Automatic authorship attribution for Croatian texts

Igor Belša Tomislav Reicher Ivan Krišto Artur Šilić
Faculty of Electrical and Computing Engineering, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
{igor.belsa, tomislav.reicher, ivan.kristo, artur.silic}@fer.hr

Abstract. *Automatic authorship attribution is useful in automatic text classification which becomes quite complex task if it's applied in large data sets (such as application at web search engines). Most of reported methods are language specific, therefore we have developed methods specific for Croatian language which range from simple, based on stylistic measures and functional words, to complex which require syntactic analysis of Croatian language. We have found that various combinations of methods are very successful in solving the authorship attribution problem.*

Keywords. text classification, SVM, functional words, MSD tagging.

1. Introduction

Automatic authorship attribution can be interpreted as a problem of text classification based on linguistic features that are specific to certain authors. Problems similar to authorship attribution are detection of age, region and gender of the author [11]. Main concern of computer–assisted authorship attribution is to define an appropriate characterization of text that captures the writing style of authors [4].

Authorship attribution can help in document indexing, document filtering and hierarchical categorization of web pages [11]. These applications are important information retrieval tasks. It is important to note that authorship attribution differs from plagiarism detection in that the latter attempts to detect similarities between two substantially different pieces of work but is unable to determine if they were produced by the same author [5].

The problem can be divided in three categories [17]: binary, multi–class and single–class (or one–class) classification. Binary classification solves the problem when each text from data set is written by one of two authors. Multi–class classification is generalization of binary classification in which we have more than two authors. One–class classification is applied when some of the texts from the data set are written by a particular author while the authorship of the other texts is unspecified. This classification answers the question does a given text belong to a single known author or not.

This paper presents a study of multi–class classification for texts in Croatian language, oriented on evaluation of different text representations.

The rest of the paper is organized as follows. Section 2 discusses related work on authorship attribution and similar problems. Section 3 describes the used data set. Section 4 introduces methods of text representation we have used. Section 5 describes classification and Section 6 presents evaluation methods and experiment results. Conclusion and future work are given in Section 7.

2. Related work

Approaches to text representation and classification for authorship attribution include: stylistic measures [4], syntactic clues [13, 15], word-based features [1, 15], compression algorithms [10, 17], grammatical error lookup [9], language modeling [4, 12], and vocabulary diversity [13]. Following paragraphs describe these approaches in more depth and relate our work with the existing research.

Coyotl-Morales et al. [4] group methods of authorship attribution into three main approaches: *stylistic measures*, *syntactic cues* and *word–based* text features.

Stylistic measures based text features is characterized by features that take into account length of words and sentences and the richness of the vocabulary. Various types of style markers can be found in [11]. In [4] it is noted that such features are not sufficient to resolve problems that depend on the genre of the text and that they lose their meaning when applied over short texts.

When using the *syntactic cues as text features* – informations related to structure of the language which are obtained by an in–depth syntactic analysis of texts, a single text is characterized by the presence and frequency of certain syntactic structures. This characterization is detailed and relevant. Unfortunately, it is computationally expensive and even impossible to build for languages lacking the text–processing resources (e.g. POS tagger, syntactic parser, etc.). It is clearly influenced by the length of texts. Description of authorship attribution process by using syntactic elements can be found in a very influential work by Stamatos et al. [13].

Approach which uses *Word–based* text features branches in three different methods. *The first one* characterizes texts using a set of functional words (their presence and frequency), ignoring the content words since they tend to be highly correlated with the text topics. *Second branch* of methods observes a text as a bag–of–words and uses single content words as text features. This method can be applied only when there is a noticeable correlation between the authors and the topics. The *third branch* considers word *n*–grams as features, i.e. features consisting of sequences of consecutive words.

The success of functional words over collocations (certain pairs of words occurring within a given threshold distance of each other) is shown in [1] – however, because of information reduction when using functional words as features, one might conclude the opposite – which would be wrong. Argamon and Levitan [1] believe that most of the discriminating power of collocations is due to the frequent words they contain and not the collocations themselves. They also note that using more training texts than features seriously reduces the likelihood of overfitting the model to the training data, improving the reliability of results. Similar claim can be found in [2] which deals with influence the of corpus size in general natural language processing.

A similar comparison of feature types is shown in [15] where the functional words and syntactic elements are compared in order to identify the author of text. It is concluded that the syntactic elements of expressions are useful as functional words in solving the problem.

Kukushkina et al. [10] explain a method which uses algorithms for data compression to identify the author. The appendix of this paper shows evaluation results for authorship attribution with different compression algorithms. The idea behind the method is to divide texts by authors, compress every set with selected algorithm and write the size of the archive. To classify text of an unknown author, text is added to each set and than the same data compression algorithm is applied. The author of the texts in archive which records the smallest increase in size is declared as the author of the new text. Review of a similar method can be found in [17], where it is noted that the method has obvious omissions. Also, a citation on another work that proves it is given.

Koppel and Schler [9] show the use of grammatical errors and informal styling (e.g. writing sentences in capital letters) as text features in order to identify the author. Method is applicable only to unedited texts (blogs, Internet forums, newsgroups, e–mail messages, etc.).

Peng et al. [12] suggest the use of *n*–gram language model to identify the author. A similar method is shown in [4].

Stamatatos et al. [13] suggest vocabulary diversity of the authors as feature for identifying the author. It is measured by the ratio of unique words and the total size of the text or counting words which occur only once (*hapax legomena*), or twice (*dis legomena*), etc. These measures are closely related to length of the text. They also note that functions defined by Yule (1944) and Honore (1979) should solve this problem, i.e. they should be constant regarding the length of the text. Analysis of similar function can be found in the [14] which claims that most of these functions are not independent regarding the length of the text. It is concluded that the vocabulary diversity is an unstable feature for texts shorter than 1000 words.

Our work is a mix of syntactic clues, word–based features, and stylistic measures, adapted for Croatian language (prosiriti).

3. Data set

We used an online archive of columns from a daily Croatian newspaper “Jutarnji list” available at http://www.jutarnji.hr/komentari/. The data set consists of 4571 texts written by 25 authors. The lowest average number of words in text per author is 315 words, and the highest is 1347 words. An average number of words in text per author for this data set is 717 words. Number of texts per author in the set is shown on Figure 1.

Because topics in columns tend to be time–specific and in order to avoid the learning of topic, the data set is split by dates – 20% of newest texts of each author are taken for testing (hold–out method). Therefore, training data set contains 3425 texts and testing data set 1146 texts.

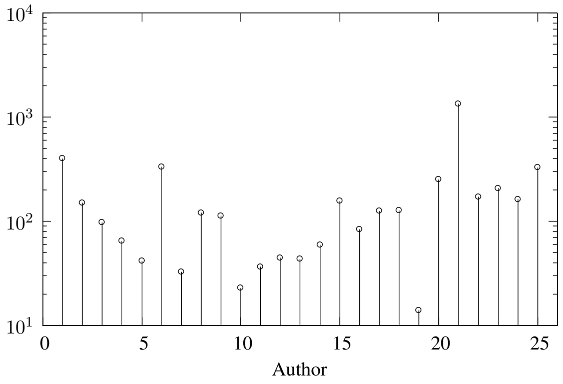

Figure 1: **Number of texts per author.**

4. Text features

When constructing an authorship attribution system, the central issuse is the selection of sufficiently discriminative features. A feature is discriminative if it is common for one author and rare for all the others. Due to large number of authors complex features are very useful if their distribution is specific for each author.

During the creation of the feature vector, it is neccessary to make the features independent of the length or of the content of each text. That dependency reduces generality of system’s application and can lead to a decrease of accurancy (e.g. relating author to concrete topic or terms).

Since features are expressed as real numbers, each text is expressed as a real vector of appropriate number of dimensions. We combine features simply by creating unions of basic feature sets.

4.1. Functional words frequency ($\mathcal{F}$)

Functional words are words that have few or none semantic content of their own, such as adverbs, prepositions, conjunctions, or interjections. They usually indicate a grammatical relationship or generic property [17].

Some of the less known functional words, such as prepositions *onkraj*, *namjesto* and *zavrh* are rarely used and may very well suggest the author. However, even the frequency of frequent functional words can distinguish the author very well. Due to high frequency of use of functional words and their roles in the grammar, the author usually has no conscious control over their use in a particular text [1].

Functional words are topic–independent, authors automatically (without conscious control) use functional words which indicate their own style. It is difficult to predict whether the functional words will give equally good results for different languages. Despite the size of research on this topic, due to various languages, the type and the size of the texts, it is difficult to conclude if these methods are generally effective [17].

In addition to functional words, auxiliary verbs and pronouns are considered. Their frequencies might be representative for different author styles.

Building the text feature vector is done by counting the appearances of functional words, auxiliary verbs and pronouns in the text. The results are written as vectors where each component corresponds to the number of occurences of related words divided by the total number of the words in the text. This is done to remove the dependency on the length of the text. For the functional words that have not appeared in a text, the resulting component always equals zero.

Marker $\mathcal{F}$ denotes evaluation result of this feature in table 1.

4.2. Lexical categories frequency ($\mathcal{C}$)

Building of the feature vector is done by counting the number of appearances of different lexical categories where the categories considered are adverbs, adpositions, conjunctions, particles, interjections, nouns, verbs, adjectives and pronouns which are obtained by syntactic analysis of Croatian language. Result is written as a vector (nine–dimensional vector) and each component of the vector is divided by the number of words in the text.

4.3. Idf weighted functional words frequency ($\mathcal{I}$)

The problem of dependency on the length of the text for the classification using a SVM (Support Vector Machine) is explained in [6]. To avoid the dependency, a combination of L_p normalization of the length and transformation of terms occurrence frequency such as *idf* (inverse document frequency) measure was used.

Idf measure is defined as [6]

$$F_{idf}(t_k) = \log \frac{n_d}{n_d(t_k)}, \tag{1}$$

where $n_d(t_k)$ is number of documents which contain term t_k and n_d total number of documents. Shown measure gives high values for terms which appear in a small number of documents and thus it is very discriminatory.

The feature vector is built by multiplying components of vector created by method defined in 4.1 and its associated *idf* weight.

Shown measure discriminate documents that contain functional words that are used in small number of other documents. The disadvantage of *idf* measure is that it records only presence of certain term in the document, term counting in the document is ignored. Therefore, word which appears many times in one text and one time in the others gets the same value as the one which appears once in all text—the word which would very well separate one document from the others is ignored.

4.4. Punctation marks ($\mathcal{P}$), vowels ($\mathcal{V}$), word length ($\mathcal{L}$) and sentence length frequency ($\mathcal{S}$)

A set of following punctuation marks is used: “.”, “,”, “!”, “?”, ““””, “"”, “-”, “:”, “;”, “+”, “*” and their number of appearance in text is counted. Result is written as 11–dimensional vector and every component is divided by total sum of characters in the text.

A feature vector based on the frequency of vowel occurence (a, e, i, o, u) is obtained in equal procedure as for the punctuation marks.

Frequency of word lengths are obtained by counting the words that have equal length. It is important to note that this procedure can lead to vectors of different features dimensions (e.g. a text has a word with length of 11, but some other text does not). The issue is solved by limiting the maximum length of words at length of 10. All words longer than 10 characters are counted to the 10th group. It is necessary to divide components of resulting a feature vector with the number of words in the text in order to diminish the dependency of features with the text's length.

The sentence length frequency is obtained in equal procedure as for the word length frequency. An different vectors dimension issue is solved by limiting sentence length to 20.

Suggested features have weak discriminatory power on their own, but they are very useful in combination with other features as shown in Section 6.

4.5. Word part–of–speech *n*–grams frequency ($\mathcal{N}_1$ & $\mathcal{N}_2$)

The two proposed features are based on word part–of–speech *n*–grams frequency. Word parts–of–speech and their morphosyntactic descriptors are obtained by POS (Part–Of–Speech) and MSD (morphosyntactic) tagging for Croatian language [16]. Having the corresponding part–of–speech, for each word in the text, makes the idea of using *n*–grams as features possible. As *n*–grams features can produce very large dimensionality only 3–grams are considered. In addition POS tagging used is not perfect, the method doesn't use context information and therefore cannot distinguish between different homographs—words with same spelling but with different meaning and probably different POS too—so all possible POS tags for given word are considered.

First proposed feature uses the words parts–of-speech (nouns, verbs, adjectives, pronouns, conjunctions, interjections and prepositions) to form various *n*–grams and count their frequencies. For example, word 3–gram “Adam i Eva” (“Adam and Eve”) forms “noun conjunction noun” trigram. Second proposed feature uses only words parts–of-speech information for nouns, verbs and adjectives and for other word parts–of-speech it uses words as they are, therefore “Adam i Eva” transforms to “noun i noun”. Due to many different pronouns, conjunctions, interjections and prepositions that make many different *n*–grams, frequency filtering is applied—only frequency of

500 most frequent 3–grams in training data set is considered. Used dimension reduction method is not optimal, therefore in future work other methods should be evaluated, such as information gain, χ^2 test, mutual information, maximum relevance, minimum redundancy or classification with sparse SVM, logistic regression or naïve Bayes.

4.6. Word morphologic categories ($\mathcal{M}$)

Building of feature vector is done by counting appearances of morphologic categories for every word in text and dividing them by number of words in text. Counted morphologic categories are *case*, *degree*,*form*, *gender*, *number* and *person*.

Table 1: **Evaluation of different features**

Method	Accuracy	C	γ
$\mathcal{F}$	88.39%	8192	0.125
$\mathcal{I}$	87.96%	8192	0.125
$\mathcal{C}$	44.50%	512	2.0
$\mathcal{P}$	57.50%	8192	0.125
$\mathcal{V}$	30.54%	128	0.125
$\mathcal{L}$	43.19%	128	0.125
$\mathcal{S}$	42.32%	128	0.125
$\mathcal{N}_1$	71.29%	512	0.125
$\mathcal{N}_2$	76.09%	512	0.125
$\mathcal{M}$	61.17%	512	0.125
$\mathcal{C}, \mathcal{M}$	63.17%	8192	0.03125
$\mathcal{P}, \mathcal{F}$	91.71%	8	0.03125
$\mathcal{F}, \mathcal{M}$	91.18%	128	0.03125
$\mathcal{F}, \mathcal{N}_1$	89.44%	128	0.03125
$\mathcal{F}, \mathcal{N}_2$	88.48%	128	0.03125
$\mathcal{I}, \mathcal{M}$	90.84%	128	0.03125
$\mathcal{N}_1, \mathcal{M}$	71.38%	128	0.03125
$\mathcal{I}, \mathcal{M}, \mathcal{C}$	91.36%	128	0.03125
$\mathcal{P}, \mathcal{F}, \mathcal{L}$	**93.11**%	128	0.03125
$\mathcal{P}, \mathcal{F}, \mathcal{L}, \mathcal{M}$	**93.46**%	32768	0.03125
$\mathcal{F}, \mathcal{M}, \mathcal{C}, \mathcal{N}_1$	89.62%	128	0.03125
$\mathcal{S}, \mathcal{P}, \mathcal{F}, \mathcal{V}, \mathcal{L}, \mathcal{M}$	**93.19**%	128	0.03125
$\mathcal{S}, \mathcal{P}, \mathcal{F}, \mathcal{V}, \mathcal{L}, \mathcal{M}, \mathcal{N}_1$	92.41%	128	0.03125

5. Classification

The classifier used is SVM (Support Vector Machine) with radial basis function as the kernel. It is shown that, with the use of parameter selection, linear SVM is special case of SVM with RBF kernel [8] what removes the need to consider linear SVM as potential classifier.

Before the use of SVM, it is required to scale the data to ensure equal contribution of every attribute to the classification. Components of every feature vector are scaled to interval $[0, 1]$.

Finding appropriate paramters (C, γ) is done by the means of cross-validation: using the 5-fold cross-validation on the learning set parameters (C, γ) which give the highest accuracy are selected. Parameters that were considered are: $C \in \{2^{-5}, 2^{-4}, \dots, 2^{15}\}$, $\gamma \in \{2^{-15}, 2^{-14}, \dots, 2^{3}\}$ [3]. The accuracy of classification is measured by the expression:

$$acc = \frac{n_c}{N}, \tag{2}$$

where n_c is number of correctly classified texts, N is total number of texts. After the best parameters (C, γ) are found for which the system achives the highest accuracy, the classifier is learned using those parameters.

6. Evaluation

Classification quality is measured by ratio of correctly classified texts and total number of texts (accuracy). The evaluation results are shown in table 1. Columns “C” and “γ” denote parameters of SVM classifier used for selected model. For each set of features, the best parameters are recorded. Not all combinations of features are shown.

7. Conclusion

It is shown that authorship attribution problem can be successfully solved by relatively simple methods, although for state–of–the–art methods, we believe that syntactic analysis of texts is required. Our results with 93% accuracy fit very well in interval of reported results which range from 70% to 97% [4, 7, 11, 13].

Result comparision is difficult due to different types of data sets (e.g. poems, newspaper articles, e–mails) and the problems (binary, multi–class or single–class classifications). There are no relevant data sets for comparision [17], but there are references on paper [13] and their “Greek data set” (e.g. [7]). There is a significant affected of data set type (number and variety of text samples) on the complexity of the problem.

In future work, methods based on word and character *n*–grams suggested in [4, 7, 12] have to be evaluated. Also, evaluation has to be performed on different types of data sets such as poems, newspaper articles or books data sets.

Acknowledgement

8. References

[1] S. Argamon and S. Levitan. Measuring the usefulness of function words for authorship attribution. *Proceedings of ACH/ALLC 2005*, 2005.

[2] M. Banko and E. Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, page 33. Association for Computational Linguistics, 2001.

[3] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.

[4] R.M. Coyotl-Morales, L. Villaseñor-Pineda, M. Montes-y Gómez, P. Rosso, and L.T. del Lenguaje. Authorship attribution using word sequences. *Lecture Notes in Computer Science*, 4225:844, 2006.

[5] O. De Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64, 2001.

[6] J. Diederich, J. Kindermann, E. Leopold, and G. Paass. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1):109–123, 2003.

[7] V. Kešelj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING*, volume 3, pages 255–264. Citeseer, 2003.

[8] S.S. Keerthi and C.J. Lin. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*, 15(7): 1667–1689, 2003.

[9] M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI*, volume 3, pages 69–72. Citeseer, 2003.

[10] OV Kukushkina, AA Polikarpov, and DV Khmelev. Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2): 172–184, 2001.

[11] Kim Luyckx and Walter Daelemans. Shallow text analysis and machine learning for authorship attribution. In *Proceedings of the fifteenth meeting of Computational Linguistics in the Netherlands (CLIN 2004)*, pages 149–160, 2005.

[12] F. Peng, D. Schuurmans, S. Wang, and V. Keselj. Language independent authorship attribution using character level language models. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, page 274. Association for Computational Linguis-

tics, 2003.

[13] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214, 2001.

[14] F.J. Tweedie and R.H. Baayen. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352, 1998.

[15] O. Uzuner and B. Katz. A comparative study of language models for book and author recognition. *Lecture Notes in Computer Science*, 3651:969, 2005.

[16] Jan Šnajder, Bojana Dalbelo Bašić, and Marko Tadić. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing and Management*, 44(5):1720–1731, 2008.

[17] Y. Zhao and J. Zobel. Effective and scalable authorship attribution using function words. *Lecture Notes in Computer Science*, 3689: 174, 2005.