Verbs and Pronouns for Authorship Attribution

Paulo Varela, Edson Justino
Pontifical Catholic University of Parana
PPGIa
Curitiba, Brazil
justino@ppgia.pucpr.br

Luiz S. Oliveira
Federal University of Parana
Department of Informatics
Curitiba, Brazil
lesoliveira@inf.ufpr.br

Abstract— In this paper we discuss the use of verbs and pronouns as features for authorship attribution on texts written in Portuguese. We demonstrate through experiments that verbs and pronouns when used with other features such as adverbs and conjunctions can bring an important improvement in terms of recognition rate. The experimental protocol takes into account Support Vector Machines as classifiers and a database composed of short articles from 20 authors. An improvement of about 4% was achieved using these features.

Keywords-stylometry, authorship attribution, feature extraction.

I. INTRODUCTION

Authorship attribution can be defined as the task of inferring characteristics of a document's author, including but not limited to identity, from the textual characteristics of the document itself. There exists a long history of linguistic and stylistic investigation into authorship attribution which goes back to the late nineteenth century, with the pioneering studies of Mascol [1] and Mendenhall [2].

Chaski [3] has published some case studies where the authorship of specific digital documents has been the key issue. In one case, a software "diary" provided crucial exculpatory evidence against the claims of its author. But were the entries genuine, or had they been planted? In another case, an employee was dismissed on the basis of emails admittedly written on her computer. But in an open plan office, anyone can use any unprotected computer. Did she really write the relevant emails, or was she wrongfully dismissed? In a third case, an investigation of a death turned up a suicide note "written" on a computer. Was this note genuinely written by the decedent, or had it been written by the murderer to cover his tracks?

In all these cases, the question is who was at the keyboard and not the computer that created the document. Besides, the traditional handwriting analysis does not apply here since the documents were electronic. In the last decades, practical applications for authorship attribution have grown in several different areas such as, criminal and civil law as well as in computer security, e.g., mining email content.

Authorship attribution can be formulated as a pattern recognition problem, therefore, one must count on features with good discrimination power. In this context, the stylometry, which can be defined as the study of the linguistic style, offers a strong support to define a discriminative feature set. Alike other forensic techniques such as handwriting analysis, the stylometry assumes that people have individual, persistent, and uncontrollable habits that can be reliable identifiable by experts.

The literature reports several stylometry-based features which generally are classified into qualitative and quantitative. The qualitative approach assesses errors and personal behavior of the authors, also known as idiosyncrasies, based on the examiner's experience. According to Chaski [3], this approach could be quantified through databasing, but it is quite difficult to develop the required databases. Without such databases to ground the significance of features, the examiner's intuition about the significance of a feature can lead to methodological subjectivity and bias. Examples of this approach can be found in [8] where the authors proposed 99 features to train different classifiers such as SVM and decision trees. The best result reported was about 72% of recognition rate.

The quantitative approach focus on readily computable and countable language features [4,5,6,7,9], e.g. word length, phrase length, sentence length, vocabulary frequency, distribution of words of different lengths. It uses standard syntactic analysis from the dominant paradigm in theoretical linguistics over the past forty years. Examples of this approach can be found in [10,11,12]. Experimental results show that usually this approach provides better results than the qualitative one tracks.

In this work we extend the research presented in [12] by Pavelec et al. In their work the authors proposed a qualitative approach using a stylometric feature set based on 77 conjunctions and 94 adverbs to perform author identification on a database of short articles written in Portuguese. After an extensive series of experiments we found out that some verbs and pronouns together with adverbs and conjunctions proposed in [12] could increase the overall performance of the system. In this paper we present a set of verbs and pronouns that can be used as features in the context of authorship attribution. Experiments on a database composed of

short articles from 20 different authors and Support Vector Machine (SVM) as classifier demonstrate that the extended feature set can improve the results in about 4% in both writer-dependent and writer-independent approaches.

II. WRITER-DEPENDENT AND WRITER-INDEPENDENT

To make this paper self-contained, in this section we briefly introduce the concepts of writer-dependent and writer-independent. For more details, please refer to [12]. The writer-dependent or personal model is based on one model per author. Usually it yields good results but its drawbacks are the need of learning the model each time a new author should be included in the system and the great number of genuine samples necessary to build a reliable model. In real applications, usually a limited number of samples per author is available to train a classifier, which leads the class statistics estimation errors to be significant, hence, resulting in unsatisfactory verification performance. It can be implemented using either one-against-all or pairwise strategy. This kind of approach has been largely used for authorship attribution.

An alternative to the personal approach is the global approach or writer-independent model. It is based on the forensic questioned document examination approach and classifies the writing, in terms of authenticity, into genuine and forgery, using for that one global model. In the case of author identification, the experts use a set of n genuine articles Sk_i, (i = 1, 2, 3,…, n) as references and then compare each Sk with a questioned sample Sq. The idea is to verify the discrepancies among Sk and Sq. Let V_i be the stylometric feature vectors extracted from the reference articles and Q the stylometric feature vector extracted from the questioned article. Then, the dissimilarity feature vectors $Z_i = \|V_i - Q\|$ are computed to feed m different instances of the classifier C, which provide a partial decision. The final decision D depends on the fusion of these partial decisions, which are usually obtained through the majority vote rule.

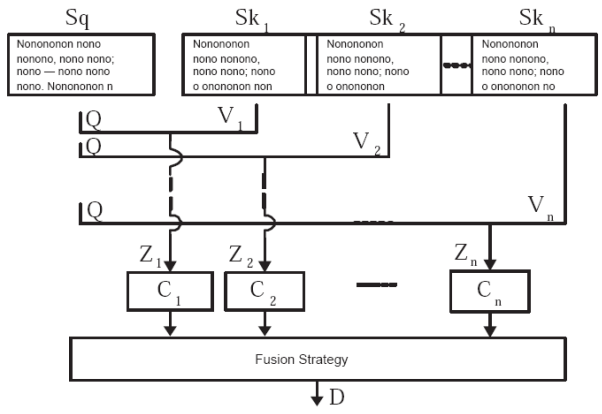

Figure 1. Architectur of the writer-independent model

Figure 1 depicts the global approach. Note that when a dissimilarity measure is used, the components of the feature vector Z tends to be close to zero when both the reference Sk and the questioned Q comes from the same author. Otherwise, the feature vector Z tends to be far from zero.

III. FEATURES

As stated before, the goal of this work is to extend the research initiated in [12] by Pavelec et al. In their work, they have proposed a feature set based on 77 conjunctions and 94 adverbs of the Portuguese language.

Just like other language, Portuguese has a large set of conjunctions that can be used to link words, phrases, and clauses. Such conjunctions can be used in different ways without modifying the meaning of the text. For example, the sentence "Ele tal qual seu pai" (He is like his father), could be written is several different ways using other conjunctions, for example, "Ele tal e qual seu pai", "Ele tal como seu pai", "Ele que nem seu pai", "Ele assim como seu pai". Such diversity has been proved to be helpful for authorship attribution.

Another feature set used with success for authorship attribution is based on adverbs of the Portuguese language. The way an author uses adverbs is somehow persistent and can be used by experts for authorship attribution. In this same vein, in this work we argue that pronouns and verbs also can bring some discriminative information. To demonstrate that we have selected the 50 most used verbs of the Portuguese language and 91 pronouns [14]. Tables I and II describe the verbs and pronouns, respectively.

TABLE I. 50 VERBS IN THE INFINITIVE FORM

escrever, achar, abrir, efetuar, pagar, falar, colar, acabar, atingir, distribuir, jogar, estar, declarar, melhorar, ligar, andar, dizer, completar, achar, usar, ver, dar, visitar, realizar, projetar, ser, escolher, encerrar, haver, desenvolver, cantar, fechar, comer, viver, poder, pular, entender, beber, aplicar, implantar, ler, fazer, pensar, gerar, trazer, ter, trocar, possuir, melhorar, iniciar

TABLE II. 91 PRONOUNS

Type	Pronouns
Relatives	quem, o qual,a qual, os quais, as quais,onde, em que, quanto, quanta, quantos, quantas, cujo, cuja, cujos, cujas
Possessives	meu, minha, meus, minhas, teu, tua, teus, tuas, seu, sua, seus, suas, nosso, nossa, nossos, vosso, vossa, vossos, vossas
Demonstrative	este, esta, estes, estas, isto, esse, esses, essa, essas, isso, aquele, aquela, aqueles, aquelas, aquilo,nessa, desta, daquela, cujo, cuja,cujos, cujas
Subjective Personal	eu, tu, ele, nós, vós, eles, me, te, se, lhe, o, a, nos, vos, lhes, os, as, mim, comigo, conosco, ti, contigo, convosco, si, consigo
Objective Personal	você, vocês, senhor, senhores, senhora, senhoras, senhorita, senhoritas, vossa senhoria, vossas senhorias

Table I shows the verbs in the infinitive form, but all the forms are considered during feature extraction.

IV. DATABASE

To build the database we have collected articles available in the Internet from 20 different authors with profiles ranging from sports to economics. Our sources were two different Brazilian newspapers, Gazeta do Povo (http://www.gazetadopovo.com.br) and Tribuna do Paraná (http://www.paranaonline.com.br) .

We have chosen 30 short articles from each author. The articles usually deal with polemic subjects and express the author's personal opinion. In average, the articles have 600 tokens (words) and 350 Hapax (words occurring once). One aspect worth of remark is that this kind of articles can go through some revision process, which can remove some personal characteristics of the texts. Besides, authorship attribution using short articles poses an extra challenge since the number of features that can be extracted are directly related to the size of the text.

V. EXPERIMENTAL PROTOCOL

This section describes how both strategies, writer-dependent and writer-independent, have been implemented. In order to extract the features, first the text is segmented into tokens. Spaces and end-of-line characters are not considered. All hyphenised words are considered as two words. In the example, the sentence "eu vou dar-te um pula-pula e também dar-te-ei um beijo, meu amor!" has 16 tokens and 12 Hapax. Punctuation, special characters, and numbers are not considered as tokens. There is no distinction between upper case and lower case.

Regarding the writer-dependent approach, a single model with n outputs is trained, where n is the number of authors enrolled into the system. The machine learning model used in our experiments is the SVM. There are two basic approaches to solve n-class problems with SVMs: pairwise and one-against-others. In this work we have used the former, which arranges the classifiers in trees, where each tree node represents a SVM. For a given test sample, it is compared with each two pairs, and the winner will be tested in an upper level until the top of the tree. In this strategy, the number of classifiers we have to train is $n(n-1)/2$. From the database described previously, we have used 20 authors ($n = 20$, consequently 190 models). From each author 10 documents were used for training and 15 documents for testing.

Differently of the writer-dependent approach, this strategy consists in training just one global model which should discriminate between author (ω_1) and not author (ω_2). To generate the samples of ω_1, we have used three articles (A_i) for each author. Based on the concept of dissimilarity, we extract features for each article and then compute the dissimilarities among them as shown in Section II. In this way, for each author we have 10 feature vectors, summing up 100 samples for training (10 authors). The samples of ω_2 were created by computing the dissimilarities of the articles written by different authors, which were chosen randomly. As stated before, the proposed protocol takes into consideration a set of references (Sk). In this case we have used 20 authors (the same 20 used for the writer-dependent), five articles per author as references and 15 as questioned (Sq - testing set).

Following the protocol introduced previously, a feature vector is extracted from the questioned (Sq) and references (Sk_i) documents as well. This produces the aforementioned stylometric feature vectors V_i and Q. Once those vectors are generated, the next step consists in computing the dissimilarity feature vector $Z_i = |V_i - Q|$, which will feed the SVM classifiers. Since we have five ($n = 5$) reference articles, the questioned article Sq will be compared five times (the SVM classifier is called five times), yielding five votes or scores. When using discrete SVM, it produces discrete outputs $\{-1,+1\}$, which can be interpreted as votes. To generate scores, we have used the probabilistic framework proposed by Platt in [15]. Finally, the final decision can be taken based on different fusion strategies, but usually majority voting is used.

VI. RESULTS

In both strategies, different parameters and kernels for the SVM were tried out but the better results were yielded using a linear kernel.

In the first experiment, we have used only the features based on verbs and pronouns (141 features). The idea was to assess the discrimination power of these features alone. In both cases we are not able to surpass the results reported in [12], which are reproduced in Table III.

TABLE III. COMPARATIVE RESULTS

Features	Recognition Rate (%)	
	W-Depend.	W-Independ.
Conjunctions + Adverbs [13]	72.5	83.2
Compression [17]		84.3
Conjunctions + Adverbs + Verbs + Pronouns	76.5	87.0

However, when we added these features to the original feature set we got an improvement of about 4% for both writer-dependent and writer-independent approaches. This corroborates to our initial hypothesis that verbs and pronouns bring discriminative information in the context of authorship attribution.

Few works have been done in the field of author identification for documents written in Portuguese. For this reason is quite difficult to make any kind of direct

comparison. Coutinho et al [15] extract features using a compression algorithm and achieve a recognition rate of 78%. However, the size of the texts used for feature extraction is about 10 times bigger. Pavelec et al [16] used the technique proposed in [15] on the database described in this work but using only the writer-independent approach. In that case the best result achieved was 84.3% of recognition rate. As we can see, the use of verbs and pronouns in the original feature set also surpassed the results achieved by the compression algorithm proposed in [16].

As one could observe, the main disadvantage of the writer-dependent model is the huge number of models necessary. This approach is unfeasible as the number of authors gets bigger. One alternative to surpass this problem is the writer-independent model, which does not depend on the number of author. Using this approach the best result we got was 76.5%.

In spite of the fact that the writer-independent approach achieves worse results, we argue that it should be considered as an alternative because of its lower computational complexity. Besides, we believe that the writer-independent can be improved if we investigate different types of features.

Another aspect that should be investigated is feature selection. Based on the vector size, it is fair to assume that there exist correlated or even unnecessary features that can be removed so that the final performance could be further improved.

VII. CONCLUSIONS

In this paper we have proposed two new feature set for authorship attribution for texts written Portuguese. The first feature set is based on 91 pronouns and the second one takes into account the 50 most used verbs of the Brazilian Portuguese language.

Through a series of experiments on a database composed of 20 authors, we have demonstrated that these feature sets, when combined with adverbs and conjunctions proposed in [12] can improve the recognition rates in about 4%.

As discussed before, feature selection will be investigated to find a smaller and possibly more discriminative subset of features. In parallel, strategies to combine different classifiers will be considered to train each classifier with a different feature set. In this way, instead of merging all the features into one feature vector, the fusion would occur at the decision level.

ACKNOWLEDGMENT

This research has been supported by The National Council for Scientific and Technological Development (CNPq) grant 306358/2008-5.

REFERENCES

[1] C. Mascol, “Curves of pauline and pseudo-pauline style”, Unitarian Review, 30:453–460, 1888.

[2] T. Mendenhall, “The characteristic curves of composition”, Science, 214:237–249, 1887.

[3] C. Chaski, “Who's at the keyboard: Authorship attribution in digital evidence investigations”, International Journal of Digital Evidence, 4(1), 2005.

[4] D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, and L. Ye. “Author identification on the large scale”. In Joint Annual Meeting of the Interface and the Classification Society of North America (CSNA), 2005.

[5] P. Juola, “Future and Trends in Authorship Attribution”, International Federation for Information Processing, Vol 242, Advances in Digital Forensics III; eds. P. Craiger and S Shenoi, pp. 119-132, 2007.

[6] D. Hoover, “Delta prime?”, Literary and Linguistic Computing, vol. 19(4):477-495, 2004.

[7] P. Juola, “On composership attribution. Proceedings of the Joint Int. Conf. of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities”, 2004.

[8] M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In Workshop on Computational Approaches to Style Analysis and Synthesis, 2003.

[9] S. Argamon, M. Aric, and S. S. Stein. Style mining of electronic messages for multiple author discrimination. In ACM Conference on Knowledge Discovery and Data Mining, 2003.

[10] G. Tambouratzis, S. Markantonatou, N. Hairetakis, M. Vassiliou, G. Carayannis, D. Tambouratzis, “Discriminating the Registers and Styles in the Modern Greek Language – Part 2: Extending the feature Vector to Optimize Author Discrimination”, *Literary* and Linguistic Computing 19(2):221-242, 2004.

[11] T. Tas, A. K. Gorur, “Author Identification for Turkish Texts”, *Journal of Arts and Sciences*, 7:151–161, 2007.

[12] D. Pavelec, L. S. Oliveira, E. Justino, L. V. Batista, “Using Conjunctions and Adverbs for Author Verification”, Journal of Universal Computer Science, 14(18):2967-2981, 2008.

[13] M. A. Ryan, Conjugação dos Verbos em Português – Prático e Eficiente, 17th Ed. Ática, São Paulo, pp. 176, 2006.

[14] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. Smola et al, editor, Advances in Large Margin Classifiers, pp. 61–74. MIT Press, 1999.

[15] B. C. Coutinho, L. M. Macedo, A. Rique-JR, and L. V. Batista, “Atribuição de autoria usando PPM”. In XXV Congress of the Brazilian Computer Society, pp. 2208–2217, 2004.

[16] D. Pavelec, L. S. Oliveira, E. Justino, F. D. Nobre Neto, L. V. Batista, “Compression and Stylometry For Author Identification”, Internation Joint Conference on Neural Networks, pp. 2445-2450, 2009.