Original scientific paper

Language Identification of Web Data for Building Linguistic Corpora

Marija Stupar
Department of Information Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
mstupar@ffzg.hr

Tereza Jurić
Department of Information Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
tjuric2@ffzg.hr

Nikola Ljubešić
Department of Information Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
nljubesi@ffzg.hr

Summary

In this paper we inspect a series of methods for language identification on web data. We start from the standard two methods based on function word frequencies and Markov chains. We investigate the problem on both the document and the paragraph level. After obtaining an insight in the strengths and weaknesses of these basic methods, we propose two hybrid methods where the more complex one outperforms or performs equally well as the best basic method. Identifying language on each paragraph of more than three million documents collected for the Croatian Web Corpus hrWaC shows that around 96% of the documents are monolingual and that the language distribution, as expected, follows a power-law distribution.

Key words: language identification, Web data, Croatian Web corpus, Markov model, function words

Introduction

The Web represents a freely available and rich source of linguistic material. With a disparate nature of contained sources, it can be used to conduct various types of linguistic research. There is a high possibility of finding texts in more

than one natural language within such sources. The problem of multilinguality therefore presents a challenge.
Language identification is a process which aims to label textual documents according to the language they are written in, and it is often applied to many fields of natural language processing since multilinguality is nowadays a frequent phenomenon.
The goal of this paper is to define a method for identifying the language which the documents collected from the Web are written in. The described method is primarily developed for building hrWac, the Croatian Web corpus, although it can be applied to other problems such as finding texts in different languages on the Web that are translations of each other (building thereby automatically parallel corpora).
In this paper we compare the two main approaches for language identification - the linguistic and the statistical one. The linguistic approach is based on function word distributions while the statistical approach is based on second-order Markov models trained on small language samples of all anticipated languages. After obtaining an insight in the weaknesses and strengths of each approach, we propose two hybrid approaches combining these two methods.

Related work

Using Web resources can be a useful basis for constructing corpora in fields of linguistics, language technologies and translation. Projects such as the WaCky initiative (Baroni et al, 2009.) aim to provide a set of tools to process, index and search the data gathered from the Web.
Language identification is a widely studied field, and many different approaches have been introduced to solve this problem. The methods vary between using special characters, information about short words, frequency of n-grams, Markov models, etc. There are also approaches that combine various methods in order to achieve better results.
The basic approaches used in this research are described in following papers: the linguistic approach based on function words has been studied by (Ingle, 1976) and (Kulikowski, 1991), and some of the Markov model approaches have been presented in (Schmitt, 1991) and (Dunning, 1994).

Experimental setup

The twelve languages observed in the research are Czech (cs), German (de), English (en), Spanish (es), French (fr), Croatian (hr), Hungarian (hu), Italian (it), Polish (pl), Slovak (sk), Slovenian (sl), and Swedish (sv). They were chosen upon their incidence in hrWac obtained through a corpus concordancer.
To have a sense of how hard our problem will be we first studied the similarity of the languages chosen for our experiment. Therefore we used the data found in „The Language Table" by (Crúbadán, 2007). The table shows the cosine similarity between the 3-gram profile vectors for each language. The data for

our 12 languages is given in Table 1. We expect to find it harder to distinguish between similar languages, such as Czech and Slovak, Croatian and Slovenian, or Spanish and French. Hungarian, on the other hand, seems very different from other languages, and therefore has a high possibility of being correctly identified. It should be noted that we did not take Serbian or Bosnian into consideration in this research for two reasons:

1. It is not too likely to find significant amount of such material on the Croatian Internet domain
2. Distinguishing these languages should be regarded a separate problem as described in (Ljubešić, Boras, Mikelić, 2007) which should follow the first language identification phase we investigate here

Table 1: A snippet from "The language table"

	cs	**de**	**en**	**es**	**fr**	**hr**	**hu**	**it**	**pl**	**sk**	**sl**	**sv**
cs	-	18	22	26	22	53	25	31	42	70	54	23
de	18	-	34	34	35	12	17	31	20	17	18	53
en	22	34	-	27	33	16	16	35	15	17	19	35
es	26	34	27	-	62	22	18	56	18	23	28	38
fr	22	35	33	62	-	18	15	48	15	18	22	35
hr	53	12	16	22	18	-	11	31	39	51	74	24
hu	25	17	16	18	15	11	-	14	10	22	13	21
it	31	31	35	56	48	31	14	-	22	28	38	32
pl	42	20	15	18	15	39	10	22	-	50	40	18
sk	70	17	17	23	18	51	22	28	50	-	55	22
sl	54	18	19	28	22	74	13	38	40	55	-	26
sv	23	53	35	38	35	24	21	32	18	22	26	-

We distinguish two phases in our experiment: 1. identifying language on the document level and 2. identifying language on the paragraph level. In each phase of the experiment we evaluated both approaches – the linguistic and the statistical one. After obtaining an insight into the strengths and weaknesses of every approach on both levels, we propose two hybrid approaches and evaluate these on both the document and paragraph level.

The linguistic approach uses lists of function words from all languages in question and picks that language for which the highest percentage of words could be identified as function words of the respective language.

The statistical approach uses second-order Markov models, i.e. conditional probabilities of a character regarding the two previous characters for which distributions of bigram and trigram characters is calculated on a training set. A detailed overview of the method used is given in (Ljubešić, Boras, Mikelić, 2007).

The data necessary for building both methods was collected by hand. The number of collected function words, i.e. the amount of training data for building Markov models is given in Table 2.

For evaluation purposes we built two gold standards from documents collected for purposes of building hrWaC, one gold standard for each level.
For the document level we collected 20 documents per language. The documents were also obtained with help of a concordancer. We are aware of the fact that this uniform distribution does not follow the actual language distribution on the Croatian web. Since it would be very hard, if not impossible to build a labeled sample of random documents with enough examples for all 12 languages, we were forced to ignore the real distribution and approximate the uniform one. Since a significant amount of documents on the Web is written in more than one language, we included in the sample also documents written in more than one language. To keep the complexity of the task under control, our rule of thumb was to label a document with a specific language label if at least 70% of the document was written in that language. We consider the documents containing less than 70% of any language unsolvable on the document level.

Table 2: Amount of data collected for each basic method (the number of function words per language, and character count as training data for building the Markov model)

Language	Function words	Character count
Czech	210	150601
German	334	150156
English	230	150041
Spanish	217	150926
French	260	150083
Croatian	204	157366
Hungarian	223	152202
Italian	219	150459
Polish	268	150198
Slovak	168	150046
Slovenian	256	143841
Swedish	256	150762

For evaluating the methods on paragraph level we labeled paragraphs in 50 documents by language they are written in. Thereby we labeled 750 paragraphs in total.
Our evaluation measure is accuracy (a+d/a+b+c+d), where the nominator contains the number of correct decisions and the denominator the overall number of decisions made.

Results and discussion

The results of the evaluation of the two standard methods are given in Table 3. All the results are rather high, but it is obvious that Markov model consistently achieves better results. Markov model was identically accurate on both document and paragraph level, while the method using function words achieved

better results on the paragraph level which could be considered rather strange. It is our opinion that this is because of different languages present in a number of documents and the inability of the method to deal with mixed content. On the paragraph level this was not an issue since most of the paragraphs are written in one language only.

A manual evaluation of the results showed the strengths and weaknesses of every method. Markov models are prone to making wrong decisions if a segment in the string contains characters characteristic for another language (a document written in English was recognized as Croatian due to frequent occurrence of the named entity "Sveučilište u Zadru - Odjel za njemački jezik i književnost"). On the other hand, the function words method tends to make wrong decisions in case of an overlap in function words between more languages (a document written in Hungarian was recognized as English, due to occurrence of the same function words with different usage such as "a" in English meaning "the" in Hungarian, or "is" in English meaning "also" in Hungarian) and in case of shorter texts. The function words method, as shown in the automatic evaluation, is in general more prone to errors.

Table 3: Results of the evaluation of the traditional methods

	Function words	**Markov model**	**Function words**	**Markov model**
	Document level		**Paragraph level**	
Positive	234	239	745	747
Negative	6	1	5	3
Accuracy	0.975	0.996	0.993	0.996

Finally, we propose a hybrid approach that combines the two methods evaluated above having in mind that these methods are erroneous in different situations. The first method calculates the harmonic mean of the certainty of the function words method and the Markov model method (certainty is calculated as a/(a+b) where a is the first result, and b the second best result). The more sophisticated hybrid method takes into account the strengths of each method and thereby does the following:

- If the Markov model and function words method give the same results, the result is accepted
- In case the results of both models are not the same, but the second best result of the Markov model method is identical to the first result of the function words method and its certainty is over 0.6, the result of the function word method is accepted
- Otherwise the result of the Markov model method is accepted

Thereby we change the decision made by the Markov model method only in case the second-best guess of the Markov model method and the first guess of the function word method are identical with a safety margin of 0.1.

The results of the automatic evaluation of the hybrid approaches are given in Table 4.
The results show a small, but consistent improvement. What is more interesting, these hybrid methods obviously handle significantly better the case where a mixture of languages is present in the string. On the paragraph level there is no visible improvement when comparing the results to the results obtained by the Markov model method. The question that arises here is if a difference would become visible on a larger (more representative) sample.

Table 4: Results of the evaluation of hybrid methods

	Harmonic balance	**Sophisticated method**	**Harmonic balance**	**Sophisticated method**
	Document level		**Paragraph level**	
Positive	239	240	746	747
Negative	1	0	4	3
Accuracy	0.996	1.0	0.995	0.996

Because of the improvements shown in specific situations shown by the sophisticated hybrid method, we decided to use this method for the task of identifying languages in our emerging web corpus. We analyzed 3,924,189 documents by each paragraph. In 95.9% of the documents all paragraphs were identified as being written in same language. From all documents containing a paragraph identified as Croatian, 95.8% of the documents were pure Croatian. In Table 5 we give a distribution of languages as identified on paragraph level. The data show a power-lawish distribution where 90% of the paragraphs are written in Croatian. Second-best, as expected, is English with 8%, Slovene with 1% and the remainder of languages making only 1% of the paragraphs.

Table 5: Distribution of languages as identified on paragraph level

Language	**Number of paragraphs**	**Paragraph percentage**
Croatian	25347696	89.9%
English	2195590	7.8%
Slovene	288829	1%
German	111078	0.4%
Italian	64268	0.2%
Spanish	39515	0.1%
Swedish	34388	0.1%
French	33817	0.1%
Czech	31812	0.1%
Slovak	22313	0.1%
Polish	18791	0.1%
Hungarian	15404	0.1%

Conclusion

In this paper we have compared the two mostly used language identification methods on web data – the function words method and the Markov model method. We have shown that in general Markov model outperforms the function words method. A case where Markov model fails is if a sequence of characters specific for another language, like a named entity, is found in the data. On the other hand, the function words method underperforms on shorter texts and suffers from collisions of function words between languages. These methods perform very well on paragraph level as well, even outranking the document level results on some occasions since web documents tend to contain mixed language content.

We proposed two hybrid approaches that showed to be more efficient on the document level, i.e. on data containing mixed language content. It is our belief that on a larger gold standard the hybrid methods would outperform the standard methods on paragraph level as well.

In the end we identified the language on paragraph level in documents collected for the Croatian Web Corpus hrWaC and showed that around 96% of documents are written in only one language where the remaining 4% have mixed content. Additionally, we showed that the distribution of languages is powerlawish where Croatian, English and Slovene make 99% of the data.

References

Martins, B; Silva, M.J. Language identification in web pages. // The 20th ACM SAC Symposium on Applied Computing. Document Engeneering Track. / L. M. Liebrock (ed.). 2005, 773 - 777.

Dunning T. Statistical identification of language. // Technical Report MCCS. New Mexico: New Mexico State University, 1994.

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

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

Baroni, M. et.al. The WaCky Wide Web : A Collection of Very Large Linguistically Processed Web-Crawled Corpora. // Language Resources and Evaluation 43(2009), 3, 209-226.

Ljubesic, N; Mikelic, N; Boras, D. Language identification : How to distinguish similar languages? // Proceedings of the 29th International Conference on Information Technology Interfaces. 2007, 541-546.

Vojtek, P; Bielikova, M. Comparing Natural Language Identification Methods based on Markov Processes. // Slovko - International Seminar on Computer Treatment of Slavic and East European Languages. 2007.

Schmitt J.C. Trigram-based method of language identification. US Patent 5062143. 1991.