Text Classification of Technical Papers Focusing on Title and Important Segments

Thien Hai Nguyen[1,a)] Kiyoaki Shirai[1,b)]

Abstract: The goal of this research is to design a multi-label classification model which determines the research topics of a given technical paper. Based on the idea that papers are well organized and some parts of papers are more important than others for text classification, segments such as title, abstract, introduction and conclusion are intensively used in text representation. In addition, new features called Title Bi-Gram and Title SigNoun are used to improve the performance. Title Bi-Gram is bi-gram in the title, while Title SigNoun is a noun in a head phrase in the title. The results of the experiments indicate that feature selection based on text segmentation and these two features are effective. Furthermore, we proposed a new model for text classification based on the structure of papers, called Back-off model, which achieves 60.45% Exact Match Ratio and 68.75% F-measure. It was also shown that the back-off model outperformed two existing methods, ML-kNN and Binary Approach.

Keywords: Text Classification, Multi-label Classification, Text Segmentation, Supervised Learning

1. Introduction

In many research fields, a lot of papers are published every year. When researchers look for technical papers by a search engine, only papers including user's keywords are retrieved, and some of them might be irrelevant to the research topics that users want to know. Therefore, a survey of past researches is hard and difficult. Automatic identification of the research topics of the technical papers would be helpful for the survey. It is a kind of text classification problem.

The first step that must be considered in text classification is how to represent texts. Typical features are words in documents. It is also known as bag-of-word approach. However, using all words in the text might be inappropriate. Some of features may be noisy and cause negative impact. We believe that considering the structure of the documents and selecting words only in important parts of documents would achieve better results. Scientific papers are well-organized and tend to follow a consistent sequential structure: title, abstract, introduction, methods, evaluation, conclusions and references. Among various sections in the paper, words in title, abstract, introduction and conclusion might be more useful for text representation than the others. Especially, the title of the paper often represents the research topics clearly.

Our goal is to design an effective model which determines the categories of a given technical paper about natural language processing. In our approach, the model will consider the text segments in the paper. Several models with different feature sets from different segments are trained and combined. Furthermore, new features associated with the title of the paper are introduced.

The rest of the paper is organized as follows. Section 2 introduces some previous approaches on text classification and multi-label classification. Section 3 describes our dataset. Section 4 investigates the effectiveness of feature selections based on text segmentation in multi-label classification. We also propose a novel model in multi-label classification of technical papers based on the structures of papers. Section 5 assesses the results of the experiments. Finally, Section 6 concludes our contribution.

2. Background

2.1 Text Classification

Text classification has a long history. Many techniques have been studied to improve the performance. The commonly used text representation is bag-of-words [1]. Not words but phrases, word sequences or N-grams [2] are sometimes used. Most of them focused on words or N-grams extracted from the whole document with feature selection or feature weighting scheme.

Some of the previous work aimed at the integration of document contents and citation structure to improve the accuracy of categorization of technical papers [3, 4]. They first use the content-based classifier. Both words and phrases are used for text representation. Then the output of the first classifier will be updated by using citation-based classifier. However, these researches use entire document as features in the content-based classifier. On the other hand, in our method, features for text classification are extracted only from the limited segments in the paper.

Nomoto supposes the structure of the document as follows: the nucleus appears at the beginning of the text, followed by any number of supplementary adjuncts [5]. Then keywords for text classification are extracted only from the nucleus. Identification of nucleus and adjuncts is as a kind of text segmentation, but our

[1] Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1211, Japan
[a)] nhthien@jaist.ac.jp
[b)] kshirai@jaist.ac.jp

Table 1 Distribution of Papers in Terms of Number of Categories

# of categories in a paper	# of papers
1	1701 (86.3%)
2	259 (13.1%)
3	11 (0.6%)
4	1

text segmentation is fit for technical papers.

Larkey proposed a method to extract words only from the title, abstract, the first twenty lines of summary and the section containing the claims of novelty for a patent categorization application [6]. His method is similar to our research, but he classifies the patent documents, not technical papers. Furthermore, we proposed a novel method called back-off model as described in Subsection 4.4.

2.2 Multi-label Classification

Many researches in text classification deal with single-labeled data, where each training example is associated with one label. However, in many applications, single-label classifications are not appropriate and helpful. For example, categorization of research topics for technical papers, music categorization by emotions, semantic annotation for image or video etc. are examples of applications requiring multi-label classification.

There are many approaches for multi-label classification. However, they can be categorized into two groups: problem transformation and algorithm adaptation [7]. The former group is based on any algorithms for single-label classification. They transform the multi-label classification task into one or more single-label classification. On the other hand, the latter group extends traditional learning algorithms to deal with multi-label data directly.

3. Dataset

We collect technical papers in proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) from 2000 to 2011. To determine the categories (research topics) of the papers, we first refer the category list used for paper submission to the Language Resources and Evaluation Conference (LREC). Categories are coarse grained research topics such as syntactic parsing, semantic analysis, machine translation and so on. Some less frequent categories are removed from the list, while some categories are added when a considerable number of papers are related to them. Categories for each paper in the collection are annotated by authors.

The total number of papers in the collection is 1,972, while the total number of categories is 38. **Table 1** summarizes the statistics of our paper collection. Our dataset is available on the git repository [*1].

4. Multi-label Classification of Technical Papers

4.1 Text Segmentation

Most scientific papers are subdivided into the following sections: abstract, introduction, methods, experiments, results, conclusion and references. In general, abstract summarizes papers and allow the reader to judge whether papers are related to his or her own research interests. Introduction describes backgrounds of the paper. Conclusion summarizes papers and may refer the research topic again. On the other hand, other sections discuss the details of papers and might not so helpful for classification of research topics.

As the preprocessing of text classification, the following segments in the paper are automatically identified: title, author information (authors' names, affiliations, e-mail addresses etc.), abstract, introduction, conclusion and reference. Title is gotten from the database of papers shown in Section 3. A segment from the beginning of the paper to abstract is supposed to be an author information section. Abstract, introduction, conclusion and reference sections are identified by keywords in the papers. Hence, the identification of these sections is not always correct.

4.2 Title Feature

Document representation is one of the most important issues in text processing, especially in text categorization. As usual, we represent a document as a feature vector. Basically, words in the paper are used as features. Stop words and numbers are removed from the features, since they are ineffective for text classification. Furthermore, words in author information and reference are also removed. Then, the content words are lemmatized by the Stanford CoreNLP. All lemmatized forms of content words are used as features. Weights in the feature vector are defined as binary (1 or 0) or TF-IDF.

In addition to the above bag-of-word features, we propose new types of feature derived from the title of the paper. Words in the title seem the most important for paper classification. However, not all words in the title may be effective features. Some words represent the details of the contents and not represent the topic of the paper. If we give high weights for all words in the title, some noisy words would give negative impact for classification. In this paper, 'Title Bi-Gram' and 'Title SigNoun' are proposed to overcome this problem.

The title is usually not a complete sentence but just a phrase. In addition, the main words in the title tend to appear in noun phrases (NPs). While words in phrases other than NPs are not so relevant to the research topic. Considering above, 'Title Bi-Gram' is defined as bi-gram in noun phrases in the title. The motivation of 'Title Bi-Gram' feature is that the noun phrases in the title represent research topic clearly. Furthermore, the topics are often represented by not a single word but a phrase.

Another title feature is 'Title SigNoun', which is defined as significant nouns in the title. Two types of significant nouns are used as this feature. One is a noun in a head NP. The other is a noun in a prepositional phrase (PP). This feature is represented in the form of 'p+n', where n and p is a noun in PP and a head preposition of PP, respectively. The motivation of 'Title SigNoun' feature is that not only the nouns in the head NP but also in some cases the words in the prepositional phrase describe topics of papers. For example, a prepositional phrase "for information retrieval" strongly indicates that the paper tends to belong to "Information Retrieval" category. However, not all prepositional phrases in the title are effective features. For instance, "with bilingual lexicon"

[*1] https://github.com/nhthien/CorpusACL

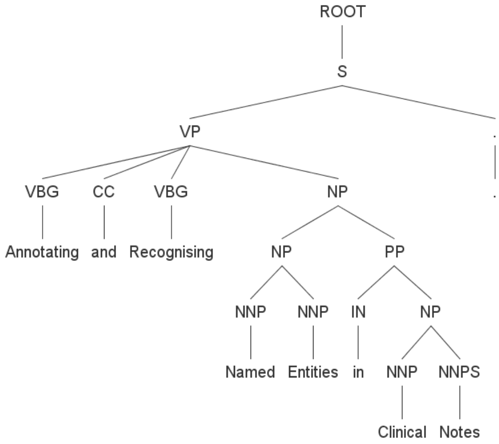

Fig. 1 Sample Parse Tree of a Title

is not useful because it might not help to identify topics of papers. The feature represented as the combination of the noun with the preposition, such as ‘for+retrieval’ or ‘with+lexicon’, enables us to distinguish effective and ineffective prepositional phrases.

For example, if the title is “Annotating and Recognising Named Entities in Clinical Notes”, Stanford parser outputs the parse tree in **Fig. 1**. Then ‘Named Entities’ and ‘Clinical Notes’ are extracted as Title Bi-Gram, while ‘Named’, ‘Entities’ and ‘in+Notes’ are extracted as Title SigNoun feature.

4.3 Feature Selection

We propose a method of feature selection based on the segments of the paper. Only words in useful segments such as title, abstract, introduction and conclusion are selected as features. We consider the five feature sets as follows:

(1) The whole content of paper: all of the words will be selected as features.

(2) Title, Abstract, Introduction, Conclusion: only words in these parts of the papers will be selected as features.

(3) Title, Abstract, Introduction, Conclusion + Title Bi-Gram: words in these parts as well as Title Bi-gram are used as features.

(4) Title, Abstract, Introduction, Conclusion + Title SigNoun: words in these parts as well as Title SigNoun are used as features.

(5) Title, Abstract, Introduction, Conclusion + Title Bi-Gram + Title SigNoun: words in these parts, Title Bi-Gram and Title SigNoun are used as features.

Furtheremore, to reduce the feature spaces, the features whose document frequency is less than 2 are removed. Note that dimensionality reduction based on document frequency and feature selection based on the text segments can be used simultaneously.

4.4 Classification Models

As discussed in Subsection 2.2, there are two approaches for multi-label classification: algorithm adaptation and problem transformation. We choose ML-kNN as the former and binary approach as the latter. Then we propose a novel model based on text segmentation of the paper.

ML-kNN [8] (Multi-label Learning K-Nearest Neighbors) is a multi-label lazy learning approach. It is derived from the traditional k-Nearest Neighbor algorithm. We used MULAN [9] as ML-kNN implementation in our experiments.

Table 2 Back-off Models

	Basic Feature Set	DF
BM1	Title + Title Bi-Gram	No
BM2	Title + Title Bi-Gram	Yes
BM3	Title + Title Sig-Gram	No
BM4	Title + Title Sig-Gram	Yes
BM5	Title + Title Bi-Gram + Title Sig-Noun	No
BM6	Title + Title Bi-Gram + Title Sig-Noun	Yes

Binary Approach [7] is a popular problem transformation method that learns $|C|$ binary classifiers for each different label in a label set C. It transforms the original data set into $|C|$ data sets $D_{C_i}(i = 1 \ldots |C|)$ that contain all examples of the original data set, where labeled as positive if the label set of the original example contained C_i and negative otherwise. To train each binary classifier, any kinds of traditional classifier can be utilized. Support Vector Machine (SVM) was used as a binary classifier in this paper. We used the tool called LibSVM [10] with linear kernel for training SVM. For the classification of a new instance, binary approach outputs the union of the labels C_i that are positively predicted by the classifiers. When no class is chosen, the system outputs one class whose posterior probability is the highest.

Based on the structure of papers, we propose a new model ‘Back-off Model’ derived from the binary approach. To improve the precision, only categories with high posterior probability from different perspectives are selected. Here the perspectives are binary approach methods with different feature sets. **Figure 2** shows an architecture of the back-off model. At first, a model with a basic feature set judges categories for the paper. As mentioned later, the basic features are words in the title and so on. The results of the model 1 are a list of categories with their posterior probabilities $\{(C_i, P_{i1})\}$. The system outputs categories C_i where P_{i1} are greater than a threshold T_1. When no class is chosen, the model 2 using words in the abstract as well as basic features is applied. Similarly, the model 3 (using words in introduction as well) and the model 4 (using words in conclusion as well) are applied in turn. When no class is chosen by the model 4, all categories whose probabilities P_{ik} are greater than 0.5 are chosen. If no P_{ik} is greater than 0.5, the system chooses one class with the highest probability. The threshold T_k for the model k is set smaller than that of the previous step. We investigate several sets of thresholds in the experiments in Section 5.

A variety of back-off models are investigated in the experiment. Three basic feature sets are considered: words in the title with and without Title Bi-Gram and Sig-Noun (described in Subsection 4.2). Furthermore, we consider the model with and without dimensionality reduction based on document frequency. **Table 2** summarizes our six back-off models.

5. Evaluation of Multi-label Classification

5.1 Experiment Setup

The proposed methods are evaluated by 10-fold cross validation on the collection of the papers described in Section 3. Evaluation criteria are summarized as follows:

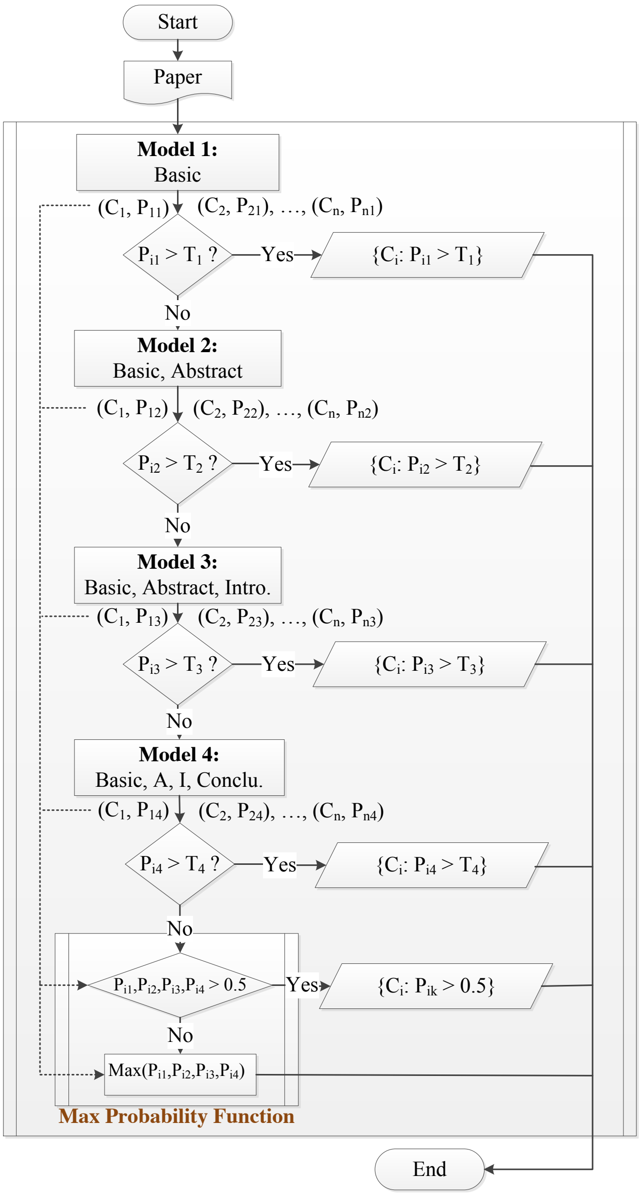
Start
Paper
Model 1: Basic
(C_1, P_{11}) (C_2, P_{21}), …, (C_n, P_{n1})
$P_{i1} > T_1$?
Yes
$\{C_i: P_{i1} > T_1\}$
No
Model 2: Basic, Abstract
(C_1, P_{12}) (C_2, P_{22}), …, (C_n, P_{n2})
$P_{i2} > T_2$?
Yes
$\{C_i: P_{i2} > T_2\}$
No
Model 3: Basic, Abstract, Intro.
(C_1, P_{13}) (C_2, P_{23}), …, (C_n, P_{n3})
$P_{i3} > T_3$?
Yes
$\{C_i: P_{i3} > T_3\}$
No
Model 4: Basic, A, I, Conclu.
(C_1, P_{14}) (C_2, P_{24}), …, (C_n, P_{n4})
$P_{i4} > T_4$?
Yes
$\{C_i: P_{i4} > T_4\}$
No
$P_{i1},P_{i2},P_{i3},P_{i4} > 0.5$
Yes
$\{C_i: P_{ik} > 0.5\}$
No
$Max(P_{i1},P_{i2},P_{i3},P_{i4})$
Max Probability Function
End

Fig. 2 Architecture of Back-off Model

- **Instance-based Metrics**: Exact Match Ratio (EMR), Accuracy, Precision and Recall.
  EMR is a proportion of instances (papers) where the gold and predicted set of categories are exactly same. While others evaluate the predicted categories for individual instances.
- **Category-based Metrics**: Micro-Precision, Micro-Recall, Micro-F, Macro-Precision, Macro-Recall, and Macro-F.
  First, we measured precision, recall and F-measure on the prediction of individual categories, then calculated micro and macro averages of all categories.

Differences of several methods are verified by a statistical test. We use the randomization test of paired sample [11] because it does not require additional assumption about population of outputs. Algorithm 1 summarizes it. In this paper, we choose $R = 100000$ to conduct this test.

Algorithm 1 Psuedocode for Randomization Test of Paired Sample

```
Let $o_A = \{o_A^1, \cdots, o_A^n\}$ and $o_B = \{o_B^1, \cdots, o_B^n\}$ be the output of the two systems on the same input.
Let $t(o_A, o_B)$ is the difference between outputs of two systems.
Start with $X = o_A$ and $Y = o_B$.
Repeat $R$ times: randomly flip each $o_A^i$, $o_B^j$ between $X$ and $Y$ with probability $\frac{1}{2}$. Calculate $t(X, Y)$.
Let $r$ be the number of times that $t(X, Y) \geq t(o_A, o_B)$
As $R \to \infty$, $\frac{r+1}{R+1}$ approached the p-value.
```

5.2 Results

5.2.1 ML-kNN:

Table 3 reveals results of ML-kNN with several feature sets. TAIC stands for the feature set derived only from title, abstract, introduction and conclusion. Binary and TF-IDF as term weighting are also evaluated. Since TF-IDF was better than binary weighting in TAIC model, we only evaluated the last three feature sets (TAIC + title features) with TF-IDF weighting. The value of the best system for each evaluation metrics is represented in bold. In the binary weighting method, feature selection by text segmentation gives worse results than use of all contents. However, in TF-IDF weighting, TAIC model shows better results than All. Since results of binary weighting are better than those of TF-IDF Weighting, we can conclude that feature selection based on text segmentation is not so effective in ML-kNN.

5.2.2 Binary Approach:

Table 4 reveals results of binary approach. **Figure 3** compares All and TAIC to show the effectiveness of feature selection based on text segmentation. Similar to ML-kNN model, feature selection by text segmentation does not work well in binary weighting, but does in TF-IDF. Unlike ML-kNN, however, the performance of TF-IDF is better than binary weighting and the performance of the binary approach is much better than ML-kNN. Therefore, we can conclude that our approach to select features based on the structure of the paper is effective. As shown in Fig. 3, EMR, micro-F and macro-F of TAIC are improved by 2-5% compared

[*2] DF: dimensionality reduction by Document Frequency.
[*3] It indicates term weighting: BW (binary weighting) or TF-IDF.

Table 3 Results of ML-kNN with K = 100

Feature Set	**DF**[*2]	**TW**[*3]	**Metrics**									
			Instance-based				**Category-based**					
			EMR	**A**	**P**	**R**	**Mi-P**	**Mi-R**	**Mi-F**	**Ma-P**	**Ma-R**	**Ma-F**
All	No	BW	39.91	44.32	48.63	44.55	48.73	42.90	45.63	42.72	26.02	37.39
		TF-IDF	34.23	37.68	41.15	37.76	41.18	36.08	38.46	33.90	20.70	29.26
	Yes	BW	41.68	46.06	50.43	46.24	50.48	44.47	47.28	46.29	26.63	38.08
		TF-IDF	40.41	44.27	47.99	44.54	48.04	42.45	45.07	43.15	28.21	35.83
TAIC	No	BW	36.81	40.94	45.00	41.12	45.06	39.61	42.16	42.75	22.75	33.67
		TF-IDF	38.44	42.40	46.30	42.59	46.37	40.84	43.43	44.29	27.32	36.48
	Yes	BW	37.98	42.14	46.35	42.23	46.33	40.62	43.29	44.80	24.13	35.23
		TF-IDF	41.43	45.86	50.02	46.27	50.23	44.57	47.22	47.82	29.75	**40.33**
TAIC + Title SigNoun	No	TF-IDF	39.04	42.62	46.22	42.73	46.30	40.74	43.34	41.02	26.16	35.35
	Yes	TF-IDF	41.48	45.96	50.29	46.27	50.39	44.56	47.30	46.73	29.00	39.24
TAIC + Title Bi-Gram	No	TF-IDF	39.30	43.33	47.29	43.56	47.34	41.77	44.38	43.01	26.60	36.40
	Yes	TF-IDF	42.54	46.69	50.59	47.05	50.87	45.17	47.85	**49.03**	**30.49**	40.09
TAIC + Title Bi-Gram + Title SigNoun	No	TF-IDF	40.46	44.38	48.21	44.56	48.35	42.70	45.35	42.28	27.44	36.47
	Yes	TF-IDF	**43.36**	**47.20**	**50.79**	**47.58**	**51.11**	**45.43**	**48.10**	46.24	30.22	39.56

Table 4 Results of Binary Approach

Feature Set	**DF**	**TW**	**Metrics**									
			Instance-based				**Category-based**					
			EMR	**A**	**P**	**R**	**Mi-P**	**Mi-R**	**Mi-F**	**Ma-P**	**Ma-R**	**Ma-F**
All	No	BW	47.57	53.74	57.60	56.08	57.53	54.44	55.93	52.69	43.79	48.58
		TF-IDF	46.55	59.00	62.51	68.89	57.57	67.44	62.10	49.22	58.83	53.55
	Yes	BW	46.60	52.71	56.58	54.99	56.60	53.42	54.96	52.03	42.55	47.40
		TF-IDF	48.22	59.17	62.83	67.45	57.67	65.79	61.45	49.05	57.00	52.22
TAIC	No	BW	44.17	50.41	53.88	53.25	53.97	51.69	52.80	48.35	41.73	44.77
		TF-IDF	51.72	61.80	65.10	68.93	62.15	67.26	64.58	55.59	59.35	56.74
	Yes	BW	43.36	49.59	53.02	52.47	53.13	50.93	52.00	47.45	41.04	43.85
		TF-IDF	51.01	60.57	64.37	66.69	61.66	65.08	63.31	56.16	57.24	56.59
TAIC + Title SigNoun	No	TF-IDF	52.84	62.95	66.27	69.98	63.40	68.41	65.79	56.52	59.62	57.79
	Yes	TF-IDF	52.23	61.61	65.14	67.73	62.59	66.32	64.39	57.08	58.25	57.36
TAIC + Title Bi-Gram	No	TF-IDF	52.94	63.37	66.90	70.57	63.91	68.89	66.29	57.59	61.27	58.59
	Yes	TF-IDF	52.18	61.75	65.54	67.87	62.83	66.24	64.47	57.66	58.38	57.68
TAIC + Title Bi-Gram + Title SigNoun	No	TF-IDF	**53.80**	**64.05**	**67.38**	**71.20**	**64.57**	**69.65**	**66.99**	**58.17**	**61.36**	**59.72**
	Yes	TF-IDF	53.35	62.77	66.32	68.94	63.58	67.47	65.45	57.88	59.14	58.40

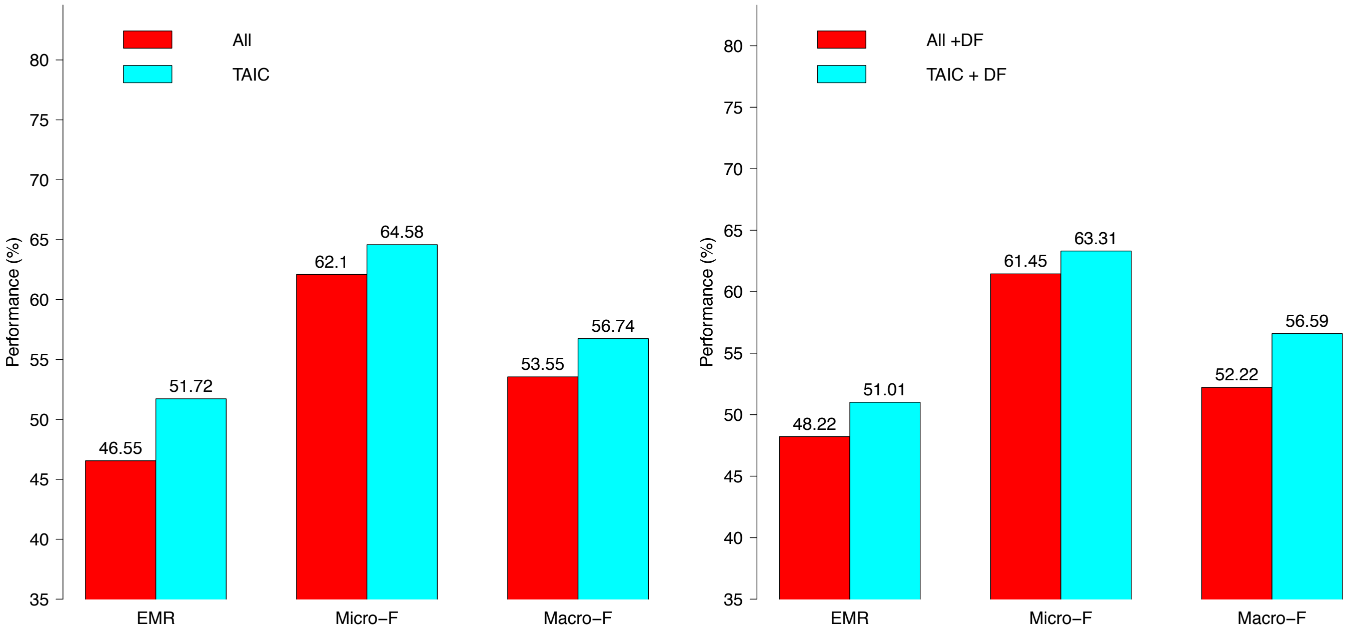

Fig. 3 Effectiveness of Feature Selection based on Text Segmentation (Binary Approach, TF-IDF Weighting)

Table 5 Statistical Test for Effectiveness of Feature Selection (Binary Approach, TF-IDF Weighting)

First Feature	Second Feature	P-value		
		EMR	Mi-F	Ma-F
All	TAIC	0	0.001	0.0019
All, DF	TAIC, DF	0.009	0.0231	0.0061

with All.

Table 5 shows results of statistical tests to verify the effectiveness of TAIC model. EMR and macro-F of TAIC is better than All at 0.01 significant level, and micro-F at 0.05.

Figure 4 compares TAIC, TAIC + Title SigNoun and TAIC + Title Bi-Gram, with and without dimentionality reduction, to evaluate the contribution of title features. It indicates that use of Title Bi-Gram and Title SigNoun improves the performance on three metrics. **Table 6** shows results of statistical tests, indicating that models with title features are better at 0.01 significant level in most cases. Therefore, we can conclude that our new features derived from the title are effective.

5.2.3 Back-off Model:

Table 7 shows results of back-off model. Threshold T_1, T_2, T_3, T_4 ($100\% \geq T_1 \geq T_2 \geq T_3 \geq T_4 \geq 50\%$) are chosen based on our intuition. We found that BM1 was better than other back-off models in most criteria.

To compare the performance of three models (ML-kNN, binary approach and back-off model) in detail, we plot the highest values of all metrics for these methods in **Fig. 5**. It indicates that ML-kNN performs much worse than binary approach and back-off model on all metrics. Binary approach method outperformed back-off model on recall, micro-Recall and macro-Recall metrics. In contrast, back-off model tends to achieve better results on EMR, accuracy, precision, micro-Precision, macro-Precision, micro-F and macro-F. Therefore, back-off model is the best among three approaches. Results of statistical tests shown in **Table 8** indicate that binary approach is better than ML-kNN at 0.01 significant level. Furthermore, back-off model is obviously better than binary approach, although the differences are not statistically significant in some cases.

6. Conclusion

Identification of research topics of papers is a challenging and difficult multi-label classification problem. To solve it, we proposed a feature selection method based on the structure of the paper and new feature derived from the title. We also proposed the back-off model, which combines classifiers with different feature sets from different segments of the papers. Experimental results indicates that our methods are effective for text categorization of technical papers.

In the future, we will explore more effective methods of feature selection and feature weighting to improve the accuracy of text classification. For example, combining Topic Modeling such as Latent Dirichlet Allocation [12] to exploit the semantics of the contents of the paper will be considered.

[*4] Insignificant cases at 0.05 significance level are indicated in italic.

References

[1] Sebastiani, F.: Machine learning in automated text categorization, *ACM Comput. Surv.*, Vol. 34, No. 1, pp. 1–47 (2002).

[2] Rahmoun, A. and Elberrichi, Z.: Experimenting N-Grams in Text Categorization, *Int. Arab J. Inf. Technol.*, pp. 377–385 (2007).

[3] Cao, M. D. and Gao, X.: Combining Contents and Citations for Scientific Document Classification., *Australian Conference on Artificial Intelligence'05*, pp. 143–152 (2005).

[4] Zhang, M., Gao, X., Cao, M. D. and Ma, Y.: Modelling citation networks for improving scientific paper classification performance, *Proceedings of the 9th Pacific Rim international conference on Artificial intelligence*, PRICAI'06, Berlin, Heidelberg, Springer-Verlag, pp. 413–422 (2006).

[5] Nomoto, T. and Matsumoto, Y.: Exploiting Text Structure for Topic Identification, *Proceedings of the 4th Workshop on Very Large Corpora*, pp. 101–112 (1996).

[6] Larkey, L. S.: A patent search and classification system, *Proceedings of the fourth ACM conference on Digital libraries*, DL '99, New York, NY, USA, ACM, pp. 179–187 (1999).

[7] Tsoumakas, G., Katakis, I. and Vlahavas, I.: Mining Multi-label Data, *Data Mining and Knowledge Discovery Handbook* (Maimon, O. and Rokach, L., eds.), Springer US, pp. 667–685 (2010).

[8] Zhang, M.-L. and Zhou, Z.-H.: ML-KNN: A lazy learning approach to multi-label learning, *Pattern Recognition*, Vol. 40, No. 7, pp. 2038 – 2048 (2007).

[9] Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J. and Vlahavas, I.: Mulan: A Java Library for Multi-Label Learning, *Journal of Machine Learning Research*, Vol. 12, pp. 2411–2414 (2011).

[10] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp. 27:1–27:27 (2011). Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[11] Morgan, W.: Statistical Hypothesis Tests for NLP. `http://cs.stanford.edu/people/wmorgan/sigtest.pdf`.

[12] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation, *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022 (2003).

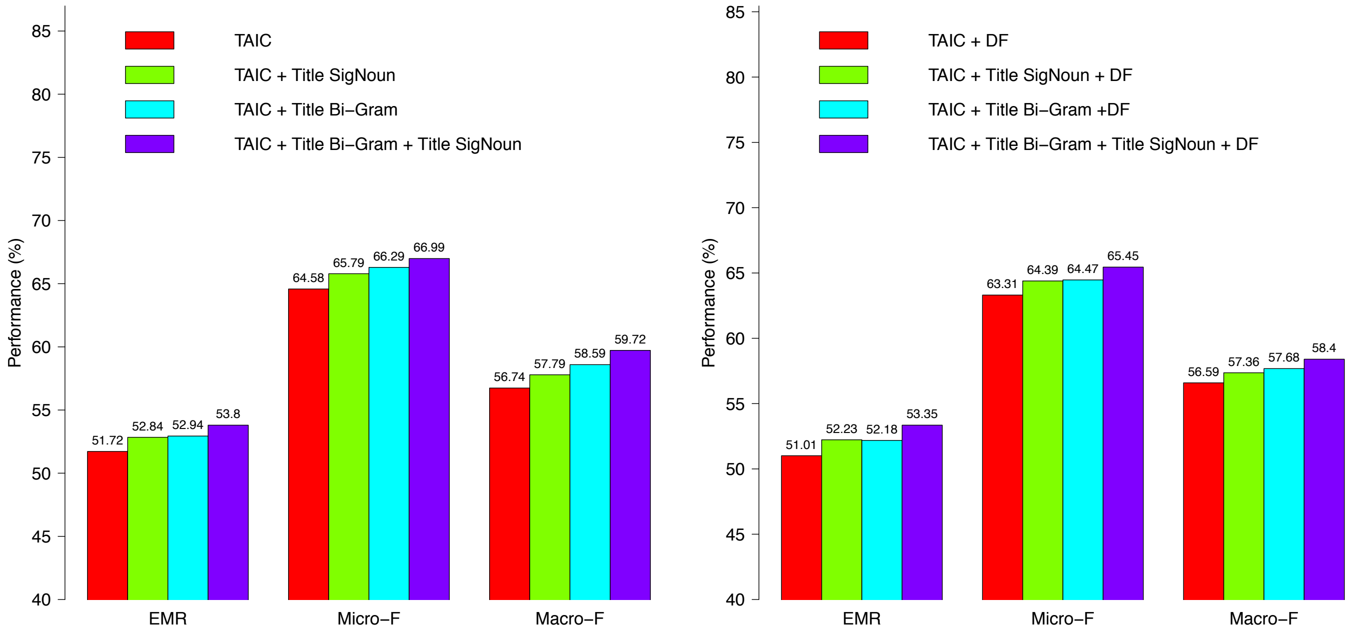

Fig. 4 Effectiveness of Title Features (Binary Approach, TF-IDF Weighting)

Table 6 Statistical Test for Effectiveness of Title Features (Binary Approach, TF-IDF Weighting)

First Feature	**Second Feature**	**P-value**		
		EMR	**Mi-F**	**Ma-F**
TAIC	TAIC + Title SigNoun	0.0181	0.0004	0.022
TAIC	TAIC + Title Bi-Gram	0.0046	0	0
TAIC	TAIC + Title Bi-Gram + Title SigNoun	0.0002	0	0
TAIC, DF	TAIC + Title SigNoun, DF	0.0131	0.0019	0.0087
TAIC, DF	TAIC + Title Bi-Gram, DF	0.0076	0.0004	0.0051
TAIC, DF	TAIC + Title Bi-Gram + Title SigNoun, DF	0	0	0

Table 7 Results of Back-off Model

Model	**Thresholds** T_1-T_2-T_3-T_4	**Metrics**									
		Instance-based				**Category-based**					
		EMR	**A**	**P**	**R**	**Mi-P**	**Mi-R**	**Mi-F**	**Ma-P**	**Ma-R**	**Ma-F**
BM1	80-80-50-50	60.04	67.21	72.01	**69.75**	70.20	67.39	**68.76**	65.66	**59.90**	61.97
	80-80-70-50	60.14	67.09	71.97	69.32	70.43	66.91	68.61	65.80	59.43	61.73
	80-80-80-50	**60.45**	**67.25**	**72.07**	69.44	**70.58**	67.04	68.75	**66.33**	59.85	**62.16**
BM2	80-80-50-50	58.27	65.29	70.25	67.65	68.09	65.30	66.66	62.59	56.88	59.65
	80-80-70-50	59.23	65.77	70.81	67.50	69.21	65.13	67.10	63.58	56.88	60.09
	80-80-80-50	59.18	65.72	70.77	67.48	69.24	65.04	67.07	63.76	56.88	60.30
BM3	80-80-50-50	58.42	65.58	70.19	68.40	68.03	66.20	67.09	61.71	57.74	59.57
	80-80-70-50	58.36	65.31	70.02	67.69	68.43	65.53	66.94	62.47	57.04	59.57
	80-80-80-50	58.42	65.32	70.05	67.69	68.47	65.49	66.94	62.70	57.33	59.73
BM4	80-80-50-50	57.45	65.07	70.07	68.02	67.50	65.57	66.51	61.22	57.06	59.23
	80-80-70-50	58.31	65.24	70.29	67.34	68.57	64.91	66.68	62.24	56.62	59.80
	80-80-80-50	58.21	65.12	70.18	67.29	68.21	64.87	66.49	62.06	56.61	59.69
BM5	80-80-50-50	59.79	66.97	71.67	69.72	69.59	**67.40**	68.47	63.48	59.23	61.24
	80-80-70-50	60.09	67.06	71.81	69.57	70.24	67.17	68.67	64.49	59.34	61.69
	80-80-80-50	60.29	67.15	71.95	69.52	70.45	67.09	68.72	64.56	59.22	61.67
BM6	80-80-50-50	57.00	64.78	69.98	67.67	67.64	65.14	66.34	62.14	56.73	58.92
	80-80-70-50	57.86	65.13	70.40	67.34	68.87	64.78	66.75	63.40	56.40	59.64
	80-80-80-50	58.21	65.34	70.56	67.55	68.71	64.96	66.76	63.46	56.90	60.15

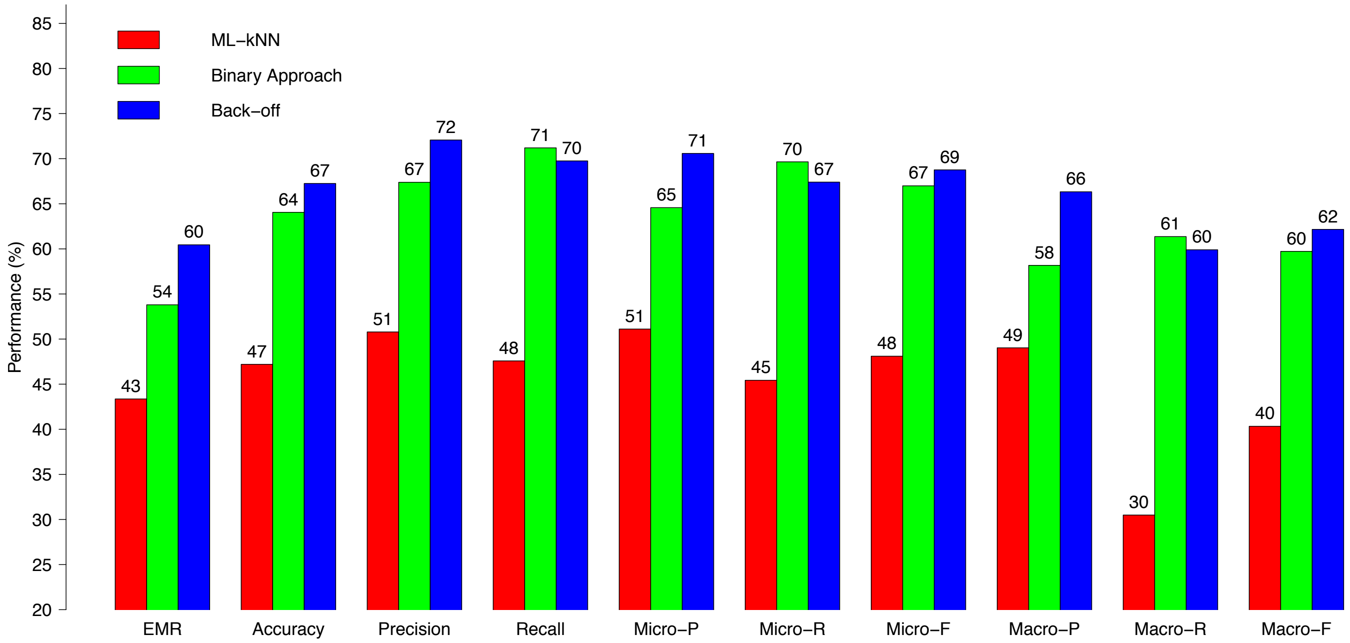

Fig. 5 Best Performance of Three Models

Table 8 Statistical Test for Comparison of Three Models

First Model	**Second Model**	**P-value**[*4]		
		EMR	**Mi-F**	**Ma-F**
ML-kNN (TAIC + TitleBiGram + TitleSigNoun)	BinaryApproach (same feature)	0	0	0
ML-kNN (TAIC + TitleBiGram + TitleSigNoun, DF)	BinaryApproach (same feature)	0	0	0
BinaryApproach (TAIC + TitleBiGram + TitleSigNoun)	BM5 (80-80-50-50)	0	0.0445	*0.0659*
BinaryApproach (TAIC + TitleBiGram + TitleSigNoun)	BM5 (80-80-70-50)	0	0.0191	0.0493
BinaryApproach (TAIC + TitleBiGram + TitleSigNoun)	BM5 (80-80-80-50)	0	0.0164	0.0459
BinaryApproach (TAIC + TitleBiGram + TitleSigNoun, DF)	BM6 (80-80-50-50)	0.0001	*0.2205*	*0.4758*
BinaryApproach (TAIC + TitleBiGram + TitleSigNoun, DF)	BM6 (80-80-70-50)	0	*0.0721*	*0.3054*
BinaryApproach (TAIC + TitleBiGram + TitleSigNoun, DF)	BM6 (80-80-80-50)	0	*0.067*	*0.258*
BinaryApproach (TAIC + TitleBiGram + TitleSigNoun)	BM1 (80-80-50-50)	0	0.0151	0.0107
BinaryApproach (TAIC + TitleBiGram + TitleSigNoun)	BM1 (80-80-70-50)	0	0.0241	0.0193
BinaryApproach (TAIC + TitleBiGram + TitleSigNoun)	BM1 (80-80-80-50)	0	0.0148	0.0118

Note: Term weights are defined as TF-IDF in all models.