

Implementation and Evaluation of a Language Identification System for Mono- and Multi-lingual Texts

Olga Artemenko, Thomas Mandl, Margaryta Shramko, Christa Womser-Hacker

Information Science, University of Hildesheim

31141 Hildesheim, Germany

mandl@uni-hildesheim.de

Abstract

Language identification is a classification task between a pre-defined model and a text in an unknown language. This paper presents the implementation of a tool for language identification for mono- and multi-lingual documents. The tool includes four algorithms for language identification. An evaluation for eight languages including Ukrainian and Russian and various text lengths is presented. It could be shown that n-gram-based approaches outperform word-based algorithms for short texts. For longer texts, the performance is comparable. The tool can also identify language changes within one multi-lingual document.

Keywords

Language identification, n-gram indexing, language model, evaluation

1 Introduction and Motivation

Language Identification is a research topic which became important with the success of the internet. The authors of internet pages do not always give meta data which shows in which language the text on the page is. For users, this is rarely a problem. However, when a user encounters a page in an unknown language and wants to automatically translate it with an online tool, he usually needs to specify the source language.

The problem of not knowing the language of an internet page is more serious for machines and automatic systems. Access to web pages is often provided by internet search engines which automatically crawl and index pages. Indexing methods are usually language dependent because they require knowledge about the morphology of a language [Fuhr 2005]

Even indexing methods which do not rely on linguistic knowledge like n-gram based stemming can be optimized for languages by choosing an appropriate value for n [McNamee and Mayfield 2004]. Often, web search engines focus on content in one specific language and aim at directing their crawlers to pages in that language [Martins and Silva 2005].

Language is a barrier for user access. Therefore, it is an important factor which needs to be considered during web usage mining of multilingual sites [Kralisch and Mandl 2006]. Automatic language identification can support this endeavor.

The motivation for the development of the language identification tool presented in this paper is twofold. First,

the tool has been developed to be used for language identification for the WebGOV corpus [Sigurbjörnsson et al. 2005]. This collection of web pages has been engineered for the Cross Language Evaluation Forum (CLEF, www.clef-campaign.org), an evaluation initiative for cross- and multi-lingual information retrieval tasks [Braschler and Peters 2004]. In 2005, the first multi-lingual web collection was developed for a comparative analysis of information retrieval approaches for web pages. The University of Hildesheim is working with the web corpus [Jensen 2005] and wants to develop an improved language identification tool for this purpose.

The second reason behind this work is an interest in multi-lingual documents. On the web, more and more pages contain text in more than one language. There may be short sentences like “optimized for internet explorer”, foreign language citations or even parallel text. So far, few research has been dedicated toward multi-lingual content. The current project and the tool presented in this paper aims at recognizing the extent to which multi-lingual content is present on the web and to which extent it can be automatically identified.

Language identification is closely related to the recognition of the character encoding. This aspect is not dealt with in this paper.

The remainder of the paper is organized as follows. The following section introduces research on language identification. Section 3 describes the tool LangIdent, the implemented algorithms, the interface and the language model creation. Section 4 shows the evaluation results and the last section gives an outlook to future work.

2 Related Work

Most language identification systems are either based on words or n-grams. This section provides a brief overview.

It is obvious that words often are unique for a language and that they can be used for language identification. On the other hand, for efficiency reasons, not all words of a language can be used for language identification nor are all words known. All languages integrate new words into their vocabulary frequently. Many character sequences can be words in more than one language. Therefore, most approaches are based on common or frequent words [Martino and Paulsen 2001, Cowie et al. 1999]

For short texts, word based language identification can easily fail, when a few words are present and these are not stored in the language model. Therefore, character n-grams have been used for identification as well. This approach primarily focused on the occurrence of characters or n-grams unique for a specific language [Souter et al.

1994]. Current approaches store the frequency of the most frequent n-grams and compares them to the n-grams in a text [Cavnar and Trenkle 1994].

Most of the approaches for mapping a document to one language model use traditional algorithms from machine learning which do not need to be mentioned here. Merely one algorithm for matching ranked lists of items should be mentioned here, the “out of place” method. It compares the ranks of the most frequent items in the document and the model. The distance between the rank in one list and the rank in the other list is calculated. The distances are summed up and provide a measure for the similarity between model and document. This method can be regarded as a simple approach to rank correlation. It has been applied by [Cavnar and Trenkle 1994].

Most previous experiments have been carried out for Western European languages. For Ukrainian and Russian, which have been analyzed in this paper, no publications can be found.

A more comprehensive review of previous research work is provided by [Artemenko and Shramko 2005].

3 Implementation of the Prototype LangIdent

LangIdent is the prototype for language identification. It has been developed in JAVA and has a graphical user interface, but can also be run in batch mode. Further details can be found in [Artemenko and Shramko 2005].

3.1 Algorithms

Based on previous research, the system includes four classification algorithms:

- Vector space cosine similarity between inverse document frequencies
- “out of place” similarity between rankings
- Bayesian classification
- Word based method (count of word hits between model and language)

The first three methods are based on n-grams. The prototype includes words as well as n-grams.

The multi-lingual language identification runs a window of k words through the text and matches the short window with the language models.

3.2 Language Model Development

The prototype allows the assembly of a language model from an example text. Words and n-grams are stored in the model and depending on the selection of the user during the classification phase, only one of them may be used.

Previous retrieval experiments with n-gram models showed that tri-grams work reasonably well for most languages [McNamee and Mayfield 2004]. Based on this experience, we implemented tri-gram models within LangIdent. For both the n-gram and the word based model, some parameters can be specified by the user.

Trigram-Parameters:

- absolute frequency
- relative frequency
- inverse document frequency
- transition probability

For language models based on words, the same parameters are used, except for the last one. It is replaced by the cumulative probability

The models can be explored within the prototype and even be manipulated manually. For example, if the user encounters a usually non-frequent word, a proper name or even a foreign language word which occurred often in the training corpus, this word can be deleted from the model. Figure 1 shows the interface for the language model selection and manipulation.

4 Evaluation

With the prototype LangIdent described above, models for eight language were developed (German, English, Spanish, French, Italian, Russian, Czech, Ukrainian). These models were evaluated. The text for the language model creation had a size of some 200 Kbyte from a newspaper corpus [Braschler and Peters 2004]. For word-based methods, the most frequent words with a cumulative probability of 40% were stored and for n-gram methods, the 1500 most frequent tri-grams were included into the model. The models were not further processed manually.

First, an evaluation for the word-based method was carried out in order to determine the best settings. Subsequently, the best word-based approach was compared to the other methods.

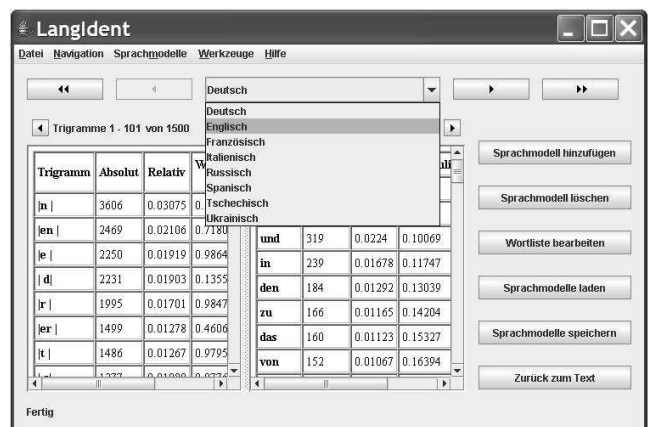


Figure 1: Language Model displayed in LangIdent

4.1 Word-based Method

There are two main approaches for the identification of a language with a word based model. Either the word hits between text and all language models are counted or the relative frequency of all word hits are added. Both methods are mentioned in the research literature.

In a preliminary test with six languages and text parts of size 250 Bytes it could be shown that the simple word count is superior. The results are presented in table 1.

Table 1: Error rates for two word-based methods

English	word frequency	0.12
	word count	0.12
French	word frequency	0.62
	word count	21.89
German	word frequency	0
	word count	0.35
Italian	word frequency	0
	word count	3.55
Russian	word frequency	0.12
	word count	0.12
Spanish	word frequency	0.48
	word count	0.12

Consequently, the main evaluation relies solely on the word count.

4.2 Eight Languages and Document Size

The quality of language identification as well as for many other classification tasks heavily depends on the amount of evidence provided. For language identification, it depends on the number of characters available. As a consequence, the system was tested with text of varying length. Newspaper documents from all eight languages were split into sections of length between 25 and 500 characters.

The recognition rate for shorter sections is important for an analysis of multi-lingual documents. The error rates for all languages can be found in table 2. The best results for are shaded. Figure 2 displays the error rates for document size 100.

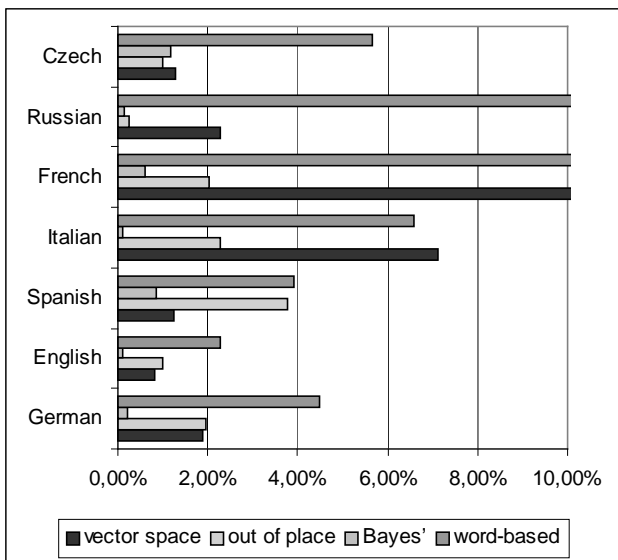


Figure 2: Error rates for Document Size 100 Bytes

It can be seen from figure 2 and table 2 that the Bayes method results in the best classification quality for most languages. Only for Czech and Ukrainian, out-of-place is superior. A look at the average performance over all languages considered confirms the assumption, that Bayes leads to the highest performance. The numbers are given in table 3.

Table 2: Detailed error rates for four classification methods

Language	document size (Bytes)	Vector space	Out of place	Bayes'	Word-based
German	50	8.30%	6.64%	2.32%	20.04%
	100	1.90%	1.96%	0.21%	4.48%
	250	0.12%	0.12%	0%	0%
	500	0%	0%	0%	0%
English	50	4.50%	6.72%	1.79%	12.40%
	100	0.82%	1.01%	0.10%	2.26%
	250	0%	0%	0%	0%
	500	0%	0%	0%	0%
Spanish	50	5.37%	10.39%	5.55%	15.76%
	100	1.25%	3.76%	0.85%	3.91%
	250	0%	0.37%	0%	0.37%
	500	0%	0%	0%	0%
Italian	50	20.26%	11.10%	1.94%	24.68%
	100	7.10%	2.27%	0.10%	6.60%
	250	0.48%	0.12%	0%	0.12%
	500	0%	0%	0%	0%
French	50	29.45%	9.15%	3.53%	31.06%
	100	14.88%	2.02%	0.62%	10.32%
	250	3.14%	0.12%	0%	0.70%
	500	0.92%	0%	0%	0%
Russian	50	5.77%	2.84%	2.16%	31.18%
	100	2.26%	0.26%	0.16%	12.55%
	250	0.61%	0%	0%	1.47%
	500	0%	0%	0%	0.24%
Czech	50	4.07%	3.51%	4.02%	20.98%
	100	1.28%	0.98%	1.18%	5.65%
	250	0.62%	0.25%	0.25%	0%
	500	0%	0%	0%	0%
Ukrainian	50	9.92%	6%	6.11%	31.32%
	100	6.46%	1.95%	2.20%	12.81%
	250	2.84%	0.49%	0.62%	1.85%
	500	1.71%	0.24%	0.73%	0.24%
1000	0%	0%	0%	0%	

Table 3: Average error rates for four classification methods

Document size (Bytes)	Vector space	Out of place	Bayes'	Word-based
0	10.95%	7.04%	3.43%	23.43%
100	4.49%	1.78%	0.68%	7.32%
250	0.98%	0.18%	0.11%	0.56%
500	0.33%	0.03%	0.09%	0.06%
1000	0%	0%	0%	0%

An informal analysis of wrongly classified text parts showed that often proper names and words in other languages led to the misclassification. However, it could be argued that in cases where a text snippet from a French

newspaper contains mainly English words, it should indeed not be classified as a French text. However, not all errors can be manually assessed.

4.3 Multi-Lingual Documents

The evaluation of language identification for multi-lingual content is ongoing. Different metrics need to be developed for this endeavor. Mainly two issues need to be considered:

- Identification of the languages present in the document
- Identification of the place of a language shift

For this evaluation, two corpora are assembled. One is a collection of real-world multi-lingual documents from the web. Some 200 documents have been found so far. Apart from this real-world data, an synthetic corpus of multi-lingual documents has been assembled from the data used for the mono-lingual experiments described above. Figure 3 shows the user interface of LangIdent for a successful recognition of multi-lingual parts of one document. The layout is modified for the languages.

5 Future Work

LangIdent allows the setting of many parameters. It enables further extense evaluation. The evaluation of LangIdent for mono-lingual documents or for documents with a dominating language will continue and will be extended to the EuroGOV corpus of web documents. We are in the

process of creating a set of manually identified pages for many languages as ground truth for the system. The list will then be compared to the one provided by the organizers of the EuroGOV corpus [Sigurbjörnsson et al. 2005]. For this corpus, several evidences for the language of a document are present. First, the top level domain provides first evidence. For example, pages of the de domain are often in German. In addition to the recognition results of LangIdent, the language of pages linking to the page under question and link label text are also available.

We intend to integrate LangIdent as one service within the RECOIN framework (REtrieval COmponent INtegrator, <http://recoin.sourceforge.net>) [Scheufen 2005]. RECOIN is an object oriented JAVA framework for information retrieval experiments. It allows the integration of heterogeneous components into an experimentation system where experiments can be carried out. RECOIN is motivated by the adaptive fusion MIMOR model for the integration of several information retrieval systems [Mandl and Womser-Hacker 2004].

For future experiments with the EuroGOV corpus, we intend to integrate advanced quality models for web documents [Mandl 2005].

The evaluation of multi-lingual documents is a great challenge which still lies ahead. During evaluation, several new approaches need to be tested. The fusion of several classifiers needs to be adapted to the language identification problem. For example, parameters like window size and fusion between word and n-gram methods need to be set based on previous knowledge like results from individual classifiers and the dominating language of the document.



Figure 3: A multi-lingual document in LangIdent

References

- [Artemenko and Shramko 2005] Olga Artemenko; Margaryta Shramko: *Entwicklung eines Werkzeugs zur Sprachidentifikation in mono- und multilingualen Texten*. Magister Thesis, University of Hildesheim, Information Science. 2005. to appear.
- [Braschler and Peters 2004] Martin Braschler, Carol Peters: Cross-Language Evaluation Forum: Objectives, Results, Achievements. *Information Retrieval*. 2004 no. 7. pp. 7-31.
- [Cavnar and Trenkle 1994] W.B. Cavnar; J. M.; Trenkle: N-Gram-Based Text Categorization. In: *Symposium on Document Analysis and Information Retrieval*. University of Nevada, Las Vegas, pp. 161-176.
- [Cowie et al. 1998] J. Cowie; E. Ludovik; R. Zacharski: An Autonomous, Web-based, Multilingual Corpus Collection Tool. In: *Proceedings International Conference on Natural Language Processing and Industrial Applications*. Moncton, pp. 142-148.
- [Fuhr 2005] Norbert Fuhr: *Information Retrieval: Skriptum zur Vorlesung*. http://www.is.informatik.uni-duisburg.de/courses/ir_ss05/fohlen/irskall.pdf
- [Jensen 2005] Niels Jensen: Mehrsprachiges Information Retrieval mit einem WEB-Korpus. In: T. Mandl, C. Womser-Hacker (eds.): *Proc. Vierter Hildesheimer Information Evaluierungs- und Retrieval Workshop (HIER)* Hildesheim, 20.7.2005. Univ.-verlag Konstanz. to appear.
- [Kralisch and Mandl 2006] Anett Kralisch, Thomas Mandl: Barriers of Information Access across Languages on the Internet: Network and Language Effects. To appear in: *Proc. Hawaii International Conference on System Sciences (HICSS-39)*
- [Mandl 2005] Thomas Mandl: The quest for the best pages on the web. *Information Service & Use*. To appear
- [Mandl and Womser-Hacker 2005] Thomas Mandl, Christa Womser-Hacker: A Framework for long-term Learning of Topical User Preferences in Information Retrieval. *New Library World* vol. 105 (5/6) pp. 184-195.
- [Martino and Paulsen 2001] Michael Martino; Robert Paulsen (2001): Natural language determination using partial words, April 2001, U.S. Patent No. 6216102 B1.
- [Martins and Silva 2005] Bruno Martins; Marió Silva: Language Identification in Web Pages. In: *Proceedings ACM SAC Symposium on Applied Computing*. Santa Fe, New Mexico, USA. March 13.-17. 2005, pp. 764-768.
- [McNamee and Mayfield 2004] Paul McNamee; J. Mayfield: Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7 (1/2) 2004. pp. 73-98.
- [Scheufen 2005] Jan-Hendrik Scheufen (2005): Das RECOIN Framework für Information Retrieval Experimente. In: T. Mandl, C. Womser-Hacker (eds.): *Proc. Vierter Hildesheimer Information Evaluierungs- und Retrieval Workshop (HIER)* Hildesheim, 20.7.2005. Univ.-verlag Konstanz. to appear.
- [Sigurbjörnsson et al. 2005] Börkur Sigurbjörnsson, Jaap Kamps, and Maarten de Rijke: Blueprint of a cross-lingual web retrieval collection. *Journal on Digital Information Management* 3 (9-13) 2005.
- [Souter et al. 1994] C. Souter, G. Churcher, J. Hayes, J. Hughes, and S. Johnson: Natural Language Identification Using Corpus-Based Models. *Hermes J. Linguistics*, 1994 vol. 13, pp. 183-203.