Text Classification of Technical Papers Using Text Segmentation

Nguyen Hai Thien (1010220)

School of Information Science,
Japan Advanced Institute of Science and Technology

August 10, 2012

Keywords: Text Classification, Multi-label Classification, Key Phrase Extraction, Text Segmentation, Supervised Learning.

When researchers look for technical papers on a search engine, only papers including user's keywords will be retrieved. Some papers are not so relevant to the research topics that users want to know. Therefore, it will be helpful if the research topics of technical papers are automatically identified. This thesis aims to help the survey of researches. The accurate classification of technical paper is the first crucial step for an intelligent search. On the other hand, the degree of generality of categories for paper classification should be considered. The fine grained categories are more useful but difficult to automatically classify papers. In contrast, the coarse grained categories are easy for classification, but the meaning of the category is too broad, i.e. the category subsumes many research topics. In our approach, research topics in Natural Language Processing (NLP) are used as coarse grained categories. In addition to text classification of research topics, we try to identify fine grained topics of papers by extracting key phrases in that papers.

In this thesis, there are two tasks to be considered: multi-label classification and subtopic key phrase extraction. The former is to design an effective model which determines one or more categories of a given technical paper in NLP. The categories are the research topics in NLP, such as syntactic parsing, semantic analysis, machine translation and so on. To

improve the performance, text segmentation will be considered to use only some important parts of papers such as Title, Abstract, Introduction and Conclusion. On the other hand, the goal of the latter task is to extract key phrases as subtopics of papers in order to exploit the specific subtopics of a paper. For example, a paper belonging to Machine Translation category can have subtopics such as Statistical Machine Translation, Alignment and so on. Text segmentation is also considered and used in the second task.

We collect technical papers in proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) in 2000-2011. Session titles in the conference programs are useful information to determine the categories (research fields) of papers. Therefore, we manually construct a set of rules to map a session title to categories. These mapping rules are just used as a hint on choosing categories. Finally, we manually check if the categories determined by the rules are correct, then revise categories if necessary. In this way, a collection of papers associated with their correct categories is constructed. It is used to develop and evaluate the proposed method. The total number of papers and categories in the paper collection is 1,972 and 38, respectively.

As preprocessing of text classification and key phrase extraction, the following text segments in the paper are identified: Title, Author Information, Abstract, Introduction, Conclusion and Reference. Title is gotten from the database of papers, which is always correct. Other segments are identified by keywords in the papers, so the accuracy of identification of these sections is not 100%.

The proposed method for text classification of technical papers is as follows. First, words in the paper are used as features. Words in Author Information, Reference, stop words and numbers are removed. Then, the content words are lemmatized by the tool of Stanford CoreNLP. In addition to bag-of-words features, we propose new two types of features derived from the title of the paper. 'Title Bi-Gram' is defined as bi-gram in noun phrases in the title. 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 nouns in a head noun phrase of the title. The other is nouns in prepositional phrases.

Next, we propose feature selection based on text segmentation. Consid-

ering a general structure of technical paper, the title, abstract, introduction and conclusion may explain the research topics of papers. While words in other sections may not be useful for classification of research topics. In this research, the following four feature sets are considered. 'All' feature set is a set of all of the words in the papers. 'TAIC' feature set is a set of only words in Title, Abstract, Introduction and Conclusion. 'TAIC + Title Bi-Gram' means bag-of-words features of TAIC and Title Bi-Gram feature, while 'TAIC + Title SigNoun' means TAIC and Title SigNoun feature.

Feature Weighting is a method assigning appropriate weights to the features in order to reflect how important features are in documents. In this research, we conducted experiments with two traditional feature weighting: Binary Weighting and TF-IDF. Term weight in Binary Weighting is 0 or 1. 1 means the term is present in the document and 0 otherwise. TF-IDF determines the weight of the feature as the product of Term Frequency and Inverse Document Frequency.

For multi-label classification, we choose ML-kNN and Binary Approach. ML-kNN which is derived from the traditional k-Nearest Neighbor algorithm is a multi-label lazy learning approach. While Binary Approach learns $|C|$ binary classifiers, each judges if a paper is categorized as each label in C (C stands for a set of categories). Furthermore, based on the structure of papers, we propose a new model 'Back-off model'. In this model, the four classification systems S_i with different feature selection algorithms are used. S_1, S_2, S_3 and S_4 are implemented by Binary Approach using features in only Title, Title+Abstract, Title+Abstract+Introduction and Title+Abstract+Introduction+Conclusion, respectively. We suppose that the order of these 4 systems may agree with the order of precision, that is, S_1 may achieve the highest precision while S_4 the lowest. In Back-off model, S_i is sequentially applied as follows: for a given paper, if S_i can find categories whose scores are high enough, choose the categories, otherwise S_{i+1} is applied. If no system can determine the category, a few categories where their scores are greater than a threshold or one category with the highest score are chosen.

Our proposed methods are evaluated by 10-fold cross validation on our paper collection. In both ML-kNN and Binary Approach, feature selection

based on text segmentation is effective in TF-IDF feature weighting, but not in Binary Weighting. However, the results in TF-IDF are better than those in Binary Weighting. Even though text segmentation is not effective in Binary Weighting, we can still conclude that it is effective in our models. Moreover, the results indicate that combining two features Title Bi-Gram and Title SigNoun improves the performance. Among three models, ML-kNN has lowest performance on all metrics than the others, while Back-off Model is the best. Back-off Model with the best parameters achieves 58.21% Exact Match Ratio and 69.78% F-measure.

Next, we will describe our method of the second task, subtopic key phrase extraction. First, acronyms will be replaced by their definitions in the papers. Then, all of words in papers are part-of-speech tagged by running Stanford CoreNLP. We investigate two models based on graph-based ranking: TextRank and SingleRank. The important words representing the topics of papers are usually adjectives or nouns. Therefore, only adjectives or nouns in Title, Abstract, Introduction and Conclusion are inputted in these models. Finally, we propose post processing to filter out overlapped or incorrect key phrases from ones obtained by TextRank or SingleRank. Key phrases are extracted from 50 papers in our collection and they are manually checked if they are appropriate as subtopics. Precision is 79% in SingleRank, which is much better than 66% in TextRank. However, with only Precision, it is not enough to say which model is better. Because we don't have the gold key phrases, we cannot evaluate Recall for these models. This will become future work of this thesis.