

Discriminating Between Similar Languages Using Large Web Corpora

Vít Suchomel

Natural Language Processing Centre
Faculty of Informatics
Botanická 68a, Brno, Czech Republic
xsuchom2@fi.muni.cz

Lexical Computing
Brno, Czech Republic

Abstract. This paper presents a method for discriminating similar languages based on wordlists from large web corpora. The main benefits of the approach are language independency, a measure of confidence of the classification and an easy-to-maintain implementation.

The method is evaluated on VarDial 2014 workshop data set. The result accuracy is comparable to other methods successfully performing at the workshop.

A tool implementing the method in Python can be obtained from web site <http://corpus.tools/>.

Keywords: language identification, discriminating similar languages, building web corpora

1 Introduction

Language identification is a procedure necessary for building monolingual text corpora from the web. For obvious reasons, discriminating similar languages is the most difficult case to deal with. Continuing in the steps of our previous work [2], our goal in corpus building is to keep documents in target languages while removing texts in other, often similar languages. The aim is to process text of billion-word sized corpora using efficient and language independent algorithms. Precision (rather than recall), processing speed and easy-to-maintain software design are of key importance to us.

Data to evaluate language discrimination methods have been created by the organisers of the workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial) since 2014 [10,11,7,9]. Various media ranging from nice newspaper articles to short social network texts full of tags were made available. Successful participants of this series of workshops have published their own approaches to the problem.

2 Method

2.1 The Aim And Desired Properties

The aim of the method presented in this paper is to provide a simple and fast way to separate a large collection of documents from the web by language. This is the use case: Millions of web pages are downloaded from the web using a web crawler. To build monolingual corpora, one has to split the data by language.

Since the set of internet national top level domains (TLDs) targeted by the crawler is usually limited and a similarity of the downloaded texts to the target languages can be easily measured using e.g. a character n-gram model [6], one can expect only a limited set of languages similar to the target languages to discriminate. The method should work with both documents in languages that have been discerned in the past as well as texts in languages never processed before.

The presented method does:

- Enable supporting new languages easily (that implies the same way for adding any language).
- Allow adding a language never worked with before, using just the web pages downloaded or a resource available for all languages (e.g. articles from Wikipedia).
- Not use language specific resources varying for each language supported (e.g. a morphological database) – since that makes supporting new languages difficult.
- Apply to any structure of text, e.g. documents, paragraphs, sentences.
- Provide a way to measure the contribution of parts of a text, e.g. paragraphs, sentences, tokens, to the final classification of the structure of the text.
- Provide a measure of confidence to allow setting a threshold and classifying documents below the threshold of minimal confidence as mixed or unknown language.
- Work fast even with collections of millions of documents.

2.2 Method Description

This method uses the initial step of the algorithm described in [2]. The reason for not including the expectation-maximisation steps is the aim to decrease the complexity of the solution, keeping the data processing time reasonably short.

The method exploits big monolingual collections of web pages downloaded in the past or even right before applying the method (i.e. using the text to identify its language as the method data source at the same time). The language of documents in such collections should be determined correctly in most cases, however some mistakes must be accepted since there are many foreign words in monolingual web corpora since e.g. foreign named entities or quotes are preserved. Even a lot of low frequency noise can be tolerable. Lists of words with relative frequency are built from these big monolingual collections of web

pages. The method uses the decimal logarithm of word count per billion words to determine the relative wordlist score of each word from the list of words according to the following formula:

$$score(w) = \log_{10} \left(\frac{f(w) \cdot 10^9}{|D|} \right)$$

Where $f(w)$ is the corpus frequency of the word (number of occurrences of the word in the collection) and $|D|$ is the corpus size (number of all occurrences of all words in the collection).

The wordlist is built for all languages to discern, prior to reading the input text. Usually, when building corpora from the web, languages similar to the target languages and languages prevalent in the region of the internet national top level domains occurring in the crawled data are considered. A big web corpus is a suitable source. To improve the list by reducing the presence of foreign words, limiting the national TLD of source web pages is advisable. E.g. using texts from TLD .cz to create a Czech word list should, intuitively, improve precision at a slight cost of recall.

The input of the method, i.e. the documents to separate by language, must be tokenised. Unitok [8] was used to tokenise text in all sources used in this work. Then, for each word in the input, the relative wordlist score is retrieved from each language wordlist. The scores of all words in a document grouped by the language are summed up to calculate the language score of a document. The same can be done for paragraphs or sentences or any corpus structure.

$$document\ score(language) = \sum_{w \in document} language\ score(w)$$

The language scores of a document are sorted and the ratio of two highest scoring languages is computed to determine the confidence of the classification. The score ratio is compared to a pre-set confidence threshold. If the ratio is below the threshold, the document is marked as a mixed language text and not included in the final collection of monolingual corpora. Otherwise the result language is the language with the highest score.

$$confidence\ ratio(document) = \frac{document\ score(top\ language)}{document\ score(second\ top\ language)}$$

According to our experience, setting the confidence threshold quite low (e.g. to 1.005) is advisable in the case of discerning very similar languages while higher values (e.g. 1.05) work for other cases (e.g. Czech vs. Slovak, Norwegian vs. Danish).

We usually understand a paragraph to be the largest structure consisting of a single language in the case of multilanguage web pages. The method presented in this work allows separating paragraphs in different languages found in a single multilingual document to multiple monolingual documents. Although code switching within a paragraph is possible, detecting that phenomenon is beyond the scope of this work.

The following sample shows the overall sentence language scores as well as particular word language scores in a sentence from VarDial 2014 test data. Words ‘scheme’, ‘council’ and ‘tenant’ contribute the most to correctly classifying the sample as British English (rather than American English). Column description: Word, en-GB score, en-US score. Punctuation was omitted from the wordlists thus getting a zero score.

```
<s lang="en-GB" confidence_ratio="1.018" en-GB="122.04" en-US="119.89">
Under    5.74    5.74
the      7.77    7.75
rent     4.70    4.59
deposit  4.56    4.40
bond     4.49    4.63
scheme   5.26    4.41
,        0.00    0.00
the      7.77    7.75
council  5.56    5.20
pays     4.20    4.26
the      7.77    7.75
deposit  4.56    4.40
for      7.06    7.07
a        7.36    7.34
tenant   4.34    3.94
so       6.34    6.31
they     6.51    6.50
can      6.53    6.54
rent     4.70    4.59
a        7.36    7.34
property 5.38    5.37
privately 4.05    3.99
.        0.00    0.00
</s>
```

3 Evaluation

The method was used to build language wordlists from sources described in the next subsection and evaluated on six groups of similar languages.

3.1 Wordlists

In this work, TenTen web corpus family [4] was used to build the language wordlists. Aranea web corpora [1] were used in addition to TenTen corpora in the case of Czech and Slovak. bsWaC, hrWaC and srWaC web corpora [5] were used in the case of Bosnian, Croatian and Serbian. All words, even hapax legomena were included in the wordlists. The source web pages were limited to the respective national TLD where possible.

Another set of wordlists to compare the method to other approaches was obtained from the DSL Corpus Collection¹ v. 1 made available at VarDial in 2014 [10].²

The last couple of wordlists for the purpose of evaluating the method was taken from corpus GloWbE comprising of 60 % blogs from various English speaking countries [3].³

The sizes and source TLDs of the wordlists are shown in Table 1. The difference of wordlist sizes is countered by using the relative counts in the algorithm.

Table 1: Sizes of wordlists used in the evaluation. Large web sources – TenTen, Aranea and WaC corpora – were limited to respective national TLDs. Other wordlists were built from the training and evaluation data of DSL Corpus Collection and parts of GloWbE corpus.

Language	Web TLD	Web wordlist	DSL wordlist	GloWbE wordlist
Bosnian	.ba	2,262,136	51,337	
Croatian	.hr	6,442,922	50,368	
Serbian	.rs	3,510,943	49,370	
Indonesian	–	860,827	48,824	
Malaysian	–	1,346,371	34,769	
Czech	.cz	26,534,728	109,635	
Slovak	.sk	5,333,581	121,550	
Portuguese, Brazilian	.br	9,298,711	52,612	
Portuguese, European	.pt	2,495,008	51,185	
Spanish, Argentine	.ar	6,376,369	52,179	
Spanish, Peninsular	.es	8,396,533	62,945	
English, Great Britain	.uk	6,738,021	42,516	1,222,292
English, United States	.us	2,814,873	42,358	1,245,821

3.2 Discriminating Similar Languages – VarDial Workshop

The evaluation of the language separation method described in this paper on DSL Corpus Collection v. 1 gold data⁴ performed by the original evaluation

¹ <http://ttg.uni-saarland.de/resources/DSLCC/>

² <http://corporavm.uni-koeln.de/wardial/sharedtask.html>

³ <http://www.corpusdata.org/>

⁴ <https://bitbucket.org/alvations/dslsharedtask2014/src/master/test-gold.txt>

script⁵ can be found in Table 2. The result overall accuracy is compared to the best result presented at VarDial 2014⁶

Table 2: Overall accuracy using large web corpus wordlists and DSL CC v. 1 training data wordlists on DSL CC v. 1 gold data. The best result achieved by participants in VarDial 2014 can be found in the last column.

Languages	Wordlist	Accuracy	DSL Best
English GB/US	Large web corpora	0.6913	0.6394
English GB/US	GloWbE	0.6956	0.6394
English GB/US	DSL training data	0.4706	0.6394
Other languages	Large web corpora	0.8565	0.8800
Other languages	DSL training data	0.9354	0.9571
Bosnian, Croatian, Serbian	DSL training data	0.8883	0.9360
Indonesian, Malaysian	DSL training data	0.9955	0.9955
Czech, Slovak	DSL training data	1.0000	1.0000
Portuguese BR/PT	DSL training data	0.9345	0.9560
Spanish AR/ES	DSL training data	0.8820	0.9095

The wordlist based language separation method performed comparably to the results of participants of VarDial 2014.

DSL data wordlists might have performed better than large web corpora wordlists on the DSL test data since DSL training sentences were more similar to test sentences than web documents. The results show that large web corpus based wordlists performed better than the DSL test data based wordlists in the case of discerning British from American English.

4 Conclusion and Future Work

A Python script implementing the method presented in this paper can be found at <http://corpus.tools/> under name *Language Filter*. In our experience with Czech and Slovak, with Norwegian and Danish, and with filtering English or languages similar to the target language out of many monolingual web corpora, the quality of the result corpus greatly benefited from applying this simple yet powerful script.

We might consider including the expectation-maximisation steps described in the original algorithm [2] in a separate version of the script in the future to evaluate discriminating language variants of Spanish (Peninsular vs. American variants), Portuguese (European vs. Brazilian), French (Hexagonal vs. Canadian).

⁵ <https://bitbucket.org/alvations/dslsharedtask2014/src/master/dslevalscript.py>

⁶ <http://htmlpreview.github.io/?https://bitbucket.org/alvations/dslsharedtask2014/downloads/dsl-results.html>

Training and test data from more recent VarDial workshops will be used to evaluate the performance on additional language groups, such as Bulgarian/Macedonian, Hexagonal/Canadian French, or Persian/Dari.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin infrastructure LM2015071. This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

References

1. Benko, V.: Aranea: Yet Another Family of (Comparable) Web Corpora. In: TSD. pp. 247–254
2. Herman, O., Suchomel, V., Baisa, V., Rychlý, P.: Dsl shared task 2016: Perfect is the enemy of good language discrimination through expectation–maximization and chunk-based language model. In: Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3). pp. 114–118 (2016)
3. Davies, M., Fuchs, R.: Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, 36(1), pp. 1–28. (2015)
4. Jakubiček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The tenten corpus family. In: 7th International Corpus Linguistics Conference CL. pp. 125–127 (2013)
5. Ljubešić, N., Klubička, F.: bs, hr, sr wac-web corpora of bosnian, croatian and serbian. In: Proceedings of the 9th Web as Corpus Workshop (WaC-9). pp. 29–35 (2014)
6. Lui, M., Baldwin, T.: langid.py: An off-the-shelf language identification tool. In: Proceedings of the ACL 2012 System Demonstrations. pp. 25–30. Association for Computational Linguistics, Jeju Island, Korea (Jul 2012), <https://www.aclweb.org/anthology/P12-3005>
7. Malmasi, S., Zampieri, M., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J.: Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In: Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3). pp. 1–14. Osaka, Japan (December 2016)
8. Michelfeit, J., Pomikálek, J., Suchomel, V.: Text tokenisation using unitok. In: RASLAN. pp. 71–75 (2014)
9. Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., Aepli, N.: Findings of the vardial evaluation campaign 2017. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). pp. 1–15. Association for Computational Linguistics, Valencia, Spain (April 2017)
10. Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J.: A report on the dsl shared task 2014. In: Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects. pp. 58–67 (2014)
11. Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., Nakov, P.: Overview of the dsl shared task 2015. In: Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects. pp. 1–9 (2015)