A Web-based Self-training Approach for Authorship Attribution[1]

Rafael Guzmán-Cabrera[1,2], Manuel Montes-y-Gómez[3],
Paolo Rosso[2], Luis Villaseñor-Pineda[3]

[1]FIMEE, Universidad de Guanajuato, México
guzmanc@salamanca.ugto.mx
[2]NLE Lab, DSIC, Universidad Politécnica de Valencia, Spain
prosso@dsic.upv.es
[3]LabTL, Instituto Nacional de Astrofísica, Óptica y Electrónica, México
{mmontesg, villasen}@inaoep.mx

Abstract. As any other text categorization task, authorship attribution requires a large number of training examples. These examples, which are easily obtained for most of the tasks, are particularly difficult to obtain for this case. Based on this fact, in this paper we investigate the possibility of using Web-based text mining methods for the identification of the author of a given poem. In particular, we propose a semi-supervised method that is specially suited to work with just few training examples in order to tackle the problem of the lack of data with the same writing style. The method considers the automatic extraction of the unlabeled examples from the Web and its iterative integration into the training data set. To the knowledge of the authors, a semi-supervised method which makes use of the Web as support lexical resource has not been previously employed in this task. The results obtained on poem categorization show that this method may improve the classification accuracy and it is appropriate to handle the attribution of short documents.

1 Introduction

Nowadays, there is a lot of information available in digital format. This situation has produced a growing need for tools that help people to find, organize and analyze all these resources. In particular, text categorization [14], the automatic assignment of free text documents to one or more predefined categories, has emerged as a very important component in many information management tasks. Most of these tasks are of thematic nature, such as newswire and spam filtering, whereas some others are non-thematically restricted, for instance, authorship attribution and sentiment classification.

The state-of-the-art approach for automatic text categorization considers the application of a number of statistical and machine learning techniques, including Bayesian classifiers, support vector machines, nearest neighbour classifiers and arti-

[1] This work was done under partial support of CONACYT-Mexico (43990, C01-39957), MCyT-Spain (TIN2006-15265-C06-04) and PROMEP (UGTO-121).

ficial neural networks [14]. A major difficulty with this kind of supervised techniques is that they commonly require a great number of labelled examples (training instances) to construct an accurate classifier. Unfortunately, because a human expert must manually label these examples, the training sets are extremely small for many application domains. In order to overcome this problem, recently many researchers have been working on semi-supervised learning algorithms (for an overview see [15]). It has been showed that by augmenting the training set with additional unlabelled information it is possible to improve the classification accuracy using different learning algorithms such as naïve Bayes [12], support vector machines [8], and nearest-neighbour algorithms [19].

In line with these current works, we have proposed a new semi-supervised method for text categorization [5, 6]. This method differs from previous approaches in two main issues. On the one hand, it does not require a predefined set of unlabelled training examples, instead it considers their automatic extraction from the Web. On the other hand, it applies a self-training approach that selects instances not only considering their labelling confidence by a base classifier, but also their correspondence with a web-based labelling[2]. This method has been applied with success in thematic text classification tasks, indicating that it is possible to automatically extract discriminative thematic information from the Web. The method was evaluated on training sets of different sizes demonstrating its usefulness for dealing with very small data sets. As an example of this fact, our method improved the categorization of natural disaster news by 26% using a naïve Bayes classifier and a small training set with 10 examples per class [5].

In this paper, we investigate the application of the proposed web-based self-training method in a non-thematic classification task, namely, authorship attribution. This task confronts the method with new challenges since an author may write about several topics as well as a topic may be treated by different authors. Therefore, in this task, words by themselves do not allow distinguishing among classes; it is necessary to take into account how words are used together (i.e., the author's writing style). In order to make harder the evaluation, we focus our experiments on poem classification where documents are usually very short and their vocabulary and structure are very different from everyday –web– language.

The rest of the paper is organized as follows. Section 2 introduces the task of authorship attribution and discusses some representative works. Section 3 describes our web-based self-training approach for text classification. Then, Section 4 presents some evaluation results on poem classification by author. Finally, Section 5 depicts our conclusions.

2 Authorship Attribution

Authorship attribution is the task of identifying the author of a given text. It can be considered as a classification problem, where a set of documents with known author-

[2] Given that each unlabelled example is downloaded from the Web using a set of automatically defined class queries, each of them has a default category or web-based label.

ship are used for training, and the aim is to automatically determine the corresponding author of an anonymous text.

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

Stylometric 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, 10]. Its results are not conclusive, but they have shown 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 stylometric measures by integrating information related to the structure of the language, which is obtained by an in depth syntactic analysis of documents [2, 4, 17]. Mainly, 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 robust 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 content words since they tend to be highly correlated with the document topics [1, 21]. This kind of methods works properly, but it is also affected by the size of documents. In this case, the document length not only influences the frequency of occurrence of the functional words but also their sole presence. The second kind of methods applies the traditional bag-of-words representation and uses single content-words as document features [9]. It is very robust and produces excellent results when there is a noticeable relation between authors and topics. Finally, a third kind of method considers word *n*-gram features, i.e., features consisting of sequences of *n* consecutive words. It attempts to capture the language structure of texts by simple word sequences instead of by complex syntactic structures [13]. Somehow, its 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 contrast to all these works, this paper does not propose another document representation for authorship attribution, it describes instead a new semi-supervised learning method that allows working with small training sets. As expected, our web-based self-training classification method may be applied along with all these kinds of features. However, given that our interest is to have a general approach for authorship attribution that allows analyzing documents of different sizes and domains, we have decided to mainly explore the use of word-based features, in particular, *n*-grams.

3 Our Text Categorization Method

Figure 1 shows the general scheme of our semi-supervised text classification method. It consists of two main processes. The first one deals with the corpora acquisition

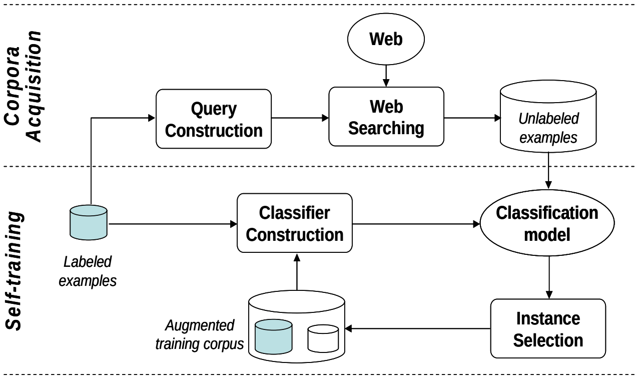

Figure 1. General overview of our classification method

from the Web, whereas the second one focuses on the self-training learning approach [11]. The following sections describe in detail these two processes.

3.1 Corpora Acquisition

This process considers the automatic extraction of unlabeled examples from the Web. In order to do this, it first constructs a number of queries by combining the most significant words for each class; then, using these queries, it looks at the Web for some additional training examples related to the given classes.

Query Construction. In order to form queries for searching the Web, it is necessary to previously determine the set of relevant words for each class in the training corpus. The criterion used for this purpose is based on a combination of frequency of occurrence and information gain of words. We consider that a word w_i is relevant for a class C if it satisfies the following two conditions:

1. The frequency of occurrence of w_i in C is greater than the average occurrence of all words (happening more than once) in that class. That is:

$$f_{w_i}^{C} > \frac{1}{|C'|} \sum_{\forall w \in C'} f_{w}^{C} \text{ , where } C' = \left\{ w \in C \middle| f_{w}^{C} > 1 \right\}$$

2. The information gain of w_i is positive, that is $IG_{w_i} > 0$.

Once obtained the set of relevant words per class, it is possible to construct the corresponding set of queries. Founded on the method by Zelikovitz and Kogan [20], we decide to construct queries of three words. This way, we create as many queries per class as all three-word combinations of its relevant words. We measure the significance of a query $q = \{w_1, w_2, w_3\}$ to the class C as indicated below:

$$\Gamma_C(q) = \sum_{i=1}^{3} f_{w_i}^{C} \times IG_{w_i}$$

Web Searching. The next action is using the defined queries to extract from the Web a set of additional unlabeled text examples. Based on the observation that most significant queries tend to retrieve the most relevant web pages, our method for searching the Web determines the number of downloaded examples per query in a direct proportion to its Γ-value. Therefore, given a set of M queries $\{q_1,\ldots, q_M\}$ for class C, and considering that we want to download a total of N additional examples per class, the number of examples to be extracted by a query q_i is determined as follows:

$$\Psi_C(q_i) = \frac{N}{\sum_{k=1}^{M} \Gamma_C(q_k)} \times \Gamma_C(q_i)$$

3.2 Self-training learning

As we previously mentioned, the purpose of this process is to increase the classification accuracy by gradually augmenting the originally small training set with the examples downloaded from the Web. Our algorithm for self-training learning is an adaptation of a method proposed elsewhere [16]. It mainly considers the following steps:

1. Build a weak classifier (C_l) using a specified learning method (l) and the training set available (T).
2. Classify the unlabeled web examples (E) using the constructed classifier (C_l). In order words, estimate the class for all downloaded examples.
3. Select the best m examples ($E_m \subseteq E$) based on the following two conditions:
    a. The estimate class of the example corresponds to the class of the query used to download it. In some way, this filter works as an ensemble of two classifiers: C_l and the Web (expressed by the set of queries).
    b. The example has one of the m-highest confidence predictions.
4. Combine the selected examples with the original training set ($T \leftarrow T \cup E_m$) in order to form a new training set. At the same time, eliminate these examples from the set of downloaded instances ($E \leftarrow E - E_m$).
5. Iterate σ times over steps 1 to 4 or repeat until $E_m = \varnothing$. In this case σ is a user specified threshold.
6. Construct the final classifier using the enriched training set.

4 Evaluation on Authorship Attribution

4.1 Experimental Setup

Corpus. Given that there is not a standard data set for evaluating authorship attribution methods, we had to assemble our own corpus. This corpus was gathered from the Web and consists of 353 poems written by five different authors [3]. Table 1 resumes some statistics about this corpus. It is important to notice that, on the one hand, the collected poems are very short texts (172 words in average), and on the other hand, that all of them correspond to contemporary Mexican poets. In particular, we were very careful in selecting modern writers in order to avoid the identification of authors by the use of anachronisms.

Table 1. Corpus Statistics

Poets	Number of documents	Word forms	Word tokens	Number of Phrases	Average Word Tokens by Document	Average Phrases by Document
Efraín Huerta	48	3831	11352	510	236.5	22.3
Jaime Sabines	80	3955	12464	717	155.8	17.4
Octavio Paz	75	3335	12195	448	162.6	27.2
Rosario Castellanos	80	4355	11944	727	149.3	16.4
Rubén Bonifaz	70	4769	12481	720	178.3	17.3

Baseline Configurations. Because of the difficulty of comparing our approach with other previous works (mainly because of 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 2 shows the results corresponding to each one of these kinds of word-based features.

Table 2. Baseline Configurations

Features	Accuracy	Macro Average Precision	Average Recall
Functional words	0.41	0.42	0.39
Content words	0.73	0.78	0.73
All kind of words	0.73	0.78	0.74
n-grams (unigrams plus bigrams)	0.78	0.84	0.79
n-grams (from unigrams to trigrams)	0.76	0.84	0.77

Our main interest in this first experiment was to determine a baseline configuration for our subsequent experiments. Because of that, we used in all cases the same classification algorithm (namely, the naïve Bayes classifier), the same technique for dimensionality reduction (information gain) as well as the same evaluation schema (a 10-cross-fold validation). In all experiments, we used the implementations facilitated by the WEKA machine-learning environment [18].

The results shown in Table 2 are very interesting since they confirm some of our major assumptions. First, functional words by themselves do not help to capture the writing style of short texts. Second, content words contain some relevant information to distinguish among 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.

4.2 Experimental Results

This section describes the application of the proposed semi-supervised method to the task of authorship attribution. The method, as depicted in Section 3, includes two main processes: the corpora acquisition from the Web and the self-training learning approach. Following, we detail some results from both of them.

The central task for corpora acquisition is the automatic construction of a set of queries that expresses the relevant content of each class. Using these queries, we collected from the Web a set of 2,400 snippets per class, obtaining 12,000 additional unlabeled examples. Then, we applied the self-training method for constructing the final poem classifier.

It is important to point out that there is not a clear criterion to determine the parameters *m* and σ of a self-training method [11]. In our case, we determined the number of unlabeled examples that must be incorporated into the training set at each iteration based on the following condition: the added information –expressed in number of words– must be proportionally small with respect to the original training data. This last condition is very important because of the small size of poems (176 words on average). In particular, we decided to incorporate 60 unlabeled examples per iteration (m = 60), approximately 10 examples per class. However, it is necessary to perform further experiments in order to determine the best value of *m* for this task.

Table 3. Training/test data sets

Poets	Training Set	Test Set	Word forms (in Training Set)
Efraín Huerta	38	10	2827
Jaime Sabines	64	16	2749
Octavio Paz	60	15	2431
Rosario Castellanos	64	16	3280
Rubén Bonifaz	56	14	3552
Total	***282***	***71***	***8377***

For this new experiment, we organized the corpus in a different way with respect to the baseline experiment described in Section 4.1 The corpus was divided in two data sets: training (with 80% of the labelled examples) and test (with 20% of the examples). The idea was to carry out the experiment in an almost-real situation, where it is not possible to know in advance all the vocabulary. This is a very important aspect to take into account in poem classification since poets tend to employ a very rich vocabulary. Table 3 shows some numbers about this collection.

Taking into account the results described in the previous section, we decided to use *n*-grams as document features. We mainly performed two different experiments. In the first one we used bigrams as features, whereas in the second one we used trigrams. Table 4 shows the results corresponding to the first five iterations of the method. As can be observed, the integration of new information improved the baseline results. In particular, the best result was obtained at the second iteration when using bigrams. We suppose this behaviour was due because bigrams are better suited to look for the most used collocations of an author from a small corpus; for trigrams –we presume– it is necessary to have more information.

Table 4 also shows the vocabulary's growing: aproximately 300 new words per iteration. Due to this increment it was possible to correctly classify more poems from the very first iteration. However, this increment was also the reason for the accuracy decrement in subsequent iterations where several non-relevant words were inserted into the training set.

Table 4. Accuracy percentages after the training corpus enrichment

***n*-grams**	**Initial Accuracy**	**Iteration**				
		1	**2**	**3**	**4**	**5**
Bigrams	78.9	80.3	**82.9**	80.3	78.9	78.9
Trigrams	74.6	74.7	78.8	**80.3**	80.3	78.7
Vocabulary Size	8377	8732	9019	9319	9676	9915

Although being preliminary results, it is surprising to verify that it is feasible to extract useful examples from the Web for the task of authorship attribution. In fact, our intuition suggested the opposite: given that poems tend to use rare and improper word combinations, the Web seemed not to be an adequate source of relevant information for this task.

5 Conclusions

This paper proposed a novel approach for authorship attribution based on a web-based self-training learning method. This method differs from others in that: (i) it is specially suited to work with few training examples, and (ii) it considers the automatic extraction of additional training knowledge from the Web.

In general, the achieved results allow us to formulate the following preliminary conclusions:

- Our web-based self-training classification method seems to be portable to non-thematic tasks. In particular, the achieved results in authorship attribution support this observation.
- The proposed method for authorship attribution, which uses *n*-gram features and a semi-supervised learning approach, could outperform most common approaches for authorship attribution. Furthermore, our method, contrary to other current approaches, is not affected by the small size of the texts, and avoids using any sophisticated linguistic analysis of documents.
- The proper identification of an author, even from a poem, must consider both stylometric and topic features of documents. Therefore, our conclusion points to use word-based features such as word *n*-grams.

Finally, it is important to comment that it is necessary to achieve a detailed analysis of current results as well as to perform further experiments in order to define better empirical criteria for selecting the values of the parameters m and σ.

References

1. Argamon S., and Levitan, S., Measuring the Usefulness of Function Words for Authorship Attribution. Association for Literary and Linguistic Computing/ Association Computer Humanities, University of Victoria, Canada, 2005.
2. Chaski C., Who's at the Keyword? Authorship Attribution in Digital Evidence Investigations. International Journal of Digital Evidence. Volume 4, Issue 1, 2005.
3. Coyotl-Morales R. M., Villaseñor-Pineda L., Montes-y-Gómez M. and Rosso P. Authorship Attribution using Word Sequences. Lecture Notes in Computer Science, vol. 4225, Springer, 2006.
4. Diederich J., Kindermann J., Leopold E., and Paas G., Authorship Attribution with Support Vector Machines. Applied Intelligence, 19(1):109-123, 2003.
5. Guzmán-Cabrera R., Montes-y-Gómez M., Rosso P., Villaseñor-Pineda L, Improving Text Classification using Web Corpora. 5th Atlantic Web Intelligence Conference, AWIC 2007. Advances in Soft Computing, Num. 43, Springer, 2007.
6. Guzmán-Cabrera R., Montes-y-Gómez M., Rosso P., Villaseñor-Pineda L., Taking Advantage of the Web for Text Classification with Imbalanced Classes. MICAI 2007. Lecture Notes in Artificial Intelligence 4827, Springer, 2007.
7. Holmes D., Authorship Attribution. Computers and the Humanities, 28:87-106. Kluwer Academic Publishers. 1995.
8. Joachims T., Transductive inference for text classification using support vector machines, Proceedings of the Sixteenth International Conference on Machine Learning, 1999.
9. Kaster A., Siersdorfer S., and Weikum G., Combining Text and Linguistic Document Representations for Authorship Attribution. Workshop Stylistic Analysis of Text for Information Access, 28th Int. SIGIR 1. MPI, Saarbrücken 2005.
10. Malyutov M. B., Authorship Attribution of Texts: a Review. Proceedings of the program "Information transfer" held in ZIF. University of Bielefeld, Germany, 2004.
11. Mihalcea R., Co-training and Self-training for Word Sense Disambiguation. Proc. of the Conference on Natural Lenguage Learning (CoNLL 2004), Boston, USA, 2004.
12. Nigam K., Mccallum A. K., Thrun S., and Mitchell T., Text classification from labeled and unlabeled documents using EM, Machine Learning, 39(2/3):103–134, 2000.
13. Peng F., Schuurmans D., Keselj V., and Wang S., Augmenting Naïve Bayes Classifiers with Statistical Languages Models. Information Retrieval, vol. 7, 317-345. Kluwer Academic Publishers. 2004.
14. Sebastiani F., Machine learning in automated text categorization, ACM Computing Surveys, 34(1):1–47, 2002.
15. Seeger M., Learning with labeled and unlabeled data. Technical report, Institute for Adaptive and Neural Computation, University of Edinburgh, Edinburgh, United Kingdom, 2001.
16. Solorio T., Using unlabeled data to improve classifier accuracy, Master Degree Thesis, Computer Science Department, INAOE, Mexico, 2002.
17. Stamatatos E., Fakotakis N., and kokkinakis G., Computer-Based Authorship Attribution Without Lexical Measures. Computers and the Humanities 35: 193-214, Kluwer Academic Publishers. 2001.
18. Witten, IH., Frank, E.: Data Mining-practical Machine Learning Tools and Techniques whit Java Implementation. Morgan Kaufmann, 2000.
19. Zelikovitz S., and Hirsh H., Integrating background knowledge into nearest-Neighbor text classification, In Advances in Case-Based Reasoning, ECCBR Proceedings, 2002.
20. Zelikovitz S., and Kogan M., Using Web Searches on Important Words to Create Background Sets for LSI Classification, 19th International FLAIRS conference, Melbourne Beach, Florida, May 2006.

21. Zhao Y., and Zobel J., Effective and Scalable Authorship Attribution Using Function Words. Lecture Notes in Computer Science, vol. 3689, 174-189. Springer Verlag. 2005.