Style-based Text Categorization: What Newspaper Am I Reading?

Shlomo Argamon-Engelson and **Moshe Koppel** and **Galit Avneri**
Dept. Mathematics and Computer Science
Bar-Ilan University
52900 Ramat Gan, Israel
Tel: 972-3-531-8407
Fax: 972-3-535-3325
Email: {argamon,koppel,avneri}@cs.biu.ac.il

Abstract

Most research on automated text categorization has focused on determining the *topic* of a given text. While topic is generally the main characteristic of an information need, there are other characteristics that are useful for information retrieval. In this paper we consider the problem of text categorization according to *style*. For example, we may wish to automatically determine if a given text is taken from a magazine or a newspaper, is an editoral or a news item, is promotional or informative, was written by a native English speaker or not, and so on. Learning to determine the style of a document is a dual to that of determining its topic, in that those document features which capture the style of a document are precisely those which are independent of its topic. We here define the features of a document to be the frequencies of each of a set of function words and parts-of-speech triples. We then use machine learning techniques to classify documents. We test our methods on four collections of newspaper and magazine articles.

1 Introduction

Most research on automated text categorization has concerned categorization according to *topic*, because topic is generally the main characteristic of an information need. In this paper we consider the problem of text categorization according to *style*. For example, we may wish to automatically determine if a given text is taken from a magazine or a newspaper, is an editoral or a news item, is promotional or informative, was written by a native English speaker and so on.

Categorization by topic and categorization by style are fundamentally similar problems in that in both cases we want to determine the proper category of a document based on analyzing features of the document. The problems are, however, orthogonal to each other in the sense that those document features which capture the style of a document are precisely the ones which are independent of its topic.

Roughly speaking, then, categorization by style requires the isolation of the stylistic features of a document. This task is more difficult than the problem of isolating topic-related features of a document, where keywords are often sufficient.

The problem of identifying stylistic features of a document has been considered in the literature on literary analysis, primarily for the purpose of attribution of classic works of disputed authorship (see the surveys by Holmes (1995) and, more recently McEnery (1998)). Some work has also been carried out in the area of forensics (Bailey 1979; McMenamin 1993).

Virtually all the early work on the authorship attribution problem focused on attempts to isolate *discriminators*, that is, individual stylistic features which are largely invariant over different passages writen by a given author, but which vary from author to author. Many types of discriminators have been considered, including sentence length (Yule 1938), vocabulary richness (Yule 1944), frequencies of selected words and other more complicated measures. A particularly important example of an attempt to combine the information obtainable from the use of several such features is the seminal work of Mosteller and Wallace (1964), which uses as features the frequencies of each of a set of selected content-free function words (i.e., conjunctions, prepositions, etc.). Although the use of such features has proved successful in some applications (McEnery & Oakes 1998), it has not been convincingly shown that any particular features are consistently effective over a wide range of authorship attribution problems (Goldberg 1995). Surely, then, there is not much hope of finding some small set of discriminators which will be prove effective for the more general problem of categorization by style.

Modern learning-based text categorization methods overcome the feature-selection problem which has stymied the literary work by considering a very large set of candidate features and allowing the data (the target documents) to drive the selection of features for a particular problem. In what follows, we will demonstrate how combining modern computational linguistic

technology with machine learning techniques can be used to advantage for style-based categorization.

2 The Algorithm

2.1 Characterizing Textual Style

The first step is to characterize a given document in terms of some very large class of stylistic features. In this paper we consider two types of features: lexical and pseudo-syntactic.

The lexical features are simply the frequency of function words, such as *and*, *about*, and *the*. Function words have been a staple of authorship attribution research since the work of Mosteller and Wallace (1964). The rationale behind their use is that the frequency of such words is presumably not driven by content and hence might be expected to remain invariant for a given author over different topics. By the same reasoning, we might expect such invariance within classes of similar documents. For example, the word *you* might appear frequently in promotional literature, the word *should* might appear frequently in editorials, and the word *today* might appear frequently in news items. In this work, we use a list of 500 function words, considerable more than previous studies have used (most used fewer than 20).

In addition to function words, we use a different set of features which more directly encode information about the syntactic structure of the text. Syntactic structure can be expected to be more revealing than simple words counts, as has been argued in the work of Baayen et al. (1996). The idea is that particular document classes will favor certain syntactic constructions. Previous work on syntactic features (Baayen, van Halteren, & Tweedie 1996) was based upon automated parsing of the input texts. However, as has been pointed out by McEnery (1998), this is a time-consuming and unreliable procedure; what is required is a method of capturing underlying syntactic structures in an efficient and reliable fashion.

In this paper we adopt the following simple and robust approach: We first tag each word in the text by its part-of-speech (POS), using Brill's rule-based part-of-speech tagger (Brill 1992) (which uses the tagset from the Brown corpus). The features we consider then are POS trigrams. Trigrams are large enough to encode useful syntactic information, while small enough to be computationally managable. For example, very frequent use of the trigram personal-pronoun; present tense-verb; verb-gerund might be indicative of a particular, possibly Germanic, syntactic style. While there are over 100,000 possible trigrams, we only consider as features those trigrams which appear in between 25 and 75 percent of the documents our document collection.

It should be noted that we exclude from consideration any content-bearing features even though these often prove useful. For example, in determining whether a particular passage was written by Arthur Conan Doyle, it would be perfectly reasonable to check the frequency of appearance of, say, "Sherlock". Nevertheless, in the larger picture such features can be deceptive. For example, given a set of New York Times documents about the Middle East and a set of Daily News articles about President Clinton, we would want to correctly categorize a Times article about Clinton as being from the Times and not the Daily News.

2.2 Learning

Once a feature set has been chosen, a given text passage can be represented by an n-dimensional vector, where n is the total number of features. Given a set of pre-categorized vectors, we can apply any machine learning algorithm to try to determine the category of a given new document. In our experiments, we used Ripper (Cohen 1995), a member of the family of decision-tree learning algorithms, which has been used successfully on topic-based text classification problems (Cohen & Singer 1996). Since Ripper's bias towards small conjunctive rules is not appropriate for every text categorization problem, other learning algorithms might also be considered. One interesting alternative to Ripper might be an algorithm such as Winnow (Littlestone 1988) which learns linear separators.

3 Experiments

3.1 Methodology

We tested our method using four text collections[1]:

NY Times news 200 national (U.S.) news stories from the month of January 1998

NY Times editorial 200 editorial and op-ed articles from the month of January 1998

NY Daily News 200 mostly[2] national news stories from the month of January 1998

Newsweek 200 magazine articles on domestic U.S. issues from July 1997 through the end of January 1998

[1] This data was obtained from the Nexus database.

[2] Nexus does not allow selecting only national stories from the Daily News. We chose articles from the first seven pages of the paper which are the pages usually reserved for national news.

Comparison	FW	POS	Both
Times news vs. editorial	78	69.3	79.5
Times news vs. Daily News	82.3	63.1	84.3
Times news vs. Newsweek	80.5	79.3	83.8
Times editorial vs. Newsweek	61.3	70	68.5
Daily News vs. Times editorial	77.6	67.3	78.1
Daily News vs. Newsweek	78.5	79.6	80.6

Table 1: Results of 5-fold validation experiments for different data sets and feature sets.

All the documents contained between 300 and 1300 words. During the chosen periods, the number of available documents in each category was slightly above 200, so the first 200 were chosen. Since the newspaper articles covered the same dates, they also covered roughly the same topics (lots of Clinton-Lewinsky).

The collections were tested for pairwise distinguishability. In each experiment, two collections were selected. The 400 documents were transformed into feature vectors and divided into training and test sets using five-fold cross validation. Each of these experiments was run using three different feature sets:

FW The frequencies of each of 500 function words,

POS The frequencies of each of 685 POS trigrams, and

Both Both the function word and trigrams frequencies together.

3.2 Results

Table 1 summarizes the accuracy results of five-fold cross validation for each of the six pairwise experiments and each feature set.

A careful review of the rules produced by Ripper adds insight into the meaning of these results. Consider for example the following rules for distinguishing Times news stories from Daily News stories using only function words.

```
NYT_NEWS :- today>=2, yesterday<=2 (104/5).
NYT_NEWS :- that>=25, her<=10, will<=3 (30/10).
default DAILY_NEWS (148/24).
```

The numbers in parentheses after each rule reflect the number of documents (correctly/incorrectly) clasified by that rule. The error rate on the 80 test documents is 15%.

The picture that emerges is that there are often a relatively small number of signature features which very decisively mark a document as belonging to a particular collection. Difficulties arise when a document does not have any of the signature features.

Let us consider some of the signature features which turn up. The word *today* is a signature feature of Times news stories, while the word *yesterday* is a signature feature of Daily News stories. Over 70% of Times stories and Daily News stories contain the words *today* and *yesterday*, respectively, while the reverse holds for less than 20% of the stories.

Similarly, there are some interesting signature features among the POS triples. A signature triple for Newsweek magazine is end-quote; present-tense-verb; proper-noun as in '... yada, yada, " opines Costanza'. This triple appears in only 1.3% of the newspaper articles, whereas it appears in 32% of Newsweek articles. This indicates that Newsweek often uses present tense verbs where newspapers would use a past tense verb.

A signature triple for Times news is determiner; proper-noun; pausal-punctuation as in: 'But Mr. Gore has used previously scheduled stops — like his appearances with the President on Wednesday in the Midwest — to lend his energetic support.' This reflects the fact that the Times uses more complex sentence structure.

The results in the above table simply reflect the pervasiveness of the various signature features. The news stories (in both newspapers) are most easily distinguished from other sources and from each other by their respective use of certain words, while Newsweek is most easily distinguished from the others by its distinctive syntax.

One immediate conclusion from this is that in each classification task considered here it is not difficult to isolate features that work very reliably for some classes of documents wheras for other classes these methods are inconclusive.

4 Discussion

In this paper we have introduced the problem of style-based text categorization. This problem subsumes the problem of authorship attribution and goes well beyond it to a broad range of applications in information retrieval.

This is definitely work in progress; much more experimentation is required. Learning algorithms other than Ripper might be considered, especially those which have already been successfully used on content-based text categorization problems, such as Winnow (Dagan, Karov, & Roth 1997; Lewis *et al.* 1996) and Naive Bayes (Lang 1995). Also classes of features other than function words and POS trigrams might be tried. Finally, and perhaps most importantly, these methods must be tested on a much wider variety of style-based categorization problems.

References

Baayen, H.; van Halteren, H.; and Tweedie, F. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11.

Bailey, R. 1979. Authorship attribution in a forensic setting. In Aeger, D.; Knowles, F.; and Smith, J., eds., *Advances in Computer-Aided Literary and Linguistic Research.*

Brill, E. 1992. A simple rule-based part of speech tagger. In *Proc. of ACL Conference on Applied Natural Language Processing.*

Cohen, W., and Singer, Y. 1996. Context-sensitive methods for text categorization. In *Proceedings of SIGIR.*

Cohen, W. 1995. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning.*

Dagan, I.; Karov, Y.; and Roth, D. 1997. Mistake-driven learning in text categorization. In *Proceedings of Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2).*

Goldberg, J. 1995. Cdm: An approach to learning in text recognition. In *Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence.*

Holmes, D. 1995. Authorship attribution. *Computers and the Humanities* 28.

Lang, K. 1995. Newsweeder: learning to filter net news. In *Proceedings of the Twelfth International Conference on Machine Learning.*

Lewis, D.; Schapire, R. E.; Callan, J. P.; and Papka, R. 1996. Training algorithms for linear text classifiers. In *SIGIR '96: Proc. of the 19th Int. Conference on Research and Development in Information Retrieval, 1996.*

Littlestone, N. 1988. Learning quickly when irrelevant features abound: A new linear-threshold algorithm. *Machine Learning* 2.

McEnery, A. M., and Oakes, M. P. 1998. Authorship studies/textual statistics. In Dale, R.; Moisl, H.; and Somers, H., eds., *Handbook of Natural Language Processing.* forthcoming.

McMenamin, G. 1993. *Forensic Stylistics.* Elsevier.

Mosteller, F., and Wallace, D. L. 1964. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers.* Springer.

Yule, G. 1938. On sentence length as a statistical characteristic of style in prose. *Biometrika* 30.

Yule, G. 1944. *The Statistical Study of Literary Vocabulary.* Cambridge University Press.