Language identification: how to distinguish similar languages?

Nikola Ljubešić, Nives Mikelić, Damir Boras
Department of Information Sciences, Faculty of Philosophy, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
E-mail: nljubesi@ffzg.hr, nmikelic@ffzg.hr, dboras@ffzg.hr

Abstract. *The goal of this paper is to discuss the language identification problem of Croatian, language that even state-of-the-art language identification tools find hard to distinguish from similar languages, such as Serbian, Slovenian or Slovak language. We developed the tool that implements the list of Croatian most frequent words with the threshold that each document needs to satisfy, we added the specific characters elimination rule, applied second-order Markov model classification and a rule of forbidden words. Finally, we built up the tool that overperforms current tools in discriminating between these similar languages.*

Keywords. Written language identification, Croatian language, second-order Markov model, web-corpus, most frequent words method, forbidden words method.

1. Introduction

Without the basic knowledge of the language the document is written in, applications such as information retrieval and text mining are not able to accurately process the data, potentially leading to a loss of critical information. The problem of written language identification is attempted to be solved for long time and various feature-based models were developed for written language identification.

Some authors used the presence of diacritics and special characters [17], some used syllable characteristics [16] and some used information about morphology and syntax [26].

Some of them used information about short words [10, 9, 13, 3, 20], while some authors used the frequency of n-grams of characters [4, 5, 6, 20, 15]. Some techniques used Markov models [23, 18, 7], while some used information theoretic measures of entropy and document similarity [19, 1]. The application of support vector machines and kernel methods to the language identification task has been considered relatively recently [22, 14, 12].

Sibun & Reynar [19] applied relative entropy to language identification. Their work is important for us because they were first to provide the scientific results for Croatian, Serbian and Slovak. For Croatian, they got recall rate of 94%, while precision was 91.74%. Interesting fact is that Sibun & Reynar's tool made error by identified Croatian as Slovak language, but it never confused Croatian and Serbian. On the other hand, Serbian and Slovak were likely to be identified as Croatian.

The improvement from Sibun & Reynar work was Elworthy's algorithm [8], which achieved recall rate of 96%, and precision rate of 97.96%, because Serbian and sometimes Slovenian were identified as Croatian.

Automated written language identification tools are nowadays widely used, such as the best known van Noord's TextCat [21], Basis Tech's Rosette Language Identifier [2], and web based language identification services such as Xerox Language Identifier [25]. TextCat is an implementation of the text categorization algorithm presented in[5]. Both TextCat and Xerox Language Identifier are freely available and commonly used and do language identification for Croatian and similar languages (Slovak, Serbian, Slovenian, Czech) as well. Basis Tech's Rosette Language Identifier also includes all these languages, but is available only when purchased.

Since Croatian and Serbian are similar languages that were considered as Serbo-Croatian language for almost a century, lan-

guage identifiers such as TextCat and Xerox do get confused and are likely identifying Croatian documents as Serbian and vice versa.

Moreover, the TextCat algorithm introduces Bosnian language that makes identification even harder. The Bosnian is spoken by Bosniaks in Bosnia and Herzegovina and the region of Sandžak (in Serbia and Montenegro), it is based on the Western variant of the Štokavian dialect, it has Latin alphabet and has its vocabulary and grammar based on Croatian and Serbian language as well. The differences that are important for distinguishing Croatian from Serbian in the process of language identification are useless when dealing with Bosnian documents, since Bosnian language accepts all of them. Generally, it prefers Croatian, but vocabulary and grammar are both a mix of Serbian and Croatian, apart from some turcisms that are frequently in use only in Bosnian. For instance, with modal verbs such as *ht(j)eti* (*want*) or *moći* (*can*), the infinitive is prescribed in Croatian, while the construction *da* (*that*/*to*) + present tense is preferred in Serbian. Both alternatives are present and allowed in Bosnian.

Therefore, in our research we did not try to distinguish Bosnian from Croatian, since it is hard even for native speakers to notice the difference between the two of them at a glance.

2. The first step in developing the language identifier for Croatian

Since Croatian, Serbian and Slovenian are proved to be most difficult languages to distinguish, we collected our training and test corpora from three most popular news portals in Croatia, Serbia and Slovenia [24].

We collected 67244 documents in Croatian, 30076 documents in Serbian and 5295 documents in Slovenian. Since Slovenian corpus was the smallest, considering the need for training corpora in the steps that follow, we took parts of the Croatian and Serbian corpus and built a smooth Croatian-Serbian-Slovenian test corpus consisting of 4364 documents for each language (13092 in total).

Firstly, we extracted the list of most frequent words from our remaining Croatian corpus. We measured the frequency distribution of the documents in our test corpus (4364 documents in each of 3 languages) regarding the percentage of N most frequent Croatian words each document contains. Since the experimental data proved obvious normality, we presented these distributions in figure 1 as normal distributions. The normality of the three distributions was proved by the Shapiro-Wilk test with the largest p-value of $9.12 * 10^{-11}$ for the Slovenian documents distribution. From figure 1 it is obvious that this approach is not capable of distinguishing between these three languages, especially not between Croatian and Serbian since their distributions overlap significantly. Nevertheless, this method is capable of distinguishing between these three and all other languages with the exception of western Slavic languages.

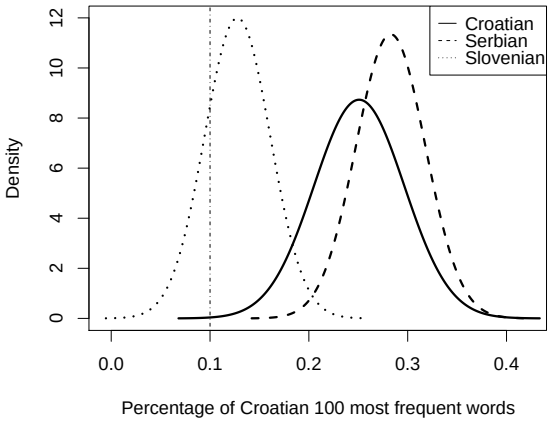

Figure 1. Normal distributions of the documents for Croatian, Serbian and Slovenian regarding the percentage of 100 most frequent Croatian words

One can notice that the distribution of Slovenian documents is moved leftwards comparing to the distribution of Croatian documents, which shows that Croatian and Slovenian are different languages. On the other hand, the Serbian distribution is moved rightwards. The reason for this is the frequency of construction *da* (*that*/*to*) + present tense in Serbian (*da* is one of the 100 most frequent words), that is replaced by infinitive in Croa-

tian. Although all figures show differences between three languages, the overlapping between them is still very high, especially between Croatian and Serbian.

Two values had to be chosen during the first step - the threshold of percentage of N most frequent Croatian words T and the number of Croatian most frequent words N. Table 1 shows recall concerning these two values. T=15% never yielded in satisfiable recall. On the other hand, choosing T as 10% for N=100 yielded in satisfiable recall which is 0.13% lower than recall with N=200, but with N=200 we would have exposed ourselves to the danger of introducing corpus-specific most frequent words. Therefore we decided to choose N=100 with T=10%.

Table 1. Change in recall when discriminating languages using T=15% and 10% (rows) for the N={25,50,75,100,200} (columns)

	25	50	75	100	200
15	.9175	.9552	.9681	.9718	.9830
10	.9853	.9913	.9931	.9943	.9956

Table 2 shows the number of documents below and above the threshold in our sample concerning the chosen T and N. All quantities that are reported in these tables are derived from the samples and not the normal distributions that underlie our data.

Table 2. Number of documents for Croatian, Serbian and Slovenian that are above or below the 10% threshold for 100 Croatian most frequent words.

	above	below
Croatian	4339	25
Serbian	4364	0
Slovenian	3986	378

Since western Slavic languages share the same alphabet and similar most frequent words with southern Slavic languages, we had to realize on a mini-corpus of 10 documents of Chech, Polish and Slovak language (30 documents in total) that the average percentage of 100 most frequent Croatian words in this documents was 10.64%. If we used T=15%, the number of Chech, Polish or Slovak documents classified as potentially Croatian would be lower, but we would irreversibly lose many Croatian documents as well. We solved that problem by introducing a special character elimination rule. The maximum percentage of the 20 most frequent characters in the mini-corpus that are not part of the Croatian alphabet in Croatian documents was 0.26%. The average percentage was 0.00181%. On the other hand, in the small corpus of Czech, Polish and Slovak texts (30 documents), the smallest percentage of these characters was 4.50%. The average one was 6.7%. Since the Croatian sample is much bigger and therefore much trustworthier, we composed a rule that eliminates all documents whose special character percentage in documents exceeds the threshold of 1%.

We can conclude that the method that uses 100 common words with the threshold of 10% gives good results in distinguishing Croatian and languages very similar to Croatian (Serbian and Slovenian) from all other languages, with one additional rule: eliminating documents whose percentage of 20 most frequent specific characters of Czech, Polish and Slovak exceeds 1%.

Since this method eliminated only 8.66% of Slovenian documents and none of the Serbian ones, we needed to apply additional classification methods that were more efficient in distinguishing between similar languages.

3. The second step in developing the language identifier for Croatian

The second step involved a simple method of supervised machine learning. We developed a set of trigram character level language models for each of the three languages (Croatian, Serbian and Slovenian), trained them and used them to estimate the probability of generation of a particular string by one of those three models.

We used a Markov model, which is a probabilistic model for sequential data that defines the probability distribution of a sequence of random variables $P(X_1, X_2, ...X_n)$ making N-th order Markov assumptions that the value

of a specific random variable depends only on values of N prior random variables. In case of a second-order Markov model the probability of a sequence and conditional probabilities are calculated as

$$p(x_1, x_2, ...x_n) = \prod_{i=1}^{n} p(x_i|x_{i-2}^{i-1})$$

$$p(x_i|x_{i-2}^{i-1}) = \frac{c(x_{i-2}^{i-1})}{c(x_{i-2}^{i})}$$

where $c(x_j^k)$ is the number of times the sequence of random variables $X_j...X_k$ takes values $x_j...x_k$.

We used a second-order Markov model since although higher order Markov models make less assumptions, they have to fight with the data sparseness, especially in the case of language identification. Namely, as we will show, Markov models for language identification achieve optimal results on relatively small amounts of training data. In our case, we solved the data sparseness problem by the simplest smoothing method defining probability of unseen data as $1 * 10^{-10}$. We also calculated the sum of logs of probabilities rather than the product of probabilities to avoid zero underflow.

Since the distinction between Croatian and Serbian is a much harder task than the one between Croatian and Slovenian, we first tried to distinguish Croatian and Serbian. Firstly, we observed the relationship between the size of the training corpus and the recall and precision measures. We decided to move the size of the training corpus to 1.000.000 characters by steps of 100 characters. We used 4588 documents from each language as the training corpus and 21124 documents from each language as the validation corpus (since we had to estimate the optimal size of the training corpus, we used the validation corpus that did not overlap with our test corpus). We realized, as shown in figure 2, that second-order Markov models trained on 350.000 characters give optimal results.

If trained on 240.000 characters for each language, precision reaches its peak, but recall decreases drastically (our results vary too much at this point to be regarded as good predictors of future performance) and if trained on 400.000 characters, precision starts decreasing more rapidly than recall increases (as shown by the decline of the F1 measure).

The precision of distinguishing Croatian documents in the Croatian-Serbian corpus using 350.000 characters as a training corpus is 99.08%, while the recall is 92.89%.

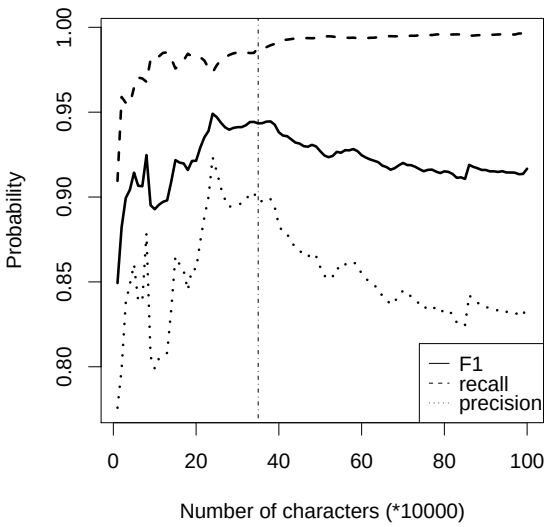

Figure 2. Relationship between the size of the training corpus and precision/recall measures in differentiating Croatian documents in a Croatian-Serbian validation corpus

Now that we obtained an optimal size of the training corpus for distinguishing Croatian from Serbian, we trained language models for all three languages (Croatian, Serbian and Slovenian) and tested them on the test corpus of 4364 documents for each language (13092 documents in total). The confusion matrix of the results on the distinguishing between three languages in the three-language-corpus is shown in table 3.

Table 3. Confusion matrix for 13092 documents of Croatian-Slovenian-Serbian test corpus (columns are the language identified)

	Croatian	Serbian	Slovenian
Croatian	4321	38	5
Serbian	309	4055	0
Slovenian	5	0	4359

The final result for distinguishing all 3 languages concerning Croatian is a recall of 99.01% and precision of 93.23%.

4. The final step in developing the language identifier for Croatian

The aim of the final step was to improve the precision of identifying Croatian documents which was primarily low because of misclassifications of Croatian and Serbian documents. The additional classification was done with the list of forbidden words for Croatian and Serbian. Both Serbian and Croatian lists consisted of words that appear at least 5 or more times in one corpus, but do not exist in the other one at all. Therefore, if the document, identified as Serbian after the second step, contained one or more words from the Croatian list and none from the Serbian one, the decision was changed and the document was identified as Croatian. There were 79827 such words in the Croatian corpus and 18733 in the Serbian one. The difference between these numbers lies primarily in the fact that the remaining part of the Croatian corpus was much bigger than the one of the Serbian corpus.

If the list of Croatian specific types is tailored down to 18733, the precision improves up to 99.84%. Since the danger of overfitting in this case is very high, we decided to take just the 1000 most frequent words from both lists and improved the precision to 99.18%. Regardless the number of words taken, recall improved up to 99.31%. The recall/precision measures through all the three steps where each step follows up on the results of the previous one are shown in table 4.

Table 4. Recall/precision measures for identifying Croatian documents in the 13092 documents test corpus through all three language identification steps

	Recall	Precision
First step	.9943	.3419
Second step	.9846	.9329
Third step	.9931	.9918

In the first step, Croatian, Serbian and Slovenian were efficiently distinguished from all other languages with a Croatian most frequent words threshold rule and a special characters threshold rule. In the second step these three languages were distinguished between themselves with a character-based second-order Markov model. In the third step, the classification between Croatian and Serbian was improved with a forbidden word list rule.

5. Conclusion

In this paper we presented the tool for language identification that overperforms existing tools in differentiating Croatian from Serbian and Slovenian.

The method of most frequent words proved to be most usable in differentiating similar from all other languages where a special character constraint also proved to be very handy. The character n-gram models proved to be quite efficient in distinguishing similar languages. The combination of these two methods proved to work best since the n-gram method requires a language model for every possible language and the most frequent words method efficiently strips the number of remaining languages to a few. The method of forbidden words proved to improve results in distinguishing very similar languages.

Although some of the state-of-the-art language guessers distinguish Bosnian as a language, in our research we did not try to distinguish Bosnian from Croatian, since it is hard for native speakers to notice the difference between the two of them at a glance.

References

[1] Aslam JA, Frost M. An information-theoretic measure for document similarity. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2003.

[2] Basis Techis Rosette Language Identifier. http://www.basistech.com/language-identification/ [07/01/2006]

[3] Batchelder EO. A learning experience: Training an artificial neural network to discriminate languages. Technical Report, 1992.

[4] Beesley KR. Language identifier: A com-

puter program for automatic natural language identification on on-line text. In Proceedings of the 29th Annual Conference of the American Translators Association, 1988. p.47-54.

[5] Cavnar WB, Trenkle JM. N-gram-based text categorization. In Proceedings of SDAIR-94, the 3rd Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, Nevada, USA, 1994. p. 161-175.

[6] Damashek M. Gauging similarity with n-grams: language independent categorization of text. Science 1995; 267(5199):843–848.

[7] Dunning T. Statistical identification of language. Technical Report MCCS. New Mexico State University, 1994. p.94-273.

[8] Elworthy, D. Language Identification With Confidence Limits. In CoRR: Computation and Language Journal, 1999.

[9] Henrich P. Language identification for the automatic grapheme-to-phoneme conversion of foreign words in a german text-to-speech system. In Proceedings of Eurospeech 1989, European Speech Communication and Technology, 1989. p. 220-223.

[10] Ingle N. A language identification table. In The Incorporated Linguist 1976, 15(4).

[11] Johnson, S. Solving the problem of language recognition. Technical report. School of Computer Studies, University of Leeds, 1993.

[12] Kruengkrai C, Srichaivattana P, Sornlertlamvanich V, Isahara H. Language identification based on string kernels. In Proceedings of the 5th International Symposium on Communications and Information Technologies (ISCIT), 2005.

[13] Kulikowski S. Using short words: a language identification algorithm. Unpublished technical report, 1991.

[14] Lodhi H, Shawe-Taylor J, Cristianini N, Watkins CJCH. Text classification using string kernels. Journal of Machine Learning Research, 2:419-444.

[15] McNamee P, Mayfield J. Character N-gram Tokenization for European Language Text Retrieval. Information Retrieval 2004; 7:73-97.

[16] Mustonen S. Multiple discriminant analysis in linguistic problems. In Statistical Methods in Linguistics. Skriptor Fack, Stockholm, 1965;(4).

[17] Newman P. Foreign language identification - first step in the translation process. In K. Kummer (editor), Proceedings of the 28th Annual Conference of the American Translators Association, 1987. p.509-516.

[18] Schmitt JC. Trigram-based method of language identification, October 1991. U.S. Patent number: 5062143.

[19] Sibun, P. and Reynar, J. C. Language Determination: Examining the Issues. In Proceedings of the 5th Annual Symposium on Document Analysis and Information Retrieval, pp. 125-135, Las Vegas, Nevada, 1996.

[20] Souter et al. Natural Language Identification Using Corpus-Based Models. Hermes J. Linguistics 1994; 13:183-203.

[21] TextCat Language Guesser Demo. http://www.let.rug.nl/ vannord/TextCat/Demo/ [07/01/2006]

[22] Teytaud O, Jalam R. Kernel-based text categorization. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2001.

[23] Ueda Y, Nakagawa S. Prediction for phoneme/syllable/word-category and identification of language using HMM. In Proceedings of the 1990 International Conference on Spoken Language Processing; November 1990; Kobe, Japan. Volume 2, p.1209-1212.

[24] News portals: http://www.net.hr, http://www.b92.net, http://novice.siol.net [4/26/2007]

[25] Xerox Language Identifier. http://www.xrce.xerox.com/competencies/content-analysis/tools/guesser-ISO-8859-1.en.html [07/01/2006]

[26] Ziegler DV. The automatic identification of languages using linguistic recognition signals. State University of New York at Buffalo, Buffalo, NY, 1992.