Text Classification of Technical Papers Based on Text Segmentation

Thien Hai Nguyen and Kiyoaki Shirai

Japan Advanced Institute of Science and Technology
{nhthien,kshirai}@jaist.ac.jp

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. 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 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.

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.

2 Background

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 [3] [4].

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 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.

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. 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. The average number of the categories per a paper is 1.144. Our dataset is available on the git repository [1].

4 Multi-label Classification of Technical Papers

4.1 Text Segmentation

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.

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

4.2 Title Feature

In addition to the ordinary 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. In this paper, 'Title Bi-Gram' and 'Title SigNoun' are proposed to overcome this problem. '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. Another title feature is 'Title SigNoun', which is defined as a noun in a head NP and 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, while "with bilingual lexicon" might not be helpful in identifying 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, from the title "Annotating and Recognising Named Entities in Clinical Notes", '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. Words in title, abstract, introduction and conclusion (TAIC).
3. Words in TAIC and Title Bi-Gram.
4. Words in TAIC and Title SigNoun.
5. Words in TAIC, Title Bi-Gram and Title SigNoun.

4.4 Classification Models

As discussed in Section 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. ML-kNN [8] is a multi-label lazy learning approach. We used MULAN [9] as ML-kNN implementation in our experiments. Binary Approach [7] is a model that determines categories from results of $|C|$ binary classifiers for each different label, where C is a label set. We used LibSVM [10] with linear kernel to train each binary classifier.

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 1 shows an architecture of back-off model. At first, a model with a basic feature set judges categories for the paper. The basic feature set is a set of words in the title with Title Bi-Gram and/or Title SigNoun feature [2]. The results of 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, model 2 using words in the abstract as well as basic features is applied. Similarly, model 3 (using words in introduction as well) and model 4 (using words in conclusion as well) are applied in turn. When no class is chosen by 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.

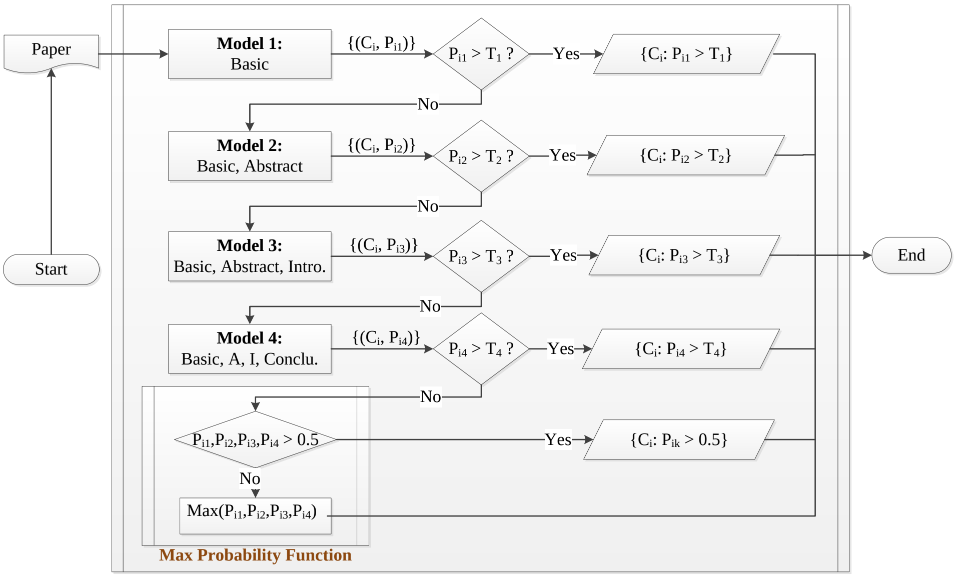

Fig. 1. Architecture of Back-off Model

5 Evaluation

The proposed methods are evaluated by 10-fold cross validation on the collection of the papers described in Section 3. We used exact match ratio (EMR), accuracy, precision and recall as the instance-based metrics [3], and micro-Precision, micro-

[2] Three basic feature sets were investigated: Title + Title Bi-Gram (BF_1), Title + Title SigNoun (BF_2) and Title + Title Bi-gram + Title SigNoun (BF_3). In our experiments, BF_1 achieved the best.

[3] 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.

Recall, micro-F, macro-Precision, macro-Recall, and macro-F as the category-based metrics [4]. Although we have evaluated various feature sets and parameters for ML-kNN, binary approach and back-off model, only some of the results will be shown in this paper due to the lack of space.

Table 1 reveals results of binary approach. It shows that using feature selection by text segmentation gives better results than using all content of the paper [5]. In addition, combining Title Bi-Gram and Title SigNoun improves the performance [6]. Table 2 shows the results of back-off model with some combinations of thresholds $T_1 \sim T_4$. We found that the performance of back-off model did not highly depend on the thresholds.

Table 1. Results of Binary Approach

Feature Set	**Instance-based Metrics**				**Category-based Metrics**					
	EMR	**A**	**P**	**R**	**Mi-P**	**Mi-R**	**Mi-F**	**Ma-P**	**Ma-R**	**Ma-F**
All	46.55	59.00	62.51	68.89	57.57	67.44	62.10	49.22	58.83	53.55
TAIC	51.72	61.80	65.10	68.93	62.15	67.26	64.58	55.59	59.35	56.74
TAIC + Title SigNoun	52.84	62.95	66.27	69.98	63.40	68.41	65.79	56.52	59.62	57.79
TAIC + Title Bi-Gram	52.94	63.37	66.90	70.57	63.91	68.89	66.29	57.59	61.27	58.59
TAIC + Title Bi-Gram + Title SigNoun	**53.80**	**64.05**	**67.38**	**71.20**	**64.57**	**69.65**	**66.99**	**58.17**	**61.36**	**59.72**

Table 2. Best Results of Back-off Model

Thresholds	**Instance-based Metrics**				**Category-based Metrics**					
T_1-T_2-T_3-T_4	**EMR**	**A**	**P**	**R**	**Mi-P**	**Mi-R**	**Mi-F**	**Ma-P**	**Ma-R**	**Ma-F**
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**

To compare the performance of ML-kNN, binary approach and back-off model, the highest values among various feature sets and parameters for three models are shown in Figure 2. 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.

6 Conclusion

To identify research topics of papers, we proposed a feature selection method based on the structure of the paper and new features derived from the title. We also proposed back-off model, which combines classifiers with different feature sets from different segments of the papers. Experimental results indicate that our

[4] They are averages of prediction of individual categories.

[5] Differences between All and TAIC are verified by a statistical test called randomization test of paired sample [11]. They are statistically significant.

[6] Differences between models with and without Title Bi-Gram/SigNoun were statistically significant.

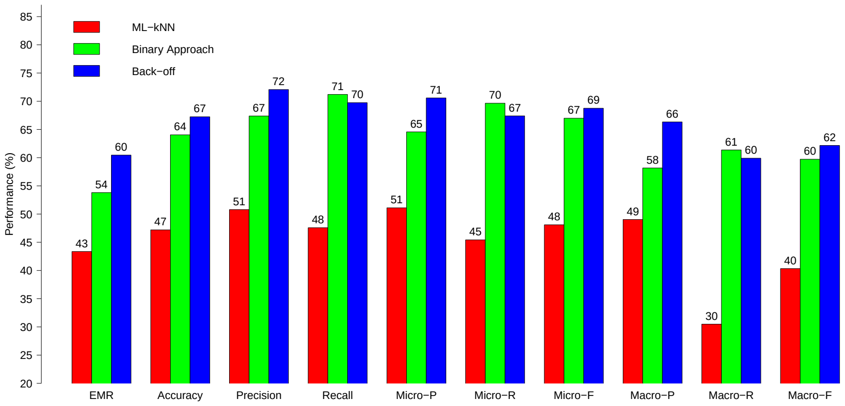

Fig. 2. Best Performance of Three Models

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.

References

1. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1) (March 2002) 1–47
2. Rahmoun, A., Elberrichi, Z.: Experimenting n-grams in text categorization. Int. Arab J. Inf. Technol. (2007) 377–385
3. Cao, M.D., Gao, X.: Combining contents and citations for scientific document classification. In: Australian Conference on Artificial Intelligence. (2005) 143–152
4. Zhang, M., Gao, X., Cao, M.D., Ma, Y.: Modelling citation networks for improving scientific paper classification performance. In: Proceedings of the 9th Pacific Rim international conference on Artificial intelligence. PRICAI'06, Berlin, Heidelberg, Springer-Verlag (2006) 413–422
5. Nomoto, T., Matsumoto, Y.: Exploiting text structure for topic identification. In: Proceedings of the 4th Workshop on Very Large Corpora. (1996) 101–112
6. Larkey, L.S.: A patent search and classification system. In: Proceedings of the fourth ACM conference on Digital libraries. DL '99, New York, NY, USA, ACM (1999) 179–187
7. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In Maimon, O., Rokach, L., eds.: Data Mining and Knowledge Discovery Handbook. Springer US (2010) 667–685
8. Zhang, M.L., Zhou, Z.H.: Ml-knn: A lazy learning approach to multi-label learning. Pattern Recognition **40**(7) (2007) 2038 – 2048
9. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., Vlahavas, I.: Mulan: A java library for multi-label learning. Journal of Machine Learning Research **12** (2011) 2411–2414
10. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27 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`.