Authorship Attribution using Word Sequences

Rosa María Coyotl-Morales[1], Luis Villaseñor-Pineda[1],
Manuel Montes-y-Gómez[1] and Paolo Rosso[2]

Laboratorio de Tecnologías del Lenguaje,
Instituto Nacional de Astrofísica, Óptica y Electrónica, México.
{mcoyotl, villasen, mmontesg}@inaoep.mx
Departamento de Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia, España.
prosso@dsic.upv.es

Abstract. Authorship attribution is the task of identifying the author of a given text. The main concern of this task is to define an appropriate characterization of documents that captures the writing style of authors. This paper proposes a new method for authorship attribution supported on the idea that a proper identification of authors must consider both stylistic and topic features of texts. This method characterizes documents by a set of word sequences that combine functional and content words. The experimental results on poem classification demonstrated that this method outperforms most current state-of-the-art approaches, and that it is appropriate to handle the attribution of short documents.

1 Introduction

Authorship attribution is the task of identifying the author of a given text. It can be considered as a typical classification problem, where a set of documents with known authorship are used for training and the aim is to automatically determine the corresponding author of an anonymous text. In contrast to other classification tasks, it is not clear which features of a text should be used to classify an author. Consequently, the main concern of computer-assisted authorship attribution is to define an appropriate characterization of documents that captures the writing style of authors.

There are several methods for authorship attribution, ranging from those using stylistic non-topic features such as the vocabulary richness of the author and the frequency of occurrence of some functional words[1] [12], to those based on the traditional bag-of-words representation that consider all content words of documents [5, 8]. In this paper, we propose a new method for authorship attribution. This method relies on the hypothesis that a proper identification of authors must consider both stylistic and topic features of texts. Therefore, an adequate characterization of documents must effectively combine functional and content words. Our proposal is to construct this characterization by means of word sequences.

[1] Words having little semantic content of their own, such as prepositions, conjunctions, and articles. In information retrieval, they are also known as stopwords.

It is important to mention that word sequences (specially, fixed-length word *n*-grams) have been applied without much success in topic-based text classification [3]. Nevertheless, there are not enough studies on their application to non-topic-based classification, and in particular to the task of authorship attribution [10].

On the other hand, other less studied difficulty is the impact of the document size on the classification accuracy. It is known that some approaches for authorship attribution are very sensible to the length of documents. Specially, the methods based on stylistic features tend to fail when confront short documents [11]. This behavior motivates us to apply our method on the classification of poems by authors. Given that poems are very short documents, our experiments not only contribute to evaluate the usefulness of word sequence features for authorship attribution, but also allow analyzing their appropriateness to handle difficult classification scenarios.

The rest of the paper is organized as follows. Section 2 discusses some previous works related to the task of authorship attribution. Section 3 introduces the proposed method. Section 4 describes the experimental setup. Section 5 presents some experimental results on the use of word sequences features. Finally, section 6 depicts our conclusions and future work.

2 Related Work

The analysis of style for authorship attribution is mainly based on the assumption that each author has habits in wording (i.e., in the use of words) that make their writing unique. However, this assumption is not completely true, since the style of an author may be variable depending on the target audience, or may change because of differences in topics or genre. For this reason, it is difficult to determine a set of features stable to these variations but adequate to distinguish between writings of different authors.

There are several methods for authorship attribution. These methods may be cluster in the following three main approaches:

Stylistic measures as document features. This approach considers features such as the length of words and sentences as well as the richness of the vocabulary [7, 9]. It results are not conclusive, but have demonstrated that these features are not sufficient for the task. It seems that they vary depending on the genre of the text, and that they lost most of their meaning when dealing with short texts.

Syntactic cues as document features. This approach uses a set of style markers. These markers go beyond the stylistic measures by integrating information related to structure of the language, which is obtained by an in depth syntactic analysis of documents [4, 5, 11]. Basically, texts are characterized by the presence and frequency of certain syntactic structures. This characterization is very detailed and relevant; unfortunately, it is computationally expensive and even impossible to build for languages lacking of text-processing resources (e.g. POS tagger, syntactic parser, etc.). Besides, it is also clearly influenced by the length of documents.

Word-based document features. This approach includes at least three different kinds of methods. The first one characterizes documents using a set of functional words, ignoring the content words since they tend to be highly correlated with the document topics [2, 12]. This method works properly, but it is also affected by the size of documents. In this case, the document length not only influences the fre-

quency of occurrence of the functional words but also their sole presence. The second method applies the traditional bag-of-words representation and uses single content-words as document features [5, 8]. It is very robust and produces excellent results when there is a noticeable relation between authors and topics. Finally, a third method considers word *n*-gram features, i.e., features consisting of sequences of *n* consecutive words. This method attempts to capture the language structure of texts by simple word sequences instead of by complex syntactic structures [10]. Somehow, it purpose is to obtain a rich characterization of texts without performing an expensive syntactic analysis. Nevertheless, due to the feature explosion, it tends to use only *n*-grams up to three words.

In general, our method is very similar to the *n*-gram based approach. In both cases, documents are characterized by a combination of function and content words. However, ours considers a special kind of word sequences (namely, maximal frequent word sequences), which are determined by their frequency of occurrence instead of by their length. Using this strategy, it selects the most relevant word sequences, and indirectly tackles the problem of feature explosion. The following section describes in detailed the proposed method.

3 Our Method

As we previously mentioned, this paper presents a new method for authorship attribution. This method characterizes documents by a set of relevant sequences that combine functional and content words. The idea is to use these sequences to classify the documents in view that they express the more significant lexical collocations[2] used by an author. Traditionally, these sequences are extracted by applying a general *n*-gram calculus. In contrast, we propose to discover them by means of a process for mining maximal frequent word sequences.

The following subsections define the maximal frequent word sequences, the process for their extraction, as well as a classification algorithm using them as document features.

3.1 Mining Maximal Frequent Word Sequences

Assume that D is a set of texts (a text may represent a complete document or even just a single sentence), and each text consists of a sequence of words. Then, we have the following definitions [1].

Definition 1. A sequence $p = a_1 \ldots a_k$ is a *subsequence* of a sequence q if all the items a_i, $1 \leq i \leq k$, occur in q and they occur in the same order as in p. If a sequence p is a subsequence of a sequence q, we also say that p occurs in q.

Definition 2. A sequence p is *frequent* in D if p is a subsequence of at least σ texts of D, where σ is a given frequency threshold.

Definition 3. A sequence p is a *maximal frequent sequence* in D if there does not exist any sequence p' in D such that p is a subsequence of p' and p' is frequent in D.

---

[2] A collocation is defined as a sequence of words or terms that co-occur more often than would be expected by chance.

Once introduced the maximal frequent word sequences, the problem of mining them can formally state as follows: Given a text collection D and an arbitrary integer value σ such that $1 \leq \sigma \leq |D|$, enumerate all maximal frequent word sequences in D.[3]

It is important to mention that the implementation of a method for sequence mining is not a trivial task because of its computational complexity. The algorithm used in our experiments is described in [6].

3.2 Classification Algorithms

Authorship attribution is a classification problem, where a set of documents with known authorship are used for training and the aim is to automatically determine the corresponding author of an anonymous text. Table 1 shows a direct classification algorithm based on the use of maximal frequent word sequences as document features.

Table 1. Direct Algorithm

Let D_T be the set of labeled documents that will be used for training
Let d be an anonymous document

TRAINING

1. Set the value of the frequency threshold σ
2. Enumerate all maximal frequent word sequences in D_T corresponding to the given frequency threshold
3. Build the training instances using the discovered word sequences as Boolean features
4. Give the learning algorithm the training instances and perform training

CLASSIFICATION

1. Build the representation of d in accordance to the training feature space
2. Let the trained classifier label the new instance

The proposed direct algorithm is conceptually simple and appropriate. However, it greatly depends on the adequate definition of the frequency threshold σ. It is expected that different values of σ generate different sets of word sequences, and consequently produce different performance rates. For instance, low σ-values allow extracting large sequences and favor the precision rate, while high σ-values tend to generate many short sequences that support the recall percentage. Unfortunately, the most adequate σ-value is influenced by the size of the given document collection, and therefore it need to be empirically determined for each particular situation.

In order to reduce the dependency of the classification performance to the used frequency threshold, we propose to construct the feature set by combining the maximal frequent sequences extracted by different σ-values. The idea is to construct the feature set by an iterative process, incrementing the σ-value at each step. This process starts with the inclusion of sequences corresponding to the frequency threshold $\sigma = 2$, and ends when there are not more lexical collocations (sequences of at least two words) to aggregate to the feature set. Table 2 describes the enhanced algorithm.

[3] It is important to notice that a maximal frequent sequence may consist of only one single word.

Table 2. Enhanced Algorithm

Let D_T be the set of labeled documents that will be used for training
Let d be an anonymous document

TRAINING

1. Set the value of the frequency threshold $\sigma = 2$
2. Set the feature set $F_1 = \{\varnothing\}$
3. DO
   a. Enumerate all maximal frequent word sequences in D_T corresponding to the frequency threshold σ. Name the set of sequences S_σ
   b. Integrate new sequences to the feature set, i.e., $F_\sigma = F_{\sigma\text{-}1} \cup S_\sigma$
   c. Increment the frequency threshold; i.e., $\sigma = \sigma + 1$

   WHILE ($S_{\sigma\text{-}1}$ contain at least one sequence of two or more words not included in $F_{\sigma\text{-}2}$)
4. Build the training instances using the discovered Boolean features
5. Give the learning algorithm the training instances and perform training

CLASSIFICATION

1. Build the representation of d in accordance to the training feature space
2. Let the trained classifier label the new instance

4 Experimental Setup

4.1 Corpus

Unfortunately, there is not a standard data set for evaluating authorship attribution methods. Therefore, we had to assemble our own corpus. This corpus was gathered from the Web. It consists of 353 poems writing by five different authors. Table 3 resumes some statistics about this corpus. It is important to notice that, on the one hand, the collected poems are very short documents (176 words in average), and on the other hand, that all of them correspond to contemporary Mexican poets. In particular, we were very careful on selecting modern writers in order to avoid the identification of authors by the use of anachronisms.

Table 3. Corpus Statistics

Poets	Number of documents	Size of Vocabulary	Number of Phrases	Average Words by Documents	Average Phrases by Documents
Efraín Huerta	48	3831	510	236.5	22.3
Jaime Sabines	80	3955	717	155.8	17.4
Octavio Paz	75	3335	448	162.6	27.2
Rosario Castellanos	80	4355	727	149.3	16.4
Rubén Bonifaz	70	4769	720	178.3	17.3

4.2 Classifier

The Naïve Bayes classifier has proved to be quite competitive for most text processing tasks including text classification. This fact supported our decision to use it as main classifier for our experiments. It basically computes the probability of a document d to belong to a category c_i given the set of features $F = \{f_1, f_2, \ldots, f_{|F|}\}$.[4] This probability can be expressed using Bayes' rule as follows:

$$P(c_i \mid d) = \frac{P(d \mid c_i)P(c_i)}{P(d)}$$

Simplifying and assuming statistical independence of the features:

$$P(c_i \mid d) = P(c_i)\prod_{j=1}^{|F|} P(f_j \mid c_i)$$

These probabilities can be estimated directly from the training set as follows:

$$P(c_i) = \frac{N_i}{N}, \quad P(f_j \mid c_i) = \frac{1 + N_{ji}}{|F| + \sum_{k=1}^{|F|} N_{ki}}$$

where N is the number of documents in the whole collection, N_i the number of documents of category c_i, and N_{ji} the number of documents from category c_i having the feature f_j. Finally, $|F|$ indicates the number of features.

4.3 Baseline Configurations

Because of the difficulty of comparing our approach with other previous works –mainly caused by the absence of a standard evaluation corpus–, we performed several experiments in order to establish a baseline. These experiments consider the use of four different kinds of word-based features: (i) functional words, (ii) content words, (iii) the combination of functional and content words, and (iv) word *n*-grams. Table 4 shows the results corresponding to each one of these approaches.

Table 4. Baseline Configurations

Features	Accuracy	Average Precision	Average Recall
Functional words	41.0%	0.42	0.39
Content words	73.0%	0.78	0.73
All kind of words	73.0%	0.78	0.74
n-grams (unigrams plus bigrams)	78.8%	0.84	0.79
n-grams (from unigrams to trigrams)	76.8%	0.84	0.77

It is important to mention that because our main interest was to determine an appropriate document characterization for authorship attribution, we used in all cases the same classification algorithm, namely, the naïve Bayes classifier. As well, we applied the same technique for dimensionality reduction (information gain) and the same evaluation schema (a 10-cross-fold validation).

[4] Text classification is the problem of assigning a document d to one of a set of $|C|$ predefined categories $C = \{c_1, c_2, \ldots, c_{|C|}\}$.

The results shown in table 4 are very interesting since they confirm some of our major assumptions. First, functional words by themselves do not help to capture the writing style from short documents. Second, content words contain some relevant information to distinguish between authors, even when all documents correspond to the same genre and discuss similar topics. Third, the lexical collocations, captured by word *n*-gram sequences, are useful for the task of authorship attribution. Fourth, due to the feature explosion and the small size of the corpus, the use of higher *n*-gram sequences not necessarily improves the classification performance.

5 Experimental Results

In this paper, we have proposed the use of *maximal frequent word sequences* as document features for authorship attribution. This section presents the results of two basic experiments. The first one evaluates the classification performance of the direct algorithm using different frequency thresholds (σ). The second experiment applies the enhanced algorithm. It goal is to evaluate the impact of using a feature set that combines maximal sequences extracted by different σ-values.

In these experiments, as in the baseline generation, we used sequences considering not only content words, but also function words as well as punctuation marks. In the same way, we used the naïve Bayes classifier, the information gain technique for dimensionality reduction[5], and a 10-cross-fold validation schema.

5.1 Experiments with the Direct Algorithm

Table 5 shows the results obtained using different frequency threshold values. It can be noticed that for all σ-values our results were worst than those obtained using the *n*-gram features (combining unigrams and bigrams). However, it is interesting to point out that number of sequences –for the best case– was much less than the number of *n*-grams, 4276 and 45245 respectively. Moreover, after the dimensionality reduction, the number of sequences was less than the number of *n*-grams, 203 and 455 respectively. This condition indicates that even when our method did not outperform the *n*-gram based approach, it could obtain a reduced set of features with better discrimination capacity.

Table 5. Results of the Direct Algorithm

σ	Number of Sequences	Average Words per Sequence	Accuracy	Average Precision	Average Recall
2	141	2.59	68.60%	0.76	0.69
3	203	2.32	77.30%	0.82	0.77
4	225	2.26	77.30%	0.82	0.77
5	195	1.67	77.10%	0.81	0.77
6	156	1.59	75.40%	0.79	0.75
7	129	1.57	74.80%	0.78	0.74
8	124	1.50	74.20%	0.76	0.74
9	105	1.46	71.40%	0.73	0.71
10	94	1.45	70.50%	0.72	0.70

[5] In particular, we selected all attributes with information gain greater than 1.

In addition, the results of table 5 demonstrate the great influence of the frequency threshold on the classification process. It is clear that the σ-value determines the number and kind of discovered sequences, and therefore, it has a direct effect on the overall classification performance. In particular, it is noticeable that the accuracy decreases while increasing the frequency threshold. This is because high σ-values tend to fragment sequences, losing several relevant lexical collocations.

5.2 Experiment using the Enhanced Algorithm

The enhanced algorithm (refer to section 3.2) constructs the feature set by combining maximal frequent sequences corresponding to different σ-values. In this way, it attempts diminishing the dependency of the classification performance on the used frequency threshold. Table 6 gives some data on the construction of the feature set. This process started with the inclusion of large sequences (those having more discriminatory capacity) and ended with the insertion of short sequences (those having more coverage). In total, we assembled a set of 425 features.

Tabla 6. Construction of the Enhanced Feature Set

σ	Extracted Sequences	Added Sequences	Average Length of Added Sequences	Number of Features
2	141	141	2.58	141
3	203	100	1.71	241
4	225	80	1.76	321
5	195	53	1.74	374
6	156	23	1.35	397
7	129	13	1.46	410
8	124	12	1.25	422
9	105	3	1	425

Table 7 shows the results related to the enhanced algorithm. From these results, it is clear that the enhanced algorithm not only does better than the direct algorithm, but also that it outperforms all baseline configurations. Furthermore, given that the resultant feature set is comparable in size to the *n*-gram set, the obtained results validate our hypothesis that determining the word sequences by their frequency of occurrence instead of by their length is a good strategy, which allows to select the most relevant word sequences and to tackle the problem of feature explosion.

Table 7. Results of the Enhanced Algorithm

Poets	Precision	Recall
Efraín Huerta	1.00	0.75
Jaime Sabines	0.83	0.83
Octavio Paz	0.95	0.75
Rosario Castellanos	0.65	0.91
Ruben Bonifaz	0.94	0.87
Average Rates	0.87	0.82
Overall Accuracy	83%	

6 Conclusions

In this paper, we proposed a new method for authorship attribution. This method is supported on the idea that a proper identification of author must consider both stylistic and topic features of documents. In particular, it characterizes the documents by a set of word sequences that combine functional and content words.

Other previous approaches for authorship attribution also characterized documents by word sequences. Specifically, they used word *n*-gram features, that is, word sequences of a fixed predefined size. In contrast to these approaches, our method considers a special kind of word sequences (namely, maximal frequent word sequences), which are determined by their frequency of occurrence instead of by their length. The experimental results demonstrated that this kind of sequences are superior to the *n*-grams, since they allow capturing the more significant lexical collocations used by an author.

It is also important to mention that our method, without using any sophisticated linguistic analysis of texts, could outperform most of the state-of-the-art approaches for authorship attribution. Furthermore, our method, contrary to other current approaches, is not very sensitive to the size of documents and the document collection.

As future work, we plan to apply the proposed method (document characterization) to other problems of text classification. In particular, we want to investigate the contribution of function words to topic-based text classification.

Acknowledgements

This work was done under partial support of CONACYT (project grants 43990 and U39957-Y, SEPSEBYN-C01-40), R2D2 CICYT (TIC2003-07158-C04-03) and ICT EU-India (ALA/95/23/2003/077-054).

References

1. Ahonen-Myka H. (2002). *Discovery of Frequent Word Sequences in Text Source*. Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery. London, UK, 2002.
2. Argamon, S. & Levitan, S. (2005). *Measuring the Usefulness of Function Words for Authorship Attribution.* Association for Literary and Linguistic Computing/ Association Computer Humanities, University Of Victoria, Canada.
3. Bekkerman, R. & Allan, J. Using (2004). *Bigrams in Text Categorization*. CIIR Technical Report IR-408 Center for Intelligent Information Retrieval, University of Massachusetts Amherst.
4. Chaski, C. (2005). Who's at the Keyword? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence*. Volume 4, Issue 1.
5. Diederich, J., Kindermann, J., Leopold, E. & Paas, G. (2003). *Authorship Attribution* with Support Vector Machines. Applied Intelligence, 19(1):109-123, 2003.
6. García-Hernández, R., Martínez-Trinidad F., and Carrasco-Ochoa A. (2006). *A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection*. International Conference on Computational Linguistics and text Processing, CICLing-2006. Mexico City, Mexico, 2006.
7. Holmes, D. (1995). Authorship Attribution. *Computers and the Humanities*, 28:87-106. Kluwer Academic Publishers. 1995.

8. Kaster, A., Siersdorfer, S., & Weikum, G. (2005). *Combining Text and Linguistic Document Representations for Authorship Attribution*. Workshop Stylistic Analysis of Text for Information Access, 28th Int. SIGIR 1. MPI, Saarbrücken 2005, 27-35.
9. Malyutov, M.B. (2004). Authorship Attribution of Texts: a Review. *Proceedings of the program "Information transfer" held in ZIF*. University of Bielefeld, Germany. 2004. 17 pages.
10. Peng, F., Schuurmans, D., Keselj, V. & Wang, S. (2004). Augmenting Naïve Bayes Classifiers with Statistical Languages Models. *Information Retrieval*, vol. 7, 317-345. Kluwer Academic Publishers. 2004.
11. Stamatatos, E., Fakotakis, N. & Kokkinakis,G. Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities* 35: 193-214, 2001. Kluwer Academic Publishers. 2001
12. Zhao, Y. & Zobel, J. (2005). Effective and Scalable Authorship Attribution Using Function Words. *Lecture Notes in Computer Science*, vol. 3689, 174-189. Springer Verlag. 2005.