# Feature Selection in SVM Text Categorization

**Hirotoshi Taira**
NTT Communication Science Labs.
2-4 Hikaridai Seika-cho Soraku-gun
Kyoto 619-0237 Japan
taira@cslab.kecl.ntt.co.jp

**Masahiko Haruno**
ATR Human Information Processing Research Labs.
2-2 Hikaridai Seika-cho Soraku-gun
Kyoto 619-0231 Japan
mharuno@hip.atr.co.jp

## Abstract

This paper investigates the effect of prior feature selection in Support Vector Machine (SVM) text categorization. The input space was gradually increased by using mutual information (MI) filtering and part-of-speech (POS) filtering, which determine the portion of words that are appropriate for learning from the information-theoretic and the linguistic perspectives, respectively. We tested the two filtering methods on SVMs as well as a decision tree algorithm C4.5. The SVMs' results common to both filtering are that 1) the optimal number of features differed completely across categories, and 2) the average performance for all categories was best when all of the words were used. In addition, a comparison of the two filtering methods clarified that POS filtering on SVMs consistently outperformed MI filtering, which indicates that SVMs cannot find irrelevant parts of speech. These results suggest a simple strategy for the SVM text categorization: use a full number of words found through a rough filtering technique like part-of-speech tagging.

## Introduction

With the rapid growth of the Internet and online information, automatic text categorization has attracted much attention among researchers and companies. Some machine learning methods have been applied to text categorization, which include, for example, k-nearest-neighbor (Yang 1994), decision trees (Lewis & Ringuette 1994) and Naive-Bayes (Lewis & Ringuette 1994). The huge number of words in these data, which can potentially contribute to the overall task, challenges machine learning approaches. More specifically, the difficulties in handling the large input space are twofold: the learning machine used and the portion of words effective for the classification depend on each other (Yang & Pederson 1997). We have to find learning machines with a feature selection criteria because the best set of words greatly differs with the learning machine used (Lewis & Ringuette 1994).



Support Vector Machines (SVMs) (Vapnik 1995; Cortes & Vapnik 1995) construct the optimal hyperplane that separates a set of positive examples from a set of negative examples with a maximum margin [1]. SVMs have been shown to yield good generalization performances on a wide variety of classification problems that require large-scale input space, such as handwritten character recognition (Vapnik 1995) and face detection (Osuna, Freund, & Girosi 1998) problems.

Recently, two groups have explored the use of SVMs for text categorization (Joachims 1998; Dumais *et al.* 1998). Although they both achieved promising performances, they used completely different feature (word) selection strategies. In (Joachims 1998), words are considered features only if they occur in the training data at least three times and if they are not stop words such as 'and' and 'or.' Then the inverse document frequency (IDF) (Salton & Buckley 1988) is employed as a value for each feature. In contrast, (Dumais *et al.* 1998) considers only 300 words for each category, which are handled by a threshold for high mutual information (Cover & Thomas 1991). The feature value in this case is assigned as a binary to indicate whether a word appears in a text. A natural question about SVM text categorization occurs to us: how much influence do different feature selection strategies have? Does there exist one best strategy for choosing appropriate words?

Feature selection becomes especially delicate in agglutinative languages such as Japanese and Chinese because in these languages, word identification itself is not a straight-forward task. Unknown words output by word-segmentation and part-of-speech tagging systems contain both important keywords (like personal and company names) and useless portions of words. The selection of these unknown words is crucial to these languages.

To address these questions, this paper investigates the effect of prior feature selection in SVM text categorization by using Japanese newspaper articles. In our experiments, the number of input spaces was gradually increased by two distinct criteria: mutual informa-

[1] A margin is intuitively the distance from a data point to the classification boundary.

tion (MI) filtering and part-of-speech (POS) filtering. MI selects discriminating words for a particular category from an information-theoretical viewpoint. Words with higher mutual information are more highly representative in a specific text category. In contrast, POS filtering constructs word input space based on part of speech.

Our first experiment addresses how many words are appropriate for each category and to what extent the numbers differ between categories in SVM text categorization. The results are: 1) the optimal number of features differed completely across categories, and 2) the average performance for all categories was best when all of the words were used. The additional comparison between SVMs and a decision tree induction algorithm C4.5 (Quinlan 1993) clarifies that C4.5 achieves the best performance for each category at much smaller number of words, and the SVMs significantly outperforms C4.5. These results indicate that SVMs are more appropriate to make the best use of the huge number of input words.

Our second experiment changes input space in the following order without further thresholding: 1) common nouns, 2) step1+proper nouns, 3) step2+verbal nouns, 4) step3+unknown words, 5) step4+verbs. This experiment aims to investigate general tendencies in increasing the number of input spaces and the effect of each part-of-speech on SVM text categorizations. The result was similar to that was seen in the first experiment.

A comparison of the two experiments clarified that POS filtering consistently outperformed MI filtering, which indicates that SVMs cannot find irrelevant parts of speech. These results suggest a simple strategy in SVM text categorization: use a full number of words found through a rough filtering technique like part-of-speech tagging.

The rest of the paper is organized as follows. The next section introduces SVMs and gives a rough theoretical sketch of why SVMs can avoid overfitting even if the input space is sufficiently large. We then report our experimental results on MI filtering and POS filtering. The last section discusses the results of the two filtering methods and concludes the paper.

## Support Vector Machines

SVMs are based on *Structural Risk Minimization* (Vapnik 1995). The idea of structural risk minimization is to find a hypothesis $h$ for which we can guarantee the lowest generalization error. The following upper bound (1) connects $error_g(h)$, the generalization error of a hypothesis $h$ with the error of $h$ on the training set $error_t(h)$ and the complexity of $h$ (Vapnik 1995). This bound holds with a probability of at least $1-\eta$. In the second term on the right hand side, $n$ denotes the number of training examples and $\lambda$ is the *VC dimension*, which is a property of the hypothesis space and indicates its complexity.

$$error_g(h) \leq error_t(h) + 2\sqrt{\frac{\lambda(\ln\frac{2n}{\lambda}+1) - \ln\frac{\eta}{4}}{n}} \quad (1)$$

Equation (1) reflects the well-known trade-off between the training error and the complexity of the hypothesis space. A simple hypothesis (small $\lambda$) would probably not contain good approximating functions and would lead to a high training (and true) error. On the other hand, a too-rich hypothesis space (high $\lambda$) would lead to a small training error, but the second term on the right hand side of (1) will be large (overfitting). The right complexity is crucial to achieving good generalization. In the following, we assume that the linear threshold functions represent a hypothesis space in which $\boldsymbol{w}$ and $b$ are parameters of a hyperplane and $\boldsymbol{x}$ is an input vector:

$$h(\boldsymbol{x}) = sign\{\boldsymbol{w}\cdot\boldsymbol{x}+b\} = \begin{cases} +1, & \text{if } \boldsymbol{w}\cdot\boldsymbol{x}+b>0 \\ -1, & \text{else.} \end{cases}$$

Lemma 1 sheds light on the relationship between the dimension of the input space $\boldsymbol{x}$ of a set of hyperplanes and its VC dimension $\lambda$.

**Lemma 1 (Vapnik)** *Consider hyperplanes $h(\boldsymbol{x}) = sign\{\boldsymbol{w}\cdot\boldsymbol{x}+b\}$ as a hypothesis. If all example vectors $x_i$ are contained in a ball of radius $R$ and the following is required such that for all examples $x_i$:*

$$|\boldsymbol{w}\cdot\boldsymbol{x}+b| \geq 1, \quad with \quad ||\boldsymbol{w}|| = A,$$

*then this set of hyperplanes has a VC dimension $\lambda$ bounded by*

$$\lambda \leq min([R^2A^2], n) + 1. \quad (2)$$

Note here that the VC dimension of these hyperplanes does not always depend on the number of input features. Instead, the VC dimension depends on the Euclidean length $||\boldsymbol{w}||$ of the weight vector $\boldsymbol{w}$. Equation (2) supports the possibility that SVM text categorization achieves good generalization even if a huge number of words are given as an input space. Further experimental evaluations are required because Equations (1) and (2) both give us only a loose bound.

Basically, SVM finds the hyperplane that separates the training data with the shortest weight vector (i.e., $min||\boldsymbol{w}||$). The hyperplane maximizes the margin between the positive and negative samples. Since the optimization problem is difficult to handle numerically, Lagrange multipliers are introduced to translate the problem into an equivalent quadratic optimization problem. For this kind of optimization problem, efficient algorithms exist that are guaranteed to find the global optimum. The result of the optimization process is a set of coefficients $\alpha_i^*$ for which (3) is minimum. These coefficients can be used to construct the hyperplane satisfying the maximum margin requirement.

$$Minimize: \quad -\sum_{i=1}^{n}\alpha_i + \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \mathrm{x}\cdot\mathrm{x} \quad (3)$$

$$so \quad that : \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \forall i : \alpha_i \geq 0.$$

SVMs can handle nonlinear hypotheses by simply substituting every occurrence of the inner product in equation (3) with any Kernel function $K(x_1, x_2)$ [2]. Among the many types of Kernel functions available, we will focus on the $d$th polynomial functions (Equation (4)):

$$K_{poly}(x_1, x_2) = (x_1 \cdot x_2 + 1)^d. \tag{4}$$

## Experimental Results

This section describes our experimental results for two feature selection methods in SVM text categorization: mutual information filtering and part-of-speech filtering. For comparison, we also tested a decision tree induction algorithm C4.5 (Quinlan 1993) with default parameters. Before going into the details of the results, we first explain the experimental setting.

### Experimental Setting

Table 1: RWCP corpus for training and test.

| Category | training sets | test sets |
|---|---|---|
| sports | 161 | 147 |
| criminal law | 156 | 148 |
| government | 135 | 142 |
| educational system | 110 | 124 |
| traffic | 112 | 103 |
| military affairs | 110 | 118 |
| international relations | 96 | 97 |
| communications | 76 | 83 |
| theater | 86 | 95 |
| agriculture | 78 | 72 |

We performed our experiments using the RWCP corpus (Toyoura *et al.* 1996), which contains 30,207 newspaper articles taken from the Mainichi Shinbun Newspaper published in 1994 (Mainichi 1995). Each article was assigned multiple UDC codes, each of which represented a category of texts. In the rest of this paper, we will focus on the ten categories that appeared most often in the corpus [3]: sports, criminal law, government, education, traffic, military affairs, international relations, communications, theater and agriculture. The total number of articles used for both training and test were 1,000. Table 1 summarizes the number of training and test articles in each category.

[2]More precisely, the Mercer's condition (Vapnik 1995) should be satisfied.

[3]The results for other categories were very similar to these 10 categories.

These articles were word-segmented and POS tagged by the Japanese morphological analyzing system Chasen (Matsumoto *et al.* 1997). The process generated 20,490 different words. We used all types of parts of speech in the mutual information filtering and used only common nouns, proper nouns, verbal nouns, unknown words and verbs (total number of subset words was 18,111) in the part-of-speech filtering. Throughout our experiments, various subsets of these extracted words were used as input feature spaces, and the value for each feature was a binary value that indicated whether a word appeared in a document or not. A binary value was adopted to study the pure effects of each word.

### Mutual Information Filtering

The mutual information (MI) between a word $t_i$ and a category $c$ is defined in equation (5). MI becomes large when the occurrence of $t_i$ is biased to one side between a category $c$ and other categories. Consequently, it can be expected that the words with high mutual information in category $c$ are keywords in the category. The question we would like to discuss here is whether words with a fixed number of high mutual information can achieve a good generalization over all text categories.

$$MI(t_i, c) = \sum_{t_i \in \{0,1\}} \sum_{c \in \{+,-\}} P(t_i, c) \log \frac{P(t_i, c)}{P(t_i)P(c)}. \tag{5}$$

Table 2 shows the words at the points of the 300th, 500th, 1,000th, 5,000th and 10,000th mutual information in each category. In general, up to the 500th or 1,000th term, the words were specific to each category. For example, 'screwball' and 'golfer,' 'peace' and 'Moscow' are specific to sports and military, respectively. It is also interesting to note that 'Kazakhstan' is an unknown word that plays an important role in the category of military affairs. In contrast, after the 1,000th term, words do not seem to be specialized to any specific category.

Table 3 and 4 show the average of the recall and precision on SVMs and C4.5, respectively, when the number of words is changed with various MI thresholds. The order of polynomial $d$ (See Equation (4)) used is 1 and 2. The boldface values in the tables represent the best performance for each category. It is easily understood that the best number of words differs greatly from category to category in SVMs, while the best performance in C4.5 is achieved at much smaller number of words. The average performance is best for SVM when the number of words is 15,000; it improves continuously although in C4.5 abrupt drop is seen at 500 words. It is also notable that in average SVMs significantly outperform C4.5, which indicates that SVMs are more appropriate to make the best use of the huge number of input words.

Let us now look more closely at the recall and precision on SVMs for the same data. Figure 1 plots the

Table 2: Words selected with MI.

| | words | | | | |
|---|---|---|---|---|---|
| Feature | 300th | 500th | 1000th | 5000th | 10000th |
| sports | 変化球 (screwball) | 応援 (cheering) | ゴルファー (golfer) | アンケート (questionnaire) | 目安 (standard) |
| criminal law | 疑惑 (suspicion) | 送検 (commit for trial) | 地下 (underground) | 売る (sell) | 増進 (increase) |
| government | 議会 (parliament) | 運輸省 (The Ministry of Transport) | 約束 (promise) | 根幹 (basis) | さえぎる (interrupt) |
| education | 塾 (cram school) | 文相 (Minister of Education) | 理想的だ (ideal) | 涙 (tear) | 即 (immediately) |
| traffic | 大型車 (large-size car) | 配達 (delivery) | 速さ (speed) | 池 (pond) | 双方向 (bi-direction) |
| military | 平和 (peace) | モスクワ (Moscow) | カザフスタン (Kazakhstan) | 実際 (practical) | 降下 (descend) |
| international | 有事 (emergency) | 各国 (countries) | 大筋 (outline) | 年内 (within the year) | 裁く (judge) |
| communications | 会議 (meeting) | 衛星通信 (satellite communications) | 伝送 (transmission) | 正常 (normal) | 慎重 (careful) |
| theater | 台本 (play script) | 終演 (the end of a show) | 賞 (prize) | 要素 (element) | ロイ (Roy) |
| agriculture | イモ (potato) | 砂糖 (sugar) | 飼料 (livestock feed) | 改善 (improvement) | 変貌 (look different) |

Table 3: Average of recall and precision with MI on SVMs.

| | poly degree $d = 1/d = 2$ | | | | | |
|---|---|---|---|---|---|---|
| Feature | 300 | 500 | 1000 | 5000 | 10000 | 15000 |
| sports | **91.9**/91.9 | 89.5/89.5 | 90.9/90.9 | 90.8/90.0 | 90.0/89.6 | 90.4/89.6 |
| criminal law | 71.5/70.7 | 69.2/71.0 | 68.2/70.3 | 72.2/73.0 | 74.3/74.1 | 75.5/**76.4** |
| government | 66.6/66.1 | 68.4/68.2 | 74.4/76.4 | 79.3/79.0 | 76.8/78.0 | 78.2/**79.8** |
| education | 68.4/68.2 | 69.1/69.7 | 71.7/73.5 | 78.1/77.8 | 80.0/79.8 | **80.1**/79.6 |
| traffic | 66.6/66.6 | 70.5/71.6 | **72.1**/71.8 | 70.7/68.3 | 71.0/69.1 | 71.0/71.1 |
| military affairs | 66.3/68.3 | 71.3/71.9 | 74.5/75.7 | 74.6/74.7 | 75.6/75.9 | **77.1**/76.3 |
| international relations | 54.3/56.9 | 60.1/61.9 | 62.9/**63.5** | 61.6/60.4 | 61.0/59.2 | 57.1/58.9 |
| communications | 64.0/64.9 | 65.7/**66.6** | 59.3/59.3 | 55.7/53.3 | 53.6/50.0 | 58.2/50.0 |
| theater | 83.9/84.0 | **88.7**/83.9 | 86.2/88.2 | 83.6/86.2 | 83.8/82.2 | 83.8/82.4 |
| agriculture | 85.9/85.2 | **87.5**/86.6 | 85.7/85.7 | 85.0/83.2 | 85.9/85.0 | 84.1/84.1 |
| avg. | 71.9/72.2 | 74.0/74.0 | 74.5/75.5 | 75.1/74.5 | 75.2/74.2 | **75.5**/74.8 |

recall and precision of $d = 1$. Overall, recall tends to improve as the number of words increase except for the 'international relations' category, which monotonically decreases. Thus, we can safely say that increasing the number of words improves recall.

In contrast to recall, the change in precision is more complicated. For the five categories with the highest precision, the curves decline continuously but only slightly. This is a reasonable phenomenon because the excessive amount of key words may extract irrelevant documents. The other five categories with middle precision differ greatly. Two curves increase monotonically and two others drift, while the remaining one has a peak at 10,000 features. The point here is that the increase in features does not involve a large decrease in precision. In other words, the feature selection ability inherent in SVM can prevent precision from dropping abruptly with an increase in feature space. These results show that good generalization performance with a large number of features (15,000) depends on achieving good precision.

## Part-of-Speech Filtering

We tested the following five feature sets. The number of different words in each part of speech is summarized in Table 5. The total number of different words of these parts of speech is 18,111.

1. common nouns
2. 1 + proper nouns
3. 2 + verbal nouns
4. 3 + unknown words
5. 4 + verbs

Table 6 and 7 show the averages of recall and precision on SVMs and C4.5, respectively, when each of the above five features are used. Boldface numbers represent the best value in each category. It is clear that the best feature set greatly differs from category to category in both cases. The best average performance is achieved in SVMs when all of the words are used (Feature Set 5).

What are the contributions of each part of speech in SVM text categorization? In Table 6, common nouns are so powerful that near-optimal performance can be achieved only with one part of speech. Proper nouns, verbal nouns and verbs improve results in more than half of the categories, while unknown words contribute to only three categories. This is probably because the unknown words contain irrelevant portions of a word as well as important keywords for a category. Note that there is no abrupt drop in performance as a result of incrementally adding any parts of speech.

Table 4: Average of recall and precision with MI on C4.5.

| Feature | 300 | 500 | 1000 | 5000 | 10000 | 15000 |
|---|---|---|---|---|---|---|
| sports | **87.5** | 86.2 | 85.2 | 83.6 | 83.6 | 83.6 |
| criminal law | 67.9 | **70.8** | 68.9 | 68.8 | 68.8 | 68.8 |
| government | **65.5** | 63.0 | 58.0 | 57.9 | 57.9 | 57.9 |
| education | **72.0** | 69.2 | 70.1 | 70.1 | 70.1 | 70.1 |
| traffic | **63.0** | 61.0 | 61.0 | 61.0 | 61.0 | 61.0 |
| military affairs | **75.9** | 73.3 | 69.1 | 68.8 | 68.8 | 68.8 |
| international relations | **50.0** | 45.6 | 42.4 | 42.4 | 42.4 | 42.4 |
| communications | **52.7** | 50.3 | 50.3 | 50.3 | 50.3 | 50.3 |
| theater | **80.9** | 80.9 | 79.5 | 79.5 | 79.5 | 79.5 |
| agriculture | **84.4** | 84.4 | 84.4 | 83.8 | 83.8 | 83.8 |
| avg. | **70.0** | 68.5 | 66.9 | 66.6 | 66.6 | 66.6 |

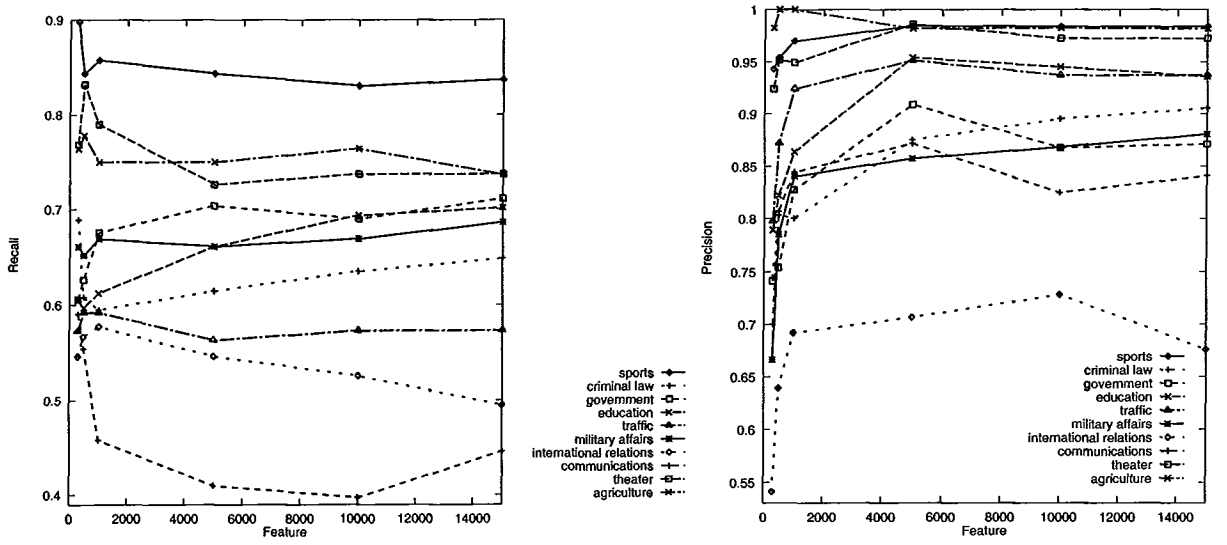


Figure 1: Recall and precision with MI features on SVMs.

Let us now consider the recall and precision results for the same data. Figures 2 plots the recall and precision on SVMs. The situation looks completely different from that of the MI filtering experiment because the number of Feature Set 1 (common nouns) reaches 8,629. Looking at Figures 1, only after 10,000 features, do we notice that POS filtering and MI filtering have the same tendencies: precision curves increase monotonically while recall curves differ among categories. This monotonic increase in the precision curve shows that every part of speech contains powerful keywords specific to one category. The recall curve shows that the increase in features above 10,000 words does not always improve recall but also that the drop in recall is not serious.

## Discussion

We have described the feature selection in SVM text categorization by focusing on two distinct experiments: MI filtering and POS filtering. The results on SVMs for each filtering technique can be summarized as follows and coincide on two points: One, the best feature set for categories changes and two, the best average performance is achieved when all of the features are given to SVMs. These are distinct characteristics of SVMs when compared with a decision tree learning algorithm C4.5.

**MI filtering** : The best number of words selected differs greatly among categories and is difficult to be determined a priori. The average performance is best when all of the words are used.

**POS filtering** The best feature set changes depending

Table 5: POS distribution of training data.

| | POS (part of speech) | | | | |
|---|---|---|---|---|---|
| | common noun | proper noun | verbal noun | unknown | verb |
| Number of words | 8629 | 2725 | 2829 | 1634 | 2294 |
| Percentage (%) | 47.6 | 15.0 | 16.0 | 7.4 | 12.7 |

Table 6: Average of recall and precision with POS filtering on SVMs.

| | poly degree $d = 1/d = 2$ | | | | |
|---|---|---|---|---|---|
| Feature | 1 | 2 | 3 | 4 | 5 |
| sports | 92.2/91.4 | **93.2**/92.4 | 92.9/92.4 | 92.0/92.0 | 90.5/90.8 |
| criminal law | 74.0/73.0 | 73.3/73.6 | 72.5/73.3 | 73.0/73.3 | **75.2**/74.9 |
| government | 76.9/76.7 | 78.4/78.7 | 79.3/79.0 | 78.9/78.2 | **79.6**/79.2 |
| education | 81.4/81.4 | 80.8/80.3 | 81.4/80.9 | 81.4/**81.8** | 81.2/80.3 |
| traffic | 72.8/72.0 | 76.0/74.5 | 74.8/**76.0** | 74.8/75.2 | 73.0/72.2 |
| military affairs | 80.1/77.3 | 76.1/76.1 | 78.8/76.2 | 77.0/77.0 | **80.1**/77.9 |
| international relations | 54.5/54.6 | 59.2/60.2 | 60.7/61.4 | 61.2/62.2 | **64.0**/64.0 |
| communications | 65.7/67.6 | 63.8/62.3 | **69.3**/68.0 | 68.9/67.1 | 65.7/63.2 |
| theater | 83.8/83.8 | 82.4/82.4 | 85.2/85.2 | 85.2/85.2 | **87.0**/85.0 |
| agriculture | 87.5/87.5 | **88.3**/88.3 | 87.5/86.6 | 86.6/86.6 | 85.0/84.8 |
| avg. | 76.9/76.5 | 77.5/77.2 | 78.0/77.9 | 77.9/77.8 | **78.1**/77.2 |

on the category. The best average performance is achieved when all of the words are used. Each part of speech makes a contribution but differs in influence among categories.

It is also important to note that POS filtering consistently outperforms MI filtering on SVMs (see 15,000 words in Table 3 and Feature Set 5 in Table 6). In MI filtering, every part of speech is adopted, including postpositional particles, conjunctions, etc. POS filtering on the other hand, selects only five parts of speech by using a natural language processing (NLP) technology (POS tagger). These include common nouns, proper nouns, verbal nouns, unknown words and verbs. The less crucial parts of speech are expected to be harmful to MI filtering. This comparison clearly shows that SVM text categorization has a limitation on its feature selection ability: it cannot detect irrelevant parts of speech.

Finally, we will briefly refer to the kernel functions we used. Changing the polynomial orders 1 and 2 makes no distinct difference in performance, although more significant influence was seen in lower level tasks like image processing (Cortes & Vapnik 1995).

## Conclusion

This paper has described various aspects of feature selection in SVM text categorization. Our experimental results clearly show that SVM text categorization handles large-scale word vectors well but is a limited in finding irrelevant parts of speech. This suggests a simple and strongly practical strategy for organizing a large number of words found through a rough filtering technique like part-of-speech tagging. SVMs and other large margin classifiers should play more important roles in handling complex and real-world NLP tasks (Haruno, Shirai, & Ooyama 1999).

## References


Cortes, C., and Vapnik, V. 1995. Support vector networks. *Machine Learning* 20:273–297.

Cover, T., and Thomas, J. 1991. *Elements of Information Theory.* John Wiley & Sons.

Dumais, S.; Platt, J.; Heckerman, D.; and Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. In *Proc. of 7th International Conference on Information and Knowledge Management.*

Haruno, M.; Shirai, S.; and Ooyama, Y. 1999. Using decision trees to construct a practical parser. *Machine Learning* 131–150. Special Issue on Natural Language Learning.

Joachims, T. 1998. Text categorization with support vector machines. In *Proc. of European Conference on Machine Learning(ECML).*

Lewis, D., and Ringuette, M. 1994. A comparison of two learning algorithms for text categorization. In *Proc. of Third Annual Symposium on Document Analysis and Information Retrieval,* 81–93.

Mainichi. 1995. *CD Mainichi Shinbun 94.* Nichigai Associates Co.

Table 7: Average of recall and precision with POS filtering on C4.5.

| Feature | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| sports | **84.7** | 82.9 | 83.4 | 83.0 | 83.4 |
| criminal law | 61.5 | 59.3 | **71.3** | 71.3 | 71.3 |
| government | 58.0 | **62.8** | 62.7 | 62.7 | 60.4 |
| education | 60.2 | 63.5 | 62.8 | **70.2** | 70.2 |
| traffic | 58.2 | 56.4 | 58.1 | 58.1 | **59.4** |
| military affairs | **75.5** | 71.8 | 71.8 | 71.8 | 71.8 |
| international relations | **49.3** | 44.1 | 48.9 | 46.4 | 46.4 |
| communications | 49.6 | 48.5 | **51.0** | 44.6 | 44.6 |
| theater | **79.7** | 71.3 | 79.2 | 79.2 | 79.2 |
| agriculture | 81.2 | **81.4** | 81.4 | 81.4 | 81.4 |
| avg. | 65.8 | 63.9 | **67.1** | 66.9 | 66.8 |

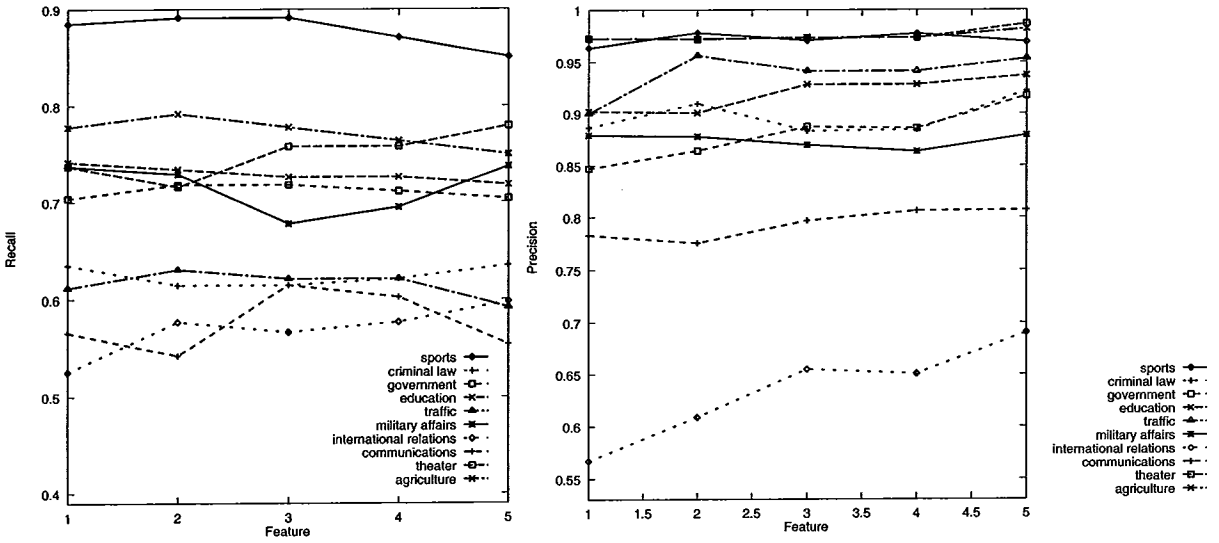


Figure 2: Recall and precision with POS features on SVMs.


Matsumoto, Y.; Kitauchi, A.; Yamashita, T.; Hirano, Y.; Imaichi, O.; and Imamura, T. 1997. *Japanese Morphological Analysis System Chasen Manual.* NAIST Technical Report NAIST-IS-TR97007.

Osuna, E.; Freund, R.; and Girosi, F. 1998. Training support vector machines: An application to face detection. In *Proc. of Computer Vision and Pattern Recognition '97*, 130–136.

Quinlan, J. 1993. *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

Salton, G., and Buckley, C. 1988. Term weighting approaches in automatic text retrieval. *Information Proceedings and Management* 24(5):513–523.

Toyoura, J.; Tokunaga, T.; Isahara, H.; and Oka, R. 1996. Development of a RWC text database tagged with classification code(in Japanese). In *NLC96-13. IEICE*, 89–96.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory.* New York: Springer-Verlag.

Yang, Y., and Pederson, J. 1997. A comparative study on feature selection in text categorization. In *Machine Learning: Proc. of the 14th International Conference (ICML'97)*, 412–420.

Yang, Y. 1994. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 13–22.