An Authorship Attribution for Serbian

Anđelka Zečević
Faculty of Mathematics
Studentski trg 16
Belgrade, Serbia
andjelkaz@matf.bg.ac.rs

Miloš Utvić
Faculty of Philology
Studentski trg 3
Belgrade, Serbia
misko@matf.bg.ac.rs

ABSTRACT

An authorship attribution is a problem of identifying the author of an anonymous or disputed text if there is a closed set of candidate authors. Due to the richness of natural languages and numerous ways of expressing individuality in a writing process, this task employs all the sources of language knowledge: lexis, syntax, semantics, orthography, etc. The impressive results of n-gram based algorithms have been presented in many papers for many languages so far. The goal of our research was to test if this group of algorithms works equally well on Serbian and if it is a case, to calculate the optimal values for the parameters appearing in the algorithms. Also, we wanted to test if a syllable based word decomposition, which represents a more human like word decomposition in comparison to n-grams, can be useful in an authorship attribution. Our results confirm good performance of an n-gram based approach (accuracy up to 96%) and show the potential usefulness of a syllable based approach (accuracy from 81% to 89%).

Keywords

Authorship attribution, classification, n-grams, syllables

1. INTRODUCTION

By definition, an authorship attribution is a problem related to identifying the author of an anonymous or disputed text if there is a closed set of candidate authors. One of the first studies concerning this topic was published in 1787 by Edmond Malone [12] who argued that Shakespeare did not write some parts of *Henry VI*. His evidences was based on the analyses of meter and rhyme and there was highly disagreement between Shakespeare's and the real author's style. Probably the most influential study is done by Mosteller and Wallace [15] in 1964 on the authorship of *The federalist papers*, a series of 85 essays written by John Jay, Alexander Hamilton and James Madison on promotion of the ratification of the United States Constitution. Nowadays, a focus of an authorship is put on modern text forms such as e-mail messages [10], SMS text messages [14], source codes [1] or blog posts [11].

All the approaches to an authorship attribution problem are based on the fact that the author's individuality impacts on his or her writing in a unique and recognisable manner. Stylometry is a field that deals with defining and analysing relevant text features (so called style markers) that can serve as an author's fingerprint. So far, numerous text features have been considered [7, 19, 5, 8]. Some of them exploit text surface and take into account an average word length or vocabulary richness while there are more complex ones dealing with text semantics or syntax trees. This large set of features influences the choice of algorithms as well as methods for a text comparison.

From machine learning point of view, an authorship attribution problem is considered as a classification task [17]: a text of unknown authorship is assigned to one of the authors from the given set of candidate authors. This treatment put at researchers' disposal all algorithms developed by the machine learning community (neural networks, support vector machines, memory based learning algorithms, Bayesian learning, etc.) and enables them to present their data and results in a mathematically well founded manner.

The remainder of the paper is organized as follows. In Sections 2 and 3 we introduce byte level n-grams and syllables as text features. In Section 4 we define two author profiles: first one is based on n-grams and the second is based on syllables. Distance measures for comparing profiles are introduced in Section 5. In section 6 we propose the structure of a profile based approach and discuss important steps of the algorithm. Measures for estimating the effectiveness of the classification are presented in Section 7. Section 8 summarizes obtained results, and finally, Section 9 presents some conclusions and future directions.

2. N-GRAMS

An n-gram is a continuous sequence of n bytes or n characters or n words of a longer portion of a text. Therefore, we distinguish *byte level*, *character level* and *word level* n-grams. Our focus is on byte level n-grams which representation depends on character encoding. For instance, if we consider standard ASCII encoding and a portion of a text *abc*, all byte level 2-grams are *01100001 01100010* and *01100010 01100011* where the code values 01100001, 01100010 and 01100011 correspond to the characters *a*, *b* and *c* respectively.

The general strengths of a byte level n-gram approach are a language independent processing and a computational simplicity. Further more, for different values of a parameter n, n-grams afford tracking of lexical, contextual or formatting information. N-gram approaches are tolerant of noise too, and behave more robustly in presence of different kind of textual errors. On the other side, adjacent n-grams overlap and

BCI'12, September 16–20, 2012, Novi Sad, Serbia.

Local Proceedings also appeared in ISBN 978-86-7031-200-5, Faculty of Sciences, University of Novi Sad.

contain redundant information so the memory requirements are more intensive in comparison to the other methods. If a portion of a text is k bytes long, the number of byte level n-grams is $k + 1 - n$, so the total size of storing memory is $(k + 1 - n) \cdot n$.

3. SYLLABLES

A syllable is defined[1] as a unit of pronunciation having one vowel sound, with or without surrounding consonants, and forming all or part of a word. Decomposition of words into syllables is not always easy and unique. Generally, every syllable requires a nucleus. Syllable nuclei in Serbian are vowels and sonorants like 'r', 'l' and 'n'. Serbian syllables can be open (if they end with a vowel) or closed (if they end with a consonant). The boundary between subsequent syllables in a word in Serbian is usually placed after a vowel. The rules of syllabication in Serbian are based on phonetic and semantic characteristics [16].

Although there are software packages and resources available for automatic syllabication of Serbian (RAS,[2] Hunspell[3] dictionaries and hyphen patterns for OpenOffice,[4]) in the first stage of our experiment we used a "naive" algorithm which sets syllable boundary after vowels and sonorant 'r'.

4. AN AUTHOR PROFILE

To study an author's style we require some operative representation based on his or her writings. This representation is called an author profile and consists of selected text features. The set of the features does not need to be homogeneous, which means that numerous features can be combined in order to obtain qualitative representation able to capture all inter-author style variations. On the other hand, the set of features should be able to distinguish authors among themselves and should be something specific for a concrete author.

4.1 N-gram Based Profiles

First author profile we used treats byte level n-grams as most relevant text features. It is defined as a set of pairs

$$P_A = \{(x_1, f_1), (x_2, f_2), \dots (x_M, f_M)\}$$

where x_i denotes an n-gram value and f_i its relative frequency. The relative frequency is calculated as the total number of the n-gram occurrences divided by the total number of n-grams. Pairs in the profile are ordered in respect to a relative frequency: from the highest to the lowest values. The number of pairs M is called a profile size and represents a very important parameter of n-gram based algorithms.

This profile is originally proposed by Keselj et al. [9] and has been applied on many languages with great success.

4.2 Syllable Based Profiles

There is a number of papers authored or co-authored by Wilhelm Fucks [2, 3, 4] on a syllables' role in an author identification process. He considered an average number of syllables per word, a word length frequency distribution in syllables (the number of monosyllabic words, the number of disyllabic words and so on) and the average distance between i-syllable words ($i \geq 1$). In a later studies [7], it is concluded that frequency distribution of syllables per word discriminates different languages more than specific authors as well as that the overall distribution of syllable counts changes from one kind of writing to another.

A profile based on syllables we used in our research consists of most frequent syllables in respect to their absolute frequency. The form of the profile is

$$P_A = \{(s_1, F_1), (s_2, F_2), \dots (s_M, F_M)\}$$

where s_i denotes a syllable and F_i its absolute frequency (the total number of its occurrences). A parameter M still represents a profile size.

The main motivation for the use of these profiles relies on the fact that for small values of the parameter n n-grams are able to represent syllable-like information. These profiles can be also observed as variable-length n-gram profiles and used in cases when the optimal value of the parameter n is unknown.

5. DISTANCE MEASURES

5.1 N-gram Based Profiles

The measure we used to compare the profile P_{A_i} of the i-th author and the profile P_a of an anonymous or disputed text is defined by formula

$$d(P_{A_i}, P_a) = \sum_{x \in P_a} \left(\frac{2 \cdot (f_{A_i}(x) - f_a(x))}{f_{A_i}(x) + f_a(x)} \right)^2$$

where x is a byte-level n-gram and $f_{A_i}(x)$ and $f_a(x)$ are the relative frequencies of the n-gram x in the author's profile and the profile of the text of unknown authorship respectively. This measure is originally proposed by Stamatatos [18] and represents the combination of measures proposed by Keselj et al. [9]

$$d(P_{A_i}, P_a) = \sum_{x \in P_a \cup P_{A_i}} \left(\frac{2 \cdot (f_{A_i}(x) - f_a(x))}{f_{A_i}(x) + f_a(x)} \right)^2$$

and Frantzeskou et al. [1]

$$d(P_{A_i}, P_a) = |P_{A_i} \cap P_a|$$

in order to improve measures' tolerance to a class imbalance problem. The class imbalance problem [6] appears when at least one profile is smaller or larger than the others. This is a very realistic situation in author identification problems since there might be only a few text samples for one candidate author and many more text samples for the other authors, or vice versa. The measure proposed by Keselj et al. [9] favours authors with shorter profiles because the union of the profiles is taken into account. On the other hand, the measure proposed by Frantzeskou et al. [1] favours authors with longer profiles since the size of the intersection of two profiles is considered.

The presented measure is actually a pseudo measure because it leaks a symmetry property - the values P_{A_i} and P_a cannot be switched. The results obtained in an experimental testing [18] are very promising and encourage researchers to manipulate with it in spite of its drawback.

[1] http://oxforddictionaries.com/
[2] http://www.rasprog.com/
[3] http://hunspell.sourceforge.net/
[4] http://ooo.matf.bg.ac.rs/dict-sr/

5.2 Syllable Based Profiles

For comparing syllable based profiles we used measure proposed by Frantzeskou et al. [1] except we used syllables instead of n-grams. The measure

$$d(P_{A_i}, P_a) = |P_{A_i} \cap P_a|$$

counts the total number of common syllables in the profile P_{A_i} of i-th author and the profile P_a of an anonymous or disputed text.

6. PROFILE BASED LEARNING

The scheme of our algorithm is depicted in Figure 1 and represents a classical profile-based algorithm.

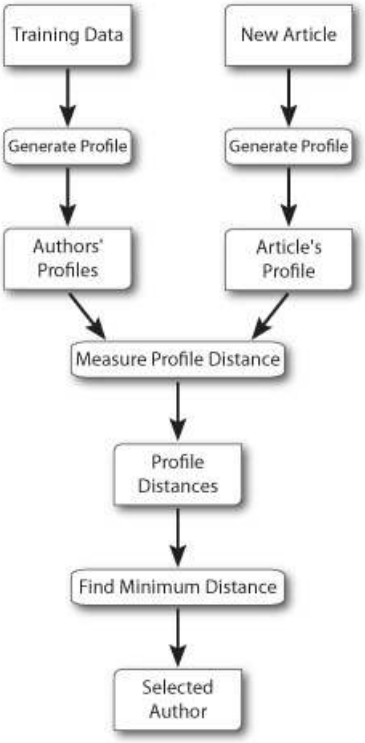

Figure 1: The algorithm scheme.

Step 1: The training data set consists of undisputed text samples of authors. All text samples per author are concatenated in one large text file and then the set of M most relevant n-grams or syllables is extracted to obtain the author profile.

Step 2: When a text of unknown authorship should be classified, the set of its M most relevant n-grams or syllables is extracted. The values of the parameters M and n are the same as the values used in Step 1.

Step 3: The profile of the text of unknown authorship is compared to the all authors' profiles in respect to the measures defined in the previous section.

Step 4: The obtained values are analysed by the system and the smallest value is picked.

Step 5: The author we treat as the writer of the unclassified text is the one who's index corresponds to the index of the selected value.

In the background of the authorship attribution algorithm is a *k Nearest Neighbour* classification algorithm [13] with the parameter k set to 1. It represents memory based classification algorithms and assigns an unclassified instance to one of the given classes according to minimum-distance principle.

7. CLASSIFICATION EFFECTIVENESS

For estimating the effectiveness [17] of a single class C_i classification we have used accuracy

$$A_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$

Values TP_i, TN_i, FP_i and FN_i are values from a confusion matrix (Table 1) and represent, respectively, the number of *yes-yes*, *no-no*, *yes-no* and *no-yes* labeled instances.

Table 1: A Confusion Matrix

class C_i		actual class	
		yes	no
predicted class	yes	TP_i	FP_i
	no	FN_i	TN_i

For overall estimation of effectiveness we used macroaverage of individual values:

$$A = \frac{\sum_{i=1}^{c} A_i}{c}$$

where c denotes the total number of classes.

The choice of a measure was strongly influenced by the current state-of-the-art results which are presented in respect to accuracy. In order to make our results comparable, we have chosen the same measure.

8. RESULTS

We experimented with a set of newspapers articles[5] written independently by six authors. In order to achieve the authorship is the most important discriminatory feature among the authors, the selected articles meet a number of specific criteria. For the purpose of avoiding an author's style change over time, all articles per author are written in the same period (within one year). To minimize the topic influence, we have only chosen articles that describe political situation in the country. All the texts (newspaper articles) are of the same genre, too. The number of articles per author and the total size of the training set is presented in Table 2.

The test set consists of non-overlapping articles and follows the distribution of the training set. The number of articles per author and the total size of the test set is presented in Table 3.

8.1 N-gram Based Profiles

The tested values of the parameter n are in the interval from 1 to 10 and the tested values for the parameter

[5]http://www.danas.rs

Table 2: Authors in Training Set

	Author name	*number of articles*	*train size (in bytes)*
A_1	Safeta Bišević	20	103,761
A_2	Zoran Panović	17	100,706
A_3	Aleksandar Roknić	27	101,809
A_4	Snežana Čongradin	28	102,756
A_5	Svetislav Basara	25	78,891
A_5	Miloš Vasić	18	102,875

Table 3: Authors in Test Set

	Author name	*number of articles*	*test size (in bytes)*
A_1	Safeta Bišević	10	82,945
A_2	Zoran Panović	9	55,415
A_3	Aleksandar Roknić	13	50,193
A_4	Snežana Čongradin	14	56,558
A_5	Svetislav Basara	12	64,684
A_5	Miloš Vasić	9	47,655

M are 20, 100, 500, 1,000, 2,000, 3,000, 4,000 and 5,000. The system achieves accuracy over 80% for all n-gram sizes greater then 3 and the profile sizes greater then 500. The best achieved results are for the parameter n between 4 and 7 and for the profile size M between 1,000 and 4,000. The best achieved accuracy at all is 0.96 for $n = 5$ and $M = 3,000$.

8.2 Syllable Based Profiles

The algorithm is tested for the parameter M with values from 100 to 1,200 by step 100. The values were limited by the maximal number of syllables per author. The results are presented in Table 4 in respect to accuracy.

Table 4: The Results for Syllable Based Profiles

M	*accuracy*	*M*	*accuracy*
100	0.81	700	0.86
200	0.85	800	0.84
300	0.88	900	0.87
400	0.88	1,000	0.85
500	**0.89**	1,100	0.85
600	**0.89**	1,200	0.86

9. CONCLUSIONS

This paper presents some insights into an authorship attribution problem for Serbian. The n-gram based approach proved its good performance and achieved accuracy from 80% up to 96% for the parameter $4 \leq n \leq 7$, as well as the syllable based approach with accuracy between 81% and 89%.

In the future, both n-gram based and syllable approaches, combined with the wider set of measures, should be tested on expanded corpora and longer list of authors. We also plan to improve a syllabication phase since the results of syllable based approach are promising.

10. ACKNOWLEDGMENTS

This research was supported by the Serbian Ministry of Education and Science under the grant 178006 (Serbian Language and its Resources).

11. REFERENCES

[1] G. Frantzeskou, E. Stamatatos, S. Gritzalis, C. Chaski, and B. Howald. Identifying authorship by byte-level n-grams: The SCAP method. *International Journal of Digital Evidence*, 6(1), 2007.

[2] W. Fucks. On mathematical analysis of style. *Biometrika*, (39):122–129, 1952.

[3] W. Fucks. On nahordnung and fernordnung in samples of literary texts. *Biometrika*, (41):116–132, 1954.

[4] W. Fucks and J. Lauter. Mathematische analyse des literarischen stils. *Mathematik und Dichtung*, pages 107–123, 1965.

[5] J. Grieve. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistics Computing*, 22(3):251–270, 2007.

[6] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou. On the class imbalance problem. In *Proc. of Fourth Int. Conf. on Natural Computation*, pages 192–201, 2008.

[7] D. Holmes. Authorship attribution. In *Computer and the Humanities*, volume 28, pages 87–106. 1994.

[8] P. Juola. Authorship attribution. *Foundation and Trends in Information Retrieval*, 1(3):233–334, 2006.

[9] V. Keselj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the Pacific Association for Computer Linguistics*, pages 255–264, 2003.

[10] M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, pages 69–72, 2003.

[11] M. Koppel, J. Schler, S. Argamon, and E. Messeri. Authorship attribution with thousands of candidate authors. pages 659–660, 2006.

[12] E. Malone. A dissertation on parts one, two and three of Henry the Sixth tending to show that those playings were not written originally by Shakespeare, 1787.

[13] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[14] A. Mohan, I. Baggili, and M. Rogers. Authorship attribution of SMS messages using an n-grams approach. Technical report, 2010.

[15] F. Mosteller and D. Wallace. Inference and disputed authorship: The Federalist. 1964.

[16] M. Pešikan, J. Jerković, and M. Pižurica. *Pravopis srpskoga jezika*. Matica srpska, 2009.

[17] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[18] E. Stamatatos. Author identification using imbalanced and limited training texts. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications*, pages 237–241, 2007.

[19] E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.