TEXT AUTHORSHIP DETECTION USING DECISION TREES AND ASSOCIATION RULES OVER N-GRAM

Hiroshi Sugimura, Yuta Taniguchi, Ryosuke Saga and Kazunori Matsumoto
Course of Information and Computer Sciences, Graduate School of Kanagawa Institute of Technology
1030 Shimo-ogino, Atsugi-shi, Kanagawa 243-0292, JAPAN

ABSTRACT

This paper shows methods for detecting Japanese texts authorship using decision trees and association rules, both of which are constructed over N-grams. Authorship detection of Japanese texts often requires additional grammatical information that is obtained from the morphological analysis. Thus the performance of a morphological tool used in the analysis influences the entire ability of the detection. To avoid this problem, we in this study use a set of N-gram that are sequences of N letters simply cut out from texts. In the first part of the study, we investigate a use of decision tree learning over N-gram. Since the size of possible N-gram becomes combinatory large, we use forward and backward selection approach to obtain the effective subset of N-gram by which the expected prediction accuracy of the decision tree becomes optimal. In the latter part, we state how association rules are used for detections, and compare the results of both approaches.

KEYWORDS

Authorship detection, N-gram, decision tree, association rule

1. INTRODUCTION

Many studies provide methods for text authorship detection problem [3,4,5,6,7,8,9,10]. Distribution of words, length of words or sentences, patterns of punctuations, etc. becomes effective information for solving this problem [5]. In the case of English or other European languages, sentences consists of words that are clearly separated by spaces or punctuation symbols. Thus it is relatively easy to collect and use such measurable information. On the other hand, in the case of Japanese or Chinese, the concept of words and punctuations are completely different from the European's. Extracting words from Japanese texts requires an extra grammatical investigation, which is called the morphological analysis. A morphological tool finds further information, including a separation into words, by using a predefined dictionary whose applicability directly influences the analysis performance. Another idea that escape from the use of a dictionary is based on a distribution of N-gram, which is a sequences of N letters simply cut out from texts, so that they are obtainable without dictionary nor grammatical analysis. As a disadvantage of N-gram, the size of entire possibilities grows combinatory in accordance with both the size of N and the number of different letters in the language. Since Japanese uses more than thousands of kanji and others letters, the distribution of N-gram disperses over the huge possibilities. Then methods based on N-gram are inevitable faces the zero probability problem. For this reason, it is pointed out [6] Japanese texts authorship detection requires at least 20,000 letters to achieve high accuracy. We thus use that length of texts in this study. In Figure.1, we show distributions of N-gram changing the values of N, they are clearly obey Zipf's law [1], where only a few topmost words occurs with relatively high frequency. The top frequent bi-gram, however, occurs with about 2.1% probability in texts. As the value of N increases, frequencies of N-gram decreases in average, and the frequencies of N-gram becomes very small except the topmost ones.

Several functions, which measure the similarity of N-gram distribution, are proposed in [4,6]. The functions mathematically compute the degree of the similarity, so that a text is attributed to the author whose average N-gram distribution has highest similarity of the target text. As a main disadvantage of this method, we cannot obtain knowledge that characterize the personality of authors. A distribution of N-gram works as a tool for implicit and automatic detection of authorship, but is hard to use for a tool of explicit understanding

the process of a detection. We thus use the decision tree learning and the mining of association rules, both of them are effective tools to extract comprehendible knowledge from large data. In the following chapters, we briefly explain the study and the experiments.

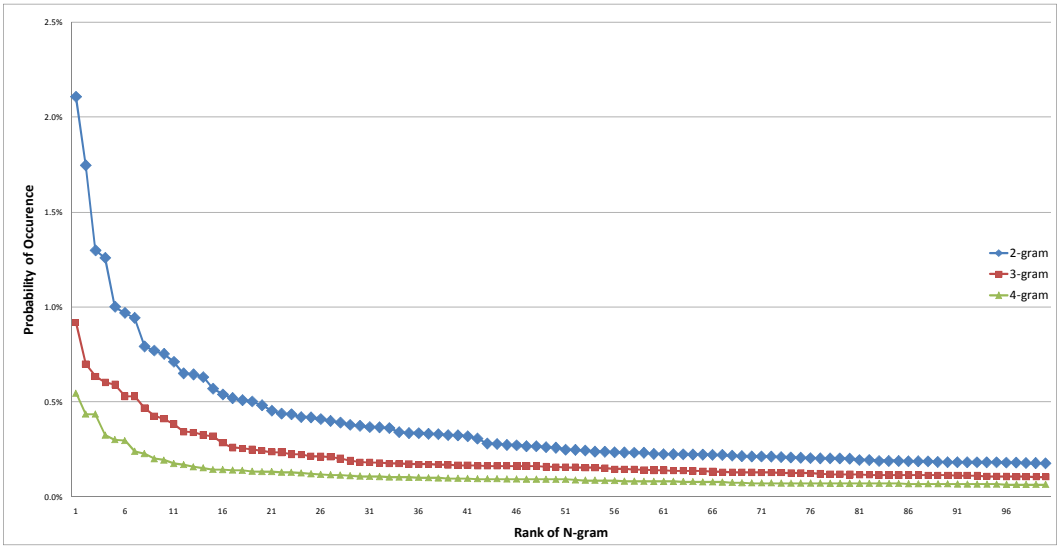

Figure 1. Distribution of N-gram

2. USING DECISION TREES

A decision tree is a powerful tool to classify objects into predefined classes. Many studies provide techniques to construct an optimal decision tree for a given set of training data. A training data is expressed as a tuple of attribute values, one of which is distinguished as a class value. Once a decision tree is obtained for the training set, the tree is also used to predict a class for data with unknown class value. The quality of the decision tree is evaluated by the accuracy of the predictions, it usually estimated by the cross-validation. The quality of a decision tree strongly depends on a selection of attributes, so that this is called a feature selection problem. There exist various kinds of studies[2] to identify a best subset of attributes among candidates. They are summarized into two types; the filter and the wrapper approaches. In the filter approach, the best subset is identified by using evaluation functions over the candidates, while the later approach actually builds decision trees and evaluates them. We investigate the applicability of the both approach in the case of authorship detection with N-grams.

The forward selection method [3] starts with a state having no attribute, and stepwisely add the attribute that has maximum effect in building a decision tree. As we stated above, the numbers of N-gram becomes huge, even in the case of N=2 we have about 70,000 different N-gram for texts of an author. In Figure 2, we show an experimental result of the forward selection. The estimated accuracies stabilize in any cases after applying top 800 N-gram. We thus apply the selection within this limit. Finally, by using only a very small numbers of N-gram we obtain a decision tree with high accuracies, the result is summarized in Table 1. The backward elimination method achieve similar result.

Table 1. Result of Selection

	2-gram	3-gram	4-gram
Max. accuracy	86.7%	80.0%	90.0%
number of N-gram	5	5	6

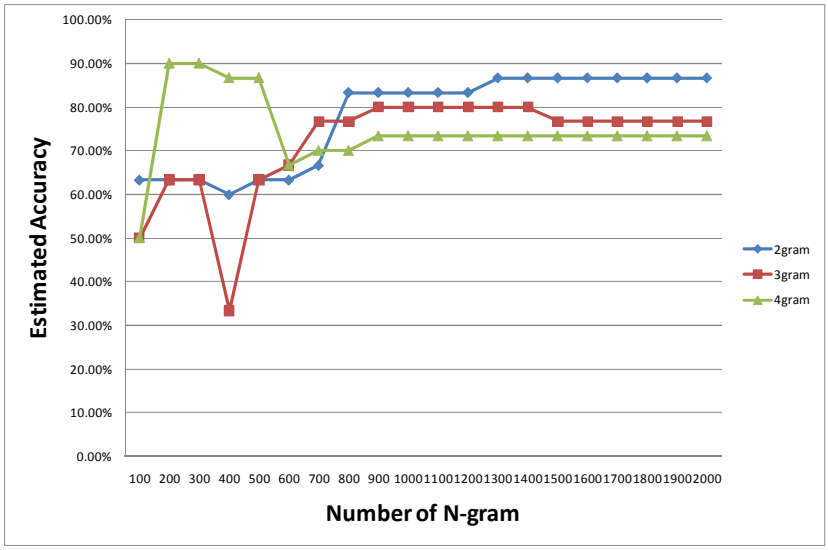

Figure 2. Result of Forward N-gram Selection

3. ASSOCIATION RULES AS AUTHORSHIP CLASSIFIER

Datamining of association rules is a well studied area in recent knowledge engineering [2]. An association rule, extracted from training texts, is an expression $X \rightarrow Y$ with support s, and confidence c, where X and Y are a N-gram or a set of N-grams sharing no common element. The support s is defined as the probability that a sentence containing both X and Y appears in the texts. The confidence c is defined as the probability that a sentence containing X also includes Y in the texts. For a given minimal support ms and confidence mc, datamining of association rule outputs all association rules having support $s' \geq ms$, and confidence $c' \geq mc$. It is typically effective to use a combination of a small minimal support and a high minimal confidence. As the value of minimal support decreases, the number of association rules grows exponentially in most of cases. We show the number of extracted association rules in Table 2, the number of rules in fact becomes huge for the minimal support=0.01. The minimal confidence is commonly set to 0.9. Note that this experiment is carried out by using 10 authors with 3 literary works, and the numbers average the values in the table.

Table 1. Average Number of Extracted Association Rules

minimal support	2-gram	3-gram	4-gram
0.01	134,685.3	14,465.8	1829.8
0.05	41.5	7.2	1.2
0.1	2.4	0.5	0-

Let XA be a set of association rules obtained from the entire set of texts of author A. We apply rules in the XA to texts of the same and different authors. Figure 3 shows one of the experimental results, the author is famous modern Japanese writer Sakaguchi Ango, and the minimal support and confidence are respectively set to 0.05 and 0.9. The circles in the figure shows the cases applying to texts of the same author, while the dashes and the crosses show the cases of different authors. Note that the XA include about 90 rules, and the result of applying a rule is show in vertically.

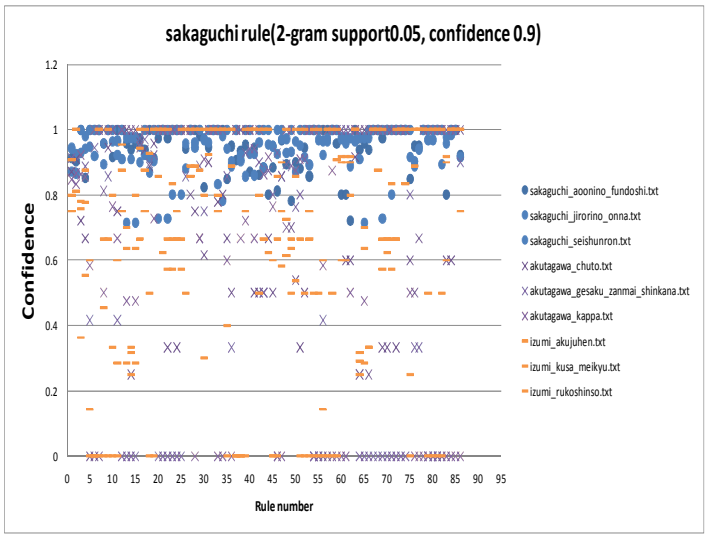

Figure 3. Applying Association Rules

4. CONCLUSIONS

This paper shows experimental results on the performance of Japanese text authorship detection ability using decision trees and association rules, which are constructed over N-gram. Both approaches obtain understandable sets of detection rules by which we can detect authorship with high accuracy. In this study, we use texts including 20,000 letters. It is obvious that the applicability of a detection method spreads widely if it can run on small size of texts. From this point of view, the current limitation of the text size need to be improved. A combinational approach, which uses both of a decision tree and association rules, is applicable for that purpose. Further experimental results with mathematical analysis will be appeared in a full paper.

REFERENCES

[1] R.Baeza-Yates, and B.Ribeiro-Neto, 1999. *Modern Information Retrieval*. Addison Wesley.

[2] J.Han, and M.Kamber, 2001. *Data Mining, Concepts and Techniques*. Morgan Kaufmann.

[3] P.Juola, and H. Baayen, 2003. *A Controlled-Corpus Experiment in Authorship Identification by Cross-entropy*. Literaryand Linguistic Computing.

[4] V. Keselj, F. Peng, N. Cercone, and C. Thomas, 2003. N-gram-based Author Profiles for Authorship Attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics (PACLING'03)*. Canada.

[5] H. Love, 2001. *Attributing Authorship: An Introduction*. Cambridge University Press.

[6] T. Matsubara, and Y. Kanada, 2000. Extraction of Authors' Characteristics from Japanese Modern Sentences via N-gram Distribution, *Lecture Notes in Computer Science,* Vol.1967. Springer.

[7] F. Peng, and D. Schuurmans, 2003. Combining Naive Bayes and N-gram Language Models for Text Classification. In *Proceedings of the 25th European conference on IR research (ECIR'03)*. Italy.

[8] F. Sebastiani, 2002. Machine Learning in Automated Text Categorisation. *ACM Computing Surveys,* Vol. 34, No. 1.

[9] E. Stamatatos, N. Fakotakis, and et.al., 1999. Automatic authorship attribution. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. Norway.

[10] Y. Zhao, and J. Zobel, 2005. *Effective and Scalable Authorship Attribution Using Function Words,* RMIT University,Melbourne, Australia.