

# Methoden zur automatischen Sprachidentifikation in mono- und multilingualen Texten

Olga Artemenko  
Margaryta Shramko



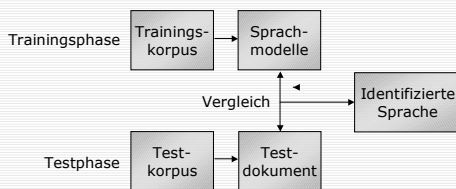
## Inhalt

- Phasen des Sprachidentifikationsprozesses
- Ansätze der automatischen Sprachidentifikation
- LangIdent
  - Sprachmodellerstellung
  - Klassifikationsmethoden
  - Evaluierung und Testergebnisse
  - Demonstration

Olga Artemenko, Margaryta Shramko



## Sprachidentifikationsprozess



Olga Artemenko, Margaryta Shramko



## Ansätze der automatischen Sprachidentifikation (1)

- N-Gramm basierter Ansatz
  - W. B. Cavnaar & J. M. Trenkle 1994
  - T. Dunning 1994
  - M. Damashek 1995, u.a.
- N-Gramm  
eine Sequenz von  $n$  ( $n=1,2,3,\dots$  Zeichen) aufeinander folgenden Zeichen eines längeren Strings bzw. Wortes  
  
Z.B.: „Information“ wird in folgende Tri-Gramme zerlegt:  
inf, nfo, for, orm, rma, mat, ati, tio, ion

Olga Artemenko, Margaryta Shramko



## Ansätze der automatischen Sprachidentifikation (2)

- Wortbasierter Ansatz
  - „frequent word“ oder „common word“ Methode  
M. J. Martino & R. C. Paulsen 1996, 1999, 2001  
C. Souter et al. 1994  
J. Cowie et al. 1999
  - „short word“ Methode  
J. M. Prager 1999
  - Geschlossene Wortklassen  
R. D. Lins & P. Gonçalves 2004

Olga Artemenko, Margaryta Shramko



## Ansätze der automatischen Sprachidentifikation (3)

- Typische Sonderzeichen: ä, ü, ç, ç, ß, ï, ...
  - P. Newman 1987
- Unikale Buchstabenkombinationen: „czy“, „sch“, ...
  - C. Souter et al. 1994 u.a.
- „word shape tokens“
  - P. Sibun & L. A. Spitz 1994
  - P. Sibun & J. C. Reynar 1996
  - C. L. Tan 1999, u.a.
- Kombinierte Methoden
  - J. M. Prager 1999
  - B. M. Schulze 2000, u.a.

Olga Artemenko, Margaryta Shramko



## LangIdent: Sprachmodellerstellung(1)

- Trainingskorpus in 8 Sprachen
  - Deutsch
  - Englisch
  - Spanisch
  - Französisch
  - Italienisch
  - Russisch
  - Tschechisch
  - Ukrainisch
- ca. 200 KBytes pro Sprache

Olga Artemenko, Margaryta Shramko



## LangIdent: Sprachmodellerstellung(2)

- Sprachmodell
  - Tri-Gramm-Liste
  - Wortliste
- Tri-Gramm-Liste = die 1500 häufigsten Tri-Gramme
  - **Tri-Gramm**: absolute Häufigkeit  
relative Häufigkeit  
inverse Dokumenthäufigkeit  
Übergangswahrscheinlichkeit
- Wortliste = die häufigsten Wörter, deren kumulative Häufigkeit 40% beträgt
  - **Wort**: absolute Häufigkeit  
relative Häufigkeit  
inverse Dokumenthäufigkeit  
kumulative Häufigkeit

Olga Artemenko, Margaryta Shramko



## LangIdent: Sprachidentifikation monolingualer Dokumente

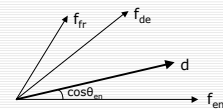
- Klassifikationsmethoden
  - Vector Space Modell
  - Ad Hoc Ranking
  - Bayes'sche Entscheidungsregel
  - Wortbasierte Methode

Olga Artemenko, Margaryta Shramko



## Vector Space Modell

$$\cos \theta_i = \frac{\vec{d} \cdot \vec{f}_i}{|\vec{d}| |\vec{f}_i|}$$



(Vgl.: J. M. Prager 1999:6)

- $d$  - Dokumentvektor
- $f$  - Sprachvektor
- Vektorenwerte sind inverse Dokumenthäufigkeiten von Tri-Grammen und Wörtern

Olga Artemenko, Margaryta Shramko



## Ad Hoc Ranking („out of place“)

	Sprachmodell	Dokumentmodell	Out of Place
most frequent	TH	TH	0
	ER	ING	3
	ON	ON	0
	LE	ER	2
	ING	AND	1
	AND	ED	No match=max
least frequent	...	...	...

Summe = Distanzwert

(Vgl.: W. B. Cavnar & J. M. Trenkle 1994: 6)

Olga Artemenko, Margaryta Shramko



## Bayes'sche Entscheidungsregel

- Der Bayes'sche Satz bezogen auf die Sprachidentifikation:

$$p(\text{Sprache}_A | \text{Dokument}_1) = \frac{p(\text{Sprache}_A) p(\text{Dokument}_1 | \text{Sprache}_A)}{p(\text{Dokument}_1)}$$

- Berechnung von Übergangswahrscheinlichkeiten

$$p(w_{k+1} | w_1 \dots w_k | A)$$

$$p(\text{der} | \text{de} | \text{Deutsch})$$

Olga Artemenko, Margaryta Shramko



## Wortbasierte Methode

- Wörter aus dem Testdokument werden in Sprachmodellen gesucht
- es wird das Sprachmodell ausgewählt, in dem die höchste Anzahl der Wörter aus dem Testdokument gefunden wird:  
z.B. „*sur la poursuite de la croissance* “  
Französisch - 3  
Spanisch - 2  
Italienisch - 1
- falls *mehrere* Sprachmodelle mit der gleichen Anzahl der gefundenen Wörter → relative Häufigkeiten der Wörter werden summiert → das Sprachmodell mit dem höchsten Wert wird ausgewählt

Olga Artemenko, Margaryta Shramko



## LangIdent: Sprachidentifikation multilingualer Dokumente

Olga Artemenko, Margaryta Shramko



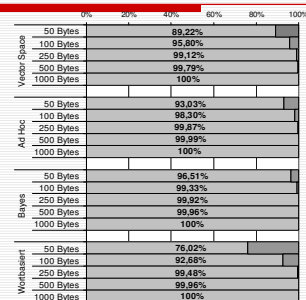
## LangIdent: Evaluierung

- Testkorpus für **monolinguale** Dokumente
  - ca. 200 KBytes Textdateien pro Sprache
  - Länge der Testdokumente:  
25 Zeichen (ca. 50 Bytes)  
50 Zeichen (ca. 100 Bytes)  
125 Zeichen (ca. 250 Bytes)  
250 Zeichen (ca. 500 Bytes)  
500 Zeichen (ca. 1000 Bytes)
- Testkorpus für **multilinguale** Dokumente

Olga Artemenko, Margaryta Shramko



## LangIdent: Testergebnisse (1)



Olga Artemenko, Margaryta Shramko



## LangIdent: Testergebnisse (2)

Dokumentgröße	Klassifikationsmethoden			
	Vector Space	Ad Hoc	Bayes'	Wortbasiert
50 Bytes	10,78%	6,97%	3,49%	23,98%
100 Bytes	4,2%	1,7%	0,67%	7,32%
250 Bytes	0,88%	0,13%	0,08%	0,52%
500 Bytes	0,21%	0,01%	0,04%	0,04%
1000 Bytes	0%	0%	0%	0%

Olga Artemenko, Margaryta Shramko



## LangIdent: Testergebnisse (3)

- Multilinguale Testdokumente

Olga Artemenko, Margaryta Shramko



## LangIdent: Demonstration

---

Olga Artemenko, Margaryta Shramko



## Fragen

---



Olga Artemenko, Margaryta Shramko

