On the Empirical Evaluation of Author Identification Hybrid Method Notebook for PAN at CLEF 2015

Seifeddine Mechti[1], Maher Jaoua[2], Rim Faiz[1, 3],Lamia Hadrich Belguith[2] and, Bassem Bsir[2]

[1]LARODEC Laboratory, ISG of Tunis B.P.1088, 2000 Le Bardo, Tunisia
mechtiseif@gmail.com, Rim.faiz@ihec.rnu.tn

[2]ANLP Group, MIRACL Laboratory, University of Sfax, 3018,Sfax Tunsia [3]IHEC of Carthage, 2016 Carthage Présidence, Tunisia
{maher.jaoua,l.belguith}@fsegs.rnu.tn, Bassem.bsir@yahoo.fr

Abstract. In this paper we focus on the identification of the author of a written text. We present a new hybrid method that combines a set of stylistic and statistical features in a machine learning process. We tested the effectiveness of the linguistic and statistical features combined with the inter-textual distance "Delta" on the PAN'@CLEF'2015 English corpus and we obtained 0.59 as c@1 precision.

Keywords: **Author Identification, machine learning, sub corpus.**

1 Introduction

The task, which consists in deciding by automatic means whether a text T was written by an author A, fits into the research field that focuses on author identification. The benefit of automating this task is substantial because of its usefulness in various fields such as forensic analysis, forensic linguistics, electronic commerce and plagiarism detection. In the latter case, the probability that a text contains plagiarism becomes important if the attribution of two parts of this text is not assigned to the same author.

In the literature, the automation of the author attribution task can draw on stylistic or statistical attributes. Currently, learning techniques are being used to infer attributes that discriminate the styles of authors. It is in this context that we propose in this paper a hybrid method that combines stylistic and statistical attributes while relying on measurements of inter-textual distances. We will present the results of our experiments, using several learning techniques.

To explain the steps of our method, its implementation, and findings, we organized this paper into five sections including this introduction. The second section is devoted to an overview of the main work in this field. The third section details the proposed method and specifies the stylistic and statistical features used. The fourth section presents the experiments and evaluations conducted on the corpus disseminated during the PAN'@CLEF'2015 conference. The conclusion reiterates the main findings and presents some prospects that are still in the experimentation stage.

2 Overview of Automatic Author Identification Methods

An overview of the main works in the field of author identification allows us to identify three types of approaches [1]. The first is based on the stylistic analysis of documents and aims to identify style invariants that help us distinguish the writings of an author from those of another author. The second approach is based on multivariate statistical analyses and aims to identify the joint distribution of some style variables making it possible to decide whether two texts show a significant correlation of style. The third approach, described as recent, is based on machine learning algorithms and seeks to build classifiers that infer the lexical and syntactic attributes that characterize the style of an author.

The basic idea of the stylistic methods is structured around the modeling of the authors from a linguistic point of view so that we can compare their writings. We cite as an example the work of Li et al. in which they focused on topographical signs [2] and the work of Zheng et al. in which they were interested in the co-occurrence of character n_grams [3]. Other works were concerned with the distribution of function words [4] or the complexity of vocabulary [5]. In another work, Raghavan et al. capitalized on the probabilistic context-free grammars to model the grammar used by an author [6]. Feng et al. based their research on the syntactic functions of words and their inter-relationships in order to discern the complex constructions used by each author [7]. Other studies focused on the semantic dependency between the words of written texts through the use of taxonomies [8]. Finally, and in a critical study, Baayen demonstrated that stylistic methods show weak performance in the analysis of short texts [9]. Moreover, he demonstrated that style can change over time or according to the literary genre of texts (poetry, novels, plays ...).

The first attempts tried to compare the occurrence frequency of certain numbers of functional words (determiners, prepositions, conjunctions, and pronouns) [10]. However, the results of the evaluation of this method prove its limitations and it is for this reason that other studies experimented with multivariate statistical indices. We cite as examples the principal component analysis [11]. Other methods use probabilistic measurements of distance such as the inter-textual distance [12], the LDA distribution [13], the KL divergence distance between the hidden Markov models [14] and the $\chi 2$ distance [15].

[16] puts forward a statistical rule called "Delta rule" which is based on the set of the most frequent terms (between 40 and 150), especially function words. It is noteworthy that this rule has been used by numerous studies in the field of author identification [17,18]. For his part, Savoy puts forward a probabilistic model for the attribution of documents addressing several topics [19]. In this framework, each document of a given corpus is modeled as a distribution of different themes, each theme representing a specific distribution of words.

The use of machine learning techniques stems from the observation that the task of author identification can be seen as a classification problem [1]. The methods which are part of this approach hinge on two stages: the first consists in representing the source texts as vectors of labeled and multivariate words. The second consists in using learning techniques to identify the boundaries of each class, meanwhile minimizing a classification loss function. To construct the classification model,

several techniques have been adopted such as the discriminating analysis [19], SVM [2], the decision trees [20], the neural networks [3], the methods of sets of classifiers [1] and the theme models [18]. It should be noted that other studies have compared the performance of some classifiers for the author identification task [20].

3 Proposed Method

Hybridization has always been considered an interesting track because it overcomes the limitations of combined approaches. It is with this objective in mind that we tried to experiment with learning techniques on all the stylistic and statistical features that have shown their efficiency in the literature. The basic idea is to create for each text T, whose belonging to an author A we want to verify, a sub corpus which includes all the texts written by this author and the texts that are close to it in terms of distance. Thus, if the text was written by author A then there is a high probability that we recognize the style via the stylistic and statistical features of author A's texts belonging to the corpus. On the other hand, if A is not the writer of T then there is a good chance that it is assigned to another author selected from the rest of the sub corpus.

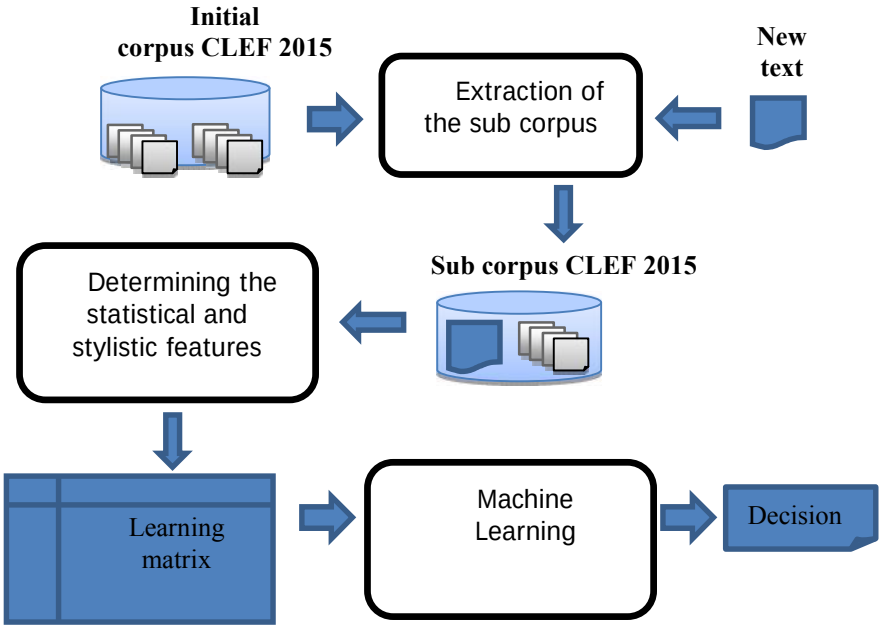

Fig. 1: Steps of the proposed method

4 Implementation of the HyTAI system

In order to implement the proposed method, we developed a system called HyTAI (Hybrid Tool for Author Identification) whose modular decomposition follows the proposed method. Thus, we used the Delta rule in the extraction module

of the sub corpus to calculate the distance between two texts. Also, we used the OpenNLP for the extraction of the stylistic and statistical features.

To calculate the distance between two documents, we used the Delta distance proposed by Burrows et al. (Burrows 2002). This distance, which takes into account the most frequent words, is characterized by the following formula:

$$\Delta(Q, A_j) = \frac{1}{m} \sum_{i=1}^{m} |Zscore(t_{iq}) - Zscore(t_{ij})|$$

where $Z\,score_{(t_{tij})} = \dfrac{tfr_{ij} - mean_i}{sd_i}$

Note that tfr_{ij} is the frequency of the term ti in the document Dj while $mean_i$ is the mean and SDI is the standard deviation.

It should be noted that if two texts are quite close, then delta tends toward 0. Similarly, the value m may vary from one corpus to another and that is why we conducted an experiment to have the value determined (see next section). For the training sub corpus, we choose the nearest texts of a document to be checked in such a way that a balance is achieved between the texts written by the author to be identified and the texts that do not belong to that author.

In order to extract the stylistic and statistical features, we used tools from the Apache OpenNLP library, which contains a set of functions that can segment texts and perform the syntactic and lexical analyses. We calculated the frequency of lexical features, the ratio V / N – where V is the hapax's size and N is the text length – and the average length of sentences. Regarding parsing, also conducted through the OpenNLP, we extract the number of nouns, the number of verbs, the number of adjectives, the number of adverbs, and the number of prepositions.

Then to extract the features related to the model of the language, we consider the text as a simple sequence of characters and determine the frequencies of the letters, the punctuation marks and the numeric characters as well as n-grams.

5 Evaluation

To evaluate the HyTAI system, we conducted a series of experiments which aim to determine the ideal parameters such as the threshold of the Delta distance as well as the most suitable algorithm for learning. We used the corpus disseminated at the PAN'@CLEF'2015 conference. This corpus consists of 200 collections of English documents (essays) which include 518 known texts and 200 unknown texts. The average length of the documents is around of 833 words per document.

To evaluate the performance of our HyTAI system, we used the c @ 1 measure adopted by the PAN'@CLEF'2015 conference and defined by Penas et al., (Penas, Rodrigo, 2011). Compared to conventional measures of precision, this measure has the advantage of taking into account the indecisions of the system, that is to say where

the system cannot decide on the authorship of the document concerned. The formula proposed for the calculation of c @ 1 is as follows:

$$c @ 1 = (1 / n) * (nc + (nu * nc / n)) \text{ [21]}$$

where n is the total number of problems; nc = number of correct decisions;
nu = number of cases of indecision

The following figure shows the c @ 1 measurement results obtained via 6 classifiers on the corpus. For an unknown text whose authorship we want to identify, we create a sub corpus containing known texts by the author and the same number of texts that are close in terms of Delta distance from unknown texts that do not belong to the author. The classifiers used in this experiment are: SVM, Bayesian Networks, Naive Bayes, Decision tables, Decision tree and KNN.

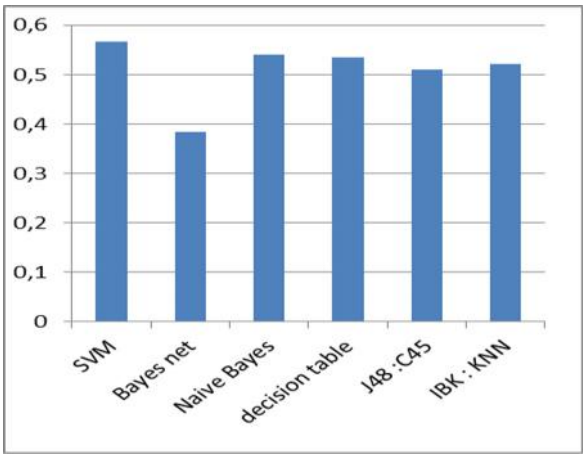

Fig. 2: Performance of the classifiers used in the experiments

In our case, indecision stems from the fact that we obtain with the classifier (in the cases of SVM, Bayes, KNN median values (close to 0.5. So for these classifiers, indecision results when the value obtained is in the range [0.4-0.6].

According to the histogram, the best results (the C @ 1 axis value are obtained through the SVM algorithm, followed by the naive Bayes classifier.

To determine the optimal threshold used by the Delta distance, we conducted an experiment with the various values of c@1 by varying the threshold from 50 to 400. In this experiment, we set the SVM algorithm as a classifier.

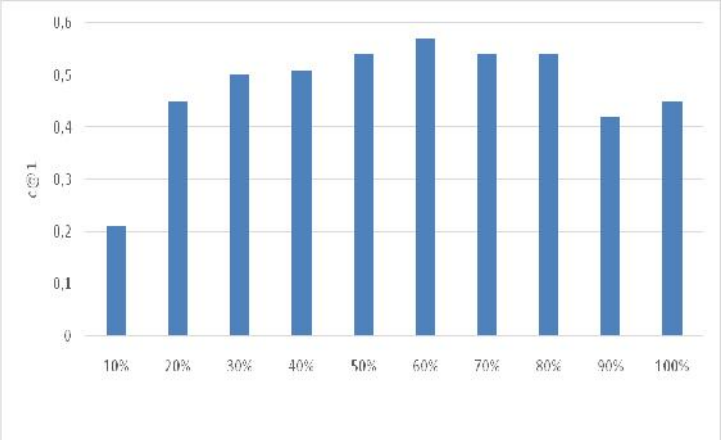

Fig. 3: System precision variation based on size of sub corpus

As shown in the previous figure, with a size of 60% of the English Corpus, the HyTai system got an accuracy rate for c@1 which is equal to 0.59.

6 Conclusion

In this paper, we presented an automatic author identification method which is based on the combination of statistical and stylistic features while relying on the SVM learning algorithm. The results obtained on the corpus of the PAN'@ CLEF'2015 conference prove the interest of hybridization, and the importance of statistical features. However, these results do not satisfy our ambitions; that is why we are planning during our next participation in the PAN conference to change the training corpus by choosing different textual forms that derive from the text. This procedure is intended to "fill" the training corpus and thus to reach more accurate decisions. The first results, that we are about to experiment in this direction, are very promising.

Also, we plan to extend our method to take into account the other languages put forward in the author identification task. Within this framework, we will focus more on the statistical features and those derived from the language model.

References

1. E. Stamatatos E. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology* 60, 538–556. 2009.
2. Li J., Zheng R., Chen H. From fingerprint to writeprint. *Communication ACM* 49(4), 76–82. 2006.
3. Zheng R., Li J., Chen H., Huang Z. A framework for authorship identification of online messages: Writing-style features and classification techniques. *American Society for Information Science and Technology* 57(3), 378-393. 2006.
4. Vartapetiance A., Gillam L. A Trinity of Trials: Surrey's 2014 Attempts at Author Verification. Proceedings of *PAN@CLEF'2014*. 2014
5. Argamon S., Whitelaw C., Chase P., Hota S.R., Garg N., Levitan S. Stylistic text classification using functional lexical features. Journal of American society of information science and technology 58(6), 802–822. 2007.
6. Raghavan S., Kovashka A., Mooney R. AUTHORSHIP ATTRIBUTION USING PROBA- BILISTIC CONTEXT-FREE GRAMMARS. PROCEEDINGS of *ACL'10*, 38–42. 2010.
7. Feng V.W., Hirst G. authorship verification with entity coherence and other rich linguistic features. proceedings of clef'13. 2013.
8. Mccarthy P.M., Lewis G.A., Dufty D.F., Mcnamara D.S. Analyzing writing styles with coh-metrix. Proceedings of FLAIRS'06, 764–769. 2006.
9. Baayen R.H*Analyzing Linguistic Data : A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press. . 2008.
10. Mosteller F., Wallace D.L. *Inference in an Authorship Problem. In Journal of the American Statistical Association* 58, 275-309. 1964.
11. Burrows J.F. Not unless you ask nicely: The interpretative nexus between analysis and information. Literary and Linguistic Computing 7(1), 91-109. 1992.
12. Labbé C. Inter-Textual Distance and Authorship Attribution : Corneille and Molière. *Journal of Quantitative Linguistics*, 213-231. 2003.

13. Blei D.M., Jordan M.I. Variational methods for the Dirichlet process. Proceedings of *the 21st international conference on Machine learning ACM*. 2004.
14. Hershey J.R., Thomas J.W., Olsen P.A., Rennie S.J. Variational Kullback-Leibler divergence for Hidden Markov models. *Proceedings of IEEE Workshop on Automatic Speech Recognition Understanding*, 323-328. 2007.
15. Grieve J. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing* 22(3), 251-270. 2007.
16. Burrows J.F. Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship. *Journal Literature Linguist Computing*. 2002.
17. Savoy J. Attribution d'auteur par ensembles de séparateurs. Acte de *la COnférence en Recherche d'Information et Applications CORIA*, 277-290. 2013.
18. Savoy J.. Etude comparative de stratégies de sélection de prédicteurs pour l'attribution d'auteur. Actes de *la COnférence en Recherche d'Information et Applications CORIA*, 215-228. 2012.
19. Stamatatos E., Fakotakis N., Kokkinakis G.. Automatic text categorization in terms of genre and author. *Computational Linguistics* 26, 471-495. 2000.
20. Zhao Y., Zobel J.. Searching with style: Authorship attribution in classic literature. Proceedings of *the Australian Computer Science Conference*, 59-68. 2007.
21. Peñas A., Rodrigo A. A Simple Measure to Assess Non response. Proceedings of *the 49th Annual Meeting of the Association for Computational Linguistics*, 1415-1424. 2011.