Structured Soft Margin Confidence Weighted Learning for Grapheme-to-Phoneme Conversion

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

In recent years, structured online discriminative learning methods using second order statistics have been shown to outperform conventional generative and discriminative models in the grapheme-to-phoneme (g2p) conversion task. However, these methods update the parameters by sequentially using N-best hypotheses predicted with the current parameters. Thus, the parameters appearing in early hypotheses are overfitted compared with those in later hypotheses. In this paper, we propose a novel method called *structured soft margin confidence weighted learning*, which extends multi-class confidence weighted learning to structured learning. The proposed method extends multi-class CW in two ways, allowing for improved robustness to overfitting: (1) regularization inspired by soft margin support vector machines, allowing for margin error, and (2) update using N-best hypotheses simultaneously and interdependently. In an evaluation experiment on the g2p conversion task, the proposed method improved over all other approaches in terms of phoneme error rate with a significant difference.

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

1. Introduction

The speech recognition and text-to-speech fields are increasingly proceeding to open-domain and the multilingual tasks [1]. In such fields, out-of-vocabulary (OOV) words which lack pronunciations are a major bottleneck [2, 3]. In order to solve the OOV problem, grapheme-to-phoneme (g2p) conversion, which predicts the pronunciation for OOV words, is an extremely important component. G2p conversion has been handled using rule-based approaches [4] and statistical approaches based on methods such as neural networks [5], decision trees [6], maximum entropy [7], generative joint sequence models [8, 9], and online discriminative learning [10, 11].

We have proposed structured online discriminative learning methods using second order statistics for g2p conversion, where second order statistics represent the confidence in the current value of a feature weight. The learning methods are methods that extend Adaptive Regularization of Weight Vectors (AROW) [12] and Narrow AROW (NAROW) [13] to structured learning, and are called as structured AROW [14] and structured NAROW [15] respectively. Especially, the performance of structured NAROW outperformed the performance of the conventional joint sequence model and online discriminative learning based on the Margin Infused Relaxed Algorithm (MIRA) [16] in our previous experiment with various g2p conversion tasks.

However, there is still room for improvement in structured NAROW. It updates the parameters by sequentially using N-best hypotheses predicted with the current parameters. Thus, the parameters that appear in early hypotheses are overfitted compared with those in later hypotheses. In addition, it updates the second order statistics independently of the amount of the margin error, which means that the score difference between a correct class and a hypothesis is insufficient. This leads to an increase the second order statistics even during the update of relatively unimportant hypotheses. This narrows the movement of a feature weights unnecessarily. We can hypothesize that the latter problem in structured NAROW can be solved by using multi-class Confidence Weighted Learning (CW) [17], which also uses second order statistics, instead of the structured NAROW. However, multi-class CW tends to overfit the current data due to non-permission of margin error, unlike structured NAROW. In addition, it does not resolve the former problem stated above.

In this paper, we propose a novel method called structured Soft Margin Confidence Weighted Learning (SMCW), which extends multi-class CW to structured learning. The learning method newly introduces two ideas, which the multi-class CW does not have, to solve the above problems: (1) regularization inspired by soft margin support vector machines, allowing us margin error, and (2) update using N-best hypotheses simultaneously and interdependently. We evaluate the effectiveness of structured SMCW on a g2p conversion task compared with the joint sequence model, structured learning based on MIRA, structured AROW, and structured NAROW.

2. G2p Conversion

2.1. Formalization

We define g2p conversion as a process of converting a grapheme sequence $\boldsymbol{x}$ into a phoneme sequence $\boldsymbol{y}$. Given a correct phoneme sequence $\boldsymbol{y}$ for a grapheme sequence $\boldsymbol{x}$, we formalize g2p conversion as

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \boldsymbol{w}^{\mathrm{T}} \Phi(\boldsymbol{x}, \boldsymbol{y}), \tag{1}$$

where $\boldsymbol{w}$ indicates the classifier's weight vector and $\Phi(\boldsymbol{x}, \boldsymbol{y})$ indicates a feature vector which consists of arbitrary values 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 programming. Structured learning can be employed to obtain a $\boldsymbol{w}$ that allows for accurate prediction of the correct phoneme sequence in this framework.

2.2. Existing Structured Learning: Structured NAROW

Structured NAROW assumes that the weight vector follows a multivariate normal distribution $\mathcal{N}(\boldsymbol{w}, \Sigma)$ which has a mean $\boldsymbol{w}$ and a variance-covariance matrix Σ. Σ, roughly speaking, is a learning rate for $\boldsymbol{w}$. Its inverse Σ^{-1} is a second order information matrix representing the confidence of each feature weight. By introducing Σ, structured NAROW avoids excessively moving the weights of the important features that have frequently been updated. This property reduces the overfitting problem over noisy data.

Given N-best hypotheses for a current data point $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and current weights $\boldsymbol{w}_i$, structured NAROW solves the following optimization problems using the hypotheses sequentially to find an optimized $\boldsymbol{w}$ and Σ in $\mathcal{N}(\boldsymbol{w}, \Sigma)$,

$$\arg\min_{\boldsymbol{w}} \sum_{t=1}^{i} \boldsymbol{z}_t^{\mathrm{T}} \boldsymbol{w} + \frac{\boldsymbol{w}^{\mathrm{T}} \Sigma_i^{-1} \boldsymbol{w}}{2}, \tag{2}$$

$$\arg\min_{\Sigma} \mathbf{D_{KL}}(\mathcal{N}(\boldsymbol{w}_i, \Sigma) || \mathcal{N}(\boldsymbol{w}_i, \Sigma_i)) + \frac{\boldsymbol{o}_{in}^{\mathrm{T}} \Sigma \boldsymbol{o}_{in}}{2 r_{in}}, \tag{3}$$

where $\boldsymbol{z}_t$ denotes a subgradient of a loss function in $\partial \ell_t(\boldsymbol{w}_t)$ on the t-th update, and $\boldsymbol{o}_{in}$ is defined as $\Phi(\boldsymbol{x}_i, \boldsymbol{y}_i) - \Phi(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_n)$ which is the difference vector between the feature vector of the correct $\boldsymbol{y}_i$ and the feature vector of the n-th hypothesis $\hat{\boldsymbol{y}}_n$. $\mathbf{D_{KL}}(\mathcal{N}(\boldsymbol{w}_i, \Sigma) || \mathcal{N}(\boldsymbol{w}_i, \Sigma_i))$ is the Kullback-Leibler (**KL**) divergence. The setting of the r_{in} in structured NAROW is $r_{in} = \frac{v_{in}}{b v_{in} - 1}$ when $b v_{in} > 1$ and $r_{in} = +\infty$ otherwise, where v_{in} is $\boldsymbol{o}_{in}^{\mathrm{T}} \Sigma_i \boldsymbol{o}_{in} > 0$, and $b > 0$ is a new hyperparameter to adjust the generalization of learning. This setting is derived based on the minimization of the mistake bound for the structured NAROW and suppresses the excessive rise of the confidence of each feature weight in order to take advantage of the subsequent learning.

In the structured NAROW, the loss function ℓ is defined as

$$\ell_t(\boldsymbol{w}_t) = \max(0, v_t d_t - \boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{o}_t), \tag{4}$$

where v_t and o_t are v_{in} and o_{in} on the t-th update respectively. Also $d_t = d(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)$ indicates the loss value incurred by incorrectly classifying $\boldsymbol{y}_t$ as $\hat{\boldsymbol{y}}_t$ on the t-th update. We define d_t as the number of phoneme errors for the g2p conversion task. As a subgradient $\boldsymbol{z}_t$ of the above loss function, we choose $-\boldsymbol{o}_t$ when margin error $\ell_t(\boldsymbol{w}_t) > 0$ and $\mathbf{0}$ otherwise.

As described above, structured NAROW updates the parameters by sequentially using N-best hypotheses. Also, from Eq.(3), it updates the second order statistics independently of the amount of margin error.

3. Multi-Class CW

Multi-class CW also assumes that a weight vector follows a multivariate normal distribution as with structured NAROW. Given N-best hypotheses for a current data point $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ with current weights $\boldsymbol{w}_i$, multi-class CW solves the following optimization problem using the hypotheses sequentially or in parallel to find an optimized $\boldsymbol{w}$ and Σ,

$$\begin{aligned} \arg\min_{\boldsymbol{w}, \Sigma} \quad & \mathbf{D_{KL}}(\mathcal{N}(\boldsymbol{w}, \Sigma) || \mathcal{N}(\boldsymbol{w}_i, \Sigma_i)) \\ \text{s.t.} \quad & \Pr_{\hat{\boldsymbol{w}} \sim \mathcal{N}(\hat{\boldsymbol{w}}, \Sigma)}[\hat{\boldsymbol{w}}^{\mathrm{T}} \boldsymbol{o}_{in} \geq 0] \geq \eta, \end{aligned} \tag{5}$$

where $\eta \in (0.5, 1]$ is a hyperparameter controlling the margin. The optimization problem in Eq.(5) finds $\boldsymbol{w}$ and Σ such that the probability that the correct answer's score is greater than or equal to the hypothesis's score using a weight vector $\hat{\boldsymbol{w}}$ generated from $\mathcal{N}(\boldsymbol{w}, \Sigma)$ (namely, the probability that $\hat{\boldsymbol{w}}^{\mathrm{T}} \boldsymbol{o}_{in}$ is positive or 0) is greater than η, while minimizing the movement of parameters in $\boldsymbol{w}$ and Σ with the KL divergence criterion.

By converting a normal distribution that regards the difference in the scores of the correct answer and a hypothesis $\hat{\boldsymbol{w}}^{\mathrm{T}} \boldsymbol{o}_{in}$ as a random variable and having a mean 0 and a variance $\boldsymbol{o}_{in}^{\mathrm{T}} \Sigma \boldsymbol{o}_{in}$ to a standard normal distribution, the constraint in Eq.(5) can be rewritten as

$$\boldsymbol{w}^{\mathrm{T}} \boldsymbol{o}_{in} \geq \phi_\eta \sqrt{\boldsymbol{o}_{in}^{\mathrm{T}} \Sigma \boldsymbol{o}_{in}}, \tag{6}$$

where ϕ_η is an output value by the inverse function of the cumulative distribution of the standard normal distribution at probability η, which is equivalent to the margin over a hypothesis. The optimization problem can not be minimized by the square root $\sqrt{\boldsymbol{o}_{in}^{\mathrm{T}} \Sigma \boldsymbol{o}_{in}}$, which is a concave function. Therefore, multi-class CW, in accordance with the CW proposed by Dredze et al. [18], approximated the above constraint as

$$\boldsymbol{w}^{\mathrm{T}} \boldsymbol{o}_{in} \geq \phi_\eta \boldsymbol{o}_{in}^{\mathrm{T}} \Sigma \boldsymbol{o}_{in}. \tag{7}$$

When the above constraint is not satisfied, a margin error occurs. Because Σ is included in the inequality for margin error, an update of second order statistics in the multi-class CW depends on the amount of the margin error, in contrast to structured NAROW.

In multi-class CW, updating parameters by solving Eq.(5) using the hypotheses sequentially like structured NAROW is called *Sequential Update*, and updating parameters by solving Eq.(5) using the hypotheses in parallel and independently and then averaging the results is called *Parallel Update*. From a previous report [17], the performance of Sequential Update and Parallel Update depends on the task. However, as described above, Sequential Update tends to overfit parameters appearing in early hypotheses compared with those in later hypotheses. Also, Parallel Update tends to avoid overfitting the current data, but does not guarantee sufficient discriminative ability over the current data due to the averaging. In addition, because multi-class CW does not allow us margin error, it tends to overfit the current data.

4. Structured SMCW

Multi-class CW can be applied to structured learning problems straightforwardly, although structured learning problems need decoding to obtain a complete feature vector. Thus, we propose structured SMCW, which inherits its basic concept from multi-class CW. In addition, structured SMCW introduces a slack variable to allow us to violate the margin for a hypothesis and updates parameters using N-best hypotheses simultaneously and interdependently, to achieve robustness to overfitting.

Given N-best hypotheses for a current data point $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ with current weights $\boldsymbol{w}_i$, structured SMCW solves the following optimization problem using all the hypotheses simultaneously to find an optimized $\boldsymbol{w}$ and Σ,

$$\begin{aligned} \arg\min_{\boldsymbol{w}, \Sigma} & -\log\det\Sigma + \tfrac{1}{2}\mathrm{Tr}(\Sigma_i^{-1}\Sigma) \\ & + \tfrac{1}{2}(\boldsymbol{w}_i - \boldsymbol{w})^{\mathrm{T}} \Sigma_i^{-1} (\boldsymbol{w}_i - \boldsymbol{w}) + \tfrac{C}{2} \textstyle\sum_n \xi_n^2 \\ \text{s.t.} \quad & \boldsymbol{w}^{\mathrm{T}} \boldsymbol{o}_{in} \geq \left(\frac{d_{in}}{\boldsymbol{o}_{in}^{\mathrm{T}} \Sigma \boldsymbol{o}_{in}} + \frac{1}{r_{in}}\right) \boldsymbol{o}_{in}^{\mathrm{T}} \Sigma \boldsymbol{o}_{in} - \xi_n; \quad \forall n, \end{aligned} \tag{8}$$

where $\boldsymbol{w}_0 = \mathbf{0}$, $\Sigma_0 = \mathrm{I}$, and the first, second, and third terms correspond to terms are related to $\boldsymbol{w}$ and Σ in the Kullback-Leibler divergence of the cost term in Eq.(5) (for the detail of

the derivation for $\mathbf{KL}$ divergence between multivariate normal distributions, refer to [19]). $\frac{C}{2}\sum_n \xi_n^2$ is a regularization term employed in L2 soft margin SVM, where ξ_n is a slack variable for the n-th hypothesis to allow us margin error, and $C > 0$ is a hyperparameter controlling the generalization for learning. Also, ϕ_η in Multi-class CW is replaced with $\frac{d_{in}}{\boldsymbol{o}_{in}^{\mathrm{T}}\Sigma\boldsymbol{o}_{in}} + \frac{1}{r_{in}}$, where d_{in} is a loss value equivalent to the number of phoneme errors in g2p conversion and r_{in} is the same as that in structured NAROW. By this setting, when C in a regularization term and b included in r_{in} are $+\infty$ and 0 respectively, this learning method can be regarded as MIRA. Also, r_{in} plays a role to suppress the excessive rise of the confidence of each feature weight as with structured NAROW.

To solve the above constrained optimization problem, we firstly obtain the dual problem by using Lagrange multipliers as

$$\begin{aligned} L = & -\log\det\Sigma + \tfrac{1}{2}\mathrm{Tr}(\Sigma_i^{-1}\Sigma) \\ & + \tfrac{1}{2}(\boldsymbol{w}_i - \boldsymbol{w})^{\mathrm{T}}\Sigma_i^{-1}(\boldsymbol{w}_i - \boldsymbol{w}) + \tfrac{C}{2}\textstyle\sum_n \xi_n^2 \\ & + \textstyle\sum_n \alpha_n(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{o}_{in} + d_{in} + \frac{1}{r_{in}}\boldsymbol{o}_{in}^{\mathrm{T}}\Sigma\boldsymbol{o}_{in} - \xi_n), \end{aligned} \tag{9}$$

where $\alpha_n \geq 0$ is a Lagrange multiplier for the n-th hypothesis. We regard Σ_i as a diagonal matrix having Σ_{ip} as diagonal elements, where p represents an index of the feature. By partially differentiating with $\boldsymbol{w}$, Σ_p, and ξ_n respectively and setting them to 0, we get

$$\boldsymbol{w} = \boldsymbol{w}_i + \Sigma_i(\sum_n \alpha_n \boldsymbol{o}_{in}), \tag{10}$$

$$\Sigma_p = \frac{\Sigma_{ip}}{1 + 2\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v}_{ip}}, \tag{11}$$

$$\xi_n = \frac{\alpha_n}{C}, \tag{12}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{v}_{ip}$ are defined as $(\alpha_1, \cdots, \alpha_N)^{\mathrm{T}}$ and $(v_{i1p}, \cdots, v_{iNp})^{\mathrm{T}} = (\frac{1}{r_{i1}}\Sigma_{ip}o_{i1p}^2, \cdots, \frac{1}{r_{iN}}\Sigma_{ip}o_{iNp}^2)^{\mathrm{T}}$ respectively. From Eq.(10) and Eq.(11), Lagrange multipliers $\boldsymbol{\alpha}$ indicate the learning rate for each hypothesis.

By substituting Eq.(10), Eq.(11), and Eq.(12) to Eq.(9), we get the following dual problem for Eq.(8),

$$\begin{aligned} \arg\max_{\boldsymbol{\alpha}} \quad & \tfrac{1}{2}\textstyle\sum_p \log(1 + 2\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v}_{ip}) + \tfrac{1}{2}\sum_p \frac{1}{1+2\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v}_{ip}} \\ & + \textstyle\sum_p \frac{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v}_{ip}}{1+2\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v}_{ip}} + \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{c} - \tfrac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}Q\boldsymbol{\alpha} \\ & \text{s.t.} \;\; \alpha_n \geq 0; \;\; \forall n, \end{aligned} \tag{13}$$

where the n-th element in vector $\hat{\boldsymbol{c}}$ and the element on the n-th row and the m-th column in matrix $\hat{Q}$ are

$$c_n = -\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{o}_{in} + d_{in}, \tag{14}$$

$$Q_{nm} = \boldsymbol{o}_{in}^{\mathrm{T}}\Sigma_i\boldsymbol{o}_{im} + \frac{\delta_{nm}}{C}, \tag{15}$$

where δ_{nm} is 1 when $n = m$, and 0 otherwise. It is difficult to find an optimal $\boldsymbol{\alpha}$ due to the first, second, and third terms in Eq.(13). However, we note that the increment of the second term and the decrement of the third term by a change in $\boldsymbol{\alpha}$ cancel each other out, and vice versa. Thus, the second and third terms can be ignored for finding an optimal $\boldsymbol{\alpha}$ without any approximation[1]. For the remaining first term, we apply a second-order Taylor series expansion at $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_N)^{\mathrm{T}}$. By applying the second-order Taylor series expansion to the first term

[1]It can be confirmed with applying the arbitrary order Taylor series expansion to the two terms and then extracting terms related to $\boldsymbol{\alpha}$ or drawing a graph for the two terms on the change of an α_n.

Table 1: *Datasets used in the g2p conversion experiment; dataset name (Dataset), the number of grapheme and phoneme symbols (g/p), vocabulary sizes of training data (Train), development data (Dev), and test data (Test) and the number of trials of cross-validation (K-fold).*

Dataset	g/p	Vocabulary size			
		Train	Dev	Test	K-fold
NETtalk	26/50	17595	1000	1000	10
Brulex	40/39	23353	1373	2747	5
CELEX English	26/53	39995	5000	15000	1
CMUdict	27/39	100886	5941	12000	2

and extracting terms related to $\boldsymbol{\alpha}$, we get the following quadratic programming problem,

$$\arg\max_{\boldsymbol{\alpha}} \;\; \boldsymbol{\alpha}^{\mathrm{T}}\hat{\boldsymbol{c}} - \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\hat{Q}\boldsymbol{\alpha} \;\; \text{s.t.} \;\; \alpha_n \geq 0; \;\; \forall n, \tag{16}$$

where the n-th element in vector $\hat{\boldsymbol{c}}$ and the element on the n-th row and the m-th column in matrix $\hat{Q}$ are

$$\hat{c}_n = -\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{o}_{in} + d_{in} + \sum_p \frac{v_{inp}(1 + 4\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{v}_{ip})}{(1 + 2\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{v}_{ip})^2}, \tag{17}$$

$$\hat{Q}_{nm} = \boldsymbol{o}_{in}^{\mathrm{T}}\Sigma_i\boldsymbol{o}_{im} + \frac{\delta_{nm}}{C} + 2\sum_p \frac{v_{inp}v_{imp}}{(1 + 2\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{v}_{ip})^2}. \tag{18}$$

Because all elements in $\hat{Q}$ are non-negative, in the range of $\alpha_n \geq 0$ for all n, the function maximized in Eq.(16) is a concave function and the quadratic programming problem guarantees that it has only one local optimal solution, which is the global optimal solution. The quadratic programming problem can be solved by the interior point methods [20].

From the function in Eq.(16), the greater $\frac{1}{C}$, the smaller the optimal α_n for all n. Thus, a greater $\frac{1}{C}$ implies better generalization for learning. Also, due to non-diagonal elements in $\hat{Q}$, when a hypothesis is similar to many other hypotheses with respect to features and the variance of its shared features is high, α_n of the hypothesis does not become unnecessarily high. Namely, α_n for all n are determined by interdependences of all the hypotheses. This avoids overfitting while guaranteeing discriminative ability over the current data.

To improve the accuracy of the approximation in the second-order Taylor series expansion of the first term, we would like to select $\hat{\alpha}_n$ close to the optimal α_n for all n. We do so by first setting arbitrary initial values of $\hat{\alpha}_n$ for all n. The second step solves the above quadratic programming problem with current $\hat{\alpha}_n$. The third step sets $\hat{\alpha}_n$ to the optimal α_n obtained in the second step. By repeating the second step and the third step, an $\hat{\alpha}_n$ that improves the approximation can be found. However, the computation costs is expensive because it is necessary to recalculate for $\hat{\boldsymbol{c}}$ and $\hat{Q}$ in Eq.(16) for new $\hat{\alpha}_n$. Thus, in this paper, we select a better $\hat{\alpha}_n$ close to the optimal α_n for all n as below, and do not repeat the update of $\hat{\alpha}_n$. We first consider an $\hat{\alpha}_n$ that can satisfy the constraint for the n-th hypothesis in Eq.(8) by itself, as a better $\hat{\alpha}_n$. If $\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{o}_{in} \geq d_{in} + \frac{1}{r_{in}}\boldsymbol{o}_{in}^{\mathrm{T}}\Sigma_i\boldsymbol{o}_{in}$ holds, such an $\hat{\alpha}_n$ is 0 in terms of the minimization of KL divergence. If it does not hold, such an $\hat{\alpha}_n$ can be derived as below by substituting Eq.(10) and Eq.(12) into the constraint in Eq.(8) and

Table 2: *Evaluation result of each method for phoneme error rate (PER) and word error rate (WER) in the g2p conversion task. Values on NETtalk and Brulex in this table are obtained by averaging results on each cross-validation.*

Dataset	Measure	Joint	MIRA	SAROW	SNAROW	SSMCW
NETtalk	PER(%)	7.71	6.70	6.75	6.53	**6.37**
	WER(%)	31.6	28.18	28.66	27.97	**27.34**
Brulex	PER(%)	1.26	1.03	1.09	**0.99**	**0.99**
	WER(%)	6.57	**5.24**	5.59	**5.14**	**5.11**
CELEX	PER(%)	2.62	2.39	2.51	2.30	**2.24**
English	WER(%)	12.15	11.07	11.81	11.17	**10.71**
CMUdict	PER(%)	6.77	6.19	**6.15**	**6.11**	**6.09**
	WER(%)	28.55	**26.35**	**26.48**	**26.46**	**26.28**

using the fact that $\boldsymbol{o}_{in}^{\mathrm{T}}\Sigma_i\boldsymbol{o}_{in} \geq \boldsymbol{o}_{in}^{\mathrm{T}}\Sigma\boldsymbol{o}_{in}$ from Eq.(11),

$$\begin{aligned}&\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{o}_{in} + \alpha_n\boldsymbol{o}_{in}^{\mathrm{T}}\Sigma_i\boldsymbol{o}_{in} - d_{in} + \frac{\alpha_n}{C}\\ &\geq \frac{1}{r_{in}}\boldsymbol{o}_{in}^{\mathrm{T}}\Sigma_i\boldsymbol{o}_{in} \geq \frac{1}{r_{in}}\boldsymbol{o}_{in}^{\mathrm{T}}\Sigma\boldsymbol{o}_{in}. \end{aligned} \tag{19}$$

Note that we consider $\hat{\alpha}_n$ that can satisfy the constraint after an updating with $\hat{\alpha}_m = 0$ for all m with $m \neq n$, and thus $\boldsymbol{w} = \boldsymbol{w}_i + \alpha_n\Sigma_i\boldsymbol{o}_{in}$ for the substitution by Eq.(10). Then, we get a lower bound for such an $\hat{\alpha}_n$ by obtaining α_n that satisfies the inequality of the left part and the middle part for the above as

$$\alpha_n \geq \frac{-\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{o}_{in} + d_{in} + \frac{1}{r_{in}}\boldsymbol{o}_{in}^{\mathrm{T}}\Sigma_i\boldsymbol{o}_{in}}{\boldsymbol{o}_{in}^{\mathrm{T}}\Sigma_i\boldsymbol{o}_{in} + \frac{1}{C}}. \tag{20}$$

We secondly consider the value of N for N-best hypotheses. The greater the value of N, the smaller the optimal α_n for all n will become. Thus, we define $\hat{\alpha}_n$ for all n as

$$\begin{cases} 0 & \text{if } \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{o}_{in} \\ & \quad \geq d_{in} + \frac{1}{r_{in}}\boldsymbol{o}_{in}^{\mathrm{T}}\Sigma_i\boldsymbol{o}_{in} \\ \frac{1}{N}\frac{-\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{o}_{in} + d_{in} + \frac{\boldsymbol{o}_{in}^{\mathrm{T}}\Sigma_i\boldsymbol{o}_{in}}{r_{in}}}{\boldsymbol{o}_{in}^{\mathrm{T}}\Sigma_i\boldsymbol{o}_{in} + \frac{1}{C}} & \text{otherwise.} \end{cases} \tag{21}$$

5. Experiment and Result

We evaluated our structured SMCW on the g2p conversion task. Table 1 shows g2p conversion task datasets employed in the experiment. The development data is employed to determine the optimal number of training iterations and parameters for each method. For the datasets in Table 1, NETtalk (English) and Brulex (French) were obtained from the Pascal Letter-to-Phoneme Conversion Challenge[2]. CMUdict (English) and CELEX (English) were also obtained from their corresponding Web pages[3][4]. We attempted to faithfully follow the convention in [9] in terms of data exclusion and data split, except extracting development data from training data[5].

We employed the joint sequence model (Joint) implemented in Sequitur[6], MIRA implemented in DirecTL+[7], structured AROW (SAROW), and structured NAROW (SNAROW) implemented in slearp[8] as baseline g2p conversion tools. MIRA, SAROW, SNAROW, and structured SMCW (SSMCW), which is our proposed approach and implemented in slearp, 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 used in MIRA, SAROW, SNAROW and SSMCW, we used the unconstrained many-to-many alignment method of [21] as implemented in mpaligner[9]. For the context window size, joint n-gram size, hyperparameter r for SAROW, hyperparameter b for SNAROW, and N-best hypotheses for training, we refer to our previous work [14] and [15], except hyperparameter b and C for SSMCW. 0.0075, 0.01, and 0.0125 for b and 10, 100, and 1000 for C were tried as the hyperparameters for SSMCW. The training iterations and the hyperparameters for each method are uniquely determined by phoneme error rate on the development data.

Table 2 shows the evaluation result on the g2p conversion task. The best performance and performances that have no significant difference according to Paired Bootstrap Resampling [22] at a level of 0.05 over the best performance are written in bold. From Table 2, for NETtalk and CELEX, SSMCW improved over all other approaches with a significant difference in terms of both PER and WER. The error rate reductions (ERR) for PER in NETtalk and CELEX were 17.4% and 14.5% over Joint, 4.9% and 6.3% over MIRA, 5.6% and 10.8% over SAROW, 2.5% over 2.6% SNAROW. Also, in Brulex and CMUdict, SSMCW obtains top-class performances compared with other approach. The result implies that our proposed method significantly contribute to the prediction performance.

6. Conclusion

We proposed structured SMCW, enhancing multi-class CW with regularization inspired by soft margin support vector machine and update using N-best hypotheses simultaneously and interdependently. We evaluated the proposed approach on g2p conversion tasks with our proposed approach significantly improving over all other approaches, obtaining 2.5%-17.4% ERR for PER in NETtalk and CELEX.

7. Acknowledgments

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.

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

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[4] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC96L14

[5] Dev and Test in NETtalk and other datasets are selected from the dataset at regular intervals in order, to be different for each fold. We termed this *cross-validation* in accordance with [9].

[6] http://sequitur.info/

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

[8] http://sourceforge.jp/projects/slearp/

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

8. References

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

[2] L. Bahl, S. Das, P. Desouza, M. Epstein, R. Mercer, B. Merialdo, D. Nahamoo, M. Picheny, and J. Powell, "Automatic phonetic baseform determination," in *Proc. ICASSP*. IEEE, 1991, pp. 173–176.

[3] 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 *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, 2002, pp. 211–214.

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

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

[6] 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.

[7] S. F. Chen *et al.*, "Conditional and joint models for grapheme-to-phoneme conversion," in *Proc. EUROSPEECH*, 2003, pp. 2033–2036.

[8] 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.

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

[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] 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.

[13] F. Orabona and K. Crammer, "New adaptive algorithms for online classification," in *Proc. NIPS*, 2010, pp. 1840–1848.

[14] K. Kubo, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Grapheme-to-phoneme conversion based on adaptive regularization of weight vectors," in *Proc. INTERSPEECH*, 2013, pp. 1946–1950.

[15] ——, "Narrow adaptive regularization of weights for grapheme-to-phoneme conversion," in *Proc. ICASSP*, 2014, pp. 2608–2612.

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

[17] K. Crammer, M. Dredze, and A. Kulesza, "Multi-class confidence weighted algorithms," in *Proc. EMNLP*, vol. 2, 2009, pp. 496–504.

[18] M. Dredze, K. Crammer, and F. Pereira, "Confidence-weighted linear classification," in *Proc. ICML*, 2008, pp. 264–271.

[19] J. Duchi, "Derivations for linear algebra and optimization," http://www.cs.berkeley.edu/˜jduchi/projects/general_notes.pdf.

[20] R. J. Vanderbei, "Loqo: An interior point code for quadratic programming," *Optimization methods and software*, vol. 11, no. 1-4, pp. 451–484, 1999.

[21] K. Kubo, H. Kawanami, H. Saruwatari, and K. Shikano, "Unconstrained many-to-many alignment for automatic pronunciation annotation," in *Proc. APSIPA*, 2011, pp. 1–4.

[22] P. Koehn, "Statistical significance tests for machine translation evaluation." in *Proc. EMNLP*, 2004, pp. 388–395.