Using Function Words for Authorship Attribution: Bag-Of-Words vs. Sequential Rules

Mohamed Amine Boukhaled, Jean-Gabriel Ganascia

Using Function Words for Authorship Attribution: Bag-of-Words vs. Sequential Rules

Mohamed Amine Boukhaled, Jean-Gabriel Ganascia

LIP6 (Laboratoire d'Informatique de Paris 6), Université Pierre et Marie Curie and CNRS (UMR7606), ACASA Team, 4, place Jussieu,
75252-PARIS Cedex 05 (France) ,
{mohamed.boukhaled, jean-gabriel.ganascia}@lip6.fr

Abstract. Authorship attribution is the task of identifying the author of a given document. Various style markers have been proposed in the literature to deal with the authorship attribution task. Frequencies of function words have been shown to be very reliable and effective for this task. However, despite the fact that they are state-of-the-art, they basically rely on the invalid *bag-of-words* assumption, which stipulates that text is a set of independent words. In this contribution, we present a comparative study on using two different types of style marker based on function words for authorship attribution. We compare the effectiveness of using sequential rules of function words as style marker that do not relay on the *bag-of-words* assumption to that of the frequency of function words which does. Our results show that the frequencies of function words outperform the sequential rules.

1 Introduction

Authorship attribution is the task of identifying the author of a given document. The authorship attribution problem can typically be formulated as follows: given a set of candidate authors for whom samples of written text are available, the task is to assign a text of unknown authorship to one of these candidate authors (Stamatatos, 2009).
This problem has been addressed mainly as a problem of multi-class discrimination, or as a text categorization task (Sebastiani, 2002). Text categorization is a useful way to organize large document collection. Authorship attribution, as subtask of text categorization, assumes that the categorization scheme is based on the authorial information extracted from the documents. Authorship attribution is a relatively old research field. A first scientific approach to the problem was proposed in the late 19th century, in the work of Mendenhall in 1887, who studied the authorship of texts attributed to Bacon, Marlowe and Shakespeare. More recently, the problem of authorship attribution gained greater importance due to new applications in forensic analysis,

humanities scholarship (Stamatatos, 2009). Current authorship attribution methods have two key steps: (1) an indexing step based on style markers is performed on the text using some natural language processing techniques such as such as tagging, parsing, and morphological analysis; then (2) an identification step is applied using the indexed markers to determine the most likely authorship. An optional features selection step can be employed between these two key steps to determine the most relevant markers. This selection step is done by performing some statistical measures of relevance such us mutual information or Chi-square testing.
The identification step involves using methods that fall mainly into two categories: the first category includes methods that are based on statistical analysis, such as principle component analysis (Burrows, 2002) or linear discriminant analysis (Stamatatos et al., 2001); the second category includes machine learning techniques, such as simple Markov Chain (Khmelev and Tweedie, 2001), Bayesian networks, support vector machines (SVMs) (Koppel and Schler, 2004) , (Diederich et al., 2003)) and neural networks (Ramyaa and Rasheed, 2004). SVMs, which have been used successfully in text categorization and in other classification tasks, have been shown to be the most effective attribution method (Diederich et al., 2003). This is due to the fact that SVMs are less sensitive to irrelevant features in terms of degradation in accuracy, and permit one to handle high dimensional data instances more efficiently.

To achieve high authorship attribution accuracy, one should use features that are most likely to be independent from the topic of the text. Many style markers have been used for this task from early works based on features such as sentence length and vocabulary richness (Yule, 1944) to more recent and relevant works based on function words (Holmes et al., 2001), (Zhao and Zobel, 2005)), punctuation marks (Baayen et al., 2002), part-of-speech (POS) tags (Kukushkina et al., 2001), parse trees (Gamon, 2004) and character-based features (Kešelj et al., 2003).

There is an agreement among different researchers that function words are the most reliable indicator of authorship. There are two main reasons for using function words in lieu of other markers. First, because of their high frequency in a written text, function words are very difficult to consciously control, which minimizes the risk of false attribution. The second is that function words, unlike content words, are more independent from the topic or the genre of the text, so one should not expect to find great differences of frequencies across different texts written by the same authors on different topics (Chung and Pennebaker, 2007).
However, despite the fact that function word-based markers are state-of-the-art, they are basically relying on the *bag-of-words* assumption, which stipulates that text is a set of independent words. This approach completely ignores the fact that there is a syntactic structure and latent sequential information in the text. De Roeck (2004) has shown that frequent words, including function words, do not distribute homogeneously over a text. This provides evidence of the fact that the *bag-of-words* assumption is invalid. In fact, cri-

tiques have been made in the field of authorship attribution charging that many works are based on invalid assumptions (Rudman, 1997) and that researchers are focusing on attribution techniques rather than coming up with new style markers that are more precise and based on less strong assumptions. In an effort to develop more complex yet computationally feasible stylistic features that are more linguistically motivated, (Hoover, 2003) pointed out that exploiting the sequential information existing in the text could be a promising line of work. He proved that frequent word sequences and collocations can be used with high reliability for stylistic attribution.

In this contribution, we present a comparative study on using two different types of style marker based on function words for authorship attribution. Our aim is to compare the effectiveness of using a style marker that do not relay on the *bag-of-words* assumption to that of the frequency of function words which does. In this study, we used sequential rule of function words as style marker relying on the sequential information contained in structure of the text. We first give an overview of the sequential rule mining problem in section 2 and then describe our experimental setup in section 3. Finally, the results of the comparative study are presented in section 4.

2 Sequential rule extraction

Sequential data mining is a data mining subdomain introduced by (Agrawal et al., 1993) which is concerned with finding interesting characteristics and patterns in sequential databases. Sequential rule mining is one of the most important sequential data mining techniques used to extract rules describing a set of sequences. In what follows, for the sake of clarity, we will limit our definitions and annotations to those necessary to understand our experiment.

Considering a set of literals called items, denoted by $I = \{i_1, \ldots, i_n\}$, an itemset is a set of items $X \subseteq I$. A sequence S (single-item sequence) is an ordered list of items, denoted by $S = < i_1 \ldots i_n >$ where $i_1 \ldots i_n$ are items.

Table 1. Sequence database SDB.

Sequence ID	Sequence
1	< a, b, d, e >
2	< b, c, e >
3	< a, b, d, e >

A sequence database SDB is a set of tuples (id, S), where id is the sequence identifier and S a sequence. Interesting characteristics can be extracted from such databases using sequential rules and pattern mining

A sequential rule $R: X \Rightarrow Y$ is defined as a relationship between two itemsets X and Y such that $X \cap Y = \emptyset$. This rule can be interpreted as follow: if the itemset X occurs in a sequence, the itemset Y will occur afterward in the same sequence. Several algorithms have been developed to efficiently extract this type of rule, such as (Fournier-Viger and Tseng, 2011). For example, if we run this algorithm on the SDB containing the three sequences presented in Table 1, we will get as a result sequential rules, such us $a \Rightarrow d, e$ with support equal to 2, which means that this rule is respected by two sequences in the SDB (i.e., there exist two sequences of the SDB where we find the item a, we also find d and e afterward in the same sequence).

In our study, the text is first segmented into a set of sentences, and then each sentence is mapped into a sequence of function words appearing with order in that sentence. For example the sentence "J'aime ma maison où j'ai grandi." will be mapped to $< je, ma, où, je >$ as a sequence of French function words, and "$je \Rightarrow où$" ; "$ma \Rightarrow où, je$" are examples of sequential rules respected by this sequence. The whole text will produce a sequential database. The rules extracted in our study represent the cadence authors follow when using function words in their writing. This gives us more explanatory properties about the syntactic writhing style of a given author than just what frequencies of function words can do.

2.1 Classification Scheme

In the current approach, each text was segmented into a set of sentences based on splitting done using the punctuation marks of the set $\{'.', '!', '?', ':', '...'\}$, then function words were extracted from each sentence to construct a sequence. The algorithm described in (Fournier-Viger and Tseng, 2011) was then used to extract sequential rules of function words sequences from each text. Each text is then represented as a vector V_K of supports of rules, such that $V_K = \{r_1, r_2, \dots, r_K\}$ is the ordered set by decreasing normalized frequency of occurrence of the top K rules in terms of support in the training set. Each text is also represented by a vector of normalized frequencies of occurrence of function words. The normalization of the vector of frequency representing a given text was done by the size of the text.

Given the classification scheme described above, we used SVMs classifier to derive a discriminative linear model from our data. To get a reasonable estimation of the expected generalization performance, we used common measures: precision (P), recall (R), and F_1 score based on a 5-fold cross–validation as follows:

$$P = \frac{TP}{TP+FP} \qquad (1)$$

$$R = \frac{TP}{TP+FN} \qquad (2)$$

$$F_1 = \frac{2RP}{R+P} \qquad (3)$$

where TP are true positives, TN are true negatives, FN are false negatives, and FP are false positives.

3 Experimental Setup

3.1 Data Set

For the comparison experiment, we use texts written by: Balzac, Dumas, France, Gautier, Hugo, Maupassant, Proust, Sand, Sue and Zola. This choice was motivated by our special interest in studying the classic French literature of the 19th century, and the availability of electronic texts from these authors on the Gutenberg project website [1]and in the Gallica electronic library[2].

Table 2. Statistics for the data set used in our experiment

Author Name	**# of words**	**# of texts**
Balzac, Honoré de	548778	20
Dumas, Alexandre	320263	26
France, Anatole	218499	21
Gautier, Théophile	325849	19
Hugo, Victor	584502	39
Maupassant, Guy de	186598	20
Proust, Marcel	700748	38
Sand, George	560365	51
Sue, Eugène	1076843	60
Zola, Émile	581613	67

We collected 4 novels for each author, so that the total number of novels is 40. The next step was to divide these novels into smaller pieces of texts in order to have enough data instances to train the attribution algorithm.
Researchers working on authorship attribution on literature data have been using different dividing strategies. For example, Hoover (2003) decided to take just the first 10,000 words of each novel as a single text, while Argamon and Levitan (2005) treated each chapter of each book as a separate text. Since we are considering a sentence as a sequence unit, in our experiment we chose to slice novels by the size of the smallest one in the collection in terms of number

[1] http://www.gutenberg.org/
[2] http://gallica.bnf.fr/

of sentences.
More information about the data set used in the experiment is presented in Table 2.

4 Results

Results of measuring the attribution performance for the different feature sets presented in our experiment setup are summarized in Table 3. These results show in general a better performance when using function words frequencies, which achieved a nearly perfect attribution, over features based on sequential rules for our corpus.

Our study here shows that the SVMs classifier combined with features extracted using sequential data mining techniques can achieve a high attribution performance (That is, $F_1 = 0.947$ for Top 400 FW-SR). Until certain limit, adding more rules increases the attribution performance.

Table 3. 5-fold cross-validation results for our data set. FW-SR refers to **S**equential **R**ules of **F**unctions **W**ords.

Feature set	**P**	**R**	**F_1**
Top 100 FW-SR	0.901	0.886	0.893
Top 200 FW-SR	0.942	0.933	0.937
Top 300 FW-SR	0.940	0.939	0.939
Top 400 FW-SR	0,951	0,944	0,947
Top 500 FW-SR	0,947	0,941	0,943
FW frequencies	**0.990**	**0.988**	**0.988**

But contrary to common sense, function-word-frequency features, which fall under the *bag-of-word* assumption known to be blind to sequential information, outperform features extracted using sequential rule mining technique. In fact, they achieved nearly a perfect performance. We believe that this due to the presence of some parameters affecting the attribution process. These parameters, that need to be more deeply studied, depend on the linguistic character of the text, such as the syntactic and the lexical differences between narrative and dialogue texts. Finally, these results are in line with previous works that claimed that bag-of-words-based features are very effective indicator of the stylistic character of a text that can enable more accurate text attribution (Argamon and Levitan, 2005).

5 Conclusion

In this contribution, we present a comparative study on using two different types of style marker based on function words for authorship attribution. We compared the effectiveness of using sequential rules of function words as style marker that do not relay on the *bag-of-words* assumption to that of the frequency of function words which does. To evaluate the effectiveness of these markers, we conducted an experiment on a classic French corpus. Our results show that contrary to common sense, the frequencies of function words outperformed the sequential rules.

Based on the current study, we have identified several future research directions. First, we will explore the effectiveness of using probabilistic heuristics to find a minimal sequential rule set that still allows good attribution performance, which can be very useful for stylistic and psycholinguistic analysis. Second, this study will be expanded to include sequential patterns (n-grams with gaps) as sequential style markers. Third, we intend to experiment with this new type of style markers for other languages and text sizes using standard corpora employed in the field at large.

Acknowledgment

This work was supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02.

References

Agrawal, R., Imieli'nski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. *In ACM SIGMOD Record* (Vol. 22, pp. 207–216).

Argamon, S., and Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. I*n Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing.*

Baayen, H., van Halteren, H., Neijt, A., and Tweedie, F. (2002). An experiment in authorship attribution. *In 6th JADT* (pp. 29–37).

Burrows, J. (2002). "Delta": A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), 267–287.

Chung, C., and Pennebaker, J. W. (2007). The psychological functions of function words. *Social Communication*, 343–359.

De Roeck, A., Sarkar, A., and Garthwaite, P. (2004). Frequent Term Distribution Measures for Dataset Profiling. *In LREC*.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2003). Authorship attribution with support vector machines. *Applied Intelligence*, 19(1-2), 109–123.

Fournier-Viger, P., and Tseng, V. S. (2011). Mining top-k sequential rules. *In Advanced Data Mining and Applications* (pp. 180–194). Springer.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. *In Proceedings of the 20th international conference on Computational Linguistics* (p. 611).

Holmes, D. I., Robertson, M., and Paez, R. (2001). Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, 35(3), 315–331.

Hoover, D. L. (2003). Frequent collocations and authorial style. *Literary and Linguistic Computing*, 18(3), 261–286.

Kešelj, V., Peng, F., Cercone, N., and Thomas, C. (2003). N-gram-based author profiles for authorship attribution. *In Proceedings of the conference pacific association for computational linguistics, PACLING* (Vol. 3, pp. 255–264).

Khmelev, D. V, and Tweedie, F. J. (2001). Using Markov Chains for Identification of Writer. *Literary and Linguistic Computing*, 16(3), 299–307.

Koppel, M., and Schler, J. (2004). Authorship verification as a one-class classification problem. *In Proceedings of the twenty-first international conference on Machine learning* (p. 62).

Kukushkina, O. V, Polikarpov, A. A., and Khmelev, D. V. (2001). Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2), 172–184.

Ramyaa, C. H., and Rasheed, K. (2004). Using machine learning techniques for stylometry. *In Proceedings of International Conference on Machine Learning*.

Rudman, J. (1997). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4), 351–365.

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

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–556.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2), 193–214.

Yule, G. U. (1944). The statistical study of literary vocabulary. *CUP Archive*.

Zhao, Y., and Zobel, J. (2005). Effective and scalable authorship attribution using function words. *In Information Retrieval Technology* (pp. 174–189). Springer.