Identificação de Autoria de Textos através do uso de Classes Linguísticas da Língua Portuguesa

Paulo Júnior Varela[1], Edson J. R. Justino[2], Luiz E. S. Oliveira[3]

[1]Universidade Tecnológica Federal do Paraná
Linha Santa Bárbara, Francisco Beltrão, PR, Brasil

[2]Pontifícia Universidade Católica do Paraná
Rua Imaculada Conceição, 1155, Curitiba, PR, Brasil

[3]Universidade Federal do Paraná
Rua XV de Novembro, 1299, Curitiba, PR, Brasil

paulovarela@utfpr.edu.br, justino@ppgia.pucpr.br, lesoliveira@ufpr.br

Abstract. *The computational solution uses to solve problems related to the authorship identification and verification has grown progressively in areas such as computing, linguistics and law. This article aims to provide a method for the identification of authors ot text, based on a conjunct of attributes stilometry, using on the characteristics of Portuguese language.*

Resumo. *A utilização do meio computacional para a resolução de casos de identificação de autoria tem crescido progressivamente em área como a computação, a linguística e o direito. Este artigo tem por objetivo apresentar um método para identificação de autoria, com base em um conjunto de atributos estilométricos, utilizando características linguísticas do idioma português.*

1. Introdução

A linguística forense dedica-se à aplicação da estilística no contexto da identificação da autoria em documentos questionados. A identificação da autoria é realizada através da análise do estilo de escrita do autor. A estilística explora as duas premissas de variabilidade da linguagem. As mesmas estabelecem que, dois escritores de uma língua não escrevem exatamente do mesmo modo e um mesmo escritor, não escreve do mesmo modo todo o tempo [Black *et al* 1990].

A linguística forense divide a análise estilística em duas categorias, a qualitativa e a quantitativa [McMenamin 2002]. A análise qualitativa consiste em avaliar as formas usadas pelo autor, como e porque elas foram utilizadas. Sua grande limitação encontra-se no processo de inferência utilizado pelos peritos. Já o estudo quantitativo avalia a medida da variação na língua escrita. Uma das limitações reside na escassez de ferramentas de auxílio à análise. Uma abordagem quantitativa exige a mensuração dos atributos estilométricos [Johnstone 2000].

Este artigo apresenta um método para a identificação da autoria de documentos, com base em um dicionário de atributos estilométricos, com enfoque nas características linguísticas do idioma Português. A abordagem adotada é a quantitativa, tendo como objeto o desenvolvimento de métricas através de modelos computacionais.

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 174–178,
Cuiabá, MT, Brazil, October 24–26, 2011.

2. Atributos Estilométricos

A atribuição de autoria pode ser vista com uma classificação, onde documentos de autoria conhecida são utilizadas como treinamento com o objetivo de identificar autores corretos de documentos questionados (de autoria desconhecida) baseado em modelos que foram gerados [Varela *et al* 2010]. O principal problema é não ter certeza de quais características devem ser utilizadas para fazer a classificação, ou seja, para se distinguir os autores. [Pavelec *et al* 2008]

As caracteríticas estilométricas podem ser classificadas em 4 grupos (Figura 1) [Zheng *et al* 2006]. As características sintáticas utilizando palavras-funções perfazem o contexto deste artigo.

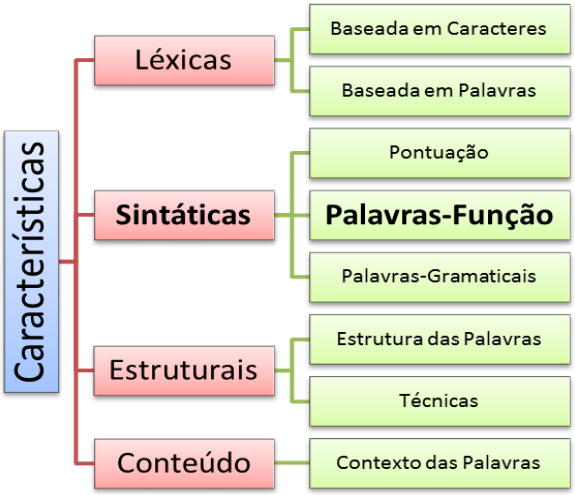

Figura 1. Grupos de Características de Estilo (Adaptado de [Zheng *et al* 2006])

O grupo das características sintáticas é formado por padrões responsáveis por formar as frases, tais como: pontuação, palavras função e palavras gramaticais. Muitos trabalhos utilizaram palavras-funções como característica discriminatória e obtiveram bons resultados na criação e na identificação de um perfil de um autor [Pavelec *et al*][Abassi e Chen 2005].

3. Base de Dados

Para os experimentos realizados foram escolhidos textos pequenos, entre 1KB (Kilobytes) e 9KB com no máximo 1200 *tokens* (número de palavras válidas pelo extrator de características) e em média 378 *tokens* por texto. Para avaliar o método proposto foram escolhidas colunas de 100 jornalistas e colunistas brasileiros de diferentes jornais, sendo que cada autor da base de dados possui uma amostra de 30 textos e pertence a uma classe de assunto: esportes, política, saúde, economia, direito, turismo, tecnologia, gastronomia, literatura e assuntos variados.

4. Método Proposto

O método proposto se baseia nos procedimentos de análise estilística forense (Figura 2). A mesma estabelece a associação ou dissociação da autoria do texto, em relação a um provável autor, como base num conjunto de atributos estilométricos previamente estabelecido. A associação indica a existência de atributos estilométricos suficiente para

garantir estatisticamente, que o texto, de autoria desconhecida, pertence ao autor avaliado. A dissociação indica que o mesmo não pertence ao autor avaliado.

Durante o processo de análise é utilizado um conjunto *n* de amostras de texto de autoria conhecida (modelos de referência), em comparação com a amostra de autoria desconhecida. Durante o processo é observado, tendo como base o conjunto de atributos estilométricos, diferenças de medidas entre as amostras conhecidas e a desconhecida e, posteriormente, é apresentado o resultado parcial da análise. O resultado final depende de regras de fusão para combinar as saídas do classificador (máximo, mínimo, média e voto majoritário), Figura 2.

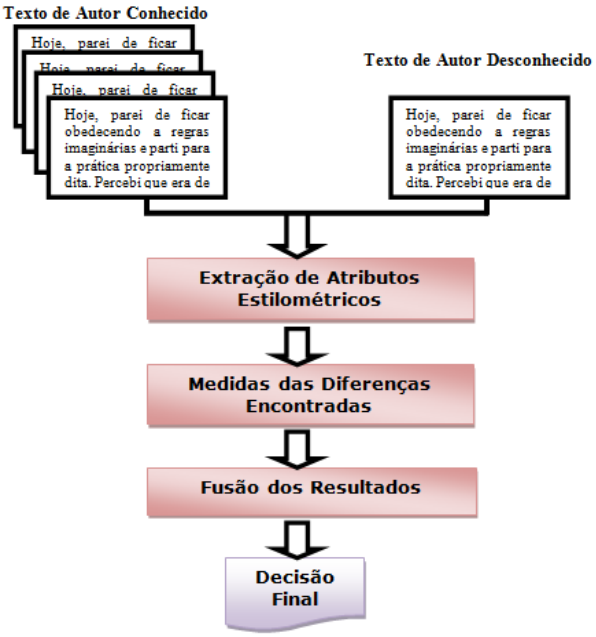

Figura 2. Diagrama do Procedimento de Análise Estilística Forense

Os atributos estilométricos usados são palavras-funções composto por verbos, pronomes, conjunções e advérbios [Varela *et al* 2010]. (Tabela 1)

Tabela 1. Palavras-Função

Tipo	Palavras-Função
Pronomes	quem, o qual, a qual, os quais, as quais, onde, em que, quanto, quanta, quantos, quantas, cujo, cuja, cujos, cujas, meu, minha, meus, minhas, teu, tua, teus, tuas, seu, sua, seus, suas, nosso, nossa, nossos, vosso, vossa, vossos, vossas, este, esta, estes, estas, isto, esse, esses, essa, essas, isso, 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, aquele, aquela, aqueles, aquelas, aquilo, nessa, desta, daquela, cujo, cuja, cujos, cujas, você, vocês, senhor, senhores, senhora, senhoras, senhorita, senhoritas, vossa senhoria, vossas senhorias.
Verbos	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.

Em outras palavras, o vetor de quantização dos atributos estilométricos é composto de 408 componentes, sendo que cada componente representa a quantidade de cada atributo estilométrico encontrado no texto. Para implementar o modelo

independente do autor, os vetores de quantização são usados para calcular os vetores de dissimilaridades, através da Equação 1.

$$V_i = |Ki - Q| \qquad \textbf{(1)}$$

4.1 Processo de Classificação

O processo de comparação é composto por duas fases, o treinamento e a verificação. No estágio de treinamento, as medidas de dissimilaridades V_i (i=1,2,3,..,n), são calculadas entre pares de textos. Quando dois textos pertencerem a um mesmo autor, o vetor de dissimilaridades é indicado com +1 (associação). Quando dois textos pertencerem a autores diferentes, o mesmo é indicado com -1 (dissociação). O vetor de dissimilaridades tenderá a possuir valores iguais a zero, quando as amostras pertencerem a um mesmo autor. Um SVM (*Support Vector Machine*) [Vapnik 1998] com *kernel* linear é então treinado para separar pequenas dissimilaridades entre atributos estilométricos (associação) e grandes dissimilaridades entre atributos estilométricos (dissociação).

5 Resultados

No protocolo de testes, foram utilizados como referência os 10 autores da classe de assuntos variados por serem mais generalistas quanto a amostras de seus textos. Foram utilizados 7 amostras de cada autor de forma aleatória, escolhida entre os 30 textos disponíveis. Na fase de teste, todos os 10 autores e suas 30 amostras de textos foram confrontados com 7 documentos de referência, gerando assim 2100 vetores de autoria e 2100 vetores de não autoria. Os resultados por classe e com as respectivas taxas de acerto (reconhecimento) são apresentados na Tabela 2.

Tabela 2. Resultados por Classes de Assuntos

Classe	Grupo	Acerto	Classe	Grupo	Acerto
Direito	Pronomes	68,2%	Política	Conjunções	75,8%
Economia	Pronomes	70,3%	Saúde	Advérbios	75,0%
Esportes	Pronomes	71,6%	Tecnologia	Advérbios	77,4%
Gastronomia	Pronomes	71,7%	Turismo	Conjunções	74,7%
Literatura	Verbos	65,6%			

6 Conclusão

O objetivo principal desse artigo foi apresentar um método computacional para a análise estilística forense, na identificação da autoria de textos. O modelo proposto mostrou-se robusto em textos com menos de 1200 palavras. O destaque ficou por conta dos pronomes que em mais de 44% das classes de assunto (direito, economia, esportes e gastronomia) foi o principal identificador, onde a especialização do mesmo permitiu um ganho médio de 5,6% em relação aos outros grupos de características, e de 4,7% em relação a todo o conjunto de características.

Como proposta para trabalhos futuros encontra-se incorporar a classe de características estruturais ao conjunto, a fim de avaliar as contribuições dessa classe de atributos no conjunto. Uma segunda proposta seria incorporar novos grupos de características estilométricas da língua portuguesa.

Referencias

Abbasi, A. Chen, H. (2005) “Applying authorship analysis to extremist group web forum messages”. IEEE Intelligent Systems, p. 67–75.

Black, H. C., Nolan, J.R., Nolan-Haley, J.M. (1990).“Black’s Law Dictionary”. West Publishing, 6 edition, St. Paul, p. 1810.

Johnstone, B. (2000) “Qualitative Methods in Socioliguistics”. Oxford University Press, New York, p. 450.

McMenamin, Gerald R. (2002) “Forensic Linguistics - Advances in Forensic Stylistics”. CRC Press, p. 330.

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

Vapnik, V.(1998) “Statistical learning theory”. Wiley, N. Y. p. 768.

Varela, P. Justino E. Oliveira, L. (2010). “Verbs and Pronouns for Authorship Attribution”. International Conference on Systems, Signals and Image Processing, Rio de Janeiro, p. 89-92.

Zheng, R. Qin, Y. Huang, Z and Chen, H. (2006) “A framework for authorship analysis of online messages: Writing-style features and techniques”. Journal of the American Society for Information Science and Technology, p. 378–393.