Grapheme-to-phoneme Conversion based on Adaptive Regularization of Weight Vectors

Keigo Kubo, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi Nakamura

Graduate School of Information Science,
Nara Institute of Science and Technology (NAIST), Japan
keigo-k@is.naist.jp, ssakti@is.naist.jp, neubig@is.naist.jp,
tomoki@is.naist.jp, s-nakamura@is.naist.jp

Abstract

The current state-of-the-art approach in grapheme-to-phoneme (g2p) conversion is structured learning based on the Margin Infused Relaxed Algorithm (MIRA), which is an online discriminative training method for multiclass classification. However, it is known that the aggressive weight update method of MIRA is prone to overfitting, even if the current example is an outlier or noisy. Adaptive Regularization of Weight Vectors (AROW) has been proposed to resolve this problem for binary classification. In addition, AROW's update rule is simpler and more efficient than that of MIRA, allowing for more efficient training. Although AROW has these advantages, it has not been applied to g2p conversion yet. In this paper, we first apply AROW to g2p conversion which is structured learning problem. In an evaluation that employed a dataset including noisy data our proposed approach achieves a 5.3% error reduction rate compared to MIRA implemented in DirecTL+ in terms of phoneme error rate while requiring only 78% the training time.

Index Terms:g2p conversion, out-of-vocabulary word, online discriminative training, structured learning, AROW

1. Introduction

Grapheme-to-phoneme (g2p) conversion is used to estimate the pronunciations of out-of-vocabulary (OOV) words, and is an essential part of large-vocabulary open-domain speech recognition systems [1] and text-to-speech systems [2]. Rule-based approaches [3] and statistical approaches based on methods such as neural networks [4], decision trees [5], and maximum entropy [6] have been proposed for the task. Recently, there are two major statistical approaches in g2p conversion: the joint sequence model [7, 8] and structured learning based on the Margin Infused Relaxed Algorithm (MIRA) [9]. The joint sequence model is a generative model employing joint n-grams for graphemes and phonemes. MIRA is an online discriminative training method for discriminative models of multiclass classification that learns parameters that correctly classify the current instance with a sufficient margin. MIRA has also been expanded to structured learning problems for which there are an extremely large number of candidate answers, such as g2p [10, 11]. Previous reports on MIRA-based g2p note that it outperforms the joint sequence model in terms of word error rate on g2p tasks. However, MIRA is also prone to overfitting, as it updates parameters to correctly classify the current example, even if the current example is an outlier or noisy.

Recently, employing pronunciaions from the World Wide Web as training data for g2p model without a cross-check of language experts has been proposed [12]. In this case, the training data is expected to include a lot of noisy data, and actually, in [12], degrades the performance of the speech recognition system in exchange for improvements of cost and time for dictionary construction. When this sort of noisy data is used to train a g2p system, it is extremely important to have an approach that is highly accurate and robust to overfitting.

Adaptive Regularization of Weight Vectors (AROW) [13] is another online discriminative training method for binary classification that has been proposed as an approach to resolve overfitting. This is achieved by gradually learning parameters to correctly classify the training data, without guaranteeing that the current example is correctly classified. In addition, AROW's update rule is simpler than that of MIRA, allowing for more efficient training. In multiple binary classification tasks, AROW has been shown to outperform the Passive-Aggressive (PA) algorithm [14] which can be regarded as the binary classification equivalent of MIRA. In this paper, we first apply AROW to g2p conversion, which is a structured learning problem. We evaluate the proposed approach on a g2p task, comparing with the joint sequence model and structured learning based on MIRA.

The rest of this paper is organized as follows. In Section 2, we describe g2p conversion based on linear classifiers, which are employed in the existing method based on MIRA and our proposed approach based on AROW. The existing structured learning approach based on MIRA is described in Section 3 and our proposed approach is described in Section 4. In Section 5, we report on a evaluation experiment for our proposed approach on a g2p task which employs a dictionary written in English. Finally, Section 6 states our conclusion.

2. G2p conversion based on linear classifiers

We define g2p conversion as a process to convert a grapheme sequence $\boldsymbol{x}$ into a phoneme sequence $\boldsymbol{y}$. To obtain a correct phoneme sequence $\boldsymbol{y}$ from a grapheme sequence $\boldsymbol{x}$, we employ a linear classifier defined as

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \boldsymbol{w} \cdot \Phi(\boldsymbol{x}, \boldsymbol{y}) \qquad (1)$$

where $\boldsymbol{w}$ indicates the classifier's weight vector and $\Phi(\boldsymbol{x}, \boldsymbol{y})$ indicates a feature vector which consists of arbitrary frequencies such as frequencies of joint n-gram features [11] on $\boldsymbol{x}$ and $\boldsymbol{y}$. In Eq.(1), $\hat{\boldsymbol{y}}$ can be efficiently obtained using dynamic programing. Structured learning can be employed to obtain a $\boldsymbol{w}$ that allows for accurate prediction of the correct phoneme sequence in this framework. In the following sections, we describe two

Algorithm 1 Online structured learning based on MIRA

```
Input:Training dataset $\boldsymbol{D} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_{|\boldsymbol{D}|}, \boldsymbol{y}_{|\boldsymbol{D}|})\}$
Output:$\boldsymbol{w}$
$\boldsymbol{w} = \boldsymbol{0}$
repeat
  for $i = 1$ to $|\boldsymbol{D}|$ do
    Predict $n$-best hypotheses $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_n$ by $\boldsymbol{w} \cdot \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}})$
    Update $\boldsymbol{w}$ by solving the constrained optimization problem of Eq.(2)
  end for
until Stop condition is met
```

structured learning algorithms: one based on MIRA and one based on AROW.

3. Online structured learning using MIRA

In this paper, we define online discriminative training extended to structured learning as online structured learning. Online structured learning based on MIRA for g2p has been proposed in [10]. When the i-th example $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and n-best hypotheses $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_n$ produced by $\boldsymbol{w}_{t-1} \cdot \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}})$ are given, it updates the current weight vector $\boldsymbol{w}_{t-1}$ by solving the constrained optimization problem defined as

$$\begin{aligned} & \min_{\boldsymbol{w}_t} \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\| \\ & \text{s.t.} \quad \forall \hat{\boldsymbol{y}} \in \{\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_n\} \\ & \boldsymbol{w}_t \cdot (\Phi(\boldsymbol{x}_i, \boldsymbol{y}_i) - \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}})) \geq d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}) \end{aligned} \qquad (2)$$

where $\boldsymbol{w}_t$ indicates the weight vector after the update, and $d(\boldsymbol{y}_i, \hat{\boldsymbol{y}})$ indicates the loss incurred by incorrectly classifying $\boldsymbol{y}_i$ as $\hat{\boldsymbol{y}}$. In g2p conversion, the source sequence $\boldsymbol{x}_i$ and the target sequence $\boldsymbol{y}_i$ are a grapheme sequence and a phoneme sequence respectively, and the phoneme error rate of prediction is used as the loss $d(\boldsymbol{y}_i, \hat{\boldsymbol{y}})$. As in Eq.(2), online structured learning based on MIRA employs n-best hypotheses $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_n$ in training. MIRA finds the updated weight vector $\boldsymbol{w}_t$ that correctly classifies the current example $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ with a sufficient margin proportional to the loss of each hypothesis $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_n$, by using a quadratic programming solver. If there are many parameters to be optimized, the quadratic programming problem is difficult to solve in terms of computation cost. For MIRA, the number of parameters to be optimized is equal to the number of hypotheses employed in update. Therefore, to decrease the computational cost, online learning is employed on MIRA instead of batch learning.

The procedure of online structured learning based on MIRA is shown in **Algorithm 1**. In [10], n-best hypotheses $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_n$ are approximately predicted by beam-search pruning based on a monotone phrasal decoder [15].

One known weakness of MIRA is that it is prone to overfitting. Even if the current example is an outlier or noisy, MIRA must classify the current example correctly, and will move the weights as much as is necessary to do so. This can degrade system performance by causing overfitting. In the next section, we describe online structured learning based on AROW, which is more robust in the face of overfitting compared with MIRA.

4. Proposed Approach based on AROW

AROW has been proposed to improve the Confidence Weighted Algorithm (CW) [16, 17], which is an online discriminative training method for binary classification. In this section, we first briefly describe CW and AROW for binary classification, then describe our proposed approach, which extends AROW to online structured learning.

4.1. CW and AROW

CW and AROW are online discriminative training methods for binary classification. Both methods assume that the weight vector $\boldsymbol{w}$ follows a multi-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, where d is the number of features in the model. During prediction, CW and AROW employ the expectation of the weight vector $\mathrm{E}[\boldsymbol{w}] = \boldsymbol{\mu}$. By considering the variance and covariance, CW and AROW control the amount each feature weight is updated after each example. Because the current weight of the features that have frequently occurred and been updated in the past has high confidence, they are not moved excessively on any update. In contrast, because the current weight of the features that have rarely been updated in the past does not have high confidence, they are widely moved on update. This property, which MIRA does not have, prevents the weights that have high confidence from widely moving in directions that degrade the system performance in the presence of outliers.

When the i-th example $(\boldsymbol{x}_i, y_i)$ is given, CW obtains an updated distribution $\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)$ for the weight vector by solving the constrained optimization problem defined as

$$\begin{aligned} (\boldsymbol{\mu}_t, \Sigma_t) = \ & \min_{\boldsymbol{\mu}_t, \Sigma_t} \mathbf{D_{KL}}(\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t) || \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) \\ & \text{s.t.} \quad \Pr_{\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)}[y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i) \geq 0] \geq \eta \end{aligned} \qquad (3)$$

where $\mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})$ is the current distribution for the weight vector, $\mathbf{D_{KL}}(\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t) || \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1}))$ indicates the Kullback-Leibler (**KL**) divergence between the updated distribution and the current distribution, and $\eta \in (0.5, 1]$ is a hyperparameter controlling the margin. Note that $\boldsymbol{x}_i$ and $y_i \in \{-1, +1\}$ here indicate the i-th input vector and the i-th correct label respectively, whereas in our description for structured learning we assume $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ to be the source sequence and the target sequence respectively. As in Eq.(3), CW finds the updated distribution that is closest to the previous distribution while satisfying the constraint that the current example $(\boldsymbol{x}_i, y_i)$ is correctly classified with at least probability $\eta \in (0.5, 1]$. The learning of CW converges quickly, as the constraint of CW forces CW to find the distribution that correctly classifies the current example $(\boldsymbol{x}_i, y_i)$ with at least probability $\eta \in (0.5, 1]$. However, like MIRA, this aggressive learning causes overfitting, since CW has the possibility to widely move even a reliable weight to satisfy this constraint.

To avoid this problem of MIRA and CW, AROW recasts terms for the constraint of CW as regularizers. The distribution found by AROW does not guarantee that the current example $(\boldsymbol{x}_i, y_i)$ is correctly classified. However, the training data comes closer to being correctly classified each time the distribution is updated, and even when an outlier appears, AROW does not widely move the reliable weights in a direction that degrades the system performance.

AROW obtains the updated distribution for the weight vector by solving the unconstrained optimization problem defined as

$$\begin{aligned} (\boldsymbol{\mu}_t, \Sigma_t) = \ & \min_{\boldsymbol{\mu}_t, \Sigma_t} \mathbf{D_{KL}}(\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t) || \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) \\ & + \tfrac{1}{2r} \ell_{h^2}(\boldsymbol{x}_i, y_i, \boldsymbol{\mu}_t) + \tfrac{1}{2r} \boldsymbol{x}_i^{\mathrm{T}} \Sigma_t \boldsymbol{x}_i \end{aligned} \qquad (4)$$

where r is a hyperparameter that has the constraint $r > 0$, and controls the update amount for $\boldsymbol{\mu}$ and Σ. Also, $\ell_{h^2}(\boldsymbol{x}_i, y_i, \boldsymbol{\mu}_t)$ is the loss function defined as

$$\ell_{h^2}(\boldsymbol{x}_i, y_i, \boldsymbol{\mu}_t) = (\max\{0, 1 - y_i(\boldsymbol{\mu}_t \cdot \boldsymbol{x}_i)\})^2. \quad (5)$$

Solving Eq.(4) is equivalent to finding the distribution that decreases the loss function value and variances of each feature that occurred, while avoiding changing the previous distribution as much as possible. In multiple binary classification tasks, AROW has been shown to outperform CW and PA [13]. We propose a method to extend AROW to structured learning in the next sub-section.

4.2. Online structured learning based on AROW

Given the i-th example $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and the hypothesis $\hat{\boldsymbol{y}}_k$, our proposed approach for online structured learning using AROW obtains a updated distribution $\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)$ to minimize the objective function defined as

$$\begin{aligned} L(\boldsymbol{\mu}_t, \Sigma_t) = \; & \mathbf{D_{KL}}(\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t) || \mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})) \\ & + \tfrac{1}{2r}\ell_{h^2}(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_k, \boldsymbol{\mu}_t) + \tfrac{1}{2r}\boldsymbol{u}_{ik}^{\mathrm{T}}\Sigma_t\boldsymbol{u}_{ik} \quad (6) \end{aligned}$$

where $\boldsymbol{u}_{ik}$ is defined as $\Phi(\boldsymbol{x}_i, \boldsymbol{y}_i) - \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_k)$, r is a hyperparameter that has the constraint $r > 0$ as before. And $\ell_{h^2}(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_k, \boldsymbol{\mu}_t)$ is the loss function defined as

$$\ell_{h^2}(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_k, \boldsymbol{\mu}_t) = (\max\{0, d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_k) - \boldsymbol{\mu}_t \cdot \boldsymbol{u}_{ik}\})^2. \quad (7)$$

By partially differentiating Eq.(6) with $\boldsymbol{\mu}_t$ and setting this derivative to 0 so that we minimize Eq.(6), the update formula for $\boldsymbol{\mu}_t$ of online structured learning based on AROW is as follows:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \frac{\max\{0, d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_k) - \boldsymbol{\mu}_t \cdot \boldsymbol{u}_{ik}\}}{\boldsymbol{u}_{ik}^{\mathrm{T}}\Sigma_{t-1}\boldsymbol{u}_{ik} + r}\Sigma_{t-1}\boldsymbol{u}_{ik}. \quad (8)$$

As the full covariance matrix can not be handled as the number of features in g2p conversion is enormous, we assume Σ_t to be a diagonal matrix, as is standard for traditional CW or AROW. We partially differentiate the objective function of Eq.(6) with the p-th diagonal element $(\Sigma_t)_{p,p}$ of Σ_t to obtain the update formula for Σ_t, and then we set the equation to be 0 as follows:

$$\begin{aligned} & \frac{\partial}{\partial(\Sigma_t)_{p,p}} L(\boldsymbol{\mu}_t, \Sigma_t) = \\ & \qquad \frac{1}{2}\left(\frac{1}{(\Sigma_{t-1})_{p,p}} - \frac{1}{(\Sigma_t)_{p,p}} + \frac{(\boldsymbol{u}_{ik})_p^2}{r}\right) = 0 \end{aligned} \quad (9)$$

where $(\boldsymbol{u}_{ik})_p$ indicates the p-th feature value of the $\boldsymbol{u}_{ik}$. We arrange the above equation to solve $(\Sigma_t)_{p,p}$ as follows:

$$(\Sigma_t)_{p,p} = \frac{r(\Sigma_{t-1})_{p,p}}{r + (\boldsymbol{u}_{ik})_p^2(\Sigma_{t-1})_{p,p}}. \quad (10)$$

Each diagonal element $(\Sigma_t)_{p,p}$ for $p = 1, ..., d$ is updated by Eq.(10). Also when $\ell_{h^2}(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_k, \boldsymbol{\mu}_{t-1})$ is equal to 0, $\boldsymbol{\mu}_{t-1}$ and Σ_{t-1} are not updated.

The procedure of online structured learning based on AROW is shown in **Algorithm 2**. $\boldsymbol{\mu}$ and Σ are initialized with the zero vector and identity matrix respectively. From $(\Sigma_0)_{p,p} = 1$, $r > 0$ and Eq.(10), $(\Sigma_{t-1})_{p,p} \geq (\Sigma_t)_{p,p}$ for all t holds. When $(\Sigma_t)_{p,p} = 0$, the p-th feature weight of the $\boldsymbol{\mu}$ is fixed. Therefore, the convergence of **Algorithm 2** is guaranteed. In **Algorithm 2**, n-best hypotheses $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_n$ are also predicted by beam-search pruning based on a monotone phrasal decoder [15], similarly to [10]. The update process for the $\boldsymbol{\mu}$ and the Σ in **Algorithm 2** is similar to sequential update proposed in Multi-Class CW [18]. The difference is that it solves the unconstrained optimization problem over each hypothesis, whereas the sequential update solves the constrained optimization problem. Also, **Algorithm 2** is an online learning algorithm in accordance with MIRA. However, our proposed approach can easily perform batch learning because it does not solve quadratic programming problem.

Algorithm 2 Online structured learning based on AROW

```
Input:Training dataset $\boldsymbol{D} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_{|\boldsymbol{D}|}, \boldsymbol{y}_{|\boldsymbol{D}|})\}$
Output:$\boldsymbol{\mu}$ as weight vector $\boldsymbol{w}$
$\boldsymbol{\mu} = \boldsymbol{0}, \Sigma = \boldsymbol{I}$
repeat
  for $i = 1$ to $|\boldsymbol{D}|$ do
    Predict $n$-best hypotheses $\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_n$ by $\boldsymbol{\mu} \cdot \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}})$
    for $k = 1$ to $n$ do
      if $\ell_{h^2}(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_k, \boldsymbol{\mu}) > 0$ then
        Update $\boldsymbol{\mu}$ and $\Sigma$ by Eq.(8) and Eq.(10) respectively
      end if
    end for
  end for
until Stop condition is met
```

5. Experiment and result

We evaluated our proposed structured learning approach using AROW on a g2p task. Table 1 shows datasets employed in this experiment; dataset name, vocabulary sizes of training data, development data, and test data and the number of trials of cross-validation. In this experiment, we employ the NETtalk dataset which is the English dictionary obtained from the Pascal Letter-to-Phoneme Conversion Challenge[1]. We attempted to faithfully follow the convention in terms of data exclusion and data split in [8], except extracting development data from training data. To confirm that the structured learning approach based on AROW is robust to overfitting, we also create a separate Noisy NETtalk dataset, for which about 10% of the training data is artificial noisy data that has been given a wrong pronunciation randomly chosen from all pronunciations in NETtalk. In Noisy NETtalk, the prediction performance of an approach that is not robust to overfitting can be expected to degrade by overfitting the noisy data.

We employed Sequitur[2] and DirecTL+[3] as baseline g2p conversion tools in this experiment. Sequitur is based on the generative model employing joint n-grams for graphemes and phonemes. DirecTL+ uses online structured learning based on MIRA. DirecTL+ and our proposed approach employed context features, chain features, and joint n-gram features in accordance with [11]. The transition feature introduced in [11] was not used, as it was found to decrease performance in the NETtalk task. For alignment, we used the unconstrained many-to-many alignment method of [19] as implemented in mpaligner[4]. All discriminative methods employ phoneme error rate as their loss

[1] http://pascallin.ecs.soton.ac.uk/Challenges/PRONALSYL/Datasets

[2] http://sequitur.info/

[3] http://code.google.com/p/directl-p/

[4] http://sourceforge.jp/projects/mpaligner/

Table 1: *Dataset used in this experiment on the g2p task. g/p indicates the number of grapheme and phoneme symbols. Noisy indicates the number of artificial noisy data, which has been given a wrong pronunciation randomly. For instance, Noisy NETtalk includes 1760 noisy data of the total vocabulary size 17595. Dev indicates development data to determine various training parameters. K-fold indicates the number of cross-validation folds.*

Dataset	g/p	Vocabulary size			
		Train (Noisy)	Dev	Test	K-fold
NETtalk	26/50	17595 (0)	1000	1000	10
Noisy NETtalk	26/50	17595 (1760)	1000	1000	10

Table 2: *Parameter settings were optimized for each method on the development data, with the parameters employed at least once in each cross-validation fold in bold.*

	Sequitur	DirecTL+	This work
joint n-gram	**5**,**6**,**7**,8,**9**,10	Follow Sequitur	Follow Sequitur
context window	-	**4**,**5**,**6**	Follow DirecTL+
n-best hypotheses	-	1,3,**5**	Follow DirecTL+
hyperparameter r	-	-	**500**,**1000**,**1500**
beam width	-	**150**	**150**

Table 3: *Evaluation result in NETtalk. PER and WER indicate phoneme error rate and word error rate respectively. Time indicates learning time for each approach. "±" indicates the 90% confidence interval.*

	PER(%)	WER(%)	Time(hr.)
Sequitur	7.63%±0.24	31.54%±0.80	**1.1h±0.3**
DirecTL+	**6.75%±0.22**	**28.15%±0.76**	8.6h±1.5
This work	**6.75%±0.20**	**28.56%±0.62**	4.7h±1.0

Table 4: *Evaluation result in Noisy NETtalk. The difference between this work and DirecTL+ for PER on this evaluation is significant according to the paired t-test at a level of 0.05.*

	PER(%)	WER(%)	Time(hr.)
Sequitur	**9.78%±0.23**	34.01%±0.85	**3.3h±1.0**
DirecTL+	10.33%±0.27	**33.52%±0.46**	100.5h±12.1
This work	**9.79%±0.45**	**33.02%±0.95**	78.1h±15.9

function. The context window size, joint n-gram size, hyperparameter r, n-best hypotheses on training, beam width for beam-search pruning, and training iterations were determined by phoneme error rate on the development data. Table 2 shows their details. Also this experiment was performed on cluster machines equipped with Intel Xeon E5649 2.53GHz.

Table 3 shows the evaluation result on NETtalk. It can be seen that the proposed approach significantly outperformed Sequitur in terms of phoneme and word error rate. Compared with DirectTL+, there was no significant difference in accuracy. On the other hand, from the point of view of learning time, the learning speed of our proposed approach was faster than DirecTL+. Since our proposed approach updates only once for each hypothesis included in the n-best, the learning speed of our proposed approach is faster than online structured learning based on MIRA, which has to iteratively seek the $\boldsymbol{w}$ that satisfies the constraints in Eq.(2) by a quadratic programming solver. This result indicates that our proposed approach is more suitable for learning from large dictionaries than online structured learning based on MIRA.

Table 4 shows the evaluation result on Noisy NETtalk. From Table 4, the performance degradation of our proposed approach on noisy data is less than that of DirecTL+. The difference between our proposed approach and DirecTL+ with regards to PER is significant according to the paired t-test at a level of 0.05. This result indicates that the structured learning approach based on AROW resolves MIRA's overfitting problems, as it does for binary classification.[5]

6. Conclusion

We extended AROW to online structured learning and evaluated it on the g2p task. In an evaluation experiment, our proposed approach achieved performance comparable to online structured learning based on MIRA offered by DirecTL+ in terms of phoneme error rate and word error rate. On a dataset including noisy data, our proposed approach outperformed DirecTL+ in terms of phoneme error rate. In addition, the learning speed of our proposed approach was faster than DirecTL+. The result revealed that our proposed approach is more suitable for learning from large dictionaries or dictionaries that are prone to overfitting, such as dictionaries including a large amount of noisy data, than online structured learning based on MIRA.

As future work, we will evaluate our approach with a larger English or other language dataset. And, to further improve our proposed approach, we will consider an approach that approximately handles the covariance between two features in Σ within the limits of memory.

7. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 24240032 and by the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

8. References

[1] L. R. Bahl, S. Das, P. V. Desouza, M. Epstein, R. L. Mercer, B. Merialdo, D. Nahamoo, M. A. Picheny, and J. Powell, "Auto-

[5] It can be noted that training time is significantly higher on noisy NETtalk. This is because artificial noisy data included in Noisy NETtalk generate new and wrong mappings between graphemes and phonemes in the alignment step due to wrong pronunciations. The mappings increase the inferable pronunciation hypotheses $\hat{\boldsymbol{y}}$, and seriously affect time for predicting n-best hypotheses for discriminative training based on MIRA and AROW. It also affects time for calculation of back-off smoothing on joint n-gram model implemented in Sequitur. However it is not a serious problem compared to that of the discriminative training methods. The problem on discriminative training can be controled by beam width in the beam-search pruning or solved a distributed training as proposed in [17].

matic phonetic baseform determination," in *Proc. ICASSP*. IEEE, 1991, pp. 173–176.

[2] J. Schroeter, A. Conkie, A. Syrdal, M. Beutnagel, M. Jilka, V. Strom, Y. J. Kim, H. G. Kang, and D. Kapilow, "A perspective on the next challenges for TTS research," in *IEEE Workshop on Speech Synthesis*, 2002.

[3] R. M. Kaplan and M. Kay, "Regular models of phonological rule systems," *Computational linguistics*, vol. 20, pp. 331–378, 1994.

[4] T. J. Sejnowski and C. R. Rosenberg, "Parallel networks that learn to pronounce English text," *Complex Syst.*, vol. 1, pp. 145–168, 1987.

[5] W. Daelemans and A. van den Bosch, "Language-independent data-oriented grapheme-to-phoneme conversion," in *Progress in Speech Processing*. Springer-Verlag, 1997, pp. 77–89.

[6] S. F. Chen, "Conditional and joint models for grapheme-to-phoneme conversion," in *Proc. EUROSPEECH*, 2003.

[7] S. Deligne and F. Bimbot, "Inference of variable-length linguistic and acoustic units by multigrams," *Speech Communication*, vol. 23, no. 3, pp. 223–241, 1997.

[8] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[9] K. Crammer and Y. Singer, "Ultraconservative online algorithms for multiclass problems," *Journal of Machine Learning Research*, vol. 3, pp. 951–991, 2003.

[10] S. Jiampojamarn and G. Kondrak, "Online discriminative training for grapheme-to-phoneme conversion," in *Proc. INTERSPEECH*, 2009, pp. 1303–1306.

[11] S. Jiampojamarn, C. Cherry, and G. Kondrak, "Integrating joint n-gram features into a discriminative training framework," in *Proc. NAACL-HLT*, 2010, pp. 697–700.

[12] T. Schlippe, S. Ochs, and T. Schultz, "Grapheme-to-phoneme model generation for Indo-European languages," in *Proc. ICASSP*, 2012, pp. 4801–4804.

[13] K. Crammer, A. Kulesza, and M. Dredze, "Adaptive regularization of weight vectors," in *Advances In Neural Information Processing Systems*, vol. 23, 2009, pp. 414–422.

[14] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, pp. 551–585, 2006.

[15] R. Zens and H. Ney., "Improvements in phrase-based statistical machine translation," in *Proc. NAACL HLT*, 2004, pp. 257–264.

[16] M. Dredze, K. Crammer, and F. Pereira, "Confidence-weighted linear classification," in *International Conference On Machine Learning (ICML)*, 2008.

[17] K. Crammer, M. Dredze, and F. Pereira, "Confidence-weighted linear classification for text categorization," *Journal of Machine Learning Research*, vol. 13, pp. 1891–1926, 2012.

[18] K. Crammer, M. Dredze, and A. Kulesza, "Multi-class confidence weighted algorithms," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2009.

[19] K. Kubo, H. Kawanami, H. Saruwatari, and K. Shikano, "Unconstrained many-to-many alignment for automatic pronunciation annotation," in *Proc. APSIPA*, 2011.