Classification of MEDLINE Abstracts

Katsutoshi Ibushi k-ibushi@is.s.u-tokyo.ac.jp
Nigel Collier nigel@is.s.u-tokyo.ac.jp
Jun'ichi Tsujii tsujii@is.s.u-tokyo.ac.jp
Department of Information Science, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan

1 Introduction

This paper provides the preliminary result in our experiments to automatically assign MeSH terms to MEDLINE abstracts. Every year about 100,000 documents are added to MEDLINE, index terms are assigned by hand to each document from a controlled vocabulary called MeSH. This is necessarily time consuming and may lead to inconsistent indexing due to the large size of MeSH. Our purpose is to explore the feasibility of automating this indexing. To achieve the purpose, we apply two documents classification methods, based on SVMV [1] and AdaBoost [4], which show good results in classification of news corpora and analyze their results.

We assumed a class consists of the abstracts which have the same MeSH term. Although MeSH terms have a hierarchical structure, each class is regarded to be independent. We used MeSH terms previously assigned by specialists as answer and compared the answer with the assigned MeSH term by application of SMVM and AdaBoost.

2 Classification Methods

2.1 Single random Variable with Multiple Values

SVMV is a probabilistic model designed for text classification and has an advantage that within-document term frequencies are made use of. In this model, a document is considered to be a set of its constituting terms. This model characterizes a document as a random sampling of a term in the document. For example, "$P(T = t_i|d)$" shows the probability that a randomly selected term from a document d is t_i.

The probability that a document d belongs to a class c is calculated as:

$$P(c|d) = P(c) \sum_{t_i} \frac{P(T = t_i|c)P(T = t_i|d)}{P(T = t_i)}$$

2.2 Text Filtering with AdaBoost

AdaBoost is a family of boosting algorithms. The main idea of boosting is to generate many, relatively weak classification rules and to combine these into a single highly accurate classification rule.

In Robert's application [3], a term t_i which supports one of the below hypothesizes is sought in each step.

- If a document d contains a term t_i then the document d belongs to the class c.
- If a document d doesn't contain a term t_i then the document d doesn't belong to the class c.

3 Experiment and Discussions

We used the Ohsumed [2] corpus of MEDLINE abstracts for experiment. There are 10,477 documents, which contain title, abstract and MeSH terms, in Ohsumed.91, a sub set of Ohsumed. We used 9,977 documents as a training set and 500 documents as a test set.

The overall results for applying the two methods are shown in Table 1. Overall the results show that MEDLINE is more difficult to classify than the news corpora used in the previously cited studies. Considering the outputs of each classification method, we could see some common cases of errors.

Table 1: Accuracy of Assignment

Method	*Precision*	*Recall*	*F-value*[2]
SVMV	0.32	0.47	0.38
AdaBoost	0.44	0.42	0.43
Combination	0.48	0.49	0.48

- Low frequent MeSH terms in training set were not assigned to documents. Especially in SVMV, this tendency was conspicuous.
- MeSH terms which appeared in a document were assigned by automatic indexing, but the MeSH term wasn't assigned by specialists.
- More general MeSH term or specific one than correct MeSH term was assigned.

To solve the first case, we combined SVMV and AdaBoost. We assigned low frequency MeSH terms by AdaBoost and high frequency ones by SVMV. The second case may indicate that the difference between automatic indexing and specialists is not always a clear error. We should consider why and when specialists ignore MeSH terms which appear in the abstract or document. The third case shows the need to consider the hierarchical structure of MeSH. This will be done in our future work.

Acknowledgments

We would like to thank members of GENIA project for their valuable comments and encouragement.

References

[1] Iwayama, M. and Tokunaga, T., A probabilistic model for text categorization: based on a single random variable with multiple values, *Proc. ANLP 1994*, 162–167, 1994.

[2] OHSUMED Test Collection, ftp://medir.ohsu.edu/pub/ohsumed/readme.

[3] Schapire, R.E., Singer, Y., and Singhal, A., Boosting and Rocchio applied to text filtering *SIGIR 1998*, 215–223, 1998.

[4] Freund, Y. and Schapire, R.E., A decision-theoretic generalization of on-line learning and an application to boosting. *J. Computer and System Sciences*, 55(1):119–139, 1997.

[2] $F\text{-}value = \frac{2 \times Recall \times Precision}{Recall + Precision}$. *F-value* is used for evaluating information retrieval system by combining *recall* and *precision*.