Text Genre Detection Using Common Word Frequencies

E. STAMATATOS, N. FAKOTAKIS, and G. KOKKINAKIS
Dept. of Electrical and Computer Engineering
University of Patras
Patras, Greece, 26500
stamatatos@wcl.ee.upatras.gr

Abstract

In this paper we present a method for detecting the text genre quickly and easily following an approach originally proposed in authorship attribution studies which uses as style markers the frequencies of occurrence of the most frequent words in a training corpus (Burrows, 1992). In contrast to this approach we use the frequencies of occurrence of the most frequent words of the entire written language. Using as testing ground a part of the *Wall Street Journal* corpus, we show that the most frequent words of the *British National Corpus*, representing the most frequent words of the written English language, are more reliable discriminators of text genre in comparison to the most frequent words of the training corpus. Moreover, the frequencies of occurrence of the most common punctuation marks play an important role in terms of accurate text categorization as well as when dealing with training data of limited size.

Introduction

The development of text databases via the Internet has given impetus to research in computational linguistics towards the automatic handling of this information. In particular, the enormous amount of texts coming from heterogeneous sources revealed the need for robust text classification tools which are able to be easily ported to other domains and natural languages and be employed with minimal computational cost.

Apart from the propositional content of the text, stylistic aspects can also be used as classificatory means. Biber studied the stylistic differences between written and spoken language (Biber, 1988) as well as the variation of registers in a cross-linguistic comparison (Biber, 1995) and presented a model for interpreting the functions of various linguistic features. Unfortunately, his model can not be easily realized using existing natural language processing tools. On the other hand, some computational models for detecting automatically the text genre have recently been available (Karlgren and Cutting, 1994; Kessler *et al.*, 1997). Kessler gives an excellent summarization of the potential applications of a text genre detector. In particular, part-of-speech tagging, parsing accuracy and word-sense disambiguation could be considerably enhanced by taking genre into account since certain grammatical constructions or word senses are closely related to specific genres. Moreover, in information retrieval the search results could be sorted according to the genre as well.

Towards the automatic detection of text genre, various types of style markers (i.e., countable linguistic features) have been proposed so far. Karlgren and Cutting (1994) use a combination of structural markers (e.g., noun count), lexical markers (e.g., "it" count), and token-level markers (e.g., words per sentence average, type/token ratio, etc.). Kessler *et al.* (1997) avoid structural markers since they require tagged or parsed text and replace them with character-level markers (e.g., punctuation mark counts) and derivative markers, i.e., ratios and variation measures derived from measures of lexical and character-level markers.

Furthermore, some interesting stylometric approaches have been followed in authorship attribution studies. Specifically, various functions that attempt to represent the vocabulary richness have been proposed (Honore 1979; Sichel, 1975). The combination

of the best vocabulary richness functions in a multivariate model can then be used for capturing the characteristics of a stylistic category (Holmes, 1992). However, recent studies have shown that the majority of these functions depend heavily on text-length (Tweedie and Baayen, 1998). Additionally, Stamatatos *et al.* (1999) attempted to take advantage of already existing text processing tools by proposing the analysis-level markers taking into account the methodology of the particular tool that has been used to analyze the text. This approach requires the availability of a robust text processing tool and the time and/or computational cost for the calculation of the style markers is proportional to the corresponding cost of the analysis of the text by this tool.

Last but not least, a stylometric approach proposed by Burrows (1987; 1992) uses as style markers the frequencies of occurrence of the most frequent words (typically the 50 most frequent words) as regards a training corpus. This method requires minimal computational cost and has achieved remarkable results for a wide variety of authors. Moreover, it is domain and language independent since it does not require the manual selection of the words that best distinguish the categories (i.e., function words). However, in order to achieve better results Burrows took into account some additional restrictions, namely:

- Expansion of the contracted forms. For example, "I'm" counts as "I" and "am".
- Separation of common homographic forms. For example, the word "to" has the infinitive and the prepositional form.
- Exception of proper names from the list of the most frequent words.
- Text-sampling so that only the narrative parts of the text contribute to the compilation of the list of the most frequent words. Note that a 'narrative' part is simply defined as 'non-dialogue'.

From a computational point of view, these restrictions (except the first one) complicate the procedure of extracting the most frequent words of the training corpus. Thus, the second restriction requires a part-of-speech tagger, the third has to be performed via a named-entity recognizer, and the last requires the development of a robust text sampling tool able to detect the narrative parts of any text.

In this paper we present a variation of this approach. Instead of extracting the most frequent word list of the training corpus, we use as style markers the frequencies of occurrence of the most frequent words of the entire written language. For English, the most frequent words of the written language component of the *British National Corpus* are considered. We show that our approach performs better than the Burrows' original method without taking into account any of the above restrictions. Moreover, we show that the frequencies of occurrence of the most frequent punctuation marks contain very useful stylistic information that can enhance the performance of an automatic text genre detector.

The paper is organized as follows. The next section describes both the corpora used in this study and the procedure of extracting the most frequent word lists. Section 2 includes the text genre detection experiments while section 3 contains experiments dealing with the role of punctuation marks. Finally, in the last section some conclusions are drawn and future work directions are given.

1 Testing Ground

1.1 Corpora

As regards the English language, in the previous work on text genre detection (Karlgren and Cutting, 1994; Kessler *et al.*, 1997) the *Brown* corpus was used as testing ground. It comprises approximately 500 samples divided into 15 categories (e.g., press editorial, press reportage, learned, etc.) that can be considered as genres. However, this corpus was not built exclusively for text genre detection purposes. Therefore, the texts included in the same category are not always stylistically homogeneous. Kessler *et al.*, (1997) underlined this fact and attempted to avoid the problem by eliminating texts that did not fall unequivocally into one of their categories. Moreover, some of the categories of the Brown corpus are either too general (e.g., general fiction) or unlikely to be considered in the framework of a practical application (e.g., belles lettres, religion, etc.). Taking all these into

1. the	11. with	21. are	31. or	41. her
2. of	12. he	22. not	32. an	42. n't
3. and	13. be	23. his	33. were	43. there
4. a	14. on	24. this	34. we	44. can
5. in	15. i	25. from	35. their	45. all
6. to	16. that	26. but	36. been	46. as
7. is	17. by	27. had	37. has	47. if
8. was	18. at	28. which	38. have	48. who
9. it	19. you	29. she	39. will	49. what
10. for	20. 's	30. they	40. would	50. said

Table 1: The 50 most frequent words of the BNC.

account we decided to use the *Wall Street Journal* (WSJ) corpus as testing ground for our approach. The texts comprising this corpus cover the majority of the press genres. Although there is no manual categorization of the WSJ documents according to their genre, there are headlines that sometimes help in predicting the corresponding text genre. The selection of the texts included in the presented corpus was performed automatically by reading the headline tag (<HL>) of each document. A typical headline tag of a WSJ document is as follows:

```
<HL> Marketing & Media:
@  RJR Nabisco Hires
@  Adviser to Study
@  Sale of ESPN Stake
@  ----
@  By Michael J. McCarthy
@  Staff Reporter of The Wall Street
Journal </HL>
```

Thus, we constructed a genre-corpus of four categories, namely: Editorials, Letters to the Editor, Reportage, and Spot news taking documents from the WSJ corpus of the year 1989. The documents containing the string "REVIEW & OUTLOOK (Editorial):" in their headline tag were classified as editorials while the documents containing the string "Letters to the Editor:" were considered as letters to the editor. The documents containing either the string "What's News -" or "Who's News:" were considered as spot news. Finally, all the documents containing one of the following strings in their headline: "International:", "Marketing & Media:", "Politics & Policy:", or "World Markets:" without including a line starting with the string "@ By " were considered as reportage. The latter assures that no signed article is considered as reportage. For example, the document of the above example was not included in any of the four genre categories.

1.2 Most Frequent Words

In order to extract the most frequent words of the acquired corpora we used equally-sized text samples (approximately 640k) from each category providing a genre-corpus of 2560k for the four categories. The genre-corpus was divided into 160 text samples of 16k (i.e., approximately 2,000 words) each, including 40 text samples from each genre. Half of the text samples from each category were used as training corpus and the rest as test corpus.

For the extraction of the most frequent words of the entire English language we used the *British National Corpus* (BNC). This corpus consists of 100M tokens covering both written and spoken language.

In this study we used the unlemmatized word frequency list of the written language component of the BNC[1] which comprises roughly 89.7M tokens. Since the homographic forms are separated, we added the frequencies of the words with more than one forms (e.g. "to") in order to attain a representative ordered list of the most frequent words of the entire written language. The resulted ordered word list of the 50 most frequent words is given in table 1.

The comparison of the most frequent word list of the genre-corpus with the one acquired by the BNC is given in figure 1. The common words (i.e., those included in both lists) constitute

[1] Available at: http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html

approximately 75% of the most frequent words of BNC.

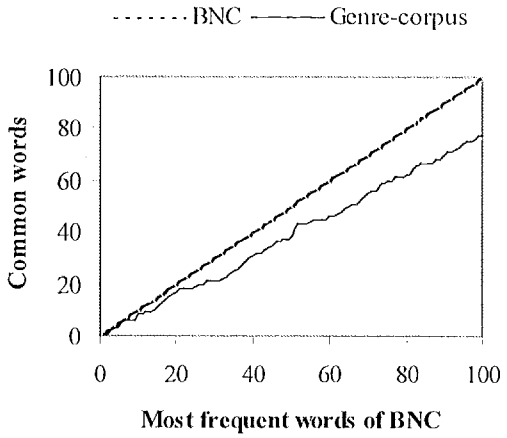

Figure 1: Comparison of the most frequent word lists.

2 Text Genre Detection

In order to detect automatically the text genre we used *discriminant analysis*, a well-known classification technique of multivariate statistics that has been used in previous work in text genre detection (Biber, 1993; Karlgren and Cutting, 1994). This methodology takes some multivariate vectors precategorized into naturally occurring groups (i.e., training data) and extracts a set of *discriminant functions* that distinguish the groups. The mathematical objective of discriminant analysis is to weight and linearly combine the discriminating variables (i.e., style markers) in some way so that the groups are forced to be as statistically distinct as possible (Eisenbeis & Avery, 1972). Then, discriminant analysis can be used for predicting the group membership of previously unseen cases (i.e., test data).

In the present case, the multivariate vectors are the frequencies of occurrence of the most frequent words of the BNC for each text sample and the naturally occurring groups are the four text genres. We applied discriminant analysis to the training genre-corpus using 5 to 75 most frequent words of BNC with a step of 5 words. The classification models were, then, cross-validated by applying them to the corresponding test corpus. The same procedure was followed using as style markers the frequencies of occurrence of the most frequent words of the training corpus (according to the original method of Burrows). Comparative results in terms of classification error rate are given in figure 2. As can been seen, the best performance achieved by our approach is 2.5% error rate (2/80) based on the 30 most frequent words of the BNC while the best performance of the Burrows' approach is 6.25% error rate (5/80) based on the 55 most frequent words of the training corpus.

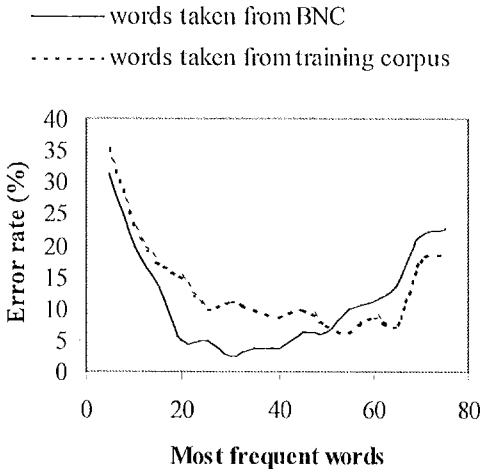

Figure 2: Comparative classification results for text genre detection.

It is worth noting that the performance of the classification model is not improved using more words beyond a certain threshold (in our approach 30 words). This is due to the training data overfitting. Figure 3 shows the training corpus in the space of the first two discriminant functions based on the 10, 30, and 70 most frequent words of the BNC.

It is obvious that 10 words are not enough for the sufficient discrimination of the genre categories. On the other hand, using the 70 most frequent words the discriminant functions are biased to the training data.

Furthermore, the genres *Editorial* and *Letters to the editor* could be grouped into a higher level genre since they share common stylistic features. Similarly, the genres *Reportage* and *Spot news* could be grouped as well. Note that the presented model managed to capture this

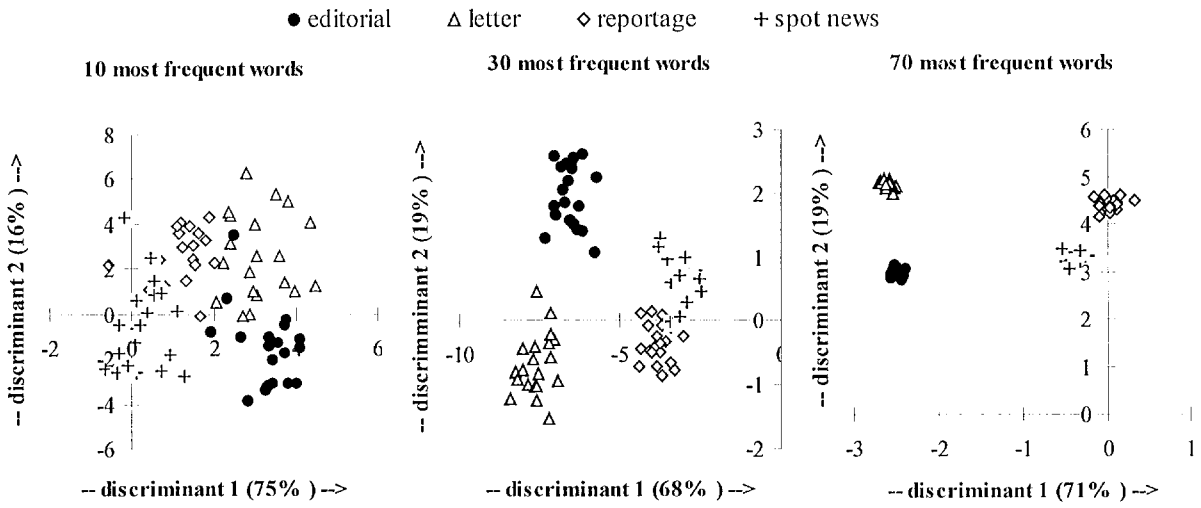

Figure 3: The training corpus in the space of the two first discriminant functions for different word lists. The numbers inside parentheses indicate the percentage of variation explained by the corresponding function.

information since in all the cases shown in figure 3 the first discriminant function (axis x), which accounts for the greatest part of the total variation, distinguishes between these high level genres. Then, the second discriminant function (axis y) attempts to discriminate each of the high level genres into genres of more specific level.

3 Punctuation Mark Frequencies

In addition to the most frequent words, the punctuation marks play an important role for discriminating reliably text genres. In fact, there are cases where the frequency of occurrence of a certain punctuation mark could be used alone for predicting a certain text genre. For example, an interview is usually characterized by an uncommonly high frequency of question marks.

In order to enhance the performance of the proposed classification models we take into account the frequencies of occurrence of the eight most frequent punctuation marks, namely: period, comma, colon, semicolon, quotes, parenthesis, question mark and hyphen. Thus, we applied discriminant analysis to the training genre-corpus taking into account the frequencies of occurrence of the above punctuation marks plus 5 to 75 most frequent words of BNC with a step of 5 words. Th cross-validated by applying them to the corresponding test corpus.

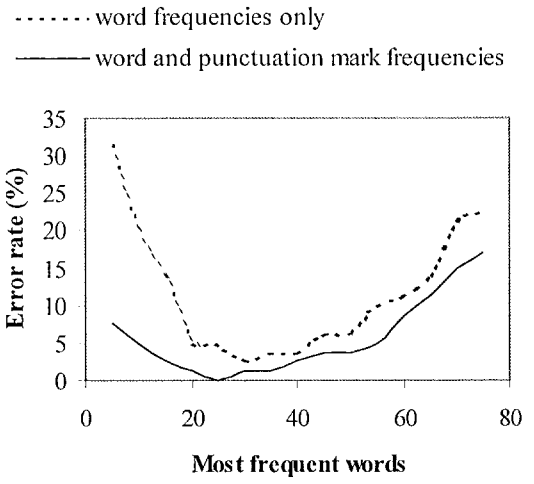

Figure 4: Classification results taking into account the punctuation marks.

The results are given in figure 4 together with the performance of the model using word frequencies only (from figure 2) for purposes of comparison. The error rate is now considerably lower and very reliable classification accuracy results (>97%) can be achieved based on a relatively small set of style markers (i.e., the frequencies of occurrence of the eight

punctuation marks plus 15 to 35 most frequent words).

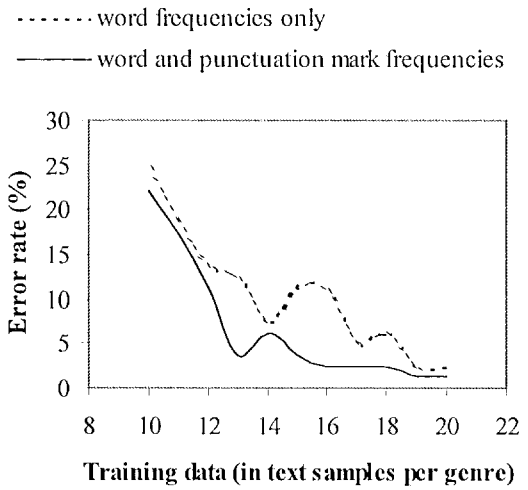

Figure 5: Error rate vs. training data size.

The role of the punctuation marks in achieving reliable classification results can be further illustrated by examining the relation between classification accuracy and training data size. Towards this end, we applied discriminant analysis to different training corpora consisting of 10 to 20 text samples from each genre taking into account the frequencies of occurrence of the 30 most frequent words of the BNC. This procedure was followed once again taking into account the eight additional style markers of the punctuation marks (i.e., totally 38 style markers). The comparative results are given in figure 5. As can been seen, the performance of the model taking into account only word frequencies is affected dramatically by the decrease of the training data. On the other hand, the performance of the model taking into account both word and punctuation mark frequencies remains satisfactory (i.e., error rate < 7%) using 13 to 20 text samples from each genre.

Conclusion

In this paper we presented a methodology for detecting automatically the text genre of unrestricted text. We followed the main idea of a stylometric approach originally proposed for attributing authorship, which uses as style markers the frequencies of occurrence of the most frequent words of a certain training corpus. In order to improve the accuracy of this model various additional restrictions have been proposed (see the introduction), which in general complicate the computational processing of the texts.

Instead of taking into account such restrictions, we considered the frequencies of occurrence of the most frequent words of the entire written language. It has been shown that they are more reliable stylistic discriminators as regards the combination of classification accuracy and the number of the required common words that have to be taken into account. Note that when dealing with multivariate models, the reduction of the required parameters is a very crucial factor for attaining reliable results and minimizing the computational cost.

As testing ground in this study we used a part of the WSJ corpus classified into four low-level genres that can be grouped into two higher-level genres. The automated classification model based on discriminant analysis applied to the frequencies of occurrence of the most frequent words of the BNC, that represent the most frequent words of the entire written English language, managed to capture the stylistic homogeneity in both levels.

Moreover, it has been shown that the frequencies of occurrence of the most frequent punctuation marks can considerably enhance the performance of the proposed model and increase the reliability of the classification results especially when training data of limited size are available.

The proposed approach meets the current trends in natural language processing since:

- it is able to deal with unrestricted text,
- it requires minimal computational cost,
- it is not based on specific characteristics of a certain domain/language.

On the other hand, any of the additional restrictions mentioned in the introduction, and especially the separation of the common homographic forms, can still be considered. The combination of this approach with style markers dealing with syntactic annotation seems to be a better solution for a general-purpose automated text genre detector. Another useful direction is

the development of a text-sampling tool able to detect different genres within the same document.

References

Biber, D. (1993). Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics*, 19(2), pp. 219-242.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge University Press.

Burrows, J. (1987). Word-patterns and Story-shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, 2(2), pp. 61-70.

Burrows, J. (1992). Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information. *Literary and Linguistic Computing*, 7(2), pp. 91-109.

Eisenbeis, R., and R. Avery (1972). *Discriminant Analysis and Classification Procedures: Theory and Applications*. Lexington, Mass.: D.C. Health and Co.

Holmes, D. (1992). A Stylometric Analysis of Mormon Scripture and Related Texts. *Journal of the Royal Statistical Society*, Series A, 155(1), pp. 91-120.

Honore, A. (1979). Some Simple Measures of Richness of Vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2), pp. 172-177.

Karlgren, J., and D. Cutting (1994). Recognizing text Genres with Simple Metrics Using Discriminant Analysis. In *Proc. of the 15th International Conference on Computational Linguistics (COLING '94)*.

Kessler, B., G. Nunberg, and H. Schutze (1997). Automatic Detection of Text Genre. In *Proc. of 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL'97)*, pp. 32-38.

Sichel, H. (1975). On a Distribution Law for Word Frequencies. *Journal of the American Statistical Associaton*, 70, pp. 542-547.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis (1999). Automatic Authorship Attribution. In *Proc. of the 9th Conf. of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pp. 158-164.

Tweedie, F. and Baayen, R. (1998). How Variable may a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, 32(5), pp.323-352.