A Prior Model of Structural SVMs for Domain Adaptation

Changki Lee and Myung-Gil Jang

In this paper, we study the problem of domain adaptation for structural support vector machines (SVMs). We consider a number of domain adaptation approaches for structural SVMs and evaluate them on named entity recognition, part-of-speech tagging, and sentiment classification problems. Finally, we show that a prior model for structural SVMs outperforms other domain adaptation approaches in most cases. Moreover, the training time for this prior model is reduced compared to other domain adaptation methods with improvements in performance.

Keywords: Domain adaptation, structural SVMs, PRIOR model for structural SVMs.

I. Introduction

Large-margin methods for structured output prediction, such as maximum-margin Markov networks [1] and structural support vector machines (SVMs) [2], have recently received substantial interest in natural language processing [3]-[6], bioinformatics [7], and information retrieval [8].

For structural SVMs, Tsochantaridis presented a cutting-plane algorithm that takes $O(1/\varepsilon^2)$ iterations to reach a desired precision ε [2]. Teo suggested a bundle method [9], while Joachims proposed a 1-slack formulation of structural SVMs, which is very similar to the bundle method [10]. The 1-slack algorithm is substantially faster than existing methods such as sequential minimal optimization (SMO) and SVM-light. The convergence rate of the 1-slack algorithm is $O(1/\varepsilon)$. However, domain adaptation methods are not exploited in structural SVMs.

The task of domain adaptation is to develop learning algorithms that can be easily ported from one domain to another, for example, a TV domain to a sports domain. This problem is particularly interesting in natural language processing and bioinformatics because we are often in situations in which we have a large collection of labeled data in one source domain but truly desire a model that performs well in a second target domain.

Two varieties of the domain adaptation problem have been addressed in the literature: supervised and semi-supervised cases. In a supervised case, we have a large annotated corpus in the source domain and a small corpus in the target domain. We want to leverage both annotated datasets to obtain a model that performs well on the target domain. The semi-supervised case is similar, but instead of having a small annotated target corpus, we have a large but unannotated target corpus. In this paper, we focus exclusively on the supervised case.

In this paper, we consider a number of domain adaptation

Manuscript received Sept. 29, 2010; revised Mar. 25, 2011; accepted Apr. 12, 2011.

Changki Lee (phone: +82 42 860 6879, email: leeck@etri.re.kr) and Myung-Gil Jang (email: mgjang@etri.re.kr) are with the Software Research Laboratory, ETRI, Daejeon, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.11.0110.0571

approaches for structural SVMs and evaluate them on named entity recognition, part-of-speech tagging, and sentiment classification problems. We show that a prior model for 1-slack structural SVMs outperforms other domain adaptation methods in most cases.

The rest of this paper is organized as follows. In section II, we give an overview of related work. Section III describes the 1-slack structural SVMs. Section IV describes our proposed prior model for 1-slack structural SVMs. Section V provides the experimental setup and results. Finally, the last section offers some concluding remarks.

II. Previous Work

There are several ways to solve the domain adaptation problem without developing new algorithms. Many of these have been presented and evaluated by Daumé and Marcu [11] as follows:

- The SRC-ONLY baseline ignores the target data and trains a single model only on the source data.
- The TGT-ONLY baseline trains a single model only on the target data.
- The ALL baseline simply trains a standard learning algorithm on the union of the two datasets.
- The PRED baseline is based on the idea of using the output of the source classifier as a feature in the target classifier. Specifically, we first train the SRC-ONLY model. We then run the SRC-ONLY model on the target data. We use the predictions made by the SRC-ONLY model as additional features and train a second model on the target data augmented with this new feature.
- In the LIN-INT baseline, we linearly interpolate the predictions of the SRC-ONLY and TGT-ONLY models. The interpolation parameter is adjusted based on target development data.

Chelba and Acero introduced the PRIOR model for maximum entropy classifiers [12]. The idea of this model is to use the SRC-ONLY model as a *prior* on the weights for a second model trained on the target data. The model trained on the target data "prefers" to have weights that are similar to the weights from the SRC-ONLY model, unless the data demands otherwise.

Daumé and Marcu presented the MEGA model for domain adaptation for maximum entropy classifiers [11]. The key idea of their approach is to learn three separate models: source-specific, target-specific, and general models. They present an expectation maximization (EM) algorithm for training the model. This model consistently outperformed all baseline approaches as well as the prior model. However, it is quite complex to implement and is roughly 10 times to 15 times slower to train than the prior model.

Daumé proposed a feature augmentation method to augment features for domain adaptation [13]. The augmented features are used to construct a kernel function for kernel methods.

Yang and others proposed adaptive support vector machine (A-SVM) for learning a new SVM classifier for a target domain, which is adapted from an existing classifier trained on the source domain [14]. However, A-SVM is slow in terms of the testing time because it employs auxiliary classifiers for the label prediction of patterns in the target domain.

Our proposed model, a prior model for structural SVMs, is different from A-SVM and feature augmentation, as our model directly adapts an existing model to new data and avoids training time overhead for existing data.

III. 1-Slack Structural SVMs

Structured classification is a problem of predicting y from x in cases in which y has a meaningful internal structure. For example, x might be a word string and y a sequence of part of speech labels. Alternatively, y might be a parse tree of x. The approach is to learn the discriminant function $f : X \times Y \to R$ over <*input, output*> pairs from which we can derive a prediction by maximizing f over the response variable for a specific given input x. Throughout this paper, we assume f to be linear in certain combined feature representations of inputs and outputs $\Psi(x, y)$, $f(x, y; \mathbf{w}) = \mathbf{w}^T \Psi(x, y)$.

The specific form of $\Psi(x, y)$ depends on the nature of the problem. An example of part of speech tagging is shown in Fig. 1.

To deal with problems in which $|Y|$ is very large, Tsochantaridis proposed two approaches; namely, slack rescaling and margin rescaling [2]. In the case of margin rescaling, which we consider in this paper, training a structural SVM amounts to solving the following quadratic program. For convenience, we define $\delta\Psi_i(x_i, y) \equiv \Psi(x_i, y_i) - \Psi(x_i, y)$, where (x_i, y_i) is the training data:

$$\min_{w,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_i^n \xi_i \quad \text{, s.t. } \forall i, \xi_i \geq 0,$$
$$\forall i, \forall y \in Y \setminus y_i : \mathbf{w}^T \delta\Psi_i(x_i, y) \geq \Delta(y_i, y) - \xi_i. \qquad (1)$$

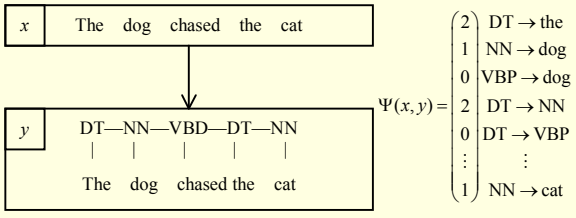

Fig. 1. Example of part of speech tagging model.

This formulation is referred to as an "*n*-slack" structural SVM because it assigns a different slack variable to each *n* training example. Tsochantaridis presented a cutting-plane algorithm that requires $O(n/\varepsilon^2)$ constraints for any desired precision ε [2].

Joachims proposed an alternative formulation of the SVM optimization problem for predicting structured outputs [10]. The key idea is to replace the *n* cutting-plane models of each hinge loss with a single cutting plane model for the sum of the hinge losses. Since there is only a single-slack variable, the new formulation is referred to as "1-slack" structural SVMs.

$$\min_{w,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\xi \quad \text{, s.t.} \quad \forall i, \xi \geq 0,$$

$$\forall (\hat{y}_1, ..., \hat{y}_n) \in Y^n : \frac{1}{n}\mathbf{w}^T \sum_{i=1}^{n} \delta\Psi_i(x_i, \hat{y}_i) \geq \frac{1}{n}\sum_{i=1}^{n} \Delta(y_i, \hat{y}_i) - \xi. \tag{2}$$

While 1-slack formulations have $|Y|^n$ constraints, one for each possible combination of labels $(\hat{y}_1, ..., \hat{y}_n) \in Y^n$, they have only one slack variable ξ, which is shared across all constraints. Interestingly, the objective functions of the *n*-slack and 1-slack formulations are equal [10].

Joachims showed that the dual form of the 1-slack formulation has a solution that is extremely sparse with the number of non-zero dual variables independent of the number of training examples and that the convergence rate is $O(1/\varepsilon)$ [10].

IV. Domain Adaptation for 1-Slack Structural SVMs

We extended structural SVMs for the domain adaptation problem using the prior model. We used the margin rescaling formula of 1-slack structural SVMs for the prior model as follows:

$$\min_{w,\xi} \frac{1}{2}\|\mathbf{w} - \mathbf{w}_0\|^2 + C\xi \quad \text{, s.t.} \quad \forall i, \xi \geq 0,$$

$$\forall (\hat{y}_1, ..., \hat{y}_n) \in Y^n : \frac{1}{n}\mathbf{w}^T \sum_{i=1}^{n} \delta\Psi_i(x_i, \hat{y}_i) \geq \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_i) - \xi, \tag{3}$$

where $\mathbf{w}_0$ is the weight vector learned in the SRC-ONLY model, (x_i, y_i) is a training example, and $L(y_i, y)$ is a real-valued loss for output y relative to the correct output y_i. Unlike regular SVMs, structural SVMs can predict complex y outputs, such as trees, sequences, or sets.

We denote the vectors as

$$\hat{\mathbf{y}} = (\hat{y}_1, ..., \hat{y}_n) \in Y^n,$$

$$L(\hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_i),$$

$$\delta\Psi(\hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n} \delta\Psi_i(x_i, \hat{y}_i).$$

We can solve the optimization problem of the prior model for 1-slack structural SVMs in (4) using standard Lagrangian duality techniques:

$$\min_{w,\xi} L_{\mathrm{P}}(\mathbf{w}, \xi) = \frac{1}{2}\|\mathbf{w} - \mathbf{w}_0\|^2 + C\xi + \sum_{\hat{\mathbf{y}} \in Y^n} \alpha_{\hat{\mathbf{y}}} (L(\hat{\mathbf{y}}) - \mathbf{w}^T \delta\Psi(\hat{\mathbf{y}}) - \xi) - \beta\xi, \tag{4}$$

$$\text{s.t.} \quad \forall \hat{\mathbf{y}}, \alpha_{\hat{\mathbf{y}}} \geq 0, \quad \beta \geq 0, \quad \xi \geq 0,$$

$$\frac{\partial L_{\mathrm{P}}}{\partial \mathbf{w}} = \mathbf{w} - \mathbf{w}_0 - \sum_{\hat{\mathbf{y}} \in Y^n} \alpha_{\hat{\mathbf{y}}} \delta\Psi(\hat{\mathbf{y}}) = 0, \tag{5}$$

$$\frac{\partial L_{\mathrm{P}}}{\partial \xi} = C - \sum_{\hat{\mathbf{y}} \in Y^n} \alpha_{\hat{\mathbf{y}}} - \beta = 0. \tag{6}$$

Substituting (5) and (6) into (4), we obtain the following dual form, which is a quadratic programming (QP) problem where the objective function L_{D} is solely dependent on a set of Lagrangian multipliers:

$$\max_{\alpha} : L_{\mathrm{D}}(\alpha) = \sum_{\hat{\mathbf{y}} \in Y^n} L(\hat{\mathbf{y}})\alpha_{\hat{\mathbf{y}}} - \frac{1}{2}\sum_{\hat{\mathbf{y}} \in Y^n}\sum_{\hat{\mathbf{y}}' \in Y^n} \alpha_{\hat{\mathbf{y}}}\alpha_{\hat{\mathbf{y}}'} \delta\Psi(\hat{\mathbf{y}})^T \delta\Psi(\hat{\mathbf{y}}') - \mathbf{w}_0{}^T \sum_{\hat{\mathbf{y}} \in Y^n} \alpha_{\hat{\mathbf{y}}} \delta\Psi(\hat{\mathbf{y}}),$$

$$\text{s.t.} \quad \forall \hat{\mathbf{y}}, \quad 0 \leq \sum_{\hat{\mathbf{y}} \in Y^n} \alpha_{\hat{\mathbf{y}}} \leq C, \quad \alpha_{\hat{\mathbf{y}}} \geq 0. \tag{7}$$

The only difference in the dual form of 1-slack structural SVMs and that of the prior model for 1-slack structural SVMs in (7) is that the latter contains an extra term $\mathbf{w}_0$. The extremum of the object function L_{D} is at

$$\frac{\partial L_{\mathrm{D}}(\alpha)}{\partial \alpha_{\hat{\mathbf{y}}}} = L(\hat{\mathbf{y}}) - \left(\mathbf{w}_0 + \sum_{\hat{\mathbf{y}}' \in Y^n} \alpha_{\hat{\mathbf{y}}'} \delta\Psi(\hat{\mathbf{y}}') \right)^T \delta\Psi(\hat{\mathbf{y}})$$

$$= L(\hat{\mathbf{y}}) - \mathbf{w}^T \delta\Psi(\hat{\mathbf{y}}) = 0, \tag{8}$$

$$\text{where} \quad \mathbf{w} = \mathbf{w}_0 + \sum_{\hat{\mathbf{y}}' \in Y^n} \alpha_{\hat{\mathbf{y}}'} \delta\Psi(\hat{\mathbf{y}}').$$

Let $\alpha_{\hat{\mathbf{y}}}^{\text{new}} = \alpha_{\hat{\mathbf{y}}} + s$ and $\mathbf{w}^{\text{new}} = \mathbf{w} + s\delta\Psi(\hat{\mathbf{y}})$. We can then obtain the following equation from (8):

$$\mathbf{w}^{\text{new}T} \delta\Psi(\hat{\mathbf{y}}) = \mathbf{w}^T \delta\Psi(\hat{\mathbf{y}}) + s\|\delta\Psi(\hat{\mathbf{y}})\|^2 = L(\hat{\mathbf{y}}), \tag{9}$$

$$s = \frac{L(\hat{\mathbf{y}}) - \mathbf{w}^T \delta\Psi(\hat{\mathbf{y}})}{\|\delta\Psi(\hat{\mathbf{y}})\|^2}. \tag{10}$$

After s is computed, it is changed to satisfy a standard box constraint, $0 \le \sum_{\hat{\mathbf{y}} \in Y^n} \alpha_{\hat{\mathbf{y}}} \le C$ and $\alpha_{\hat{\mathbf{y}}} \ge 0$.

$$s^{\text{clipped}} = \max(-\alpha_{\hat{\mathbf{y}}}, \min(C - \sum_{\hat{\mathbf{y}} \in Y^n} \alpha_{\hat{\mathbf{y}}}, s)). \quad (11)$$

To train the prior model for 1-slack structural SVMs, we use a modified 1-slack cutting plane algorithm and a modified fixed-threshold SMO (FSMO) algorithm [5]. The pseudocode of the modified 1-slack cutting algorithm is given in algorithm 1.

Algorithm 1. 1-slack cutting plane algorithm for PRIOR model.

1: Input: $(x_1, y_1), \ldots, (x_n, y_n), C, \varepsilon$
2: $S \leftarrow \emptyset$
3: repeat
$(\mathbf{w}, \xi) \leftarrow \arg\min_{\mathbf{w}, \xi > 0} \frac{1}{2}\|\mathbf{w} - \mathbf{w}_0\|^2 + C\xi$
4: s.t. $\forall (\hat{y}_1, \cdots, \hat{y}_n) \in S$:
$\frac{1}{n}\mathbf{w}^T \sum_{i=1}^{n} \delta\Psi_i(x_i, \hat{y}_i) \ge \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_i) - \xi$
5: for i=1,...,n do
6: $\hat{y}_i \leftarrow \arg\max_{\hat{y} \in Y} \{L(y_i, \hat{y}) + \mathbf{w}^T \Psi(x_i, \hat{y})\}$
7: end for
8: $S \leftarrow S \cup \{(\hat{y}_1, \cdots, \hat{y}_n)\}$
9: until $\frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_i) - \frac{1}{n}\mathbf{w}^T \sum_{i=1}^{n} \delta\Psi_i(x_i, \hat{y}_i) \le \xi + \varepsilon$
10: return $(\mathbf{w}, \xi)$

The modified FSMO uses the fact that the formulation of 1-slack structural SVMs has no bias and no linear equality constraint of binary classification SVMs [5]. Therefore, the modified FSMO breaks the QP of a structural SVM into a series of smallest QPs, each involving only one variable. By involving only one variable, the modified FSMO is advantageous in that each QP sub-problem does not require a working set selection when support vectors are unbounded.

A pseudocode of the modified FSMO for PRIOR model is depicted in algorithm 2. The algorithm is called a 1-slack cutting plane algorithm (line 4 in algorithm 1) and is used to solve the dual problem over working set S. Iterating through the constraint $\hat{\mathbf{y}} = (\hat{y}_1, \cdots, \hat{y}_n)$ in working set S, the algorithm updates individual Lagrange multipliers (that is, $\alpha_{\hat{\mathbf{y}}}$) and $\mathbf{w}$ when support vectors are unbounded (lines 4 to 9). When support vectors are bounded, the algorithm chooses two Lagrange multipliers by using the working set selection algorithm of SMO and updates two Lagrange multipliers. The algorithm stops if no $\alpha_{\hat{\mathbf{y}}}$ has changed during iteration.

Algorithm 2. Modified FSMO algorithm for PRIOR model: 1-slack cutting plane algorithm (line 4 in algorithm 1) used to solve dual problem over working set S.

1: Input: $(x_1, y_1), \ldots, (x_n, y_n), S, \alpha_S, C, \varepsilon$
2: repeat
3: if $\sum_{\hat{\mathbf{y}} \in Y^n} \alpha_{\hat{\mathbf{y}}} < C$ do /* unbounded SVs: FSMO */
4: for $\hat{\mathbf{y}} = (\hat{y}_1, \cdots, \hat{y}_n) \in S$ do
5: if $\{L(\hat{\mathbf{y}}) - \mathbf{w}^T \delta\Psi(\hat{\mathbf{y}}) > \varepsilon,\ \alpha_{\hat{\mathbf{y}}} < C\}$ or
$\{L(\hat{\mathbf{y}}) - \mathbf{w}^T \delta\Psi(\hat{\mathbf{y}}) < -\varepsilon,\ \alpha_{\hat{\mathbf{y}}} > 0\}$ do
6: calculate s and s^{clipped}
7: $\alpha_{\hat{\mathbf{y}}}^{\text{new}} \leftarrow \alpha_{\hat{\mathbf{y}}} + s^{\text{clipped}}$
8: end if
9: end for
10: else /* bounded SVs: SMO */
11: $\hat{\mathbf{y}} = \arg\max_{\hat{\mathbf{y}}' \in S} g(\hat{\mathbf{y}})$ where $g(\hat{\mathbf{y}}) = L(\hat{\mathbf{y}}) - \mathbf{w}^T \delta\Psi(\hat{\mathbf{y}})$
12: $\hat{\mathbf{y}}' = \arg\min_{\alpha_{\hat{\mathbf{y}}'} > 0, \hat{\mathbf{y}}' \in S} \frac{-(g(\hat{\mathbf{y}}) - g(\hat{\mathbf{y}}'))^2}{\|\delta\Psi(\hat{\mathbf{y}}) - \delta\Psi(\hat{\mathbf{y}}')\|^2}$
13: if $g(\hat{\mathbf{y}}) - g(\hat{\mathbf{y}}') > \varepsilon$ do
14: calculate s_2 and s_2^{clipped}
15: $\alpha_{\hat{\mathbf{y}}}^{\text{new}} = \alpha_{\hat{\mathbf{y}}} + s_2^{\text{clipped}}$, $\alpha_{\hat{\mathbf{y}}'}^{\text{new}} = \alpha_{\hat{\mathbf{y}}'} - s_2^{\text{clipped}}$
16: end if
17: end if
18: until no $\alpha_{\hat{\mathbf{y}}}$ has changed during iteration.

V. Experiments

We demonstrated the effectiveness of our domain adaptation for structural SVMs on named entity recognition (NER), part-of-speech (POS) tagging, and sentiment classification problems.

For the NER problem, we used a Korean named entity data set. We used three domains: a TV domain (TV scripts), a sports domain (sports news articles), and a celebrity domain (celebrity news articles). In the dataset, 108,984 (TV domain), 84,564 (sports domain), and 5,988 (celebrity domain) sentences were annotated into 15 NE categories: *person*, *location*, *organization*, *artifacts*, *study fields*, *theory*, *civilization*, *date*, *time*, *quantity*, *event*, *animal*, *plant*, *material,* and *term*. For the training set, we used 105,265 (TV domain), 81,829 (sports domain), and 4,804 (celebrity domain) sentences. For the test set, we used 3,719 (TV domain), 2,735 (sports domain), and 1,184 (celebrity domain) sentences.

For POS tagging, we used 18,537 Wall Street Journal (WSJ) sentences from sections 00-18 of the Penn Treebank as the source domain data, and 1,964 PubMed sentences from the ontology section of the PennBioIE corpus [15] as the target domain data [16].

For sentiment classification, we used a multidomain

Table 1. Data sets used in experiments.

Task (data set)	Source domain training set (#sent.)	Target domain training set (#sent.)	Target domain test set (#sent.)	#class	#feature
NER (Korean NER)	TV (105,265)	Sports (10,000)	Sports (2,735)	31	408,542
	Sports (81,829)	TV (30,000)	TV (3,719)	31	408,542
	Sports (81,829)	Celebrity (4,804)	Celebrity (1,184)	31	301,722
POS tagging (PubMed)	WSJ (18,537)	PubMed (1,964)	PubMed (1,065)	45	720,522
Sentiment classification (Amazon)	Books (6,065)	Kitchen (1,600)	Kitchen (400)	2	224,156
	DVD (5,186)	Kitchen (1,600)	Kitchen (400)	2	220,070
	Electronics (7,281)	Kitchen (1,600)	Kitchen (400)	2	154,126

sentiment dataset [17] containing product reviews from 4 Amazon domains (book, dvd, electronics, and kitchen) [18]. The goal in each domain is to classify a product review as either positive or negative. We used 6,065 (book), 5,186 (dvd), and 7,281 (electronics) reviews as the source domain data, and 1,600 kitchen reviews as the target domain data. Table 1 summarizes the characteristics of the data sets.

We implemented the prior model for 1-slack structural SVMs using a modified FSMO in C++ [5]. For comparison, we ran baseline domain adaptation methods using structural SVMs as a base learner. For all experiments, a linear kernel was used. Regularization constant C from {1, 3, 10, 30, 100, 300, 1,000, 3,000, 10,000} was chosen based on an optimization of the test set for all experiments. For a precise stopping condition, we set $e = 0.1$. All experiments were conducted on an Intel Core i5 CPU PC with 2.67 GHz and 8 GB of RAM.

In our first NER experiment, we used the sports domain as the target domain and the TV domain as the source domain. Figure 2 shows the accuracy of the compared methods, while Table 2 shows their F-measure. S-SVM+Src, S-SVM+Tgt, S-SVM+Lin, and S-SVM+Pred are SRC-ONLY, TGT-ONLY, LIN-INT, and Pred baselines using structural SVMs as a base learner, respectively. S-SVM+Prior is the prior model for 1-slack structural SVMs. All domain adaptation algorithms perform better than S-SVM+Tgt baseline when there is very little target data available. For the largest amount of target data, however, only S-SVM+Prior significantly outperforms S-SVM+Tgt. The improvements of S-SVM+Prior over S-SVM+Pred

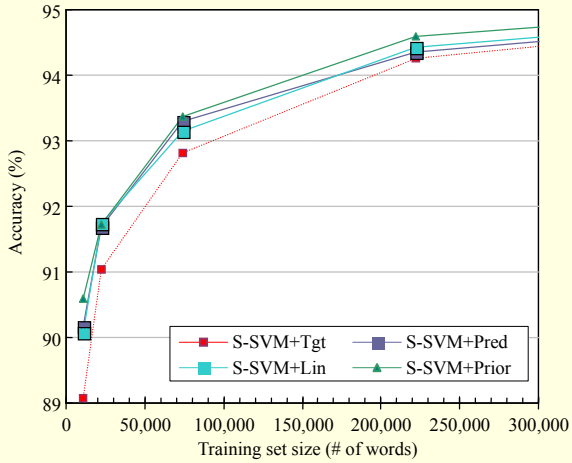

Fig. 2. Accuracy of compared methods vs. training set size (# of words) on sports domain.

Table 2. F-measure of compared methods on sports domain.

Training words	S-SVM+Tgt (baseline)	S-SVM+Pred	S-SVM+Lin	S-SVM+Prior
1.1×10^4	63.19	66.47 (+3.28)	68.14 (+4.95)	**69.10** (+5.91)
2.2×10^4	69.52	71.74 (+2.22)	**72.93** (+3.41)	72.70 (+3.18)
7.4×10^4	76.77	77.84 (+1.07)	77.40 (+0.63)	**77.99** (+1.22)
2.2×10^5	81.67	81.65 (–0.02)	82.24 (+0.57)	**82.38** (+0.71)

and S-SVM+Lin are statistically significant with a significance level less than 0.01 using paired t-tests (the two-tailed p-values are 3.9×10^{-7} and 3.1×10^{-4}, respectively).

In a second NER experiment, we used the TV domain as the target domain and the sports domain as the source domain. Figure 3 shows their accuracy, while Table 3 shows their F-measure. Similar to the first experiment, all domain adaptation algorithms perform much better than the target-only baseline when there is very little target data available. Among the domain adaptation algorithms, S-SVM+Prior performed the best in most cases, while S-SVM+Lin performed second best. For the largest amount of target data, S-SVM+Prior significantly outperforms other algorithms. The improvements of S-SVM+Prior over S-SVM+Pred and S-SVM+Lin are statistically significant with a significance level less than 0.01 using paired t-tests (the two-tailed p-values are 2.1×10^{-7} and 0.0018, respectively).

In our final NER experiment, we used the celebrity domain as the target domain and the sports domain as the source domain. Figure 4 shows their accuracy, while Table 4 shows their F-measure. S-SVM+Prior significantly outperformed the

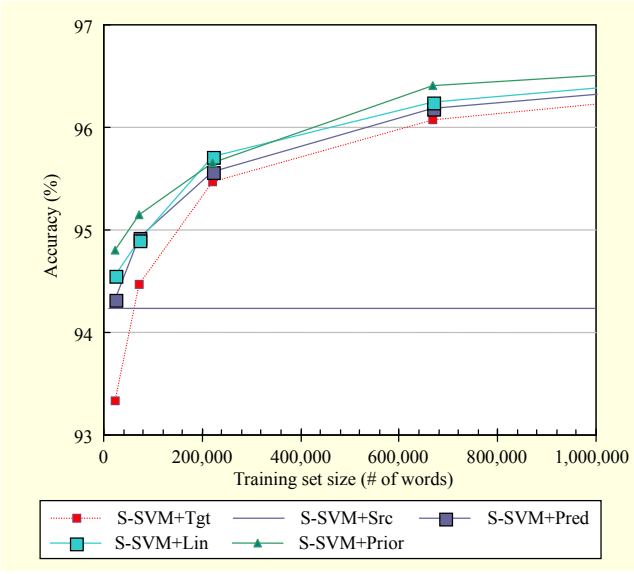

Fig. 3. Accuracy of compared methods vs. training set size (# of words) on TV domain.

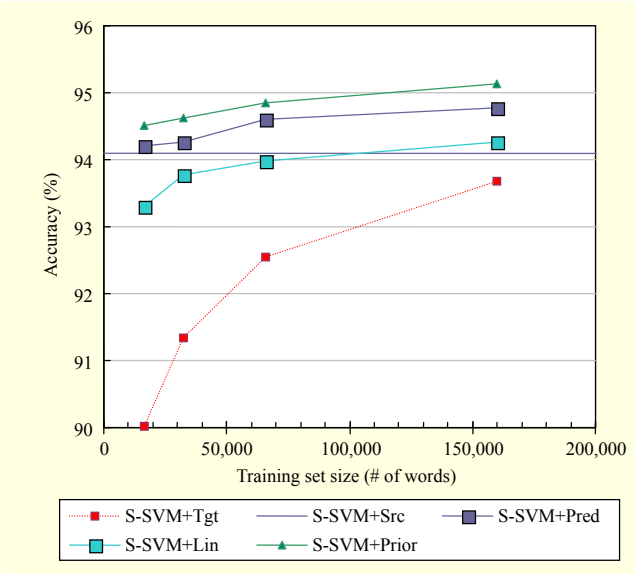

Fig. 4. Accuracy of compared methods vs. training set size (# of words) on celebrity domain.

Table 3. F-measure of compared methods on TV domain.

Training words	S-SVM+Tgt (baseline)	S-SVM+Pred	S-SVM+Lin	S-SVM+Prior
2.2×10^4	62.70	68.53 (+5.83)	71.91 (+9.21)	**73.35** (+10.7)
7.3×10^4	70.82	73.18 (+2.36)	74.07 (+3.25)	**75.52** (+4.70)
2.2×10^5	76.79	77.48 (+0.69)	**78.97** (+2.18)	78.76 (+1.97)
6.7×10^5	80.17	80.84 (+0.67)	81.26 (+1.09)	**82.07** (+1.90)

Table 4. F-measure of compared methods on celebrity domain.

Training words	S-SVM+Tgt (baseline)	S-SVM+Pred	S-SVM+Lin	S-SVM+Prior
1.7×10^4	68.51	83.02 (+14.5)	79.88 (+11.4)	**83.88** (+15.4)
3.2×10^4	72.90	83.09 (+10.2)	80.69 (+7.79)	**84.01** (+11.1)
6.6×10^4	76.90	83.76 (+6.86)	81.36 (+4.46)	**84.77** (+7.87)
1.6×10^5	80.37	84.70 (+4.33)	82.19 (+1.82)	**85.69** (+5.32)

other methods in all cases. For the largest amount of target data, the improvements of S-SVM+Prior over S-SVM+Pred and S-SVM+Lin are statistically significant with a significance level less than 0.01 using paired t-tests (the two-tailed p-values are 6.7×10^{-8} and 1.3×10^{-30}, respectively). Figure 5 shows the training time of the compared methods. The training time for S-SVM+Prior is reduced 2.5 times (from 5 minutes to 2 minutes) compared to S-SVM+Pred with an improvement in performance.

For POS tagging experiments, we used sections 00-18 of the Penn Treebank as the source domain data and ontology sections of the PennBioIE corpus as the target domain data. We additionally ran the feature augmentation method (S-SVM+FA) that Daumé proposed in [13] and used structural SVMs as a base learner. Figure 6 shows their accuracy. S-SVM+FA and S-SVM+Prior significantly outperformed the other methods. The improvement of S-SVM+Prior over S-SVM+Pred is statistically significant with a significance level less than 0.01 using a paired t-test (the two-tailed p-value is 3.5×10^{-6}). The difference of S-SVM+Prior and S-SVM+FA is not statistically

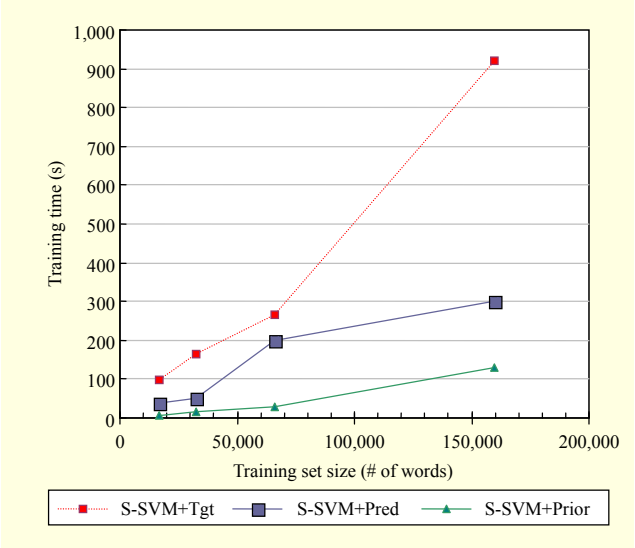

Fig. 5. Training time of compared methods vs. training set size (# of words) on celebrity domain.

significant (the two-tailed p-value is 0.57). However, the performances of S-SVM+All and S-SVM+Lin were lower

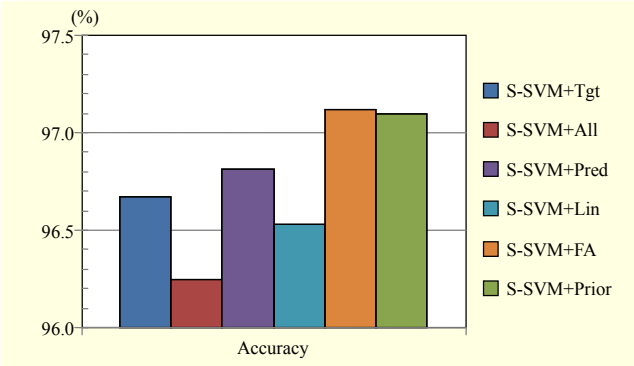

Fig. 6. Accuracy of compared methods on POS tagging task.

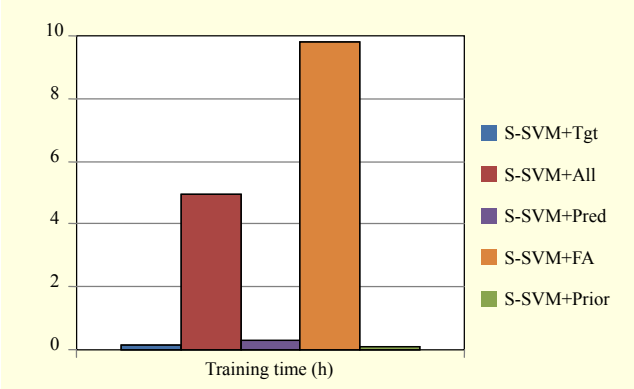

Fig. 7. Training time of compared methods on POS tagging task.

than S-SVM+Tgt. Figure 7 shows the training time of the compared methods. The training time for S-SVM+Prior is reduced by over 100 times (from 10 hours to 6 minutes) compared to S-SVM+FA.

For sentiment classification experiments, we used product reviews from 3 Amazon domains (books, dvd, and electronics) as the source domain data and kitchen reviews as the target domain data. Figure 8 shows accuracies for all pairs of domain adaptation. For each set of bars, the first letter is the source domain and the second letter is the target domain. For example, the first set of bars shows that S-SVM+All achieves 89% accuracy adapting from books domain to kitchens domain. We can observe S-SVM+All and S-SVM+Prior outperformed the other methods. For all data sets, the differences of S-SVM+Prior and S-SVM+All are not statistically significant having a significance level less than 0.01 (the two-tailed p-values are 0.66, 0.66, and 0.37, respectively). For E->K data set, the improvements of S-SVM+Prior over S-SVM+Lin and SVM+Pred are not statistically significant (the two-tailed p-values are 0.058 and 0.059, respectively), but the improvement of S-SVM+Prior over S-SVM+FA is statistically significant (the two-tailed p-value is 0.0065).

From our observations, we can conclude that S-SVM+Prior

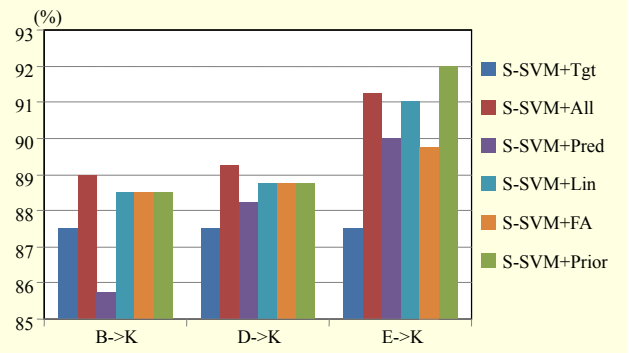

Fig. 8. Accuracy of compared methods on sentiment classification task.

outperforms the other domain adaptation methods in most cases. Moreover, the training time for S-SVM+Prior is reduced by 2.5 times and 100 times compared to S-SVM+Pred and S-SVM+FA, respectively.

VI. Conclusion

In this paper, we extended structural SVMs for domain adaptation using our prior model. We evaluated the proposed model on NER, POS tagging, and sentiment classification problems. We showed that the proposed model outperforms the other domain adaptation methods in most cases. Moreover, the training time for the proposed model is reduced by 2.5 times to 100 times compared to other domain adaptation methods, that is, the Pred baseline and feature augmentation methods.

References

[1] B. Taskar, C. Guestrin, and D. Koller, "Max Margin Markov Networks," *Proc. NIPS*, vol. 16, 2004.

[2] I. Tsochantaridis et al., "Support Vector Machine Learning for Interdependent and Structured Output Spaces," *Proc. ICML*, 2004.

[3] Ben Taskar et al., "Max-Margin Parsing," *Proc. EMNLP*, 2004.

[4] C. Lee and M. Jang, "Fast Training of Structured SVM Using Fixed-Threshold Sequential Minimal Optimization," *ETRI J.*, vol. 31, no. 2, Apr. 2009, pp. 121-128.

[5] C. Lee and M. Jang, "A Modified Fixed-Threshold SMO for 1-Slack Structural SVMs," *ETRI J.*, vol. 32, no. 1, Feb. 2010, pp. 120-128.

[6] C. Lee, S. Lim, and M. Jang, "Large-Margin Training of Dependency Parsers Using Pegasos Algorithm," *ETRI J.*, vol. 32, no. 3, June 2010, pp. 486-489.

[7] C.N. Yu et al., "Support Vector Training of Protein Alignment Models," *Proc. RECOMB*, 2007.

[8] Y. Yue et al., "A Support Vector Method for Optimization Average Precision," *Proc. SIGIR*, 2007, pp. 271-278.

[9] C.H. Teo et al., "A Scalable Modular Convex Solver for Regularized Risk Minimization," *Proc. KDD*, 2007, pp. 727-736.
[10] T. Joachims, T. Finley, and C.N. Yu, "Cutting-Plane Training of Structural SVMs," *MLJ*, vol. 77, no. 1, 2008, pp. 27-59.
[11] H. Daumé III and D. Marcu, "Domain Adaptation for Statistical Classifiers," *J. Artificial Intell. Research*, vol. 26, 2006, pp. 101-126.
[12] C. Chelba and A. Acero, "Adaptation of Maximum Entropy Capitalizer: Little Data Can Help a Lot," *Comput. Speech Language*, vol. 20, no. 4, 2006, pp. 382-399.
[13] H. Daumé III, "Frustratingly Easy Domain Adaptation," *Proc. ACL*, 2007, 2010, pp. 256-263.
[14] J. Yang et al., "Cross-Domain Video Concept Detection Using Adaptive SVMs," *Proc. 15th Int. Conf. Multimedia*, 2007, pp. 188-197.
[15] PennBioIE Corpus. http://bioie.ldc.upenn.edu/publications/latest_release/data
[16] J. Jiang and C. Zhai, "Instance Weighting for Domain Adaptation in NLP," *Proc. ACL*, 2007, pp. 264-271.
[17] Multidomain Sentiment Dataset. http://www.cs.jhu.edu/~mdredze/datasets/sentiment/
[18] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. Association of Computational Linguistics," *Proc. ACL*, 2007, pp. 440-447.

Changki Lee received the BS in computer science from KAIST, Rep. of Korea, in 1999. He received the MS and PhD in computer engineering from POSTECH, Rep. of Korea, in 2001 and 2004, respectively. Since 2004, he has been with ETRI, Rep. of Korea, as a senior member of research staff. He has served as a reviewer for international journals, such as *Information System, Information Processing & Management*, and *ETRI Journal*. His research interests are natural language processing, information retrieval, data mining, and machine learning.

Myung-Gil Jang received the BS and MS in computer science and statistics from Pusan National University, Rep. of Korea, in 1988 and 1990, repspectively. He received the PhD in information science from Chungnam National University in 2002. From 1990 to 1997, he was a researcher with System Engineering Research Institute (SERI), Rep. of Korea. Since 1998, he has been with ETRI, Rep. of Korea, as a senior/principle member of research staff. His research interests are natural language processing, information retrieval, question answering, knowledge and dialogue processing, media retrieval/management, and the semantic web.