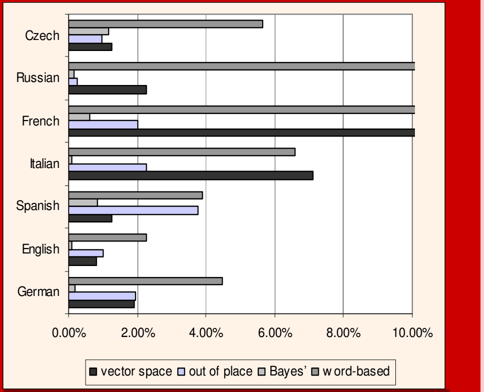
Olga Artemenko, Thomas Mandl, Margaryta Shramko, Christa Womser-Hacker
Information Science
University of Hildesheim - Germany

mandl@uni-hildesheim.de

Evaluation of a Language Identification System for Mono- and Multi-lingual Text Documents

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.



Evaluation Results for Mono-lingual Documents

LangIdent System

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 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.

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 [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

Evaluation for Multi-lingual Documents

Different metrics need to be developed for the evaluation of language identification for multi-lingual content . Mainly two issues need to be considered:

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

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:

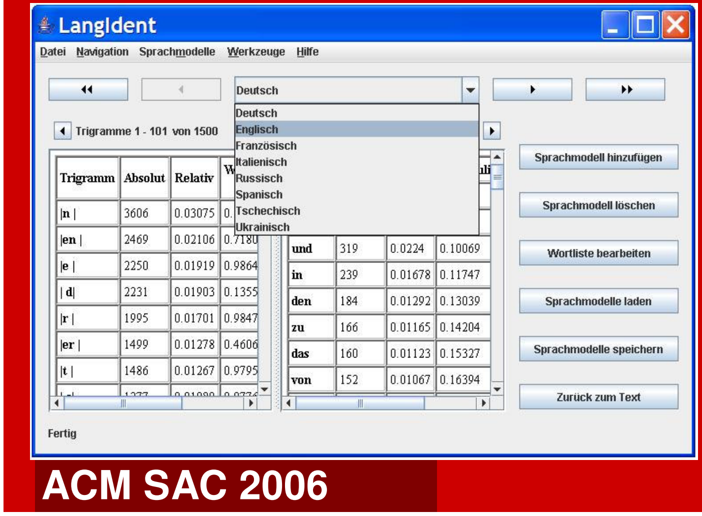
- Parallel text: the same text is present in two languages
- Citation: Text in one language contains a citation in another language

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

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

Type	*Exact posi-tion*	*1 word off*	*2 words off*	*3 words off*	*4 words off*	*Cumulative for at most 2 words off*
Inter-net	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 %

Evaluation Results for Multi-lingual Documents



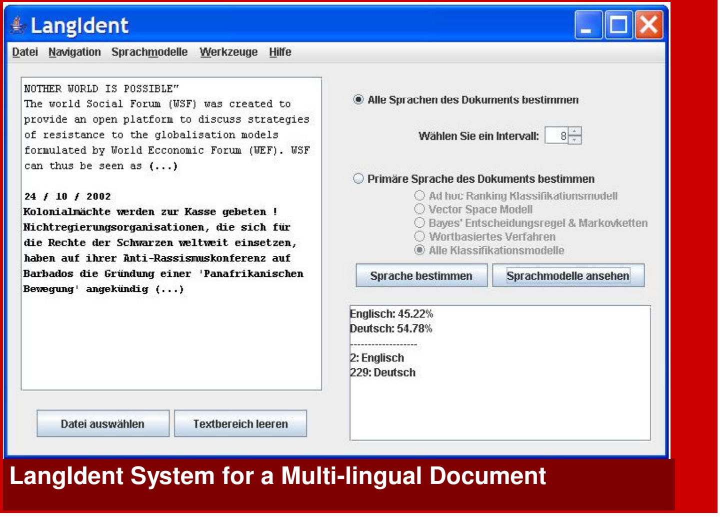
Trigramme 1 - 101 von 1500

Trigramm	Absolut	Relativ	W
|n |	3606	0.03075	0.
|en |	2469	0.02106	0.7180
|e |	2250	0.01919	0.9864
| d|	2231	0.01903	0.1355
|r |	1995	0.01701	0.9847
|er |	1499	0.01278	0.4606
|t |	1486	0.01267	0.9795

und	319	0.0224	0.10069
in	239	0.01678	0.11747
den	184	0.01292	0.13039
zu	166	0.01165	0.14204
das	160	0.01123	0.15327
von	152	0.01067	0.16394

Fertig

ACM SAC 2006



```
NOTHER WORLD IS POSSIBLE"
The world Social Forum (WSF) was created to
provide an open platform to discuss strategies
of resistance to the globalisation models
formulated by World Ecconomic Forum (WEF). WSF
can thus be seen as (...)

24 / 10 / 2002
Kolonialmächte werden zur Kasse gebeten !
Nichtregierungsorganisationen, die sich für
die Rechte der Schwarzen weltweit einsetzen,
haben auf ihrer Anti-Rassismuskonferenz auf
Barbados die Gründung einer 'Panafrikanischen
Bewegung' angekündig (...)
```

● Alle Sprachen des Dokuments bestimmen

Wählen Sie ein Intervall: 8

○ Primäre Sprache des Dokuments bestimmen
- ○ Ad hoc Ranking Klassifikationsmodell
- ○ Vector Space Modell
- ○ Bayes' Entscheidungsregel & Markovketten
- ○ Wortbasiertes Verfahren
- ● Alle Klassifikationsmodelle

Englisch: 45.22%
Deutsch: 54.78%

2: Englisch
229: Deutsch

LangIdent System for a Multi-lingual Document