

Comparative Literary Style Mining between Native and Non-native English Writers

Bei Yu

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel St., Champaign, IL 61820. Email: beiyu@uiuc.edu.

Introduction

Understanding the commonalities and differences between contrasting groups is a fundamental task in data analysis. Previous research has adapted rule-discovery techniques for contrasting data sets in social science (Bay & Pazzani, 2001) and commerce (Webb et al., 2003) domains. In the text mining area, Marx et al. (2002) and Zhai et al. (2004) have explored topicality-concerned comparative mining with clustering techniques.

Style information, usually considered as orthogonal to the topics of a document, provides a complementary channel to understand the document and thus mine useful information from it. Traditional stylistic analysis has long been used for some humanities computing tasks such as authorship attribution (Mosteller & Wallace, 1964). In recent years, computational stylistics, combining traditional stylometry, statistics and machine learning techniques, has even been used in many other applications, such as text genre detection (Biber, 1988, Karlgren & Cutting, 1994, Kessler et al., 1997), email classification (de Vel, 2000), and opinion (sentiment) categorization (Dave et al., 2003).

In this poster we describe our research of applying machine learning and computational stylistics techniques to the problem of comparative literary style mining between native and non-native English writers in academic writing. The research result is expected to contribute to style analysis and second language writing education.

We try to locate and compare the collective style characteristics between the writers from different language backgrounds based on the analysis of two data sets: a collection of electronic theses and dissertations and a collection of research papers. By transforming this problem into a categorization task, we extract various kinds of style features and feed them into a naive Bayesian text classification algorithm. The prediction accuracy for each feature set is used as an index to rank its discrimination power to differentiate the two author groups. The classification performance also provides evidence of the existence of collective style characteristics of the writers from the same language background.

The preliminary experiment results show that the two author groups are highly separable using a bi-gram common word sequences feature set. More interesting results are found in further feature analysis.

Data Preparation

Data selection and preparation plays an important role in this project. The following criteria are followed to collect appropriate contrasting data sets:

1. The authors' language background should be identifiable;
2. The authors should have comparable English proficiency;
3. The documents should reflect the collective style of the category they belong to;
4. The collected document sets should be domain independent.

Based on the above criteria, we collect two datasets. The first has 40 selected electronic theses and dissertations (ETD) with 20 contributed by Chinese students and 20 by American students from the ETD database at Virginia Tech. The second has 40 selected research articles downloaded from Microsoft Research (MSR) with 20 contributed by Chinese researchers in Beijing, China and 20 by their British colleagues in Cambridge, UK. The ETD documents are long and each strictly attributed to one author, while the MSR documents are much shorter and many are co-authored. We restrict the co-authors to the same language background.

Research Method

Style Feature Extraction

We use a set of independent features to represent the writing styles. Three types of features are used in this study. They are common word sequences (CWS), parts-of-speech (POS), and traditional style markers (TSM). We use three general style markers: average word length, average sentence length, and type/token ratio. Common words consist of function words and some content words without specific meanings (Koppel et al., 2002). It is well known that authors can not easily "control" the use of common words and that they do not carry specific domain or genre information, which ensures them as good style indicators with the least interference of other environmental variables.

For CWS, we use uni-gram, bi-gram, tri-gram and longest CWS. As an example, given a piece of text "*SW4 is among the few known wireless system tools for in-building network design.*", the following n-gram CWS features are extracted:

1. uni-gram: is, among, the, few, known, for;
2. bi-gram: is among, among the, the few, few known;

3. tri-gram: is among the, among the few, the few known;
4. longest: is among the few known, for.

The sizes of the feature sets are shown in Table 1.

Table 1: Feature set sizes.

	CWS				POS	TSM
	Uni-gram	Bi-gram	Tri-gram	Longest		
MSR	929	16578	25825	31015	42	3
ETD	467	5116	5922	7936	42	3

The Naive Bayes Algorithms

The Naive Bayes (NB) classification methods are simple, scalable to large feature sets and among the most effective algorithms to classify text documents. When applying this method, we have to choose from two kinds of NB algorithms for different feature sets. One is regular NB with the requirement of binning numeric features, and the other is text NB without binning (Mitchell, 1997). The document representations are slightly different for the two methods.

Text NB follows a kind of generative model. It assumes that the documents from different classes are generated from different word probability distributions. The process of training the classifier is to estimate these probabilities. To predict the class of a document is to determine which distribution the document is generated from. Given a text document, each word position is defined as an attribute and its value is the word that appeared in that position. The number of probabilities to be estimated is $(number_of_classes * vocabulary_size * document_length)$.

By assuming that the attributes are independent and identically distributed (bag-of-words), we greatly reduce the number of estimated probabilities to $(number_of_classes * vocabulary_size)$. For each position j , we have $P(a_j=w_k|v_j) = P(w_k|v_j)$, so the word position and document length no longer matter. In other words, the trained classifier should remain the same if we scramble the word order in a document or concatenate all the document examples in each class into one single example. In this sense, the size of each document does not affect.

For regular NB, each numeric feature is discretized into several bins before counting the frequencies to estimate the probabilities. In consequence the version of binning greatly affects the algorithm performance. In our case we uniformly bin each feature to 2 bins, so for a feature set of size n , the number of probabilities to be estimated is $2 * n * number_of_classes$. If the features are unique words, this method is sensitive to document length. So normalized frequencies have to be used for this method.

We apply text NB method to the CWS feature sets and regular NB to the POS and the TSM feature sets. Laplace (add-one) smoothing is used for both methods.

We also test the regular NB on normalized CWS features. The prediction accuracy for uni-gram is 90%, but when the gram length increases, the performance quickly goes down to 50% or even lower. Our guess for the failure reason is that the zero probability problem deteriorates with the increase of gram length because the feature set size increases and the feature table becomes sparse. Notice that in this case the number of probabilities to be estimated is twice the number in text NB. An evidence to support our guess is that the classification accuracy gets worse when each feature is discretized to more bins. Improved smoothing techniques may alleviate the problem.

Preliminary Results

Classification Results for ETD Data Set (see Table 2).

Table 2: Confusion matrices for 10-fold cross validation for Naive Bayes classification on ETD data set.

	C	E	
C	8	12	40%
E	0	20	100%
			70%
Uni-gram CWS			
	C	E	
C	18	2	90%
E	2	18	90%
			90%
Bi-gram CWS			
	C	E	
C	20	0	100%
E	8	12	60%
			80%
Tri-gram CWS			
	C	E	
C	13	7	65%
E	6	14	70%
			67.5%
POS			
	C	E	
C	13	7	65%
E	9	11	55%
			60%
TSM			

Classification Results for MSR Data Set (see Table 3).

Table 3: Confusion matrices for 10-fold cross validation for Naive Bayes classification on MSR data set.

	C	E	
C	1	19	95%
E	3	17	85%
			90%
Uni-gram CWS			
	C	E	
C	20	0	100%
E	3	17	85%
			92.5%
Bi-gram CWS			
	C	E	
C	20	0	100%
E	17	3	15%
			57.5%
Longest CWS			
	C	E	
C	20	0	100%
E	20	0	0%
			50%
Tri-gram CWS			

	C	E	
C	16	4	80%
E	8	12	60%
			70%

POS

	C	E	
C	12	8	60%
E	4	16	80%
			70%

TSM

Comparing the Experiment Results for ETD and MSR Data Sets (see Figure 1).

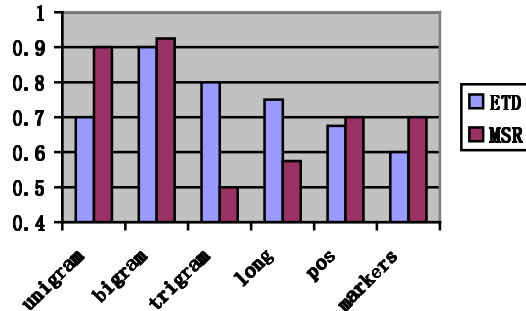


Figure 1: Classification result comparison

Interesting Findings from Feature Analysis

We did simple n-gram CWS feature comparison and discovered some interesting phenomena with some as well-known facts in second language education. To name a few, English authors tend to use longer common word sequences, and much more past and perfect tense auxiliary words than Chinese authors.

Conclusions

The experiment results in Tables 2 and 3 show that bi-gram common word sequences (CWS) are the best feature set to discriminate the two author groups. The accuracy from 10-fold cross validation on both data sets reaches 90% or higher. We need further investigation why the experiment results on both data sets do not follow the same trend as shown in Figure 1. For example, the other n-gram CWS feature sets reached fair accuracy (70%~80%) for the ETD data but failed in the MSR case. One thing to note is that the errors in extracting text from ps or pdf files hurt the accuracy of the Brill part-of-speech tagger we used, so the 70% accuracy from POS feature set may not reflect its discrimination power in this experiment. The three general style markers here do not compete with the bi-gram CWS feature sets.

Future Work

We have not achieved an effective strategy to rank the features according to their contrasting power. We tried the Relief F attribute evaluation (Sikonja & Kononenko, 1997) to rank the n-gram CWS, but tens of the features have tied

scores. Further feature analysis needs to be done to rank the importance of the discriminative features. Help from second language education experts is also needed to explain the observed differences.

ACKNOWLEDGEMENTS

This project is supported by the Automated Learning Group at National Center for Supercomputing Applications. I would like to thank Mr. Duane Searsmith, Dr. Dave Dubin and Prof. Linda Smith for their invaluable guidance and contributions to this project.

REFERENCES

- Bay, S.D., & Pazzani, M.J. (2001). Detecting group differences: mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3), 213-246.
- Biber, D. (1988). *Variations across speech and writing*. Cambridge.
- Dave, K., Lawrence, S., & Pennock, D.M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *WWW'03*.
- de Vel, O. (2000). Mining email authorship. *ACM SIGKDD'00 Workshop on Text Mining*.
- Karlgren, J., & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. *COLING'94*.
- Kessler, B., Nunberg, G., & Schutz, H. (1997). Automatic detection of text genre. *ACL/EACL'97*.
- Koppel, M., Argamon, S., & Shimoni, A.R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), 401-412.
- Marx, Z., Dagan, I., Buhmann, J., & Shamir, E. (2002). Coupled clustering: a method for detecting structural correspondence. *Journal of Machine Learning Research*, 3, 747-780.
- Mitchell, T. (1997). *Machine learning*. McGraw-Hill.
- Mosteller, F., & Wallace, D. (1964). *Inference and disputed authorship: the federalist*. Massachusetts: Addison-Wesley.
- Sikonja, M.R., & Kononenko, I. (1997). An adaptation of relief for attribute estimation on regression. *Proceedings of 14th International Conference on Machine Learning ICML'97*.
- Webb, G., Butler, S., & Newlands, D. (2003). On detecting differences between groups. *ACM SIGKDD'03*, 256-265.
- Zhai, C., Velivelli, A., & Yu, B. (2004). A cross-collection mixture model for comparative text mining. *ACM SIGKDD'04*.