

Rule-based Word Clustering for Text Classification

Hui Han¹ Eren Manavoglu¹ C. Lee Giles^{1,2} Hongyuan Zha¹

¹Department of Computer Science and Engineering

²The School of Information Sciences and Technology

The Pennsylvania State University University Park, PA, 16802

{hhan,zha,manavogl}@cse.psu.edu giles@ist.psu.edu

ABSTRACT

This paper introduces a rule-based, context-dependent word clustering method, with the rules derived from various domain databases and the word text orthographic properties. Besides significant dimensionality reduction, our experiments show that such rule-based word clustering improves by 8% the overall accuracy of extracting bibliographic fields from references, and by 18.32% on average the class-specific performance on the line classification of document headers.

Categories and Subject Descriptors

H.4 [Information Systems]: Information Search and Retrieval – Clustering

General Terms

Algorithms

Keywords

Word Clustering, Feature Dimensionality Reduction

1. INTRODUCTION

Word clustering techniques have been successfully used for text classification, with two main advantages: dimension reduction and improving classification accuracy [1, 2, 3, 6].

Information theoretic approach to word clustering considers the word distributions over categories to determine similar words. Such methods need labeled training data. Instead, we introduce a rule-based, context-dependent word clustering method, with the rules extracted from various domain databases and text orthographic properties of words. Rule-based method relies on prior knowledge embedded in domain databases.

2. RULE-BASED CLUSTERING METHOD

We have two types of rules to cluster words. First rule considers the words' text orthographic properties. For example, email addresses and numbers are converted into special tokens +email+ and +num+, using regular expressions[5].

Second rule clusters words based on their membership in domain databases. “Domain” here corresponds to the class in classification tasks. We define two types of Domain databases. **External Domain Databases** are collected

from public sources: standard on-line dictionary of Linux system, Bob Baldwin's name words collection, USA state names and Canada province names, USA city names, country names from the World Fact Book. **Constructed Domain Databases** are representative words selected from the training samples of corresponding classes, using expected entropy loss. Words belonging to the same domain database are represented by a common cluster-specific token.

Depending on the context, we use different clustering methods for the words appearing in multiple domain databases. First option uses a “specific-to-general” order. For example, if a word appears in both name word database, and standard linux word dictionary, we convert this word into the name class-specific token “+mayname+”. Second we can encode this multi-class attribute of the word using a N -digit code. For example, “1001” means a word in both domain databases “name” and “location”, but not in domain databases “affiliation” or “title”. Third, each of these words forms an independent cluster.

3. EXPERIMENTS AND RESULTS

We experiment on the following two tasks and compare the performance before and after word clustering.

3.1 Bibliographic field extraction

We use Hidden Markov Model and the dataset provided by CMU [5] to extract 13 bibliographic fields from references shown in Figure 1 and Table 1. We split the dataset into 250 training, 125 validation and 125 test samples.

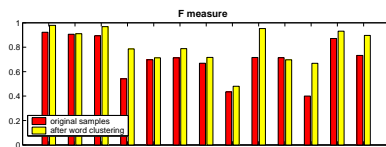


Figure 1: The F measure of reference word tagging before(left bar) and after(right bar) word clustering.

Word clustering reduces the feature dimensionality from 2300 words to 300 words, and increases the overall accuracy of bibliographic field extraction from 82.9% to 90.8%.

Table 1 shows that the class-specific extraction performance after rule-based word clustering is improved in general, especially for classes “editor”, “page”, “tech” and “volume”. A reason is that the model learned from the clustered training data is more general, less prone to noise and naturally smoothed. Calculating the emission probabilities of

Table 1: F measure of bibliographic field words tagging before and after rule-based word clustering.

data	author	title	date	editor	institution	journal	location	note	pages	publisher	tech	booktitle	volume
before	92.2%	90.6%	89.4%	54.1%	69.8%	71.3%	66.9%	43.5%	71.6%	71.4%	40.0%	87.1%	73.3%
after	97.9%	91.0%	96.8%	78.6%	71.3%	78.8%	71.6%	48.0%	95.3%	69.7%	66.8%	93.1%	89.6%

cluster-specific features rather than the original words increases probability of emitting each word in those domain databases, and even the words not seen in the training data. However, domain databases without sufficient class specific information may degrade the effect of word clustering. For example, there’s no external domain database for the class “publisher”, and the internally constructed one is far from being representative since it contains mostly names and general dictionary words. False prediction of name words and general dictionary words as “publisher” decreases the classification performance in this class compared with raw data.

3.2 Line classification of document headers

The dataset we use is provided by CMU [5] and contains labeled headers of computer science research papers, with 500 headers for training and 435 headers for testing. Each word is labeled with one of the 15 classes as shown in Table 2, and each line is marked by +L+.

Our experiment is based on SVMlight [4]. We use the Gaussian kernel of SVM [7] with an extension to multi-class classifiers using “One class versus all others” approach, to classify each line into one or more classes.

Data clustering in this experiment reduces the feature dimensionality from 8256 to 1113, and speeds up the cpu run times for classification. Table 2 also shows that the classification performance (F measure) increases after word clustering. On average, each class improves by 18.32%.

Table 2: Line classification performance of document headers before and after rule-based word clustering. F-F measure and T-classification runtime (without IO) in cpu-seconds.

class	F(Before)	T(after)	F(after)	T(after)
title	74.91%	10.76	92.45%	4.36
author	64.16%	14.32	92.70%	2.59
affiliation	89.46%	15.21	91.77%	5.11
address	80.90%	19.96	92.55%	4.13
note	66.02%	22.59	67.19%	8.52
email	26.79%	14.44	97.69%	2.30
date	88.31%	3.39	93.21%	0.85
abstract	96.94%	19.98	97.59%	7.16
introduction	95.19%	3.07	96.94%	0.78
phone	59.02%	13.33	88.31%	1.25
keyword	59.60%	10.56	71.29%	2.15
web	32.26%	8.35	96.00%	2.17
degree	65.63%	11.11	67.68%	2.79
pubnum	53.34%	6.29	81.25%	1.08
page	100.00%	5.24	100.00%	1.65

4. CONCLUSION AND DISCUSSION

Our rule-based word clustering method shows significant dimensionality reduction, and improvement on bibliographic field extraction and line classification of document headers.

In addition, rule-based method has computational advantages since we only need to search the domain databases and use simple rules to cluster words according to word text orthographic properties.

The domain databases and proper use of text orthographic properties are important to effective word clustering. Inappropriate or small domain databases may introduce bias in word clustering. Various feature selection and word clustering methods such as distributional word clustering may help construct domain databases from training samples.

The model structure or algorithms used in the classification tasks may constrain the effect of word clustering. In bibliographic field extraction with HMMs task, the HMM transition structure learned from training data is fixed, regardless of word clustering. If a transition from the class “author” to the class “title” is not present in the training samples, rule-based word clustering will not help predicting this case for test samples.

Taking single words as observations may also restrict the exploitation of the domain databases. For example, since most of the country names contain multiple words, a single word is unable to match most country names, when using first-order HMM model with unigram emission distributions.

5. ACKNOWLEDGMENTS

We acknowledge Andrew McCallum for providing the HMM code and Cheng Li for useful suggestions through the experiments. We would like to acknowledge partial support from NSF grant NSDL 0121679.

6. REFERENCES

- [1] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of SIGIR-98*, pages 96–103, 1998.
- [2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
- [3] I. Dhillon, S. Manella, and R. Kumar. A divisive information-theoretic feature clustering for text classification. *to appear in Machine Learning Research (JMLR)*, 2002.
- [4] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98*.
- [5] K. Seymore, A. McCallum, and R. Rosenfeld. Learning hidden Markov model structure for information extraction. In *AAAI 99 Workshop on Machine Learning for Information Extraction*, 1999.
- [6] N. Slonim and N. Tishby. The power of word clusters for text classification. In *ECIR*, 2001.
- [7] V. Vapnik. *Statistical Learning Theory*. 1998.