Journal of Universal Computer Science, vol. 14, no. 18 (2008), 2967-2981
submitted: 1/3/08, accepted: 29/9/08, appeared: 1/10/08 © J.UCS

Using Conjunctions and Adverbs for Author Verification

Daniel Pavelec, Luiz S. Oliveira, Edson Justino
(Pontifícia Universidade Católica do Paraná, Curitiba, PR, Brazil
{pavelec,soares,justino}@ppgia.pucpr.br)

Leonardo V. Batista
(Federal University of Paraíba, João Pessoa, PB, Brazil
leonardo@di.ufpb.br)

Abstract: Linguistics and stylistics have been investigated for author identification for quite a while, but recently, we have testified a impressive growth in the volume with which lawyers and courts have called upon the expertise of linguists in cases of disputed authorship. This motivates computer science researchers to look to the problem of author identification from a different perspective. In this work, we propose a stylometric feature set based on conjunctions and adverbs of the Portuguese language to address the problem of author identification. Two different approaches of classification were considered. The first one is called writer-independent and it reduces the pattern recognition problem to a single model and two classes, hence, makes it possible to build robust system even when few genuine samples per writer are available. The second one is called the personal model, or writer-dependent, which very often performs better but needs a bigger number of samples per writer. Experiments on a database composed of short articles from 30 different authors and Support Vector Machine (SVM) as classifier demonstrate that the proposed strategy can produced results comparable to the literature.

Key Words: Author Verification, Pattern Recognition
Category: H.3.7, H.5.4

1 Introduction

The literature shows a long history of linguistic and stylistic investigation into author identification [Mendenhall, 1887], [Mascol, 1888] but the work published by Svartvik [Svartvik, 1968] marked the birth of term forensic linguistics, i.e., the linguistic investigation of authorship for forensic purposes. In it, he analyzed four statements that Timothy Evans, executed in 1950 for the murder of his wife and baby daughter, was alleged to have made following his arrest. Using both qualitative and quantitative methods Svartvik demonstrated considerable stylistic discrepancies between the statements, thus raising serious questions about their authorship. It was later discovered that both victims had actually been murdered by Evan's landlord, John Christie. [Coulthard, 2005]

Since then, there has been a impressive growth in the volume with which lawyers and courts have called upon the expertise of linguists in cases of disputed authorship. Hence, practical applications for author verification have grown in several different areas such as, criminal law (identifying writers of ransom notes and harassing letters), civil law (copyright and estate disputes), and computer security (mining email content). Chaski [Chaski, 2005] points out that in the investigation of certain crimes involving digital evidence, when a specific machine is identified as the source of documents, a legitimate issue is to identify the author that produced the documents, in other words, “Who was at the keyboard when the relevant documents were produced?”.

Author verification is the task of verifying the author of a given text. Consequently, it can be approached as a typical classification problem, which depends on discriminant features to represent the style of an author. In this context, the stylometry (application of the study of linguistic style) offers a strong support to define a discriminative feature set. The literature shows that several stylometric features that have been applied include various measures of vocabulary richness and lexical repetition based on word frequency distributions. As observed by Madigan et al [Madigan et al, 2005], most of these measures, however, are strongly dependent on the length of the text being studied, hence, are difficult to apply reliably. Many other types of features have been investigated, including word class frequencies [Forsyth and Holmes, 1996], syntactic analysis [Baayen et al, 1996], word collocations [Smadja, 1989], grammatical errors [Koppel and Schler, 2003], number of words, sentences, clauses, and paragraph lengths [Argamon et al, 2003b], [Argamon et al, 2003a].

To deal with the problem of author verification, researchers have investigated two different approaches: writer-dependent and writer-independent. The former is the standard approach where a specific model is built for each writer. The main drawbacks of the writer-dependent approach are the need of learning the model each time a new writer 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 writer is available to train a classifier, which leads the class statistics estimation errors to be significant, hence, resulting in unsatisfactory verification performance.

Another option to the writer-dependent approach is the writer-independent, which models the probability distributions of within-class and between-class similarities. These distributions are used to determine the likelihood of whether a questioned document is authentic or forgery. The concept of dissimilarity representation for pattern recognition was introduced by Pekalska and Duin [Pekalska and Duin, 2002] and the seminal work using this concept in the field of author verification was presented by Cha and Srihari [Cha and Srihari, 2002]. Later, Santos et al [Santos et al, 2004] use the idea of dissimilarity representation for author verification. The main benefit provided by this approach is the possibility of reducing an n-class pattern recognition problem to a 2-class problem, in the case of author verification, genuine and forgery.

In this work we discuss the two aforementioned approaches for writer verification. We also propose a stylometric feature set for the Portuguese language, which is based on conjunctions and adverbs. Comprehensive experiments on a database composed of short articles and Support Vector Machine (SVM) as classifier demonstrate the advantages and drawbacks of each strategy and also that both can produce results comparable to the literature.

The remaining of this paper is divided as follows: Section 2 introduces the basic concepts of forensic stylistics and describes the linguistic features used in this work. Section 2.2 describes the basic concepts of the SVM. Section 3 describes how both writer-independent and writer-dependent approaches work. Section 3.3 presents the database used in this work. Section 4 describes both writer-dependent and writer-independent methods for author verification while Section 5 reports the experimental results. Finally, Section 6 concludes this work.

2 Forensic Stylistics

Forensic stylistics is a sub-field of forensic linguistics and it aims at applying stylistics to the context of author verification. The stylistic is based on two premisses: a) Two writers (same mother-tongue) do not write in the same way and b) The writer does not write in the same way all the time.

The stylistic can be classified into two different approaches: 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 [Chaski, 2005], this approach could be quantified through databasing, but until now the databases which would be required have not been fully developed. Without such databases to ground the significance of stylistic features, the examiner's intuition about the significance of a stylistic feature can lead to methodological subjectivity and bias. In this vein, Koppel and Schler [Koppel and Schler, 2003] proposed the use of 99 error features to feed different classifiers such as SVM and decision trees. The best result reported was about 72% of recognition rate.

The second approach, which is very often refereed as stylometry, is quantitative and computational, focusing on readily computable and countable language features, 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 Tambouratzis et al [Tambouratzis et al, 2004], Chaski [Chaski, 2005] and [Tas and Gurur, 2007]. The latter addresses the problem of author verification for Turkish texts and reports an average success rate of 80%. Experimental results show that usually this approach provides better results than the qualitative one. For this reason we have chosen this paradigm to support our work.

2.1 Linguistic Features

The literature suggests many linguistic features to be used for author verification. In [Chaski, 1998], Chaski discusses about the differences between scientific and replicable methods for author verification. Scientific methods are based on empirical, testable hypotheses, and the use of these methods can be done by anyone, i.e., it is not dependent on a special talent. In the same work, nine empirical hypotheses that have been used to identify authors in the past are reported: Vocabulary Richness, Hapax Legomena, Readability Measures, Content Analysis, Spelling Errors, Grammatical Errors, Syntactically Classified Punctuation, Sentential Complexity, Abstract Syntactic Structures.

Vocabulary Richness is given by the ratio of the number of distinct words (type) to the number of total words (token). Hapax Legomena is the ratio of the numbers of words occurring once (Hapax Legomena) to the total number of words. Readability Measures compute the supposed complexity of a document, and are calculations based on sentence length and word length. Content Analysis classifies each word in the document by semantic category, and statistically analyze the distance between documents. Spelling Errors quantifies the misspelled words. Prescriptive Grammatical Errors test errors such as sentence fragment, run-on sentence, subject-verb mismatch, tense shift, wrong verb form, and missing verb. Syntactically Classified Punctuation takes into account end-of-sentence period, comma separating main and dependent clauses, comma in list, etc. Finally, Abstract Syntactic Structures computationally analyzes syntactic patterns. It uses verb phrase structure as a differentiating feature.

In this work we propose the use of conjunctions and 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. Table 1 describes all the Portuguese conjunctions we have used in this work.

Such conjunctions can be used in different ways without modifying the meaning of the text. For example, the sentence “Ele é *tal qual* seu pai” (He is like his father), could be written is several different ways using other conjunctions, for example, “Ele é *tal e qual* seu pai”, “Ele é *tal como* seu pai”, “Ele é *que nem* seu pai”, “Ele é *assim como* seu pai”. The way of using conjunctions is a characteristic of each author, and for this reason we decided to use them in this work.

To complete the feature set, we have used adverbs of the Portuguese language. An adverb can modify a verb, an adjective, another adverb, a phrase, or a clause. Authors can use it to indicate manner, time, place, cause, or degree and answers questions such as “how”, “when”,“where”, “how much”. Table 2 reports the list of 94 adverbs we have used in this work.

2.2 Support Vector Machines

As stated before two different models for author verification are the subject of this work. In both strategies binary classifiers fit quite well. For the global approach just one model should be built while for the personal approach one binary model for each author is necessary. In light of this, Support Vector Machine (SVM) [Vapnik, 1995] seems quite suitable since it was originally developed to deal with problems with two classes. Moreover, SVM is tolerant to outliers and perform well in high dimensional data.

The concept of SVM was developed by Vapnik Let us suppose we have a given set of l samples distributed in a $\Re^n$ space, where n is the dimensionality of the sample

Table 1: Conjunctions of the Portuguese language

Group	Conjunctions (in Portuguese)
Coordinating additive	e, nem, mas também, senão também, bem como, como também, mas ainda.
Coordinating adversative	porém, todavia, mas, ao passo que, não obstante, entretanto, senão, apesar disso, em todo caso contudo, no entanto
Coordinating conclusive	logo, portanto, por isso, por conseguinte.
Coordinating explicative	porquanto, que, porque.
Subordinating comparative	tal qual, tais quais, assim como, tal e qual, tão como, tais como, mais do que, tanto como, menos do que, menos que, que nem, tanto quanto, o mesmo que, tal como, mais que.
Subordinating conformative	consoante, segundo, conforme.
Subordinating concessive	embora, ainda que, ainda quando, posto que, por muito que, se bem que, por menos que, nem que, dado que mesmo que, por mais que.
Subordinating conditional	se, caso, contanto que, salvo que, a não ser que, a menos que
Subordinating consecutive	de sorte que, de forma que, de maneira que, de modo que, sem que
Subordinating final	para que, fim de que
Subordinating proportional	a proporção que, quanto menos, quanto mais a medida que.

Table 2: Adverbs of the Portuguese language

Group	Conjunctions (in Portuguese)
Place	aqui, ali, aí, cá, lá, acolá, além, longe, perto, dentro, adiante, defronte, onde, acima, abaixo, atrás, em cima, de cima, ao lado, de fora, por fora.
Time	hoje, ontem, amanhã, atualmente, sempre, nunca, jamais, cedo, tarde, antes, depois, já, agora, então, de repente, hoje em dia.
Affirmation	certamente, com certeza, de certo, realmente, seguramente, sem dúvida, sim
Intensity	ainda, apenas, de pouco, demais, mais, menos, muito, pouca, pouco, quase, tanta, tanto
Negative	absolutamente, de jeito nenhum, de modo algum, não, tampouco
Subordinating concessive	embora, ainda que, ainda quando, posto que, por muito que, se bem que, por menos que, nem que, dado que mesmo que, por mais que.
Quantity	todo, toda
Mode	assim, depressa,bem, devagar, face a face, facilmente, frente a frente, lentamente, mal, rapidamente, algo, alguém, algum, alguma, bastante, cada, certa, certo, muita, nada, nenhum, nenhuma, ninguém, outra, outrem, outro, quaisquer, qualquer, tudo

space, and for each x_i sample there is an associated label $y_i \in \{-1, 1\}$. According to Vapnik, this sample space can be described by an hyperplane separating the samples according to their label ($\{-1, 1\}$). This hyperplane can be modeled using only a few samples from the sample space, namely the support vectors. So training an SVM is simplified to identifying the support vectors within the training samples. After that, a decision function (1) can be used to predict the label for a given unlabeled sample.

$$f(x) = \sum_i \alpha_i y_i K(x, x_i) + b \tag{1}$$

The function parameters α_i and b are found by quadratic programming, x is the unlabeled sample and x_i is a support vector. The function $K(x, x_i)$ is known as kernel function and maps the sample space to a higher dimension. In this way, samples that are not linearly separable can become linearly separable (in the higher dimensional space). The most common kernel functions are: Linear, Polynomial, Gaussian and Tangent Hyperbolic.

One of the limitations with SVMs is that they do not work in a probabilistic framework. There is several situations where would be very useful to have a classifier producing a posterior probability $P(class|input)$. In our case, particulary, we are interested in estimation of probabilities because we want to try different fusion strategies like Max, Min, Average, and Median.

Due to the benefits of having classifiers estimating probabilities, many researchers have been working on the problem of estimating probabilities with SVM classifiers. Sollich in [Sollich, 2002] proposes a Bayesian framework to obtain estimation of probabilities and to tune the hyper-parameters as well. His method interprets SVMs as maximum a posteriori solutions to inference problems with Gaussian process priors. Wahba et al [Wahba et al, 1999] use a logistic function of the form

$$P(y = 1|f(x)) = \frac{1}{1 + \exp(-f(x))} \tag{2}$$

where $f(x)$ is the SVM output and $y = \pm 1$ stands for the target of the data sample x. In the same vein, Platt [Platt, 1999] suggests a slightly modified logistic function, defined as:

$$P(y = 1|f(x)) = \frac{1}{1 + \exp(Af(x) + B))} \tag{3}$$

The difference lies in the fact that it has two parameters trained discriminatively, rather one parameter estimated from a tied variance. The parameters A and B of Equation 3 are found by minimizing the negative log likelihood of the training data, which is a cross-entropy error function.

3 Dependent vs Independent

As stated before, there are two different approaches to deal with verification problems. The former is the standard approach where a specific model is built for each writer. The later takes into account a global model and reduces any pattern recognition problem to a 2-class problem using the concept of dissimilarity. In the next subsections we discuss both strategies.

3.1 Writer Dependent

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 writer 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 writer 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 author verification.

3.2 Writer Independent and Dissimilarity

The idea of the writer-independent approach is to classify a handwriting samples, in terms of authenticity, into genuine or forgery, which means that any pattern recognition problem can be reduced to a 2-class problem. The approach we use in this work is the one employed by forensic experts, who compare the questioned samples with some references to assert whether a piece of handwriting is genuine or forgery. During this comparison, the experts extract different features to compute the level of similarity between the samples being compared.

The concepts of dissimilarity and proximity have been discussed in the literature from different perspectives [Santini and Jain, 1999, Goldfarb, 1992, Mottl et al, 2002, Pekalska and Duin, 2002]. Pekalska and Duin [Pekalska and Duin, 2002] introduce the idea of representing the relations between objects through dissimilarity, which they call dissimilarity representation. This concept describes each object by its dissimilarities to a set of prototype objects, called the representation set R. Each object x is represented by a vector of dissimilarities $D(x, R) = [d(x, p_1), d(x, p_2), \ldots, d(x, p_n)]$ to the objects $P_i \in R$.

Let R be a representation set composed of n objects. A training set T of m objects is represented as the $m \times n$ dissimilarity matrix $D(T, R)$. In this context, the usual way of classifying a new object x represented by $D(x, R)$ is by using the nearest neighbor rule. The object x is classified into the class of its nearest neighbor, that is the class of the representation object p_i given by $d(x, p_i) = \min_{p \in R} D(x, R)$. In another approach, each dimension corresponds to a dissimilarity $D(\cdot, p_i)$ to an object p_i. Hence, the dimensions convey a homogeneous type of information. The key here is that the dissimilarities should be small for similar objects (belonging to the same class) and large for distinct objects. In this way, $D(\cdot, p_i)$ can be interpreted as an attribute.

The concept of dissimilarity turns out to be very interesting when a feasible feature-based description of objects might be difficult to obtain or inefficient for learning purposes, e.g., when experts cannot define features in a straightforward way, when data are high dimensional, or when features consist of both continuous and categorical variables [Pekalska and Duin, 2002]. In the case of author verification, however, several different features have been proposed so that intra- and extra-class variation can be modeled.

In light of this, in this work we propose to combine feature-based description with the concept of dissimilarity. The idea is to extract the feature vectors from both questioned and reference texts and then compute what we call the dissimilarity feature vector. If both samples come from the same author (genuine), then all the components of such a vector should be close to 0, otherwise (forgery), the components should be far from 0.

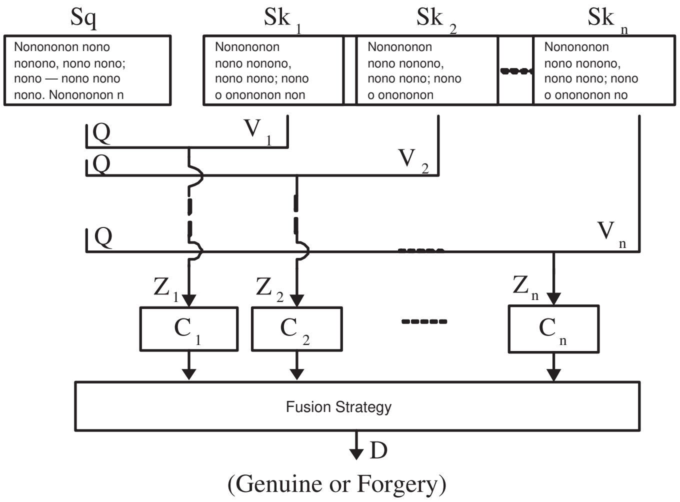

Figure 1: Architecture of the global approach.

To implement this, we use a reference set of n genuine text samples $Sk_i, (i = 1, 2, 3, \ldots, n)$ and then compare each Sk with a questioned sample Sq. Let V_i be the graphometric features extracted from the reference texts and Q the graphometric features extracted from the questioned texts. Then, the dissimilarity feature vectors $Z_i = |V_i - Q|$ are computed to feed the classifiers C_i, 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 depicts the global approach.

3.3 Database

To build the database we have collected articles available in the Internet from 30 different people with profiles ranging from sports to economics. Our sources were two different Brazilian newspapers, *Gazeta do Povo* (http://www.gazeta dopovo.com.br) and *Tribuna do Paraná* (http://www.parana-online.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 and 350 Hapax.

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. Figure 2 depicts an example of the article of our database.


COLUNISTAS

ANTÔNIO ERMÍRIO DE MORAES

Um programa que jamais deve ser esquecido

São Paulo é uma bela cidade que, infelizmente, está sendo atacada por pichadores de paredes que praticam um verdadeiro vandalismo contra o patrimônio público e privado. A eles se juntam os que utilizam as praças públicas para dormir, comer, lavar roupa e fazer necessidades. Mais triste é ver a proliferação de menores e adolescentes nos cruzamentos das ruas e pedir esmolas e vender quinquilharias.

O caso dos menores é muito triste por dois motivos. Primeiro, há a dimensão humana. O lugar de menores é na escola, em casa e em ambientes de lazer sadio. É aí que eles aprendem os valores fundamentais do convívio civilizado. Deixá-los nas ruas é submetê-los a um abandono que certamente irá deformar seu caráter.

Figure 2: An example of an article used in this work.

4 Implementation

This section describes how both strategies have been implemented. In both cases we have used a feature vector of 171 components, which is composed of 77 conjunctions and 94 adverbs. In order to extract the features, first the text is segmented into tokens. Spaces and end-of-line characters are not considered. All hyphenated 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.

4.1 Writer-Dependent

There are two basic approaches to solve q-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 $q(q-1)/2$.

From the database described previously, we have used 20 authors ($q = 20$, consequently 190 models). From each author 10 documents were used for training and 15 documents for testing.

4.2 Writer-Independent

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 3. 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 composed of 171 components is extracted from the questioned (Sq) and references (Sk_i) documents as well. This produces the aforementioned stylometric feature vectors V_i e 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 images, the questioned image 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 described in Section 2.2. Finally, the final decision can be taken based on different fusion strategies, but usually majority voting is used.

5 Results

In this section we report the experiments we have performed. In both strategies, different parameters and kernels for the SVM were tried out but the better results were yielded using a linear kernel.

Considering the writer-dependent model, the best result we got was 83.2% of recognition rate. As mentioned previously, few works have been done in the field of author verification for documents written in Portuguese. For this reason is quite difficult to make any kind of direct comparison. To the best of our knowledge, the only work dealing with author verification for documents written in Portuguese was proposed by Coutinho et al [Coutinho et al, 2004]. In this work the authors 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.

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 75.1%. Contrary to the writer-dependent approach where we have used a feature vector composed of conjunctions and adverbs, here the best results were produced using only 77 conjunction features. Table 5 summarizes the results.

Table 3: Results on the test set composed of 200 documents from 20 different authors

Strategy	Rec. Rate (%)
Writer-dependent	83.2%
Writer-independent	75.1%

To assess different fusion strategies, we have chosen the well-known ROC (Receiver Operating Characteristics). The area under the ROC (AUC) is convenient way of comparing classifiers. A random classifier has an area of 0.5, while and ideal one has an area of 1. We can observe from Figure 3 that the ROC with greatest AUC is the majority voting rule. This Figure corroborates to the choice of majority voting as fusion strategy.

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.

6 Conclusion

In this paper we have compared two different strategies for author verification using a feature set based on conjunctions and adverbs of the Portuguese language. We could observe that the writer-dependent method achieves better results but at an elevated computation cost. On the other hand, the writer-independent is quite simple as strategy and has a very accessible cost, but it has a bigger error rate. If the application has few writ-

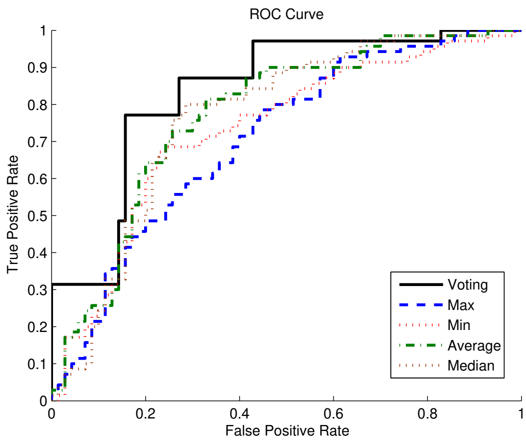

Figure 3: Comparison of different fusion strategies.

ers, the writer-dependent should the strategy to be considered. But if the number of writes gets bigger, writer-independent should be taken into account as alternative.

In spite of the fact that comparisons are always quite complicated since authors usually refer to specific databases for several different languages, we can observe that most of the works published in the literature yield results ranging from 70 to 80% of success [Coutinho et al, 2004] [Tas and Gurur, 2007], [Koppel and Schler, 2003]. In this context, the experiments we carried out using two different strategies of classification on a database composed of short articles form 30 different authors demonstrate that both strategies compares to the literature. As future work, we plan to define new features and new classification schemes so that the overall performance of the system could be improved.

Acknowledgements

This research has been supported by The National Council for Scientific and Technological Development (CNPq) grant 471496/2007-3.

References

[Argamon et al, 2003a] Argamon S., Koppel M., Fine J., and Shimony A. R.: The characteristic curves of composition; *Text*, 23,3 (2003).

[Argamon et al, 2003b] Argamon S., Saric M., and Stein S. S.: Style mining of electronic messages for multiple author discrimination; *Proc. ACM Conference on Knowledge Discovery and Data Mining* (2003).

[Baayen et al, 1996] Baayen H., van Halteren H., and Tweedie F.: Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution; *Literary and Linguistic Computing*, 11,3 (1996). 121–131.

[Cha and Srihari, 2002] Cha S-H and Srihari S. N.: On measuring the distance between histograms; *Pattern Recognition*, 35,6 (2002), 1355-1370.

[Chaski, 1998] Chaski C. E.: A daubert-inspired assessment of current techniques for language-based author identification; *ILE Technical Report* (1998).

[Chaski, 2005] Chaski C. E.: Who is At The Keyboard. Authorship Attribution in Digital Evidence Investigations; *International Journal of Digital Evidence*, 4,1 (2005).

[Coulthard, 2005] Coulthard M.: Author Identification, Idiolect, and Linguistic Uniqueness; *Applied Linguistics*, 25,4 (2004), 431-447.

[Coutinho et al, 2004] Coutinho B. C., Macedo L. M., Rique-JR A., Batista L. V.: Atribuição de Autoria usando PPM; *Proc. XXV Congress of the Brazilian Computer Society* (2004), 2208–2217.

[Forsyth and Holmes, 1996] Forsyth R. S. and Holmes D. I.: Feature finding for text classfication; *Literary and Linguistic Computing*, 11,4 (1996), 163–174.

[Goldfarb, 1992] Goldfarb L.: What is distance and why do we need the metric model for pattern learning; *Pattern Recognition* 25, 4 (1992), 431–438.

[Koppel and Schler, 2003] Koppel M. and Schler J.: Exploiting stylistic idiosyncrasies for authorship attribution; *Proc. Workshop on Computational Approaches to Style Analysis and Synthesis* (2003).

[Madigan et al, 2005] Madigan D., Genkin A., Lewis D. D., Argamon S., Fradkin D., and Ye L.: Author Identification on the Large Scale; *Proc. Joint Annual Meeting of the Interface and the Classification Society of North America* (2005).

[Mascol, 1888] Mascol C.: Curves of pauline and pseudo-pauline style I; *Unitarian Review*, 30 (1888), 453–460.

[Mendenhall, 1887] Mendenhall T.: The characteristic curves of composition; *Science*, 214 (1887), 237–249.

[Morton, 1978] Morton A.: Literary Detection; *Charles Scribners Sons* (1978).

[Mosteller and Wallace, 1964] Mosteller F. and Wallace D. L.: Inference and Disputed Authorship: The Federalist; *Series in behavioral science: Quantitative methods edition* (1964).

[Mottl et al, 2002] Mottl V., Seredin O., Dvoenko S., Kulikowski C., Muchnik I.: Featureless pattern recognition in an imaginary Hilbert space; *16th International Conference on Pattern Recognition* (2002), 88–91.

[Pekalska and Duin, 2002] Pekalska E. and Duin R. P. W.: Dissimilarity representations allow for building good classifiers; *Pattern Recognition* 23 (2002), 943–956.

[Platt, 1999] Platt J.: Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods; *Advances in Large Margin Classifiers* (1999), 61–74.

[Santini and Jain, 1999] Santini S. and Jain R.: Similarity Measures; *IEEE Trans. on Pattern Anaysis and Machine Intelligence* 21, 9 (1999), 871–883.

[Santos et al, 2004] Santos C., Justino E., Bortolozzi F., and Sabourin R.: An Off-Line SignatureVerification Method Based on Document Questioned Experts Approach and a Neural Network Classifier; *9th Int. Workshop on Frontiers in Handwriting Recognition* (2004), 498–502.

[Smadja, 1989] Smadja F.: Lexical co-occurrence: The missing link; *Journal of the Association for Literary and Linguistic Computing* 4, 3 (1989).

[Sollich, 2002] Sollich P.: Bayesian Methods for support vecotr machines: Evidence and predictive class probabilities; *Machine Learning* 46, 1-3 (2002), 21–52.

[Svartvik, 1968] Svartvik J.: The Evans statements: A case for forensic linguistics; *Acta Universitatis Gothoburgensis* (1968).

[Tambouratzis et al, 2004] Tambouratzis G., Markantonatou S., Hairetakis N., Vassiliou M., Carayannis G., Tambouratzis D.: 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 (2004), 221–242.

[Tas and Gurur, 2007] Tas T. and Gorur A. K.: Authro Identification for Turkish Texts; *Journal of Arts and Sciences*, 7 (2007), 151–161.

[Vapnik, 1995] Vapnik V.: The Nature of Statistical Learning Theory; *Springer-Verlag*, New York (1995).

[Wahba et al, 1999] Wahba G., Lin X., Gao F., Xiang D., Klein R., Klein B.: The bias-variance trade-off and the randomized GACV; *Proc. 13th Conference on Neural Information Processing Systems* (2001).