outperformed in all the domains by D-RADMM-I and D-RADMM-CI, which took an attribute indicator as an input. This indicates the effectiveness of exploiting the domain attribute information as auxiliary information.

#### 4.1.4 Visualization of Mixture Weights

To visualize how experts work in the conventional D-RADMM and our proposed DA-RADMMs, we averaged the outputs from the Mixer $g(t)$ in Eq. (5) over the all evaluation data for each of the four target domains. Figures 4, 5, 6, and 7 respectively present D-RADMM-C, D-RADMM-CI, D-RADMM-I, and D-RADMM-CI results. The attributes, which are composed of the target domains, are shown in red. The words inside the parentheses in the results of the DA-RADMMs indicate the names of domains used to train each expert, and red words indicate matches with target domain.

Figure 4 illustrates that in the conventional D-RADMM, the common LSTM was dominant and domain-specific expert LSTMs contributed less to the word prediction in (a) Academic Humanities and (d) Popular Technology. This could be because the amount of data in each domain was too small to train each expert LSTM, and the mixer tended to select the most reliable common LSTM. In addition, in the conventional D-RADMM, a different domain from the target domain was dominantly used for (b) Academic Technology and (c) Popular Humanities. This result implies that D-RADMM failed to extract domain-specific knowledge that is useful in the unknown target domain from the known multiple domains used in the training, causing the higher perplexities than the baseline LSTM as shown in Table 5.

Figure 5 indicates that the context-dependent mixer in the proposed D-RADMM-C selected the domain attribute that is not contained in the target domain. This is presumed because the experts of the attribute of the target domain were trained on a single domain, while other experts were trained on a combination of domains (i.e., larger dataset), making the latter experts more reliable than the former ones. As a result, the mixer in DA-RADMM tended to select the experts that were not contained in the target domain. This result is not what we expected but raises another perspective: while not matching the target domain, the experts trained on a combination of domains in DA-RADMM were more specific to domain attributes than the common LSTM in D-RADMM, contributing to the lower perplexities in (a) Academic Humanities and (b) Academic Technology. This hy-
Fig. 4 Average mixer outputs obtained from conventional domain-based RADMM (D-RADMM). Red words in name of domain expert indicate a match with target domain.

Fig. 5 Average mixer outputs obtained from proposed domain attribute-based RADMM with context-dependent mixer (DA-RADMM-C). The words inside the parentheses indicate the names of domains used to train each expert, and red words indicate matches with target domain.

Fig. 6 Average mixer outputs obtained from proposed domain attribute-based RADMM with a domain attribute indicator (DA-RADMM-I).

Fig. 7 Average mixer outputs obtained from proposed domain attribute-based RADMM with a context-dependent mixer using domain attribute indicator (DA-RADMM-CI).

Table 5 Test set perplexities on unseen domain data in ICE dataset obtained from conventional and proposed language models.

| Model             | Mixer Attribute | Academic Writing | Popular Writing |
|-------------------|-----------------|-------------------|-----------------|
| Baseline LSTM     | -               | 317.81            | 411.58          |
| D-RADMM           | ✓               | 331.06            | 413.90          |
| DA-RADMM-C        | ✓               | 331.67            | 416.52          |
| DA-RADMM-I        | ✓               | 328.85            | 409.30          |
| DA-RADMM-CI       | ✓               | 320.36            | 340.01          |
|                   |                 | 327.54            | 420.64          |
|                   |                 | 314.80            | 397.77          |
|                   |                 | 325.39            | 411.06          |

Inadequate, although integrating such information was effective. Note that other than in the case of (b) Academic Technology in Fig. 7, two highly weighted domain attributes are consistent with the style and topic of the evaluation target domain. The reason for the high weight of the expert for Popular in (b) Academic Technology is that the knowledge derived from the Popular writing for word prediction on this domain might be needed because a document on Technology includes many Popular expressions.

Figure 8 depicts examples of mixer output for each input word in a document selected from the evaluation set of (a) Academic Humanities. Each line shows the weight of each expert’s domain attribute. Figure 8 (b) indicates that the mixer of DA-RADMM-C generated context-dependent weights that changed word by word in accordance with the context, although one of DA-RADMM-I generated a consistent weight with the attribute indicator (Fig. 8 (c)). On the other hand, the mixer of DA-RADMM-CI successfully generated a stable context-dependent weight based on the consistent weight. Figure 9 demonstrates the detailed mixer outputs that correspond to the sentence inside the red frame.
Fig. 8  Examples of mixer outputs obtained from each model on each word in an Academic Humanities document. Area inside red frame in (d) DA-RADMM-CI is detailed in Fig. 9. Note that the weights of some experts were consistently near zero and overlapped with the y-axis.

(c) DA-RADMM-I  
(d) DA-RADMM-CI

Fig. 9  Detailed example of mixer outputs in Fig. 8 (d) DA-RADMM-CI. Red words correspond to those dominant for Humanities.

in Fig. 8 (d) DA-RADMM-CI. The domain attributes are indicated by the left and input words shown at the bottom. The words highly weighted by the Humanities expert are shown in red, indicating that those related to Humanities, such as philosophy, Gausen (name of a philosopher), and metaphor, were highly weighted by the Humanities expert.

4.1.5 Discussion of Model Size

The proposed DA-RADMMs have more parameters than the baseline LSTM since they include the baseline LSTM as a common LSTM. To show that DA-RADMM’s improvement over the baseline was not just caused by the larger model size, we compared the relationship of the model size and the performance of both the baseline LSTM and the proposed DA-RADMM. Figure 10 compares the test set perplexities as a function of the number of parameters obtained by the baseline LSTM and the proposed DA-RADMM-CI. The blue lines and orange crosses respectively show the test set perplexities obtained by the baseline LSTM with a different number of hidden layer units $H$, and orange crosses show those obtained from proposed DA-RADMM-CI with optimal number of hidden layer units.

Fig. 10  Test set perplexities as a function of number of parameters. Blue lines show test set perplexities obtained from baseline LSTMs with a different number of hidden layer units $H$, and orange crosses show those obtained from proposed DA-RADMM-CI with optimal number of hidden layer units.

achieved lower perplexities than the baseline with the same number of parameters. This result indicates that the effect of simply increasing the size of the common LSTM is limited, as mentioned in Irie et al.’s work [12], and introducing the experts for the domain attributes is effective for obtaining a better model with more parameters.

4.2 Case Study 2: Domain Attributes for Dialects and Types of Wordings

The second experiment investigated the effectiveness of using dialects and types of wordings as domain attributes.

4.2.1 Experimental Materials

We used the American national corpus (ANC) [28] and the British national corpus (BNC) [29] as text materials. ANC and BNC are composed of two subsets of spoken and written English collected in the United States and the United Kingdom. The written English subset includes such published and unpublished materials as newspapers, research journals, fiction and non-fiction books, leaflets, letters, and many other types of texts. The spoken English subset includes transcriptions of monologues and natural conversations spoken both formally and informally, such as face-to-face meetings, phone conversations, and various types of speeches. We assumed that each subset in ANC and BNC has different wordings and grammars due to the differences in dialects and types of wordings and used these written and spoken subsets from ANC and BNC as domain attributes. Here note that in the training set, no texts from the target domain were included in the training set, although texts from the different attributes of the target domain were included. The English subset was used for hyper-parameter tuning and early stopping, we used 10% of the training data as development data.
Table 6 Statistics of ANC-BNC dataset

| Dialect       | Type         | # of words (train) | # of words (dev) | # of words (test) | Vocab size |
|---------------|--------------|--------------------|------------------|-------------------|------------|
| American      | Spoken       | 19.4M              | 4.19M            | 3.59M             | 14.8k      |
| English       | Written      | 16.4M              | 4.38M            | 4.25M             | 80.9k      |
| British       | Spoken       | 23.6M              | 5.21M            | 5.56M             | 32.8k      |
| English       | Written      | 19.7M              | 2.56M            | 3.99M             | 64.4k      |

Table 7 Data used to train each expert in D-RADMM for ANC-BNC set

| Target domain        | Experts       | Common | Expert 1 | Expert 2 | Expert 3 |
|----------------------|---------------|--------|----------|----------|----------|
| American_Spoken (AS) | AS + BS + BW  | AW     | BS       | BS       | BS       |
| American_Written (AW)| AS + BS + BW  | AS     | BS       | BS       | BS       |
| British_Spoken (BS)  | AS + AW + BW  | AS     | AW       | BS       | BS       |
| British_Written (BW) | AS + AW + BS  | AS     | AW       | BS       | BS       |

Table 8 Data used to train each expert in DA-RADMM for ANC-BNC set

| Target domain        | Experts       | Common | American | British | Spoken | Written |
|----------------------|---------------|--------|----------|---------|--------|---------|
| American_Spoken (AS) | AS + BS + BW  | AW     | BS       | BS      | AW + BS| BS      |
| American_Written (AW)| AS + BS + BW  | AS     | BS       | BS      | BS     | BS      |
| British_Spoken (BS)  | AS + AW + BW  | AS     | BS       | BS      | BS     | BS      |
| British_Written (BW) | AS + AW + BS  | AS     | BS       | BS      | BS     | BS      |

Table 9 RNNLM parameters used for ANC-BNC experiment

| Parameter                  | Value           |
|----------------------------|-----------------|
| Number of projection layer units $E$ | 1024            |
| Number of hidden layers of expert LSTM $H_m$ | 1024            |
| Number of hidden layers of Mixer $H_{m}$ | 512             |
| BPTT                        | 16              |
| Initial learning rate       | 0.1             |
| $\beta$-regularization parameter | $1.0 \times 10^{-6}$ |
| Number of attributes $N$    | 4               |

due to a sufficient amount of training data. Nevertheless, the proposed domain attribute-based RADMMs (DA-RADMM-I and DA-RADMM-CI) keep improving over the baseline LSTM and the conventional D-RADMM. In addition, the DA-RADMM without the attribute indicator, i.e., DA-RADMM-C, exhibited lower performance in all the domains than DA-RADMM-I and DA-RADMM-CI, which took an attribute indicator as an input. This indicates the effectiveness of exploiting domain attribute information as auxiliary information. In this task, the proposed DA-RADMM-CI also outperformed DA-RADMM-I except for British Spoken. We conducted a paired t-test over perplexities obtained by the baseline LSTM, D-RADMM and proposed DA-RADMM-CI on 240 articles for each domain. We found statistical significances between LSTM and DA-RADMM-CI as well as between D-RADMM and DA-RADMM-CI when $p$-value $< 0.01$. This result indicates that the context-dependent weight for each domain attribute is also effective for this task.

4.2.4 Visualization of Mixture Weights

To provide a detailed analysis, we again visualize the mixer outputs from the conventional D-RADMM and the proposed DA-RADMMs. Figures 11, 12, 13, and 14 present the averaged outputs from the Mixer over all the evaluation data obtained by D-RADMM, D-RADMM-C, D-RADMM-I, and DA-RADMM-CI, respectively. The attributes, which comprise the target domain, are shown in red. The words inside the parentheses in the results of the DA-RADMMs indicate the names of domains used to train each expert, and red words indicate matches with target domain.

Unlike the result of the previous experiment in which the common LSTM was dominantly selected, Fig.11 indicates that the conventional D-RADMM tended to select domains that were comprised of one of the attributes of the target document. This could be because more appropriate experts were obtained, which attributed to more training data than the previous experiment. However, compared with our domain attribute-based approach, only a limited amount of data is available to train experts in D-RADMM, and this caused lower perplexities than in the proposed method.

In contrast, Figs. 12, 13, and 14 indicate that the proposed DA-RADMMs assigned high weights to the attribute of the target domain, supplementary using the other experts with lower weights. Specifically, Figs. 13 and 14 especially indicate that the proposed DA-RADMM-I and DA-
**Fig. 11** Average mixer output distributions obtained from conventional RADMM (D-RADMM). Red words in name of domain expert indicate matches with target domain.

**Fig. 12** Average mixer outputs obtained from proposed domain attribute-based RADMM with context-dependent mixer (DA-RADMM-C). The words inside the parentheses indicate the names of domains used to train each expert, and red words indicate matches with target domain.

**Fig. 13** Average mixer outputs obtained from proposed domain attribute-based RADMM with a domain attribute indicator (DA-RADMM-I).

**Fig. 14** Average mixer outputs obtained from proposed domain attribute-based RADMM with context-dependent mixer using domain-attribute indicator (DA-RADMM-CI)

**RADMM-CI** assigned the highest weight to the domain attributes of the target domain, except for British Written. In (d) British Written documents, high and low weights were respectively assigned to the American and Written attributes because such broad types of documents as American books and citations from the speech were included in this domain. In addition, considering that the expert of the American in **DA-RADMM-CI** are trained using both data from American_Written and American_Spoken domains, this result indicates that the expert of American in **DA-RADMM-CI** is more reliable than the expert in **DA-RADMM** trained on only the American_Spoken domain.

Figures 15 and 16 respectively show examples of mixer output and corresponding input sentences in the evaluation set of (a) American Spoken. Note that Fig. 16 shows the sentences taken from the first and latter halves of the document used in Fig. 15 (d). Its latter is composed of short sentences, such as responses and fillers, and its first half is composed of long utterances. This probably caused the relatively low weight of the spoken expert for the first half and a high weight for the latter in Fig. 15 (d). This indicates that the proposed DA-RADMM-CI assigned high weights to the appropriate domain attributes, resulting in the lower perplexities than the other models.

### 5. Conclusion

We extracted useful knowledge in word prediction for the relevant domain attributes, including styles, topics, and dialects, and integrated responses to situations to make language models robust against word prediction on unknown domains. Assuming that each linguistic domain in the text data is characterized based on a combination of its description style and topic, we extended the conventional RADMM such that experts learned a type of wording, writing style, topic, or dialect. These experts were effectively integrated based on current inputs. Experimental comparisons demonstrated that our proposed DA-RADMM yielded improvements in terms of perplexity over a conventional LSTM-LM and a conventional domain-based RADMM. We also visually explained the effect of a mixture of experts. The results demonstrated that the proposed DA-RADMM effectively combined domain attribute information, including the types of wordings, writing styles, topics, and di-
Our future work will explore three directions to expand domain-specific knowledge.

Our future work will explore three directions to expand our model. First, to evaluate our model on scaled-up experiments, we need to analyze it on a larger dataset, including more than a billion words [30]. The proposed method also needs to be analyzed concerning situations when the number of domains and defined domain attributes increases. In addition, we plan to apply the proposed model to language models for the acoustic speech recognition to adapt them between different domains composed of various speaking styles and topics such as natural conversations and lecture speeches. Second, to improve our DA-RADMM to work without prior knowledge of domain attributes, we can introduce the estimation of domain attribute indicators using other domain estimation techniques [31]. Finally, we can introduce more sophisticated techniques to train experts, including adversarial techniques [22], [24], kernel techniques [19], or sub-space approaches [20]. These are all good choices to obtain a more efficient model by generating sparser weights and experts that are more specialized to the corresponding domain attribute.

References

[1] X. Chen, T. Tan, X. Liu, P. Lanchantin, M. Wan, M.J.F. Gales, and P.C. Woodland, “Recurrent neural network language model adaptation for multi-genre broadcast speech recognition,” Proc. INTERSPEECH, pp.3511–3515, 2015.
[2] T. Morioka, N. Tawara, T. Ogawa, A. Ogawa, T. Iwata, and T. Kobayashi, “Language model domain adaptation via recurrent neural networks with domain-shared and domain-specific representations,” Proc. ICASSP, pp.6084–6088, 2018.
[3] J. Brousseau and S.A. Fox, “Dialect-dependent speech recognizers for Canadian and European French,” Proc. ICSLP, pp.1003–1006, 1992.
[4] R.C. Moore and W.D. Lewis, “Intelligent selection of language model training data,” Proc. ACL, pp.220–224, 2010.
[5] R. Masumura, S. Haehm, and A. Ito, “Training a language model using webdata for large vocabulary Japanese spontaneous speech recognition,” Proc. INTERSPEECH, pp.1465–1468, 2011.
[6] Z. T¨uske, K. Irie, R. Sch¨utler, and H. Ney, “Investigation on log-linear interpolation of multi-domain neural network language model,” Proc. ICASSP, pp.6005–6009, 2016.
[7] S. Deena, M. Hasan, M. Doulaty, O. Saz, and T. Hain, “Combining feature and model-based adaptation of RNNLMs for multi-genre broadcast speech recognition,” Proc. INTERSPEECH, pp.2334–2347, 2016.
[8] H. Daume III, “Frustratingly easy domain adaptation,” Proc. ACL, pp.256–263, 2007.
[9] J. Park, X. Liu, M.J.F. Gales, and P.C. Woodland, “Improved neural network based language modelling and adaptation,” Proc. INTERSPEECH, pp.1041–1044, 2010.
[10] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, “Domain separation networks,” Proc. NIPS, pp.343–351, 2016.
[11] K. Young-Bam, S. Karl, and K. Dongchan, “Domain attention with an ensemble of experts,” Proc. ACL, pp.643–653, 2017.
[12] K. Irie, S. Kumar, M. Nischl, and H. Liao, “RADMM: recurrent adaptive mixture model with applications to domain robust language modeling,” Proc. ICASSP, pp.6079–6083, 2018.
[13] J. Zhang, X. Wu, A. Way, and Q. Liu, “Fast gated neural domain adaptation: Language model as a case study,” Proc. COLING, pp.1386–1397, 2016.
[14] E. Garmash and C. Monz, “Ensemble learning for multi-source neural machine translation,” Proc. COLING, pp.1409–1418, 2016.
[15] S. Li and C. Zong, “Multi-domain sentiment classification,” Proc. ACL: HLT, Short Papers, pp.257–260, 2008.
[16] J. Guo, D. Shah, and R. Barzilay, “Multi-source domain adaptation with mixture of experts,” Proc. EMNLP, pp.4694–4703, 2018.
[17] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton, “Adaptive mixtures of local experts,” Neural Comput, vol.3, no.1, pp.79–87, 1991.
[18] G. Blanchard, G. Lee, and C. Scott, “Generalizing from several related classification tasks to a new unlabeled sample,” Proc. NIPS, pp.2178–2186, 2011.
[19] K. Muandet, D. Balduzzi, and B. Schölkopf, “Domain generalization via invariant feature representation,” Proc. ICMIL, pp.10–18, 2013.
[20] C. Gan, T. Yang, and B. Gong, “Learning attributes equals multi-source domain generalization,” Proc. CVPR, pp.87–97, 2016.
[21] M. Ghifary, W.B. Kleijn, M. Zhang, and D. Balduzzi, “Domain generalization for object recognition with multi-task autoencoders,” Proc. ICCV, pp.2551–2559, 2015.
Tomoharu Iwata received the B.S. degree in environmental information from Keio University in 2001, the M.S. degree in arts and sciences from the University of Tokyo in 2003, and the Ph.D. degree in informatics from Kyoto University in 2008. In 2003, he joined NTT Communication Science Laboratories, Japan. From 2012 to 2013, he was a visiting researcher at Machine Learning Group, Department of Engineering, University of Cambridge, UK. He is currently a distinguished researcher at Ueda Research Laboratory of NTT Communication Science Laboratories, Kyoto, Japan. His research interests include data mining, and machine learning.

Hiroto Ashikawa received the B.E. degree and M.E. degree in computer science and engineering from Waseda University, Tokyo, Japan, in 2017 and 2019, respectively. He joined LINE Corporation, Tokyo, Japan, in 2019 and has been engaged in research and development on spoken language processing. His current interest lies in the area of domain-robust end-to-end ASR and machine learning operations. He is a member of ASJ.

Tetsunori Kobayashi received B.E, M.E, and Dr.E. degrees from Waseda University, Japan, in 1980, 1982, and 1985, respectively. In 1985, he joined Hosei University where he served as a lecturer and then as an associate professor. In 1991, he moved to Waseda University and has been a professor there since 1997. He was a visiting researcher in MIT Laboratory for Computer Science, Advanced Telecommunication Laboratory, and NHK Science and Technical Research Laboratory. His research interests include the basics of speech recognition and synthesis and of image processing and applying them to spoken dialogue systems.

Tetsuji Ogawa received his B.S., M.S., and Ph.D. in electrical engineering from Waseda University in Tokyo, Japan, in 2000, 2002, and 2005. He was a Research Associate from 2004 to 2007 and a Visiting Lecturer in 2007 at Waseda University. He was an Assistant Professor at Waseda Institute for Advanced Study from 2007 to 2012. He was an Associate Professor at Waseda University from 2012 to 2016. He has been a Professor at Waseda University since 2019. He was an Adjunct Associate Professor at Egypt-Japan University of Science and Technology (E-JUST) from 2012 to 2015. He was a Visiting Scholar in the Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, from June to September in 2012 and from June to August in 2013, and a Visiting Scholar in Speech Processing Group, Faculty of Information Technology, Brno University of Technology, Czech Republic from June to July in 2014 and May to August in 2015. His research interests include speech and acoustic information processing, image and video information processing, and the sustainable development and operation of pattern recognition systems. He is a member of the IEEEE, IPSJ, the Japanese Society for Artificial Intelligence, the ASJ, the Japan Society of Mechanical Engineers, Japan Wind Energy Association, Japanese Society of Animal Science, and the Japanese Society of Fisheries Science.

Naohiro Tawara received his B.E., M.E., and Ph.D. degrees from Waseda University, Tokyo, Japan in 2010, 2012, and 2017, respectively. After serving as a lecturer at Waseda University, he is working as a research scientist at NTT Communication Science Laboratories, Kyoto, Japan since 2019. He was an Adjunct Associate Professor at Waseda Institute for Advanced Study from 2012 to 2015. He was a Visiting Scholar in the Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, from June to September in 2012 and from June to August in 2013, and a Visiting Scholar in Speech Processing Group, Faculty of Information Technology, Brno University of Technology, Czech Republic from June to July in 2014 and May to August in 2015. His research interests include speech and acoustic information processing, image and video information processing, and the sustainable development and operation of pattern recognition systems. He is a member of the IEEEE, IPSJ, the Japanese Society for Artificial Intelligence, the ASJ, the Japan Society of Mechanical Engineers, Japan Wind Energy Association, Japanese Society of Animal Science, and the Japanese Society of Fisheries Science.

Atsunori Ogawa received the B.E. and M.E. degrees in information engineering, and the Ph.D. degree in information science from Nagoya University, Aichi, Japan, in 1996, 1998, and 2008, respectively. Since 1998, he has been with Nippon Telegraph and Telephone (NTT) Corporation. He has been engaged in researches on speech recognition and speech enhancement at NTT Cyber Space Laboratories and NTT Communication Science Laboratories. He is a member of the IEEE, International Speech Communication Association (ISCA), IEICE, IPSJ, and ASJ. He received the ASJ Best Poster Presentation Awards in 2003 and 2006, respectively.