Text Categorization Using Latent Topics as Additional Features

Hiroshi Mizugai, Incheon Paik*, Shigeru Kanemoto**
University of Aizu
Computer Science of Engineering
Aizuwakamatu, Fukushima, Japan
{*m5121203,paikic*,kanemoto***}*@u-aizu.ac.jp*

Abstract-*In feature selection of text categorization, there are methods which handle word sense disambiguation by extracting synonymy and polysemy among words in documents. One of the methods utilizes latent topics underlying documents by using a topic model. PLSA and LDA have been proposed as representative models. In this paper, two features which include both TF-IDF and the latent topic values which extracted automatically from topic models were utilized for text categorization using AdaBoost. Then, the performances were compared with the ones of only TF-IDF features. As a result, this study evaluates effectiveness and weakness of the augmented features.*

Keywords: Machine Learning, Text Categorization, Latent Topics, AdaBoost

1. Introduction

Currently, standard feature selection of text categorization uses a base of so-called Bag-of-Words (BOW) consisting of only raw words in documents. BOW can not handle synonymy and polysemy among vocabularies. Therefore, essential meanings of words can not be handled sufficiently. To resolve the problem of BOW, two methods were mainly proposed to manage essential meanings of words. The fundamental idea of the methods is adding essential meanings of words to BOW features such as TF-IDF. One of the methods utilizes a thesaurus dictionary[1] [1][2] which could handle synonymy and polysemy. The method finds common synonyms among words in a document and uses frequency of extracted synonyms as additional text features. For example, if *Soccer* and *Baseball* occur in a document, *Sports* is extracted as a common synonym from used dictionary. However, it is very difficult to estimate the appropriate number of going up hypernyms of words. For example, although *Soccer* and *Marathon* may have *Athletics* as a common synonym generalizing the words, it can not know whether the document tends to the concept actually. Furthermore, as the performances of the categorization seriously depend on both the corpus domain and the used dictionary, this method is limited by corpus and dictionary. Another method utilizes a topic model [3]. The topic model can automatically extract latent topics from documents. The basic idea of the topic model assumes that a document has multiple topics, and a word appears on a topic. Latent topics are represented as probabilities and statistically derived from the probabilistic definition. When the latent topics are derived, a occurrence rate of each word on topics and a topic distribution of each document can be obtained. The topic model methods are robust because it can apply any corpora without the problems of the dictionary methods. Probabilistic Latent Semantic Analysis (PLSA) [4] supports the topic model. PLSA was utilized for text categorization in [3], and topic values of PLSA were used as additional features in the study. In this paper, Latent Dirichlet Allocation (LDA) [5] which extends the structure of PLSA model was also used. This paper's experiment uses the same way of [3], but our work analyzed additively the comparison of the performances between PLSA and LDA to investigate advantage of the topic models. The purpose of our study was surveying the potency of latent topics as additional features in text categorization. The augmented features consisting of the topic values and TF-IDF were applied to text categorization for two corpora. Then, the performances of the augmented features were compared with the ones of only TF-IDF features. From the results, effectiveness and weakness of the augmented features were considered.

[1] *WordNet* is often used in the dictionary method.

2. Method

2. 1. Probabilistic Latent Semantic Analysis

PLSA was proposed by Hofmman [4]. PLSA represents Latent Semantic Analysis (LSA) [6] as probabilistic model. PLSA assumes that a document d_i $(i = 1, 2, \ldots N)$ has latent topics z_k $(k = 1, 2, ..., K)$ and words w_j $(j = 1, 2, \ldots, M)$ occur on one latent topic. Figure 1 shows the graphical model of PLSA. The joint probability of d_i and w_i is defined by:

$$P(d_i, w_j) \quad = \quad P(d_i) \sum_{z=1}^{K} P(w_j \mid z_k) P(z_k \mid d_i) \quad (1)$$

and the log likelihood function for all documents is given:

$$\mathcal{L} = \log \prod_{i=1}^{N} \prod_{j=1}^{M} P(d_i,\, w_j)^{n(d_i,\, w_j)} \quad (2)$$

$n(d_i, w_j)$ denotes the observed number of the joint appearance of d_i and w_j. The parameters, $P(w_j \mid z_k)$ and $P(z_k \mid d_i)$ which maximize (2), can be solved using Expectation Maximization (EM) algorithm of maximum likelihood estimation. Concretely, the following update equations are calcuated:

E-step:

$$P(z_k \mid d_i, w_j) = \frac{P(w_j \mid z_k)P(z_k \mid d_i)}{\sum_{l=1}^{K} P(w_j \mid z_l)P(z_l \mid d_i)} \tag{3}$$

M-step:

$$P(w_m \mid z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)P(z_k \mid d_i, w_j)}{\sum_{i=1}^{N}\sum_{m=1}^{M} n(d_i, w_m)P(z_k \mid d_i, w_m)} \tag{4}$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j)P(z_k \mid d_i, w_j)}{n(d_i)} \tag{5}$$

In PLSA, the number of the estimated parameters are $kV + kM$ which grows linearly as M increases. This tends to cause overfitting to the learned documents. Indeed, to relax overfitting, Tempered EM [4] is used for the estimation. In this experiments, however, the model was not applied to unknown documents in terms of measuring the perplexity. For this reason, overfitting problem was not considered in this study.

2. 2. Latent Dirichlet Allocation

LDA which was proposed by Blei et al [5]. LDA extends the structure of PLSA model. LDA assumes that a document is represented as random variables denoting topic distribution, and a word occurs on the word probabilities of the topic corresponding to the word. The estimated parameters do not depend on the number of documents. Hence, LDA does not have the problem of overfitting of the PLSA model. The probabilistic distribution of each document follows Dirichlet distribution. The terms are defined: documents in a corpus $D = (\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_M})$, a sequence of words in a document $\boldsymbol{w} = (w_1, w_2, \ldots, w_N)$, the number of word vocabularies is V, and latent topics denote $z_k\,(k = 1, \ldots, K)$. The Dirichlet distribution is given by:

$$P(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_K^{\alpha_K - 1} \tag{6}$$

$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K)$ is the parameter of the Dirichlet distribution and $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_K)$ is the random variable as the topic distribution of each document. The word probabilities on each topic are represented as $K \times V$ matrix $\boldsymbol{\beta} = \{\beta_{ij}\} = p(w_j \mid z_i)$. Given $\boldsymbol{\alpha}, \boldsymbol{\beta}$, the joint probability of $\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}$ is:

$$P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = P(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} P(z_n \mid \boldsymbol{\theta}) P(w_n \mid z_n, \boldsymbol{\beta}) \tag{7}$$

By marginalizing $\boldsymbol{\theta}$ and summing over $\boldsymbol{z}$, a document probability is given:

$$\begin{aligned} & P(\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \int P(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \left(\prod_{n=1}^{N} \sum_{z_n} p(w_n \mid z_n, \boldsymbol{\beta}) p(z_n \mid \boldsymbol{\theta}) \right) d\boldsymbol{\theta} \\ &= \frac{\Gamma(\sum_{i=1}^{K} \alpha_k)}{\prod_{i=1}^{K} \Gamma(\alpha_k)} \int \left(\prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{K} (\theta_i \beta_{i w_n}) \right) d\boldsymbol{\theta} \end{aligned} \tag{8}$$

The equation (8) is intractable. To make it tractable, the equation is approximated by utilizing variational Bayes. In variational Bayes, the maximization of the lower bound of Jensen's inequality about $P(\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ is estimated. Concretely speaking,

$$\begin{aligned} & \log P(\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \log \int \sum_{\boldsymbol{z}} P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} \\ &= \log \int \sum_{\boldsymbol{z}} \frac{P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) Q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi})}{Q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi})} d\boldsymbol{\theta} \\ &\geq \int \sum_{\boldsymbol{z}} Q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi}) \log P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} \\ &\quad - \int \sum_{\boldsymbol{z}} Q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi}) \log Q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi}) d\boldsymbol{\theta} \end{aligned} \tag{9}$$

In (9), $Q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi})$ which maximizes the right side of the inequality is solved by utilizing maximum likelihood estimation. $Q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi})$ is represented as:

$$Q(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\gamma}) \prod_{n=1}^{N} Q(z_n \mid \phi_n) \tag{10}$$

$\boldsymbol{\gamma}$ is the parameter of the Dirichlet distribution for the document. $(\phi_1, \ldots, \phi_N)$ are the multinomial parameters determining the topic for the nth word in the document. Note that $\boldsymbol{\gamma}$ can be interpreted a set of topic frequencies for a document. By the derivative of the right side of (9) set to be zero, the update equations of $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ for a document are obtained:

$$\phi_{ni} \propto \beta_{i w_n} \exp\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K} \gamma_j \right) \right) \tag{11}$$

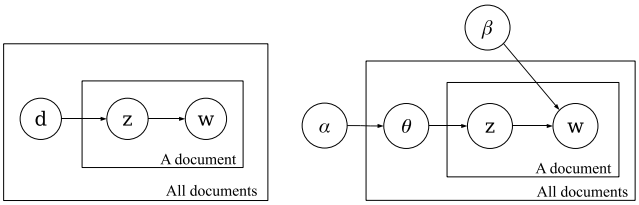
$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{12}$$

Ψ function is the first derivative of $\log \Gamma$. By using the convergence value of $\boldsymbol{\phi}$, the update equation for β is given:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j} \tag{13}$$

w_{dn}^{j} take 1 only if the nth word in the document d is w_j and the other cases take 0. The update equation with regard to

$\boldsymbol{\alpha}$ is estimated by utilizing Newton-Raphson method[2]. The graphical model of LDA is shown in Figure 1.



PLSA LDA

Figure 1: Graphical models of PLSA and LDA.

2. 3. AdaBoost

AdaBoost [7] is a representative method of ensemble learning. AdaBoost performed high accuracy compared to the other learning methods in text categorization [8]. The basic idea of AdaBoost is combining weak hypotheses into one strong hypothesis as majority voting. In this study, AdaBoost.MH which is an algorithm of AdaBoost was utilized because AdaBoost.MH marked higher performance than the other algorithms of AdaBoost. The terms are defined: training set $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in X, y_i \in Y = \{1, -1\}$. X is a set of examples. If a example belongs to the target category, Y takes 1. In the other cases, Y takes -1. T is the total number of rounds and h_t denote a weak hypothesis at a round t. The algorithm of AdaBoost.MH is defined by the following:

1 : Training set : $(x_1, y_1), \ldots, (x_m, y_m)$
where $x_i \in X$, $y_i \in Y = \{1, -1\}$.
2 : Initialize $D_1(i) = 1/m$.
3 : For $t = 1, \ldots, T$:
4 : Choose the weak hypothesis h_t following criteria.

$$h_t = \arg\min_h e_t \text{ where}$$

$$e_t = \sum_{i:h(x)\neq y_i} D_t(i).$$

5 : Compute $\alpha_t \in R$

$$\alpha_t = \frac{1}{2}\ln(\frac{1-e_t}{e_t}).$$

6 : Update the distribution :

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i\, h_t(x_i))}{Z_t}$$

where Z_t is the normalization factor.
7 : Get the final strong hypothesis :

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x).$$

In AdaBoost, each training example has $D_t(i)$ where $\sum_{i=1}^{M} D_t(i) = 1$ and $D_t(i) \geq 0$. $D_t(i)$ denotes the weight of the example at the round t. At the first round, $D_1(t)$ is initialized to uniform the distribution. In each round, the optimal weak learner is chosen and learns the weighted training examples. α_t is the weight for the weak hypothesis which grows as e_t decreases. D_{t+1} is updated as misclassified examples getting high weight. Therefore, as the rounds increases, the difficult examples to classify are focused. Finally, the strong hypothesis is obtained by combining α_t and h_t of all of the rounds. In this experiments, Boostexter [10] was used for AdaBoost.MH implementation. Boostexter can handle continuous value of features such as TF-IDF. For the domain of continuous values of the feature, the weak learner learns the optimal threshold. The threshold denotes that the examples are classified to the category when the feature value is more than the threshold, and vice versa.

[2]The update equation about $\boldsymbol{\alpha}$ can be referred in Blei's paper [5].

2. 4. Topic Features in Text Categorization

When the latent topics are applied to text categorization, for a document, the feature values of the topics are represented as $(P(z_1 \mid d), \ldots, P(z_K \mid d))$ on PLSA which denote the topic distribution and $(\gamma_1, \ldots, \gamma_K)$ on LDA which denote the set of the topic frequencies. In the parameters, if the words which have a similar meaning each other are in the document, the feature value of the topic related to the meaning is augmented. For example, if there are *Windows* and *Bill Gates* in the document, the two words contribute to the topic value such as *Microsoft*. Meanwhile, although the word including the multiple meanings, the suitale topic value corresponding to the word is augmented because the word occurred on one topic is assumed in the topic models. For example, if *Apple* appears in the document, the related topic *Computer* can be considered by according to the word in the document although the word also has *Fruit*. Thus, word sense disambiguation can be considered. In recent years, Support Vector Machine (SVM) had higher performance than the other learning methods in text categorization as well as AdaBoost [9]. In this paper, SVM was not used because the bias is not explicitly to combine the two heterogeneous dimensions including TF-IDF and topic values. In contrast, already mentioned, AdaBoost can combine the different classifiers to one strong classifier. Based on the idea of combining the weak classifiers which is consisting of TF-IDF and the topics values, AdaBoost was utilized in this experiment as like used in [3]. Note that in the past research, although text categorization using only topic features as feature reduction were experimented, it is doubtful to raise up the performance because some of the important words may be disregard. Therefore, both two features were used in this experiments.

3. Experiments

3. 1. Settings

Reuter-21578 and OHSUMED collection[11] were used to evaluate the performances in this experiments. Reuter-

21578 collection is Reuter's news articles in 1987. For the Reuter collection, Aptemod split was used to divide all documents to the training and test documents. 90 categories which were included at least once in both the training and test documents were selected. Then, 7768 training documents and 3019 test documents were obtained. OHSUMED collection is medical abstracts of Medical Subject Headings in 1991. 6286 training documents and 7463 test documents were used for the classification about the 23 disease categories. All documents of two corpora were tokenized for symbols such as punctuation and removed stop words which mean general words. To extract latent topics, all the documents consisting of the training and test documents were used. The number of the latent topics was set to $K = 10, \ldots, 200$ considering the time to extract the LDA topics[3]. The convergence threshold of the EM step was set to 0.0001 for PLSA and LDA respectively. Figure 2 shows the flowchart in this experiments. First, all features which include both TF-IDF and topic values were extracted. Second, AdaBoost learns the features of the training documents. Finally, AdaBoost classifies the test documents.

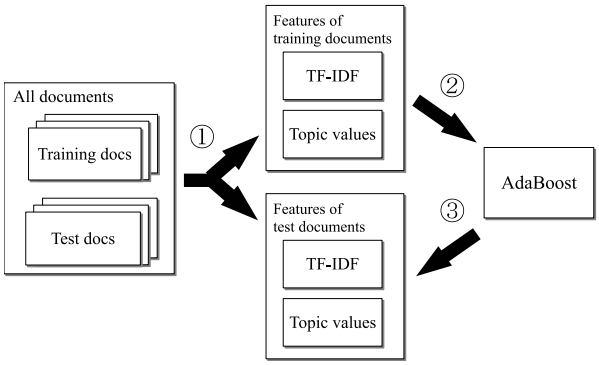


Figure 2: Flowchart in experiments.

3.2. Evaluation Metrics

One-Error was used for the evaluation. In general, a document has multiple categories. Since AdaBoost is a binary classifier, AdaBoost learns for each category and gets the final strong hypothesis corresponding to the category. Then, AdaBoost classifies a test document into one category which is the highest value of the final hypothesis function in all categories. One-Error measures the misclassified numbers that the documents was classified to the undefined categories. Terms are defined: test set $S = \{(x_1, Y_1), (x_m, Y_m)\}$ where Y denotes all of the categories of a test example x. The classifier is $H(x) = \arg\max_{l \in \mathcal{Y}} f(x)$ where $\mathcal{Y}$ denotes all of the categories in the corpus. One-Error is defined:

$$\text{One} - \text{Error} = \frac{1}{m} \sum_{i=1}^{m} \left[H(x_i) \notin Y_i \right]$$

[3]For instance, to extract 200 topics of LDA in the OHSUMED corpus, 25 hours was taken at least in 3.2 GHz CPU.

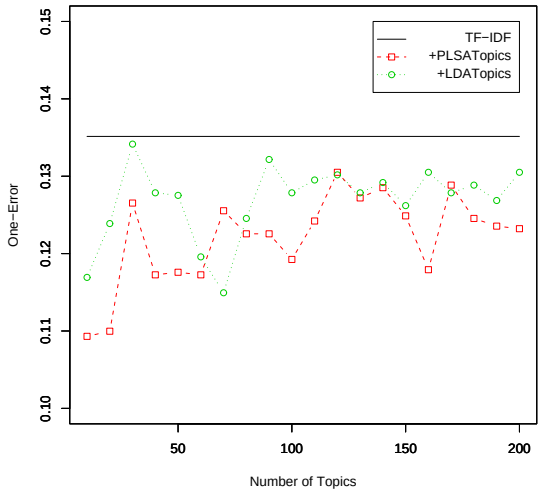


Figure 3: Comparison of number of topics of PLSA and LDA at rounds 200 of AdaBoost with only TF-IDF feautes on Reuter-21578.

3.3. Reuter-21578 Collection

One-Error of the augmented features including both the topic values and TF-IDF was compared with only TF-IDF features by adjusting the number of latent topics $K = 10, \ldots, 200$ for PLSA and LDA. Supplementarily, the words which have the highest probabilities on each topic by LDA model setting $K = 20$ were shown in Appendix.A. Figure 3 shows all of the augmented features improved the only TF-IDF features at the rounds 200 of AdaBoost. In particular, $K = 10$ in PLSA and $K = 70$ in LDA marked the highest performance respectively. Table 1 shows the augmented features improved One-Error at least 2% for the only TF-IDF features. The result shows that the augmented features for text categorization were very effective. Figure 4 shows the comparison by adjusting the rounds $t \leq 10000$. On all of the rounds, the augmented features outperformed the only TF-IDF features. Moreover, the added 70 topic features of LDA outperformed the added 10 topic features of PLSA from almost $t \geq 2000$, although the PLSA features outperformed the LDA features at the rounds $t = 200$. It means the performances are very different according to the rounds for the topics. Note that in Figure 4, One-Error of TF-IDF stop decreasing from $t \geq 8000$. However, One-Error of added topic features continue to decrease from $t \geq 8000$. Incidentally, in the same number of topics $k = 70$ on PLSA and LDA, the performances were different, although it was not explicit which one was more effective.

3.4. OHSUMED Collection

OHSUMED collection has very different contents from Reuter-21578. However, in the same way, the latent topics were automatically extracted by PLSA and LDA, One-Error of the augmented features adjusting the number of topics $k = 1, \ldots, 200$ were also compared with only TF-IDF fea-

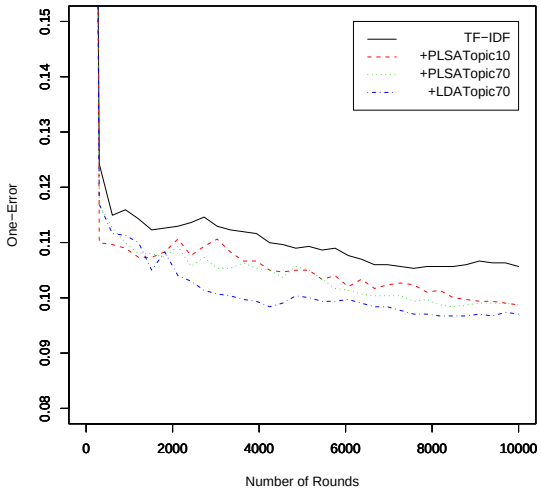


Figure 4: Comparison of number of rounds of AdaBoost in PLSA and LDA with only TF-IDF features on Reuter-21578 .

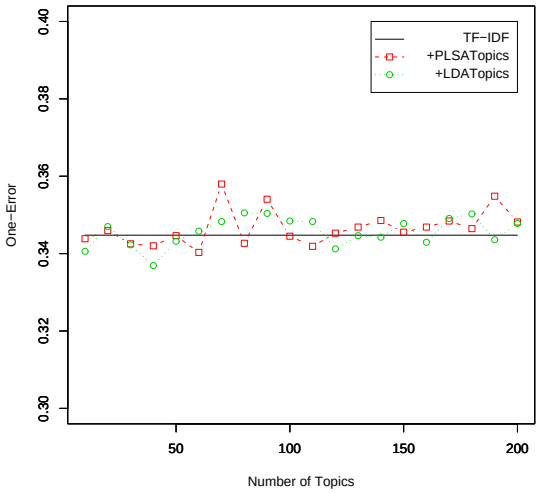


Figure 5: Comparison of number of topics of PLSA and LDA at rounds 2000 of AdaBoost on OHSUMED with only TF-IDF features.

Method	One-Error
TF-IDF	0.135
+PLSATopic10	0.109
+LDATopic70	0.114

Table 1: The highest performances of the added topic features of PLSA and LDA compared with only TF-IDF features at rounds 200 of AdaBoost on Reuter-21578.

tures. In fact, one of the optimal topic numbers for PLSA was found at $k = 400$ in the previous work [3]. In this paper, to experiment the added topic features of PLSA and LDA all together, the number of topics was limited. Therefore, the performances of the augmented features may be more worse than the 400 topic models. On the basis of the fact, Figure 5 shows the comparison of the performances by the number of topics at the round $t = 2000$. The number of topics $k = 60$ on PLSA and $k = 40$ on LDA marked the highest performance respectively shown in Table 2. Some added features improved only TF-IDF features slightly, although the other added latent topics made the performance worse. It seems that the number of the latent topics could not match the corpus successfully. From the result, it could be said that choosing the number of the topics was a critical issue of added topic features for text categorization.

Method	One-Error
TF-IDF	0.344
+PLSATopic60	0.340
+LDATopic40	0.336

Table 2: Performance of the added topic of PLSA and LDA compared with only TF-IDF at rounds 2000 of AdaBoost on OHSUMED.

4. Conclusion and Future Work

In this paper, latent topic values was used for text categorization as additional features using AdaBoost. In the results, some added topic features outperformed only TF-IDF features. Therefore, it can be concluded that adding latent topics to TF-IDF feautes was very effective to classify documents. However, there were the cases that the augmented features made the performance worse because the number of the topics might not match the corpus. It means that determining the number of the added topic features was very important issue to improve the performance. Moreover, according to the rounds on AdaBoost, the performances were different. Hence, determining the number of rounds on AdaBoost was also important. In the practice of the application, the estimation time of determining the optimal number of topics may be a bottleneck. Therefore, studying the theory of extracting automatically effective number of the topics will be the further study.

References

[1] G. Ifrim, M. Theobald, and G. Weikum, Learning Word-to-Concept Mappings for Automatic Text Classification, *Proceedings of the 22nd International Conference on Machine Learning - Learning in Web Search*, 2005

[2] E. Gabrilovich and S. Markovitch, Feature Generation for Text Categorization Using World Knowledge, *In Proceedings of The Nineteenth International Joint Conference for Artificial Intelligence*, pages 1048-1053, Edinburgh, Scotland, 2005

[3] L. Cai and T. Hofmann, Text Categorization by Boosting Automatically Extracted Concepts, *Proceedings of the 26th annual international ACM SIGIR*, 2003.

[4] T. Hofmann, Probabilistic Latent Semantic Analysis, *In Proceedings of the 15th Conference on Uncertainty in AI*, 1999.

[5] D. Blei, A. Ng, and M. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research* 3:993-1022, 2003.

[6] S. Deerwester, S. Dumais, G. Furnas, Landauter, and R. Harshman, Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, 41, 1990.

[7] R. Schapire and Y. Singer, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, *Journal of Computer and System Sciences*, 55 (1):119–139, 1997.

[8] R. Schapire and Y. Singer, BoosTexter: A Boosting-based System for Text Categorization, *Machine Learning*, 39 (2/3):135–168, 2000.

[9] T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, *Proceedings of ECML-98*, 1997.

[10] http://www.cs.princeton.edu/ schapire/boostexter.html

[11] http://dit.unitn.it/ moschitt/corpora.htm

Appendix. A

DLRS	COFFEE	GOLD	PCT
YEAR	MLN	MINE	SHARES
COMPANY	PRICES	USAIR	CANADIAN
QUARTER	OPEC	COPPER	CANADA
EARNINGS	YEAR	PCT	BRITISH
SHARE	CHINA	MINING	OFFER
MLN	TONNES	TONS	PLC
TRADE	CORP	MLN	DOLLAR
BILL	COMPANY	DLRS	YEN
GOVERNMENT	UNIT	BILLION	EXCHANGE
HOUSE	DLRS	YEAR	CURRENCY
TAIWAN	MARCH	NET	BANK
BILLION	MLN	STG	JAPAN
FOREIGN	SALE	TAX	RATES
MLN	SHARES	BILLION	PCT
CTS	PCT	BANK	BANK
NET	COMPANY	MLN	RATE
LOSS	GROUP	PCT	MONEY
SHR	STOCK	YEAR	MARKET
DLRS	STAKE	LOANS	RATES
QTR	DLRS	FRANCS	STG
GULF	MLN	TONNES	SHARES
OIL	TONNES	SUGAR	STOCK
SAUDI	CORN	WHEAT	COMPANY
IRAN	GRAIN	MARCH	SHARE
STRIKE	NIL	EXPORT	COMMON
IRANIAN	YEAR	COCOA	DLRS
ARABIA	WHEAT	APRIL	OFFER
TRADE	OIL	PCT	BLAH
JAPAN	GAS	YEAR	CTS
EC	CRUDE	JANUARY	MARCH
JAPANESE	DLRS	FEBRUARY	APRIL
STATES	PRICES	BILLION	RECORD
OFFICIALS	ENERGY	ROSE	DIVIDEND
COUNTRIES	BARRELS	RISE	DIV

Table 3: The seven words ordered by the highest word probabilities on each topic extracted by LDA setting 20 latent topics on Reuter-21578.