Language Identification in Multi-lingual Web-Documents

Thomas Mandl, Margaryta Shramko, Olga Tartakovski,
and Christa Womser-Hacker

Universität Hildesheim, Information Science
Marienburger Platz 22, D-31141 Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract. Language identification an important task for web information retrieval. This paper presents the implementation of a tool for language identification in mono- and multi-lingual documents. The tool implements four algorithms for language identification. Furthermore, we present a n-gram approach for the identification of languages in multi-lingual documents. An evaluation for monolingual texts of varied length is presented. Results for eight languages including Ukrainian and Russian are shown. It could be shown that n-gram-based approaches outperform word-based algorithms for short texts. For longer texts, the performance is comparable. The evaluation for multi-lingual documents is based on both short synthetic documents and real world web documents. Our tool is able to recognize the languages present as well as the location of the language change with reasonable accuracy.

1 Introduction

Language Identification is a research topic which is becoming increasingly important due to the success of the internet. The authors of internet pages do not always provide reliable meta data which indicates the language of the text on the page. Even worse, many internet documents contain text portions in different languages. Language identification is an important task for web search engines.

For users, it is rarely a problem to identify the language of a document as long as it is written in a language they understand. However, when a user encounters a page in an unknown language and wants to automatically translate it with an online tool, the source language usually needs to specified. 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 [7]. Consequently, automatic language identification is necessary for web usage mining studies.

Access to web pages is often guaranteed 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 [5]. 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 [11]. In 2005, the first multi-lingual web collection for a comparative analysis of information retrieval approaches for web pages was released [13]. Systems working with this corpus need accurate language identification.

C. Kop et al. (Eds.): NLDB 2006, LNCS 3999, pp. 153 – 163, 2006.

Often, web search engines focus on content in one specific language and aim at directing their crawlers to pages in that language [10].

Multilingual documents containing text in different languages are a reality on the web. There may be short sentences like “optimized for internet explorer”, foreign language citations or even parallel text. Many popular language identification methods deal inadequately with multi-lingual documents and their performance drops. So far, little 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

Language identification is a classification task between a pre-defined model and a text in an unknown language. Most language identification systems are either based on short and frequent words or character 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 [11, 3,4]. Typically, the most frequent words in a trainings corpus of a known language are determined. The number of words used varies, Souter et al. use 100 words [10] and Cowie et al. use 1000 [3]. Similar languages often share some common words and are therefore more difficult to be distinguished.

A novel and elaborated approach considers word classes and their order [8]. Closed word classes which rarely change in a language like adverbs or prepositions are listed for each language. At first the algorithm checks the presence of prepositions in the text. Depending on the result and previous empirical experience, another word class is called and tested. In that manner, a word class can be called upon which discriminates well between previous hypotheses [8].

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 and n varies between 2 and 5. This approach primarily focused on the occurrence of characters or n-grams unique for a specific language [14]. Current approaches store the frequency of the most frequent n-grams and compare them to the n-grams encountered in a text [2]. Excellent results can be achieved by combining words and n-grams [12].

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, except for 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 [2].

Most previous experiments have been carried out for Western European languages. Little research has been done for Ukrainian and Russian.

Research has focused on the identification of one language per document. The identification of multiple languages in documents has not been dealt with. The most prominent approach lies a floating window of n words on the text [10]. The language within the window is identified and all windows are projected onto the document.

3 Implementation of LangIdent

We developed LangIdent, a prototype for language identification. LangIdent allows the development of models from training data. It has been implemented in JAVA and has a graphical user interface, but can also be run in batch mode. More details can be found in [1].

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 form 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 [11]. 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

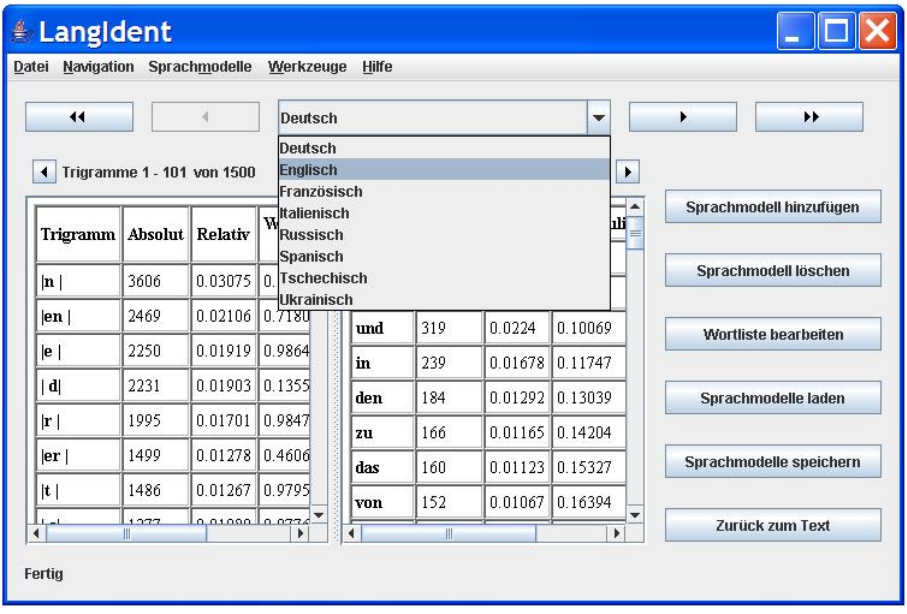

Fig. 1. Language Model displayed in LangIdent

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.

3.3 Multi-lingual Language Identification

Our algorithm for language identification in multi-lingual documents is based on the word-window approach [10]. However, it was modified to overcome some shortcomings. For short text passages in another language, word based approaches are not the optimal solution. This will also be shown by the evaluation for mono-lingual documents in the following section. As a consequence, we build a n-gram language model. We chose tri-grams and a windows size of eight words.

For each window, the most likely language is determined based on the transition probability between tri-grams. For windows in which a different language occurs, a language change is assumed. The position is determined based on the position of the first window in which the new language is encountered. The change is assumed to occur at the first position of that window plus two words.

4 Evaluation for Mono-lingual Documents

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 several online newspapers.

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	21.9
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

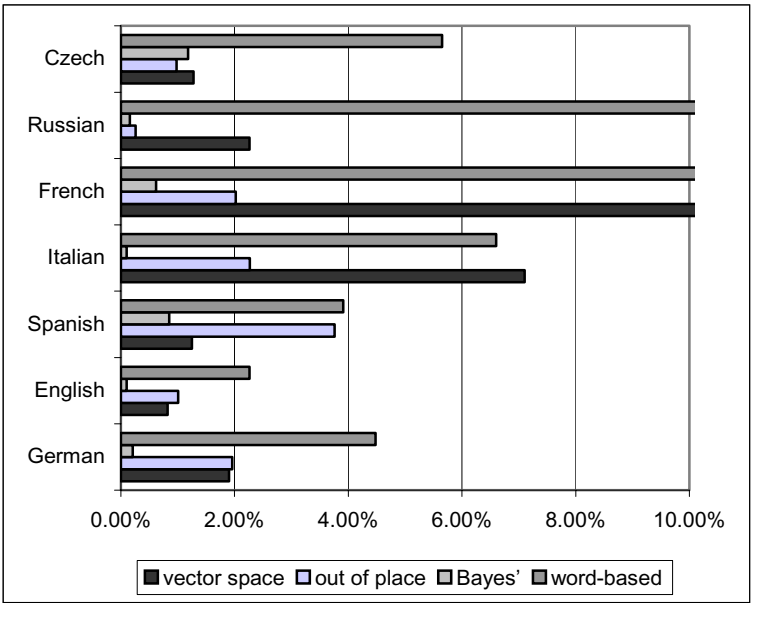

Fig. 2. Error rates for Document Size 100 Bytes

Table 2. Detailed error rates for four classification methods

Language	document size (chars)	Vector space	Out of place	Bayes'	Word-based
German	*25*	8.30%	6.64%	2.32%	20.04%
	50	1.90%	1.96%	0.21%	4.48%
	125	0.12%	0.12%	0%	0%
	250	0%	0%	0%	0%
	500	0%	0%	0%	0%
English	*25*	4.50%	6.72%	1.79%	12.40%
	50	0.82%	1.01%	0.10%	2.26%
	125	0%	0%	0%	0%
	250	0%	0%	0%	0%
	500	0%	0%	0%	0%
Spanish	*25*	5.37%	10.39%	5.55%	15.76%
	50	1.25%	3.76%	0.85%	3.91%
	125	0%	0.37%	0%	0.37%
	250	0%	0%	0%	0%
	500	0%	0%	0%	0%
Italian	*25*	20.26%	11.10%	1.94%	24.68%
	50	7.10%	2.27%	0.10%	6.60%
	125	0.48%	0.12%	0%	0.12%
	250	0%	0%	0%	0%
	500	0%	0%	0%	0%
French	*25*	29.45%	9.15%	3.53%	31.06%
	50	14.88%	2.02%	0.62%	10.32%
	125	3.14%	0.12%	0%	0.70%
	250	0.92%	0%	0%	0%
	500	0%	0%	0%	0%
Russian	*25*	5.77%	2.84%	2.16%	31.18%
	50	2.26%	0.26%	0.16%	12.55%
	125	0.61%	0%	0%	1.47%
	250	0%	0%	0%	0.24%
	500	0%	0%	0%	0%
Czech	*25*	4.07%	3.51%	4.02%	20.98%
	50	1.28%	0.98%	1.18%	5.65%
	125	0.62%	0.25%	0.25%	0%
	250	0%	0%	0%	0%
	500	0%	0%	0%	0%
Ukrainian	*25*	9.92%	6%	6.11%	31.32%
	50	6.46%	1.95%	2.20%	12.81%
	125	2.84%	0.49%	0.62%	1.85%
	250	1.71%	0.24%	0.73%	0.24%
	500	0%	0%	0%	0%

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 which would be possible within LangIdent.

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.

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. 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.

It can be seen from figure 2 and table 2 that a Bayes classifier 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 3. Average error rates for four classification methods

Document size (Chars)	Vector space	Out of place	Bayes'	Word-based
25	10.95%	7.04%	3.43%	23.43%
50	4.49%	1.78%	0.68%	7.32%
125	0.98%	0.18%	0.11%	0.56%
250	0.33%	0.03%	0.09%	0.06%
500	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 argues that in cases where a text snippet form 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. The experiments are described in more detail [1].

5 Evaluation for Multi-lingual Documents

The evaluation of language identification for multi-lingual content is difficult. 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 location of a language shift

Figure 3 shows the user interface of LangIdent for a successful recognition of multi-lingual parts of one document. The text layout is modified for the different languages.

5.1 Corpus Creation

For this evaluation, two corpora were assembled. One is a collection of real-world multi-lingual documents from the web. Some suitable 100 documents with two languages have been identified. Most multilingual texts contain more language due to the following reasons:

- Parallel text: the same text is present in two languages
- Citation: Text in one language contains a citation in another language

For the evaluation, we did not consider syntactic hints for language changes or layout changes. These texts were long on average. Length varies between 120 and 1500 characters and average length is 550 characters. In order to evaluate the performance of our approach for smaller portions of text, a synthetic corpus of short texts was created. The synthetic corpus of multi-lingual documents has been assembled from the data used for the mono-lingual experiments described above.

Three methods were used to create text which has similar features as the real world texts:

- XY: Two languages were subsequently pasted into a document (like a parallel text)
- XYX: One portion in one language is inserted into a document in another language (like a citation)
- XYZ: Three languages were subsequently pasted into a document

All eight languages mentioned above were used for the synthetic corpus. Altogether, 100 texts were created. Their average length is 130 characters for type XY and 280 for the other types.

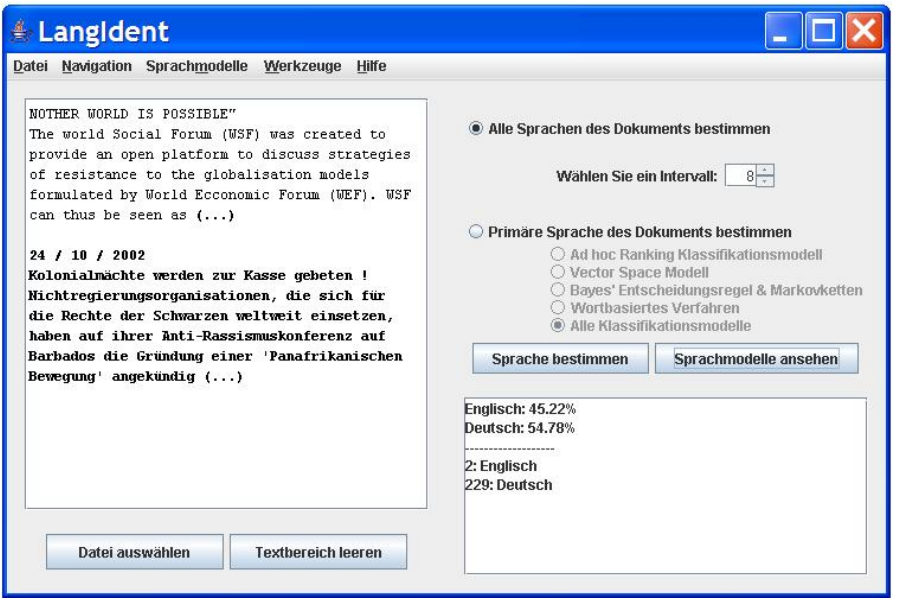

Fig. 3. A multi-lingual document in LangIdent

5.2 Evaluation Results

The languages present in the documents were generally well identified in both corpora. Results are shown in table 4.

Table 4. Accuracy of multiple language identification

Type	All languages in document correctly identified
Internet	97 %
XY	96 %
XYX	95 %
XYZ	97 %

Most errors are due to the presence of similar languages like Ukrainian and Russian or Italian and Spanish.

The identification of the location of the language change is a harder challenge. The evaluation measures the distance between actual change and detected change. Documents of type XYX and XYZ have two languages changes. The change detected with less accuracy is used as measure for the entire document. Table 5 displays the evaluation results.

Overall, the identification of language change locations in multilingual documents is possible with good accuracy. For more than 50% of the documents, language

Table 5. Accuracy of language change location detection

Type	*Exact position*	*1 word off*	*2 words off*	*3 words off*	*4 words off*	*Cumulative for at most 2 words off*
Internet	29 %	26 %	26 %	10 %	3.2 %	81 %
XY	38 %	40 %	16 %	2.0 %	-	94 %
XYX	20 %	55 %	10 %	10 %	-	85 %
XYZ	39 %	45 %	13 %	-	-	97 %

changes are detected with no or one word distance from the actual location. The recognition rate for real world documents is lower than for the synthetic corpus. This seems surprising at first, because the internet documents are longer. However, the language identification actually works on the shorter window. The method reaches a recognition rate for the change location of 81% for internet documents. This compares well to 80% for monolingual HTML documents in a recent experiment [8].

For future evaluation experiments, other evaluation measures need to be considered. Recall and precision of correctly identified sections might also be a valid measure. However, the final yardstick for the evaluation of languages identification programs depends on the application area. For information retrieval systems, the quality of the subsequent retrieval system will determine the adequacy of a language identification system.

6 Conclusion and Future Work

LangIdent allows the setting of many parameters. It enables further extensive 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.

For the EuroGOV 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.

References

1. Artemenko, O.; Shramko, M.: Entwicklung eines Werkzeugs zur Sprachidentifikation in mono- und multilingualen Texten. Master thesis. University of Hildesheim. 2005. http://web1.bib.uni-hildesheim.de/2005/artemenko_shramko.pdf
2. Cavnar, W.B., Trenkle, J. M: N-Gram-Based Text Categorization. Symposion on Document Analysis and Information Retrieval. Univ. of Nevada, Las Vegas. 161-176.

3. Cowie, J., Ludovik, E., Zacharski, R.: An Autonomous, Web-based, Multilingual Corpus Collection Tool. Proc Intl. Conference on Natural Language Processing and Industrial *Applications*. Moncton. 1998. 142-148.
4. Grefenstette, G.: Comparing two language identification schemes. In: *JADT 1995, 3rd International conference on Statistical Analysis of Textual Data*, Rome.
5. Hollink, V., Kamps, J., Monz, C., de Rijke, M.: Monolingual Document Retrieval for European Languages. In: *Information Retrieval* 7 (1-2) 2004. 33-52
6. Kikui, G. Identifying the Coding System and Language of On-line Documents on the Internet. In: *Proc 16th Conf. on Computational linguistics. Denmark*, vol. 2 (1996), 652-657.
7. Kralisch, A., Mandl T.: Barriers of Information Access across Languages on the Internet: Network and Language Effects. In: *Proc Hawaii International Conference on System Sciences (HICSS-39)* 2006.
8. Lins, R., Gonçalves, P.: Automatic Language Identification of Written Texts. *Proc ACM SAC Symposium on Applied Computing*, March 2004, Nicosia, Cyprus. 1128-1133.
9. Martino, M. and Paulsen, R.: Natural language determination using partial words, Apr. 2001, U.S. Patent No. 6216102 B1.
10. Martins, B., Silva, M.: Language Identification in Web Pages. *Proc ACM SAC Symposium on Applied Computing*. March 13.-17. 2005. Santa Fe, New Mexico, USA. 764-768.
11. McNamee, P., Mayfield, J.: Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7 (1/2) 2004. 73-98.
12. Prager, J.: Linguini: Language Identification for Multilingual Documents. In: *Proc 32nd Hawaii International Conf on System Sciences*, 1999.
13. Sigurbjörnsson, B., Kamps, J., de Rijke, M.: Blueprint of a cross-lingual web retrieval collection. *Journal on Digital Information Management* vol. 3 (9-13) 2005.
14. Souter, C., Churcher, G., Hayes, J., Hughes, J., Johnson, S.: Natural Language Identification Using Corpus-Based Models. Hermes J. Linguistics vol. 13, 1994. 183-203.