Latin Etymologies as Features on BNC Text Categorization *

Alex Chengyu Fang[a], Wanyin Li[a], and Nancy Ide[b]

[a]Department of Chinese, Translation, and Linguistics, City University of Hong Kong,
Tat Chee Avenue, Kowloon, Hong Kong
{acfang, claireli}@cityu.edu.hk

[b]Department of Computer Science, Vassar College,
124 Raymond Avenue, Poughkeepsie, NY 12604
ide@cs.vassar.edu

Abstract. This paper presents an early experimental work on BNC Text Categorization (TC) with Latin etymologies as features, emphasis on spoken and written texts. Two aims achieved in this study: (1) to explore discriminative new linguistic features rather than lots of noise-bringing “bag-of-words” (BoW). (2) to build up a base step to represent texts in distinct types of linguistic features with different weighting scheme rather than a plain feature vectors of BoW. The experiments disclose a notable distinct distribution pattern of Latin etymologies in spoken and written BNC texts. The performance of a home-made classifier based on the probability distribution ranges of Latin etymologies reaches a precision of 72.31% and recall of 73.22% on BNC spoken texts and precision of 73.31% and recall of 69.98% on BNC written texts.

Keywords: Text Categorization, Latin Etymologies, Discriminative Features.

1 Introduction

Text Categorization (TC) is the task of classifying natural language texts into a predefined set of semantic categories (Lan *et al.*, 2006). Features selection is always a bottleneck in the tasks of TC, especially the common used BoW introduces a large features space, some of the features are redundant, some of them bring noise. The current TC studies are based on features like words/phrases frequencies (Olsson and Douglas, 2006), therefore, need (1) features selection algorithms such as Information Gain (Wang *et al.*, 2004; Lee and Lee, 2006; Olsson and Douglas, 2006; Shang *et al.*, 2007), Mutual Information (Wang *et al.*, 2004; Pei *et al.*, 2007), χ^2 (Wang *et al.*, 2004; Olsson and Douglas, 2006; Shang *et al.*, 2007), Maximum Entropy (Nigam *et al.*, 1999; B. Chen *et al.*, 2008) etc. A good review on the state-of-art feature selection techniques can be found in (Liu and Yu, 2005). However, as stated by (Mukras *et al.*, 2007; Shang *et al.*, 2007), these routine feature selection methods may fail to identify discriminatory features, particularly when they are distributed over multiple ordinal classes or especially like χ^2 (Olsson and Douglas, 2006) are known to be misled by infrequent terms; (2) features transformation technique like Term clustering (Lin and Kondadadi, 2001; Beil *et al.*, 2002), Latent Semantic Indexing (LSI) (Wu and Gunopulos, 2002; Kontostathis and Pottenger, 2006) so that the texts can be represented in features vectors; (3) because of this kind of

* This work is supported by the project “A Descriptive Multi-tiered Model for Text Type Classification Based on Syntactic Complexities (Project No. 7002190); Investigating the Etymological Composition of Lexical Use in Contemporary English: A study Based on the British Notional Corpus (Project No. 7200120) & Text Genre Classification Based on Lexico-Grammatical and Syntactic Cues (Project No. 7002387), Department of Chinese, Translation and Linguistics, City University of Hong Kong.

representation of document, usually need to employ computationally expensive learning algorithms from machine learning like Naïve Bayes Classification (Wang *et al.*, 2004; J. Chen *et al.*, 2008), Support Vector Machine Classification (Wang *et al.*, 2004; Lan *et al.*, 2006; Shang *et al.*, 2007), linear classification (T. Zhang and Oles, 2001; J. Zhang and Y. Yang, 2003), KNN (Lan *et al.*, 2006; Olsson and Douglas, 2006; Shang *et al.*, 2007), Neural Network (Yu *et al.*, 2008) etc. A thorough survey can be found in (Sebastiani, 2002). As stated by (D. Zhang and W. S. Lee, 2006), the learning algorithms even the verified most de facto SVM algorithm can be neither effective nor efficient to take all selected features straightforwardly. This study wishes to explore and verify discriminative features beyond words/phrases frequencies based on linguistic analysis and have not been reached yet up to now and limit the efficient features set as small as it can be. As stated by (Rogati and Yang, 2002), "The results we obtained using only 3% of the available features are among the best reported, including results obtained with the full feature set". In addition, this study provides a base step for the future work in which we do not want to deem classified texts as simple as a feature vectors of "bag-of-words", but as different levels of linguistic information, such as the investigated one in this study, which is the probabilities of the words having Latin-etymologies in the classified texts.

The rest of this paper is organized as follows. Section 2 describes the approach proposed. Section 3 evaluates the proposed method. And Section 4 opens a discussion and presents the outline of the future works.

2 The Method

Two lexicographical resources are used in this study. The Collins English Dictionary (CED) is a collection of total 128 different languages of etymological knowledge for contemporary English, which includes Latin, French, Greek etc. CED contains 249,331 entries which are finally found 48,593 words with etymological origins. Another resource is BNC corpus which is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of Modern British English built from 1991 to 1994. The version used in this study is the *BNC XML Edition*, released in 2007.

Recall the observation that a majority of Latinate words are normally used in more formal speech and written texts. The first experiment is constructed to verify that the probabilities of Latin etymologies are distributed in distinct ranges for BNC spoken and written texts. The principles of the experiment is to find exclusive ranges of the probability distribution in Latin etymologies for spoken and written texts, therefore, we conducted the experiments on different unified size of texts. The approach has three steps: (1) Extract <word, language-etymology> pairs like <impress, Latin> from CED as the word list (named wordlist_ety) which contains total 48,593 such kind of pairs in the final version. A language list of total 1,362 valid languages is used to eliminate the pseudo languages in the above <word, language-etymology> pairs. (2) Extract headwords from each text of a given BNC category, which will be further refined against the top-2000 stop-word-list discussed in Section 2.1. The refined headwords will be processed based on the "wordlist_ety" to summarize their language etymologies which is associated with the frequencies, total number of tokens in each text, as well as the local probabilities of sub-languages (such as New Latin, Late Latin, Medieval Latin etc) and of reduced languages (such as Latin). The sample outputs like < Late Latin, 7, 123, 0.056911> and <Medieval Latin, 19, 158 0.031646>, and the reduced one like < Latin, 62, 159, 0.018868>. (3) The above statistical values are passed to a home-made range classifier (RangeClassifier) which takes a step of 0.05 as the Latin etymology probabilities to automatically seek the best ranges based on the different evaluation schemes of precision, recall and F_Score.

2.1 Preprocessing

BNC texts are originally in uneven size from the minimum of 8KB each to the maximum of 19,738KB each, in another word, the total number of words in each text is different which brings bias when assigning a local value (probability in this study) to the examined features (Latin etymologies in this study). In order to achieve a consistent weighting scheme, such as consistent probability in investigated features against a unified text size, each BNC text is firstly transferred into even sized one counted in the number of words.

Another preprocessed work for evaluating the performance of the proposed classifier is to divide the corpus into training and testing parts. This job is done by randomly selecting 80% of texts from each category and another exclusive 20% of texts from the same category.

Finally, to filter the so called stop-word features from the texts, a filter function is applied to generate the top-2000 frequent words calculated against the whole BNC corpus.

2.2 Features of Latin Etymologies in BNC Texts

Challenging tasks in text categorization: (1) which features should be selected and (2) what type of values to be assigned to the selected features. Well discussed features include single tokens/words (Nigam *et al.*, 1999; Olsson and Douglas, 2006), keywords (Wang *et al.*, 2004; Anette, 2006), bi-grams/n-grams (Mansur *et al.*, 2006; Kanaris and Stamatatos, 2007), noun phrases (Liao *et al.*, 2003; Zhang *et al.*, 2006), and syntactic patterns (Lewis, 1992; Johannes *et al.*, 1998). Most of the above feature selection requires more engineering effort such as the parsing of the texts so that the target syntactic patterns can be identified successfully. Furthermore, the classification performance relies on the qualities of the identified features.

Selected features are assigned a numerical value to show their significance to the classification task. Summary of term weighting schemes include binary feature (BI), term frequency (TF), inversed document frequency (IDF), TF.IDF, TF.χ^2, TF.RF (relevance frequency) etc borrowed from information retrieval.

Rather than the reported features, this study examines the frequencies of the words having Latin etymologies in the running texts and so far has not been reported. Assigning a so-called local probability which is calculated using the number of words with Latin etymologies divided by the total number of tokens in the same text. Rather than a feature vector representation of texts, each text is represented by the features of the words with Latin etymologies and assigned the value of their probabilities against the total number of tokens in the text. All texts under a given category contribute a probability ranges for that category which brings a possible distinct probability ranges for different categories. The experiments show the existence of such distinct ranges for the categories of BNC spoken and written (Table 6a), but multiple overlapped ranges for the sub-categories under written (Table 6b). Table 1 and Table 2 show the distribution difference of Latin etymologies (Latin-ety) on BNC spoken-written and sub-written texts.

Table 1: Distribution Difference of Latin-ety between speech and written texts.

	Conversation	OtherSpoken	Written
Anglo-Latin	21	53	343
Late Latin	879	2,798	20,222
Latin	9,126	22,707	144,410
Medieval Latin	1,053	2,722	20,578
New Latin	541	1,530	6,906
Vulgar Latin	3	0	18
Total Latin	11,623	29,810	192,477
Total Tokens	196,594	218,745	1,778,836
Latin-ety-Density (%)	5.91	13.63	10.82

Table 2: Distribution Difference of Latin-ety on sub-written texts.

	Fiction	News	Otherpub	Unpub
Anglo-Latin	102	76	129	36
Late Latin	3,301	4,149	6,249	6,523
Latin	31,987	34,448	41,226	36,749
Medieval Latin	3,446	5,089	5,302	6,741
New Latin	1,066	1,293	1,945	2,602
Vulgar Latin	15	0	2	1
Total Latin	39,917	45,055	54,853	52,652
Total Tokens	391,105	510,905	441,599	435,227
Latin-ety-Density (%)	10.20	8.82	12.42	12.10

From Table 1, the distribution probability of Latin etymologies in the informal speech texts (Conversation), which is 5.91%, is distinct lower than in the written texts, which is 10.82%, as well as in the formal speech texts (OtherSpoken), which is 13.63%.

2.3 RangeClassifier

The classifiers in the most reported literals on text categorization come from Machine Learning algorithms like Neural Networks, Support Vector Machines, Naïve Bayes, rule-based learners etc which have been proved to be successful in handling feature vectors of texts representation. In this study, each text has been reduced into a one-dimensionality feature of Latin etymologies in values of local probabilities against the total number of tokens in the running text. Thus, a task-based dedicated classification algorithm is implemented to automatically learn the best distribution pattern of the proposed features in representation of the texts. Given a range (within the wide of 0.05 difference between two ranges) of Latin etymology probabilities, RangeClassifier starts from the range (Index*Mode*) which the most number of texts in the category fall in, extends in bi-direction of the left index (Index*L*)and the right index (Index*)t,* stops when the best F-score achieved. Table 3 is the algorithm of RangeClassifier in Step three.

. **Table 3:** Algorithm of RangeClassifier.

```
Input: Training Corpus T. Latin-ety Probability Parameter p
Output: Association arrays of maximum Precision, Recall, and
        F-score with the pair of <IndexL, IndexR>.
for each class $c_i$ in T
    for each text $t_i$ in class $c_i$
        extract Latin probability p for $t_i$
        set the <$R_l$,$R_r$> of p for each $t_i$
        rangeCount++
        set IndexMode as mode of p for $t_i$
    end
end
IndexL = IndexR = IndexMode
Loop until IndexL == 0 && IndexR == rangeCount
     calculate F-score, F; Precision, P;
     Recall, R from <IndexL, IndexR>
     push F, R, P
     push <IndexL, IndexR>
     IndexL--
     IndexR++
end
Output $F_{max}$, $R_{max}$, $P_{max}$ with the associated pair of <IndexL,
IndexR>
```

3 Experiments

Table 4 contains the number of texts under each category in which each text has 4000 words**.**

Table 4: Number of texts under each Category (each text has 4000 words).

	Conversation		Otherspoken		Written
Training (80%)	3,091		2,826		11,699
Testing (20%)	816		801		3,229
	Fiction	News		Otherpub	Unpub
Training (80%)	2,986	2,856		2,946	2,911
Testing (20%)	811	826		831	761

The second set of the experiment builds up the best ranges for each category according to the training texts, the ranges are then used to classify the testing texts. To well verify the performance of RangeClassifier, a 5-fold cross validation scheme is applied. Table 5a/b, and Table 6a/b show the best ranges in average measured by precision, recall and F_Score against the text size of 4000 words, and 6000 words respectively.

Table5a: Best ranges for spoken and written (4000 words each text).

	Conversation	Otherspoken	Written
Range	0.0058-0.0658	0.0996-0.1946	0.0556-0.1206
Precision	**0.7189**	**0.2251**	**0.7552**
Recall	0.7364	0.6207	0.6628
F_Score	0.7275	0.3304	0.7059

Table 5b: Best ranges for sub-written (4000 words each text).

	Fiction	News	Unpub	Otherpub
Range	0.0456-0.1306	0.0599-0.1049	0.0387-0.1237	0.0770-0.1320
Precision	0.2772	0.3448	0.2109	0.2524
Recall	0.8391	0.6807	0.6117	0.6
F_Score	0.4168	0.4577	0.3137	0.3552

Table6a: Best ranges for spoken and written (6000 words each text).

	Conversation	Otherspoken	Written
Range	0.0134-0.0664	0.0541-0.1491	0.0590-0.1140
Precision	**0.7274**	**0.1382**	**0.7578**
Recall	0.7258	0.6337	0.6034
F_Score	0.7266	0.2270	0.6718

Table 6b: Best ranges for sub-written (6000 words each text).

	Fiction	News	Unpub	Otherpub
Range	0.07880-0.1240	0.0572-0.1022	0.0479-0.1229	0.0803-0.1353
Precision	0.2831	0.3639	0.2132	0.2653
Recall	0.6047	0.6990	0.6056	0.6349
F_Score	0.3856	0.4786	0.3154	0.3743

From the above tables, with the input texts in the size of 4000 and 6000 words, RangeClassifer performs slightly better in overall with the texts in the size of 4000 words based on F_Score.

4 Evaluation

RangeClassifier is tested on the testing data with the number of documents shown in Table 4 with the variations of the texts size from 1,000 to 11,000 in the step of 1,000. Its performance is evaluated by precision P, recall R and F_Score F as defined:

$$P = \frac{a}{a+b} \tag{1}$$

$$R = \frac{a}{a+c} \tag{2}$$

$$F = 2\frac{PR}{P+R} \tag{3}$$

where,

a: retrieved relevant documents
b: retrieved un-relevant documents
c: not-retrieved but relevant documents

Figure 1 and Figure 2 show the variations of precision and F-score with the size of texts on the testing set for spoken-written and sub-written categories. Both figures show that the precision increases with the increasing of the text sizes, but the trend moderates after the size of 8000 for spoken-written texts and 9000 for sub-written texts.

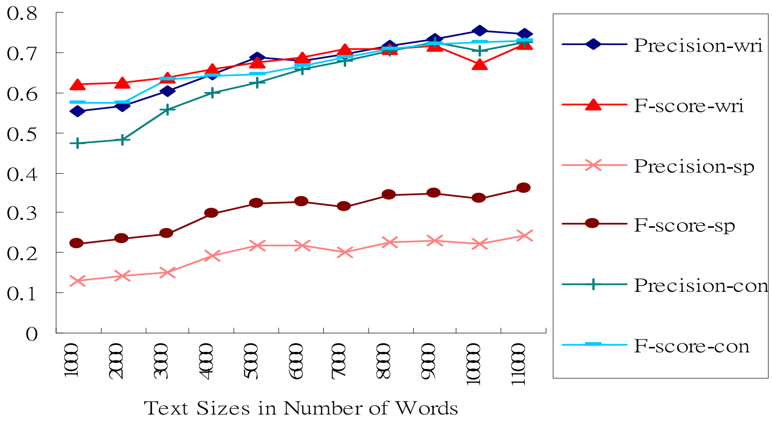

Figure 1: Performance Variation on Spoken-Written based on the different text sizes.

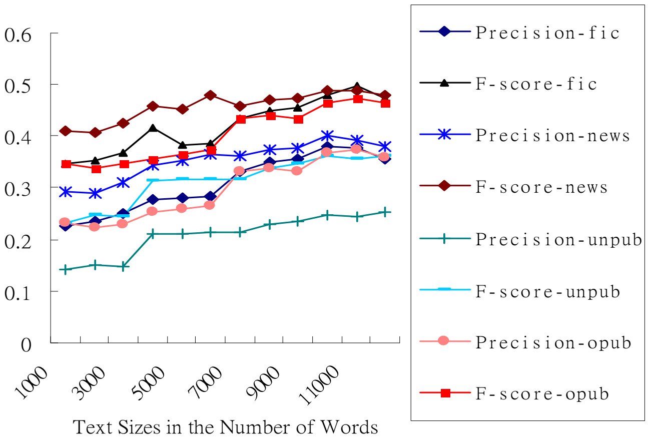

Figure 2: Performance Variation on Sub-Written based on the different text sizes.

5 Discussion and Future Work

The current experiment proves that the features of the different probability distribution on Latin etymologies can be used to classify the BNC Conversation and Written, achieves the F_Score of 72.76% on conversation and of 71.61% on written. However, the result is not good (36.02% of precision for text size of 10,000 words) on the category of OtherSpoken which is mixed with formal and informal speeches, while the conversion is consisted of unscripted informal speeches. Hence, we conclude that distribution of Latin etymologies is distinct in informal spoken and written texts which can be used as a good feature for classifying spoken in informal and written.

To the fact that the performance of this study is not as competitive as the reported studies such as around 69% in F_Score using KNN algorithm and 80% in F_Score using SVM in (Lan *et al.*, 2006), up to 90% in F_Score using SVM and 83% in F_Score using KNN in (Shang *et al.*, 2007), another possible future job is to explore other different levels of linguistic features (such as the words having Latin etymologies against the other bag-of-words), the different features may be assigned a different weighting value to identify their significance.

References

Anette, H. 2006. A study on automatically extracted keywords in text categorization. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp. 537-544.

Beil, Florian, Martin Ester and Xiaowei Xu. 2002. Frequent Term-Based Text Clustering. *ACM SIGKDD 02*.

Chen, Bo, Hui He and Jun Guo. 2008. Constructing Maximum Entropy Language Models for Movie Review Subjectivity Analysis. *Journal of Computer Science and Technology, 23(2):* 231-239.

Chen, Jingnian, Houkuan Huang, Shengfeng Tian and Youli Qu. 2008. Feature selection for text classification with naïve bayes. *Expert Systems with Applications*, In Press, Corrected Proof.

Dinsmoor, James A.. 2004. The etymology of basic concepts in the experimental analysis of behavior. *Journal of the Experimental Analysis of Behavior*, 82 (3): 311–316.

Furnkranz, Johannes, Tom Mitchell and Ellen Riloff. 1998. A case study in using linguistics phrase in text categorization on the WWW. *AAAI/ICML Workshop*.

Grzega, Joachim and Marion Schöner. For the processes and triggers of English vocabulary changes cf. *English and General Historical Lexicology*.

Kanaris, I. and E. Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. *In Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, Vol.2, Washington, DC, USA, IEEE Computer Society.

Kontostathis, A. and W. M. Pottenger. 2006. A framework for understanding LSI performance. *Information Processing and Management*, Volume 42, No. 1, pp.56-73.

Lan, M., C.L. Tan and H-B. Low. 2006. Proposing a new term weighting scheme for text categorization. *AAAI-06*.

Lee, Changki and Gary Geunbae Lee. 2006. Information gain and divergence-based feature selection for machine learning-based text categorization. *Information Processing and Management: an International Journal*, v.42 n.1, p.155-165.

Lewis, D.D. 1992. Feature selection and feature extraction for text categorization. *Proceedings of Speech and Natural Language Workshop*.

Liao, C., S. Alpha and P. Dixon. 2003. Feature Preparation in Text Categorization. *ADM03 workshop*.

Lin, King-Ip and Ravikumar Kondadadi, 2001. A Similarity-Based Soft Clustering Algorithm For Documents. *IEEE*.

Liu, H. and L. Yu. 2005. Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491-502.

Mansur, M., N. UzZaman and M. Khan. 2006. Analysis of n-gram based text categorization for bangla in a newspaper corpus. *In 9th International Conference on Computer and Information Technology (ICCIT 2006)*, Dhaka, Bangladesh.

Mukras, R., N. Wiratunga, R. Lothian, S. Chakraborti and D. Harper. 2007. Information Gain Feature Selection for Ordinal Text Classification using Probability Re-distribution. *Proceedings of the Textlink workshop at IJCAI-07*.

Nigam, K., L. John and A. McCallum. 1999. Using maximum entropy for text classification. *In IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp. 61-67.

Olsson, J., Scott Olsson and Douglas W. Oard. 2006. Combining feature selectors for text classification. *Proceedings of the 15th ACM international conference on Information and knowledge management*, November 06-11, Arlington, Virginia, USA.

Pei, Zhili, Xiaohu Shi, Maurizio Marchese and Yanchun Liang. 2007. Text Categorization Method Based on Improved Mutual Information and Characteristic Weights Evaluation Algorithms. *Fourth International Conference on Fuzzy Systems and Knowledge Discovery*, Vol.4.

Rogati, M. and Y. Yang. 2002. High-performing Feature Selection for Text Classification. *International Conference on Information and Knowledge Management-CIKM.*

Rougemont, A. de. 1987. *Review: French Etymology*., by © 1887. The Johns Hopkins University Press.

Shang, Wenqian, Houkuan Huang, Haibin Zhu, Yongmin Lin, Youli Qu and Zhihai Wang. 2007. A novel feature selection algorithm for text categorization. *Expert Systems with Applications: An International Journal*, v.33 n.1, p.1-5.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1): 1-47.

Wang, G., F.H. Lochovsky and Q. Yang. 2004. Feature selection with conditional mutual information maximin in text categorization. *In: Proceedings of the 13th ACM International Conference on Information and Knowledge Management, ACM*, Washington, DC, USA. pp. 342-349.

Williams, J. M. 1986. *Origins of the English Language: A social and linguistic history*. New York: The Free Press.

Word Etymologies: The Greek and Latin Roots of English *(CL903-2A)*, http://www.brown.edu/scs/pre-college/course-one/course-detail.php?course_code=CL903-2A.

Wu, H. and D. Gunopulos. 2002. Evaluating the Utility of Statistical Phrases and Latent Semantic Indexing for Text Classification. *IEEE International Conference on Data Mining*, pp.713-716.

Yu, Bo, Zong-ben Xu and Cheng-hua Li. 2008. Latent semantic analysis for text categorization using neural network. *Knowledge-Based Systems*, v.21 n.8, pp.900-904.

Zhang, D. and W.S. Lee. 2006. Extracting key-substring-group features for text classification. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press*, pp. 474–483.

Zhang, J. and Y. Yang. 2003. Robustness of regularized linear classification methods in text categorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference*, pp. 190–197.

Zhang, T. and F.J. Oles. 2001. Text categorization based on regularized linear classification methods. *Information Retrieval*, 4(1):5–31.

Zhang, Wei, Shuang Liu, Clement Yu, Chaojing Sun, Fang Liu and Weiyi Meng. 2006. Recognition and classification of noun phrases in queries for effective retrieval. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management,* USA.