International Journal of Fuzzy Logic and Intelligent Systems, vol. 12, no. 1, March 2012 pp. 6-14
http://dx.doi.org/10.5391/IJFIS.2012.12.1.6 pISSN 1598-2645 eISSN 2093-744X

# A Cost Sensitive Part-of-Speech Tagging: Differentiating Serious Errors from Minor Errors

**Jeong-Woo Son, Tae-Gil Noh, and Seong-Bae Park***

**School of Computer Science and Engineering,
Kyungpook National University,
Daegu, 702-701, Korea**

**Abstract**

All types of part-of-speech (POS) tagging errors have been equally treated by existing taggers. However, the errors are not equally important, since some errors affect the performance of subsequent natural language processing seriously while others do not. This paper aims to minimize these serious errors while retaining the overall performance of POS tagging. Two gradient loss functions are proposed to reflect the different types of errors. They are designed to assign a larger cost for serious errors and a smaller cost for minor errors. Through a series of experiments, it is shown that the classifier trained with the proposed loss functions not only reduces serious errors but also achieves slightly higher accuracy than ordinary classifiers.

**Key words** : POS tagging, Cost sensitive learning, Loss

## 1. Introduction

Part-of-speech (POS) tagging is needed as a preprocessing for various natural language processing (NLP) tasks such as parsing, named entity recognition (NER), and text chunking. Since POS tagging is normally performed in the early step of an NLP task, the errors in POS tagging are critical in that they affect all subsequent steps and often lower the overall performance of NLP tasks.

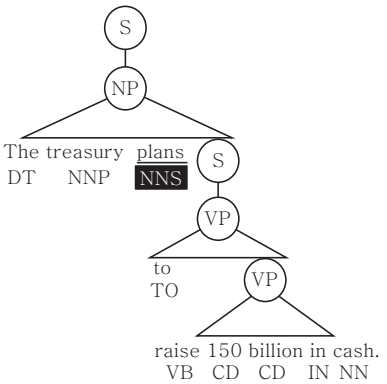
Previous studies on POS tagging have shown successful performances with machine learning techniques such as hidden markov model (HMM) [1], conditional random fields (CRF) [2], and maximum entropy model [3]. Supervised machine learning techniques were commonly used in early studies on POS tagging. These studies focused on how to apply the characteristics of a language to a machine learning technique [3, 4] or how to extract more informative features for POS tagging [5]. The state-of-the-art supervised POS tagging achieves over 97% of accuracy [6, 7]. This performance is generally regarded as the maximum performance that can be achieved by machine learning techniques, and recent studies on POS tagging aim to design unsupervised machine learning methods [8, 9]. However, there still exists a room to improve with supervised POS tagging in terms of error differentiation.

It should be noted that all errors are not equally important in POS tagging. Let us consider parse trees in Figure 1 as an example. In Figure 1(a), the word "*plans*" is mistagged as a noun where it should be a verb. This error results in a wrong parse tree that is severely different from the correct tree shown in Figure 1(b). The verb phrase of the verb "*plans*" in 1(b) is discarded in Figure 1(a) and the whole sentence is analyzed as a single noun phrase. Figure 1(c) and (d) show another tagging error and its effect. In Figure 1(c), a noun is tagged as a NNS (plural noun) where its correct tag is NN (singular or mass noun). However, the error in Figure 1(c) affects only locally in the noun phrase to which "*physics*" belongs. As a result, the general structure of the parse tree in Figure 1(c) is nearly same with the correct one in Figure 1(d). That is, a sentence analyzed with this type of error would yield a correct or near-correct result in many NLP tasks such as machine translation and text chunking.

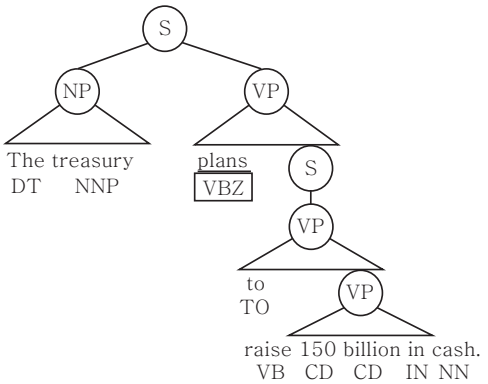
The goal of this paper is to differentiate the serious POS tagging errors from the minor errors. POS tagging is generally regarded as a classification task, and zero-one loss is commonly used in learning classifiers [10]. Since zero-one loss considers all errors equally, it can not distinguish error types. Therefore, a new loss is required to incorporate different error types into the learning machines.

This paper proposes two gradient loss functions to re-

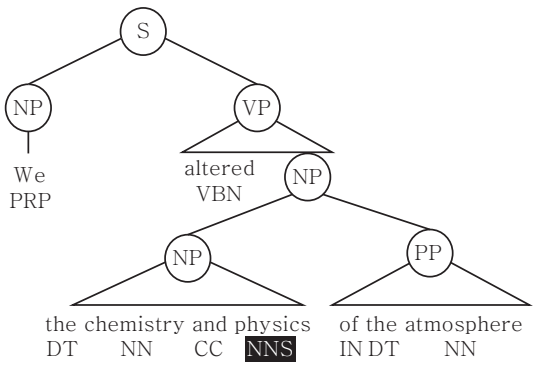
Manuscript received Jul. 26, 2011; revised Oct. 23, 2011;
accepted Mar. 07, 2012.
* Corresponding Author: Seong-Bae Park (seongbae@knu.ac.kr)
This work was supported by the Converging Research Center Program funded by the Ministry of Education, Science and Technology (2011K000659).



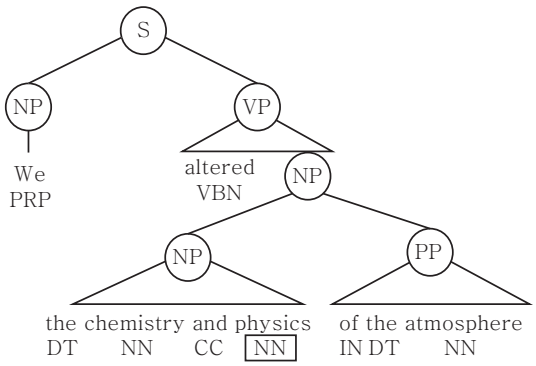

(a) A parse tree with a serious error.




(b) The correct parse tree of the sentence *"The treasury plans . . ."*.




(c) A parse tree with a minor error.




(d) The correct parse tree of the sentence *"We altered . . ."*.

Figure 1. An example of POS tagging errors

flect differences among POS tagging errors. The functions assign relatively small cost to minor errors, while larger cost is given to serious errors. They are applied to learning multiclass support vector machines [11], one of the best classifiers in POS tagging [12]. Then, this multiclass SVM is trained to minimize the serious errors. Overall accuracy of this SVM is not much improved against ordinary SVMs, but the serious errors are drastically reduced with the proposed method.

The rest of the paper is organized as follows. Section 2 reviews the related studies on POS tagging. In Section 3, serious and minor errors are defined, and it is shown that both errors are observable in a general corpus. Section 4 proposes two new loss functions for discriminating the error types in POS tagging. Experimental results are presented in Section 5. Finally, Section 6 draws conclusions.

## 2. Related Work

POS tagging problem has been generally solved by machine learning methods for sequential labeling. In early studies, rich linguistic features and supervised machine learning techniques are applied by using annotated corpora like Wall Street Journal corpus [13]. Ratnaparkhi [3] used a maximum entropy model for POS tagging. In this study, the features for rarely appearing words in a corpus are expanded to improve the overall performance. Following this direction, various studies have been proposed to extend informative features for POS tagging [6, 5]. In addition, various supervised methods such as HMMs and CRFs are widely applied to POS tagging. Lafferty et al. [2] adopted CRFs to predict POS tags. The methods based on CRFs have all the advantages of the maximum entropy models and also resolve the well-known problem of label bias. Kudo et al. [4] modified CRFs for non-segmented languages like Japanese which have a problem of word boundary ambiguity.

As a result of these efforts, the-state-of-the-art supervised POS tagging achieved over 97% of accuracy [6, 7]. Due to high accuracy of supervised approaches for POS tagging, it has been considered that there is no room to improve the performance on POS tagging in supervised manner. Thus, recent studies on POS tagging focus on unsupervised approaches [14, 15, 16]. Most previous studies on POS tagging focus on how to extract more linguistic features or how to adopt supervised or unsupervised approaches based on a single evaluation measure, *accuracy*.

Table 1. Tag categories and POS tags in Penn Tree Bank tag set

| Tag category | POS tags |
|---|---|
| Substantive | NN, NNS, NNP, NNPS, CD, PRP, PRP$ |
| Predicate | VB, VBD, VBG, VBN, VBP, VBZ, MD, JJ, JJR, JJS |
| Adverbial | RB, RBR, RBS, RP, UH, EX, WP, WP$, WRB, CC, IN, TO |
| Determiner | DT, PDT, WDT |
| Etc | FW, SYM, POS, LS |

However, with a different viewpoint for errors on POS tagging, there still exists a room to improve the performance of POS tagging for subsequent NLP tasks, even though the overall accuracy can not be much improved.

In ordinary studies on POS tagging, costs of errors are equally assigned. However, with respect to the performance of NLP tasks relying on the result of POS tagging, errors should be differently treated. In machine learning community, cost sensitive learning has been studied to differentiate costs among errors. By adopting different misclassification costs for each type of errors, a classifier is optimized to achieve the lowest expected cost [17, 18, 19].

This paper aims to reduce the serious errors which severely affect subsequent NLP tasks. For this purpose, two types of loss functions are proposed to optimize the performance of a machine learning method by considering differences of errors. As a result, even with the similar overall accuracy, the method trained with the proposed loss functions reduces serious POS tagging errors drastically than ordinary POS taggers.

## 3. Error Analysis of Existing POS Tagger

The effects of POS tagging errors to subsequent NLP tasks are different according to their type. Some errors are serious, while others are not. In this paper, the seriousness of tagging errors is defined by categorical structures of POS tags. Table 1 shows Penn tree bank POS tags and their categories. There are five categories in this table: *substantive*, *predicate*, *adverbial*, *determiner*, and *etc*. In this paper, serious tagging errors are defined as misclassifications among the categories, while minor errors are defined as misclassifications within a category. This definition follows the fact that POS tags in a same category form similar syntax structures in a sentence [20]. That is, inter-category errors are treated as serious errors, while intra-category errors are treated as minor errors.

Table 2 shows the distribution of inter-category and intra-category errors observed in the section 20 of WSJ corpus [13] that is tagged by SVMTools [12] (trained with WSJ sections 15–18). In this table, bold numbers denote inter-category errors, while other numbers show intra-category errors. The number of total errors is 1,061 from 47,585 words (2.3%). Among them, only 431 errors (40.6%) are intra-category (about 1.0%), while 630 errors (59.4%) are inter-category (about 1.3%). If we can reduce these inter-category errors under the cost of minimally increasing intra-category errors, the tagging results would be better in quality.

In general POS tagging, all tagging errors are regarded equally in importance. However, inter-category and intra-category errors should be distinguished. Since a machine learning method is optimized by a loss function, inter-category errors can be efficiently reduced if a loss function is designed to handle both types of errors with different cost. This paper proposes two loss functions for POS tagging and they are applied to multiclass Support Vector Machines.

## 4. Learning SVMs with Class Similarity

POS tagging has been solved as a sequential labeling problem which assumes dependency among words. However, by adopting sequential features such as POS tags of previous words, the dependency can be partially resolved. If it is assumed that words are independent one another, POS tagging can be regarded as a multiclass classification problem. One of the best solutions for this problem is SVM [21].

### 4.1 Learning SVMs with Loss Function

Assume that a training data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$ is given where $x_i \in \mathbf{R}^d$ is an instance vector and $y_i \in \{+1, -1\}$ is its class label. SVM finds an optimal hyperplane satisfying

$$x_i \cdot w + b \geq +1 \quad \text{for} \quad y_i = +1, \tag{1}$$

$$x_i \cdot w + b \leq -1 \quad \text{for} \quad y_i = -1, \tag{2}$$

where $w$ and $b$ are parameters to be estimated from training data $D$. To estimate the parameters, SVMs minimizes a hinge loss defined as

$$L_h(y_i, \hat{y}_i) = \max\{0, 1 - y_i \cdot \hat{y}_i\}, \tag{3}$$

where $\hat{y}_i = w \cdot x_i + b$ is a estimated value for $x_i$ by SVMs. With regularizer $||w||^2$ to control model complexity, the op-

Table 2. The distribution of tagging errors on WSJ corpus by SVMTools.

| | | Predicted category | | | | |
|---|---|---|---|---|---|---|
| | | Substantive | Predicate | Adverbial | Determiner | Etc |
| True category | Substantive | 171 | **225** | **9** | **1** | **0** |
| | Predicate | **316** | 234 | **30** | **0** | **0** |
| | Adverbial | **15** | **20** | 23 | **6** | **0** |
| | Determiner | **2** | **0** | **6** | 2 | **0** |
| | Etc | **1** | **0** | **0** | **0** | 1 |

timization problem of SVMs is defined as

$$\min_{w,L_h} \frac{1}{2}||w||^2 + C\sum_{i=1}^{l} L_h(y_i, \hat{y}_i), \tag{4}$$

subject to

$$y_i(x_i \cdot w + b) \geq 1 - L_h(y_i, w \cdot x_i + b), \quad \text{and } L_h(y_i, \hat{y}_i) \geq 0 \ \ \forall i, \tag{5}$$

where $C$ is a user parameter to penalize errors.

Crammer et al. [22] expanded the binary-class SVM for multiclass classifications. In multiclass SVMs, by considering all classes, the optimization of SVM is generalized as

$$\min_{w,L_h} \frac{1}{2}\sum_{k\in K} ||w_k||^2 + C\sum_{k\in K}\sum_{i=1}^{l} L_h(y_i, k), \tag{6}$$

with constraints

$$(w_{y_i} \cdot \phi(x_i, y_i)) - (w_k \cdot \phi(x_i, k)) \geq 1 - L_h(y_i, k), \tag{7}$$

$$L_h(y_i, k) \geq 0 \ \ \forall i, \ \ \forall k \in K \setminus y_i, \tag{8}$$

where $\phi(x_i, y_i)$ is a combined feature representation of $x_i$ and $y_i$, and $K$ is the set of classes.

Since both binary and multiclass SVMs adopt a hinge loss, the errors between classes have the same cost. To assign different cost to different errors, Tsochantaridis et al. [11] proposed an efficient way to adopt arbitrary loss function, $L(y_i, y_j)$ which returns zero if $y_i = y_j$, otherwise $L(y_i, y_j) > 0$. Then, the hinge loss $L_h(y_i, y_j)$ is re-scaled with the inverse of additional loss between two classes. By scaling slack variables with the inverse loss, margin violation with high loss $L(y_i, y_j)$ is more severely restricted than that with low loss. Then, the optimization problem with $L(y_i, y_j)$ is given as

$$\min_{w,L_h} \frac{1}{2}\sum_{k\in K} ||w_k||^2 + C\sum_{k\in K}\sum_{i=1}^{l} L_h(y_i, k), \tag{9}$$

with constraints

$$(w_{y_i} \cdot \phi(x_i, y_i)) - (w_k \cdot \phi(x_i, k)) \geq 1 - \frac{L_h(y_i, k)}{L(y_i, k)}, \tag{10}$$

$$L_h(y_i, y_k) \geq 0 \ \ \forall i, \ \ \forall k \in K \setminus y_i, \tag{11}$$

With the Lagrange multiplier $\alpha$, the optimization problem in Equation (9) is easily converted to the following dual quadratic problem.

$$\min_{\alpha} \frac{1}{2}\sum_{i,j}^{l} \sum_{k_i \in K\setminus y_i} \sum_{k_j \in K\setminus y_j} \alpha_{i,k_i}\alpha_{j,k_j} \times \tag{12}$$

$$J(x_i, y_i, k_i)J(x_j, y_j, k_j) - \sum_{i}^{l} \sum_{k_i \in K\setminus y_i} \alpha_{i,k_i}, \tag{13}$$

with constraints

$$\alpha \geq 0 \text{ and } \sum_{k_i \in K\setminus y_i} \frac{\alpha_{i,k_i}}{L(y_i, k_i)} \leq C, \ \ \forall i = 1, \cdots, l, \tag{14}$$

where $J(x_i, y_i, k_i)$ is defined as

$$J(x_i, y_i, k_i) = \phi(x_i, y_i) - \phi(x_i, k_i). \tag{15}$$

## 4.2 Loss Function for POS tagging

To design a loss function for POS tagging, this paper adopts categorical structures of POS tags. The simplest way to reflect the structure of POS tags shown in Table 1 is to assign larger cost to inter-category errors than to intra-category errors. Thus, the loss function with the categorical structure in Table 1 is defined as

$$L_c(y_i, y_j) = \begin{cases} 0 & \text{if } y_i = y_j, \\ \delta & \text{if } y_i \neq y_j \text{ but they belong} \\ & \text{to the same POS category,} \\ 1 & \text{otherwise,} \end{cases} \tag{16}$$

where $0 < \delta < 1$ is a constant to reduce the value of $L_c(y_i, y_j)$ when $y_i$ and $y_j$ are similar. As shown in this equation, inter-category errors have larger cost than intra-category errors. This loss $L_c(y_i, y_j)$ is named as *category loss*.

The loss function $L_c(y_i, y_j)$ is designed to reflect the categories in Table 1. However, the structure of POS tags can be represented as a more complex structure. Let us consider the category, **predicate**. This category has ten POS tags, and can be further categorized into two sub-categories: **verb** and **adject**. Figure 2 represents a categorical structure of POS tags as a tree with five categories of POS tags and their seven sub-categories.

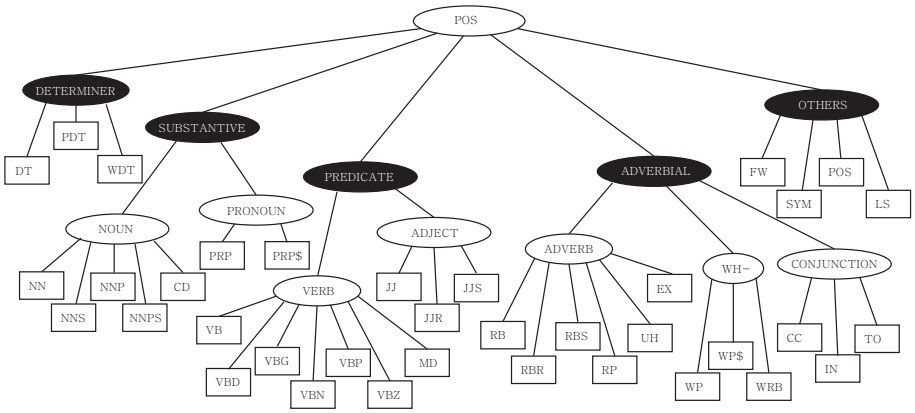
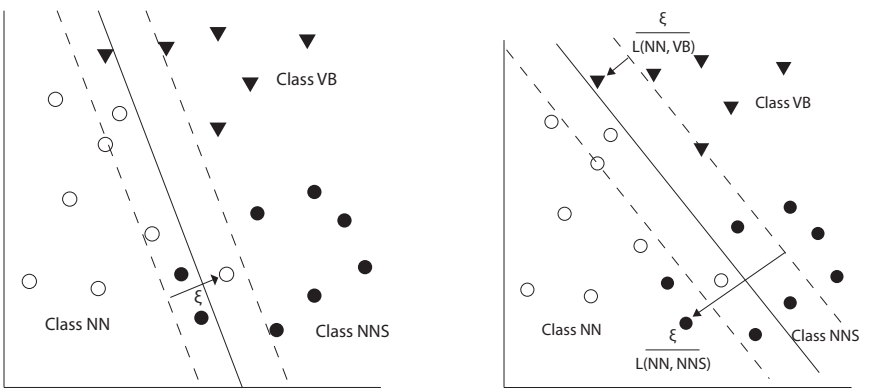



Figure 2. A tree structure of POS tags.




(e) Multiclass SVMs with hinge loss

(f) Multiclass SVMs with the proposed loss function

Figure 3. Effect of the proposed loss function in multiclass SVMs.

To represent the tree structure of Figure 2 as a loss, another loss function $L_t(y_i, y_j)$is defined as

$$L_t(y_i, y_j) = \tag{17}$$

$$\frac{1}{2}[Dist(P_{i,j}, y_i) + Dist(P_{i,j}, y_j)] \times \gamma, \tag{18}$$

where $P_{i,j}$ denotes the nearest common parent of both $y_i$ and $y_j$, and the function $Dist(P_{i,j}, y_i)$ returns the number of steps from $P_{i,j}$ to $y_i$. The user parameter $\gamma$ is a scaling factor of a unit loss for a single step. This loss $L_t(y_i, y_j)$ returns large value if the distance between $y_i$ and $y_j$ is far in the tree structure, and it is named as *tree loss*.

As shown in Equation (9), two proposed loss functions adjust margin violation between classes. They basically assign less value for intra-category errors than inter-category errors. Thus, a classifier is optimized to strictly keep inter-category errors in smaller boundary. Figure 3 shows a simple example. In this figure, there are three POS tags and two categories. NN (singular or mass noun) and NNS (plural noun) belong to the same category, while VB (verb, base form) is in another category. Figure 3(a) shows the decision boundary of NN based on hinge loss. As shown in this figure, a single $\xi$ is applied for the margin violation among all classes. Figure 3(b) also presents the decision boundary of NN, but it is determined with one of the proposed loss functions. In this figure, the margin violation is differently applied for inter-category (NN to VB) and intra-category (NN to NNS) errors. It results in reducing errors between NN and VB even if the errors between NN and NNS could be slightly increased.

## 5. Experiments

### 5.1 Experimental Setting

Experiments are performed with a well-known data, Wall Street Journal (WSJ) corpus. Among WSJ corpus, the documents from sections 15–18 are used as training data, and those from section 20 are as test data. Table 3 shows a simple statistics of the corpus. As shown in this table,

Table 3. Simple statistics of experimental data

| | Training | Test |
|---|---|---|
| Section | 15–18 | 20 |
| # of sentences | 8,936 | 2,011 |
| # of terms | 211,727 | 47,585 |

training data contains 8,936 sentences with 211,727 words. In test data, there are 2,011 sentences and 47,585 words.

Table 4 shows the feature set for our experiments. In this table, $w_i$ and $t_i$ denote the lexicons and POS tag for the $i$-th word in a sentence respectively. The POS tags for following words are obtained from a two-pass approach proposed by Nakagawa et al. [23]. The combinations of POS tags from previous words ($t_{i-2} \cdot t_{i-1}$) and those from next words ($t_{i+1} \cdot t_{i+2}$) are adopted to reflect interaction between POS tags of surrounding words.

The dimension of the feature space is over 112,000. In the experiments, two multiclass SVMs with proposed loss functions are used. One is CL-MSVM with category loss and the other is TL-MSVM with tree loss. They are compared with two base-line SVMs: one-vs-all SVM (SVM) and multiclass SVM (MSVM). A linear kernel is used for all SVMs.

## 5.2 Experimental Result

Figure 4 shows error rates of CL-MSVM, MSVM, and SVM. In this figure, both types of errors are plotted according to the values of parameter $\delta$. Figure 4(a) plots inter-category error rates while Figure 4(b) shows intra-category error rates. When $\delta = 1.0$, CL-MSVM is completely same with MSVM of which error rate is 2.667%. However, inter-category error rate of CL-MSVM is just 2.372% when $\delta$ is 0.6. $\delta = 0.6$ implies that the cost of intra-category errors is set to 60% of that of inter-category errors.

One thing to note is that the inter-category error rate is larger than 2.372% when $0 \leq \delta < 0.6$. This phenomena can be explained with Figure 4(b). With $0.6 \leq \delta \leq 1.0$, intra-category error rate of CL-MSVM is similar to MSVM. However, when $\delta$ is less than 0.6, the intra-category error rate is reciprocal to $\delta$. These intra-category errors affect inter-category errors, since mislabeled POS tags of surrounding words affect estimation of POS tag for a current word seriously. As a result, the inter-category error rate rather increases in $0 \leq \delta \leq 0.6$. However, even in this situation, CL-MSVM achieves lower inter-category error rate than both SVM and MSVM.

Similar results are observed for TL-MSVM. Figure 5 plots the error rates of TL-MSVM, SVM, and MSVM. In Figure 5(a), TL-MSVM shows the lowest inter-category error rate at $\gamma = 0.4$. However, when $\gamma$ is less than 0.4, the inter-category error rate of TL-MSVM rather increases due to high intra-category error rate in interval $0 \leq \delta \leq 0.4$ (see Figure 5(b)). The reason for this increment of error rate in TL-MSVM is same with that in CL-MSVM. One difference from CL-MSVM is that intra-category errors remain stable with $\gamma > 0.4$, while inter-category errors increase up to 2.887%. This is because TL-MSVM assigns different costs even to intra-category errors and these different costs cause more inter-category errors as $\gamma$ increases.

Overall error rates of four experimental methods are given in Figure 6. Figure 6(a) depicts overall error rate of CL-MSVM comparing with SVM and MSVM, while Figure 6(b) shows error rate of TL-MSVM. Both CL-MSVM and TL-MSVM aim to minimize inter-category errors without sacrificing many intra-category errors. As a result, their overall error rate is lower than both SVM and MSVM. CL-MSVM shows lower error rate than them in $0.4 \leq \delta \leq 0.9$, and TL-MSVM outperforms them when $\gamma \geq 0.1$.

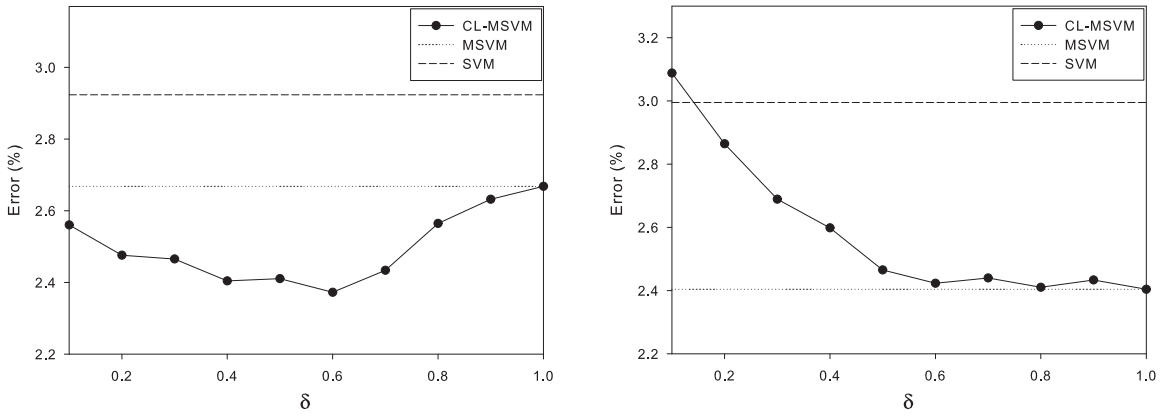
Table 5 compares four experimental methods at their best accuracies. TL-MSVM shows the best overall performance where its error rate is as low as 4.565%. Since one-vs-all approach is easily affected from skewed data, SVM shows the worst performance. The error rate of MSVM is just 5.072%, but about 53% of the errors are inter-category. On the other hand, CL-MSVM and TL-MSVM outperform both MSVM and SVM. The error rate of CL-MSVM is 4.795% which is slight improvement over MSVM. However, only about 49% of CL-MSVM are inter-category.

For inter-category error, CL-MSVM achieved the best performance. Its inter-category error rate is 2.372 %. Both TL-MSVM and CL-MSVM have significantly improved the results of SVM and MSVM in inter-category error. CL-MSVM achieves 20% of improvement over SVM and 12% over MSVM in terms of inter-category error. The improvement of TL-MSVM in inter-category error is less than CL-MSVM (18% over SVM and 11% over MSVM), but the improvement is still significant. The 10% improvement (-0.3 error rate) in inter-category errors means reduction of more than 120 serious errors. Both CL-MSVM and TL-MSVM reduce about 250 serious errors compared to base-line SVMs.

Since SVMTools, which is an ordinary POS tagger based on SVMs, uses various external knowledge repositories and heuristic processings to extract more information on sentences, we did not directly compare with SVMTools in this paper. However, without additional knowledge repositories and processings, SVMTools is exactly same with SVM in the experiment. Thus, from these results, we can conclude that SVMs trained with the proposed loss functions outperform ordinary SVMs and they successfully discriminate the serious POS tagging errors from the minor errors. Especially by adopting TL-MSVM both intra- and inter-category errors can be reduced efficiently. In case of CL-MSVM, even though it achieves the best performance in terms of inter-category error, it only reduces 0.02 inter-category errors compared with TL-MSVM at the cost of 0.26 intra-category errors. Thus, for various

Table 4. Feature set for experiments

| Feature Name | Description | Dimension |
|---|---|---|
| Lexical Feature | $w_{i-2}$,$w_{i-1}$,$w_i$,$w_{i+1}$,$w_{i+2}$ | 107,940 |
| Tag Feature | $t_{i-2}$,$t_{i-1}$,$t_{i+1}$,$t_{i+2}$ | 144 |
| Combination Feature | $t_{i-2} \cdot t_{i-1}$, $t_{i+1} \cdot t_{i+2}$ | 2,592 |




(g) Inter-category errors (h) Intra-category errors

Figure 4. Two types of errors in CL-MSVM and baseline SVMs.

downstream applications, TL-MSVM could be more efficient than not only SVM and MSVM but also CL-MSVM.

## 6. Conclusion

In this paper, we have shown that supervised POS tagging can be improved by discriminating inter-category errors from intra-category ones. An inter-category error occurs by mislabeling a word with a totally different tag, while an intra-category error is caused by a similar POS tag. Therefore, inter-category errors affect the performances of subsequent NLP tasks far more than intra-category errors. This implies that different costs should be considered in learning POS tagger according to error types.

As a solution to this problem, we have proposed two gradient loss functions which reflect different costs for two error types. The cost of an error type is set according to (i) categorical difference or (ii) distance in the tree structure of POS tags. Our experiments have shown that if these loss functions are applied to multiclass SVMs, they could significantly reduce inter-category errors. In addition, it is also shown that the multiclass SVMs trained with the proposed loss functions outperform the ordinary SVMs even in overall performance.

In this paper, we have shown that cost sensitive learning can be applied to POS tagging only with multiclass SVMs. However, the proposed loss functions are general enough to be applied to other existing POS taggers. Most supervised machine learning techniques are optimized on their loss functions. Therefore, the performance of POS taggers based on supervised machine learning techniques can be improved by applying the proposed loss functions to learn their classifiers.

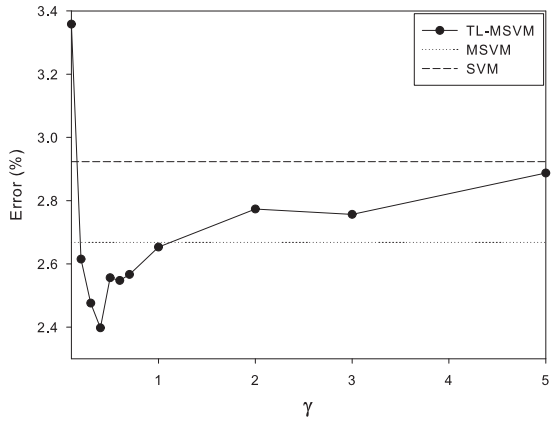
## References

[1] T. Brants, "TnT-A Statistical Part-of-Speech Tagger," In *Proceedings of the Sixth Applied Natural Language Processing Conference*, pp. 224–231, 2000.

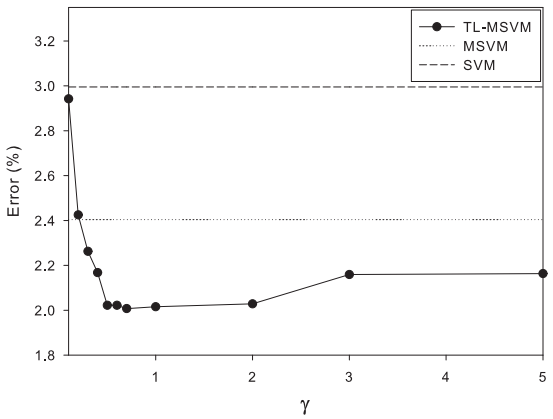
[2] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289, 2001.

[3] A. Ratnaparkhi, "A Maximum Entropy Model for Part-Of-Speech Tagging," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 133–142, 1996.

[4] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, 2004.

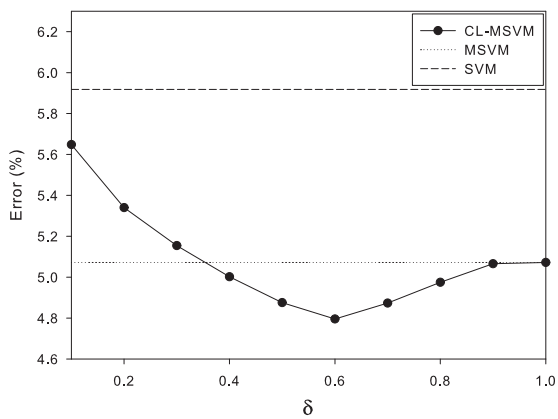



(i) Inter-category errors

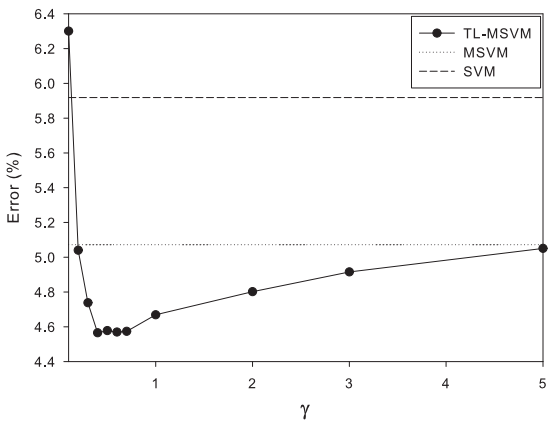



(j) Intra-category errors

Figure 5. Two types of errors in TL-MSVM and baseline SVMs




(k) Overall errors of CL-MSVM and baseline SVMs




(l) Overall errors of TL-MSVM and baseline SVMs

Figure 6. Overall performances of experimental methods

[5] K. Toutanova and C. Manning, "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 63–70, 2000.

[6] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," In *Proceedings of HLT-NAACL*, pp. 252–259, 2003.

[7] Y. Tsuruoka and J. Tsujii, "Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 467–474, 2005.

[8] S. Goldwater and T. Griffiths, "A fully Bayesian Approach to Unsupervised Part-of-Speech Tagging," In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 744–751, 2007.

[9] A. Haghighi and D. Klein, "Prototype-driven Learning for Sequence Models," In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 320–327, 2006.

[10] Y. Altun, M. Johnson, and T. Hofmann, "Investigating Loss Functions and Optimization Methods for Discriminative Learning of Label Sequences," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 145–152, 2003.

Table 5. Comparison of the proposed SVMs with ordinary SVMs

| | Overall error (%) | Intra-category error | Inter-category Error |
|---|---|---|---|
| SVM | 5.918 | 2.995 | 2.923 |
| MSVM | 5.072 | 2.404 | 2.667 |
| CL-MSVM ($\delta = 0.6$) | 4.795 | 2.423 | **2.372** |
| TL-MSVM ($\gamma = 0.4$) | **4.565** | **2.167** | 2.397 |


[11] I. Tsochantaridis, T. Hofmann, T. Joachims, and T. Altun, "Support Vector Learning for Interdependent and Structured Output Spaces," In *Proceedings of the 21st International Conference on Machine Learning*, pp. 104–111, 2004.

[12] J. Giménez and L. Màrquez, "SVMTool: A general POS tagger generator based on Support Vector Machines," In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pp. 43–46, 2004.

[13] M. Marcus, B. Santorini, and M. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics,* vol. 19, no.2, pp. 313–330, 1994.

[14] T. Berg-Kirkpatrick, A. Côté, J. DeNero, and D. Klein, "Painless Unsupervised Learning with Features," In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 582–590, 2010.

[15] J. Graca, K. Ganchev, B. Taskar, and F. Pereira, "Posterior vs Parameter Sparsity in Latent Variable Models," In *Advances in Neural Information Processing Systems 22*, pp. 664–672, 2009.

[16] M. Johnson, "Why doesn't EM find good HMM POS-taggers?," In *Proceedings of the 2007 Joint Meeting of the Conference on Empirical Methods in Natural Language Processing and the Conference on Computational Natural Language Learning*, pp. 296–305, 2007.

[17] L. Cai and T. Hofmann, "Hierarchical Document Categorization with Support Vector Machines," In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, pp. 78–87, 2004.

[18] C. Elkan, "The Foundations of Cost-Sensitive Learning," In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978, 2001.

[19] Z. Zhou and X. Liu, "On Multi-Class Cost-Sensitive Learning," In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 567–572, 2006.

[20] Q. Zhao and M. Marcus, "A Simple Unsupervised Learner for POS Disambiguation Rules Given Only a Minimal Lexicon," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 688–697, 2009.

[21] J. Sunghae, "Support Vector Machine based on Stratified Sampling," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 9, no. 2, pp. 141–146, 2009.

[22] K. Crammer, Y. Singer, N. Cristianini, J. Shawe-taylor, and B. Williamson, "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[23] T. Nakagawa, T. Kudo, and Y. Matsumoto, "Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines," In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, pp. 325–331, 2001.



**Jeong-Woo Son**
Research Associate at Kyungpook National University
Research Area: Machine Learning, Natural Language Processing, Semantic Web
E-mail : jwson@sejong.knu.ac.kr

**Tae-Gil Noh**
Research Associate at Kyungpook National University
Research Area: Natural Language Processing, Information Retrieval
E-mail : tgnoh@sejong.knu.ac.kr

**Seong-Bae Park**
Professor of Kyungpook National University
Research Area: Machine Learning, Natural Language Processing
E-mail : seongbae@knu.ac.kr