Logarithmus muss

Kritik an einer verbreiteten Methodik der Stilometrie

FELIX GOLCHER

Humboldt-Universität zu Berlin — Institut für deutsche Sprache und Linguistik — Korpuslinguistik

Worum geht's? Was ist Stilometrie?

- Stilometrie ist eine Art der Textklassifikation.
- Es wird dabei aber nicht nach *Topic* oder *Genre* klassifiziert,
- Sondern nach *stilistischen Kriterien*
- Standardbeispiel: Automatische Autorenbestimmung.
- Aussage dieses Posters: Eine verbreitete Praxis ist suboptimal.

Eine typische Arbeit zum Thema

Clement und Sharp 2003:

- 5 Autoren, je 10 Texte.
- n-Gramme bis zu $n = 25$
- Die Häufigkeit der n-Gramme werden in Wahrscheinlichkeiten umgedeutet.
- Mithilfe dieser Wahrscheinlichkeiten wird berechnet, wer am ehesten Autor eines Textes ist.

Viele mögliche Varianten existieren:

- Verwende häufige Elemente wie Funktionsworte, POS-Tags, etc.
- Verwende moderne Maschinenlernalgorithmen wie z.B. *support vector machines*
- **Gemeinsamkeit**: So gut wie immer gehen die Frequenzen linear ein.

eigene Definitionen mit einfachen Beispielen

geteilte Zeichenketten	Text $A =$ `abrax`	Text $B =$ `rabbbi`
`a`	2	1
`ra`	1	1
`b`	1	3

$F_A(s)$ ist die Häufigkeit der Zeichenkette s in Text A.
So ist $F_A(\texttt{a}) = 2$

das Maß	die Formel	in Worten	im Beispiel
linear	$S_l(A,B) = \sum_{\text{alle Zeichenketten } s} F_A(s)F_B(s)$	1. Nimm eine Zeichenkette, die in beiden Texten vorkommt. 2. Multipliziere die beiden Häufigkeiten. 3. Mach dies für alle solchen Zeichenketten. 4. Summiere.	$S_l(A,B) =$ $F_A(\texttt{a})F_B(\texttt{a}) + F_A(\texttt{ra})F_B(\texttt{ra}) + F_A(\texttt{b})F_B(\texttt{b}) =$ $2 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 =$ $2 + 1 + 3 = 6$
links logarithmisch	$S_{llog}(A,B) = \sum_{alle\ s} F_A(s)\left(\log\left(F_B(s)+1\right)\right)$	• $F(B)$ geht logarithmisch ein. • das $+1$ vermeidet mathematische Probleme.	$S_{llog}(A,B) =$ $2 \cdot \log(1+1) + 1 \cdot \log(1+1) + 1 \cdot \log(1+1) =$
rechts logarithmisch	$S_{rlog}(A,B) = \sum_{alle\ s} F_B(s)\left(\log\left(F_A(s)+1\right)\right)$	Genauso, nur vertausche A und B	$S_{rlog}(A,B) =$ $\log(2+1) \cdot 1 + \log(1+1) \cdot 1 + \log(1+1) \cdot 1 =$
logarithmisch	$S_{log}(A,B) = \sum_{alle\ s} \log\left(F_A(s)F_B(s)+1\right)$	Beide Häufigkeiten logarithmisch.	$S_{log}(A,B) =$ $\log(2 \cdot 1 + 1) + \log(1 \cdot 1 + 1) + \log(1 \cdot 1 + 1) =$

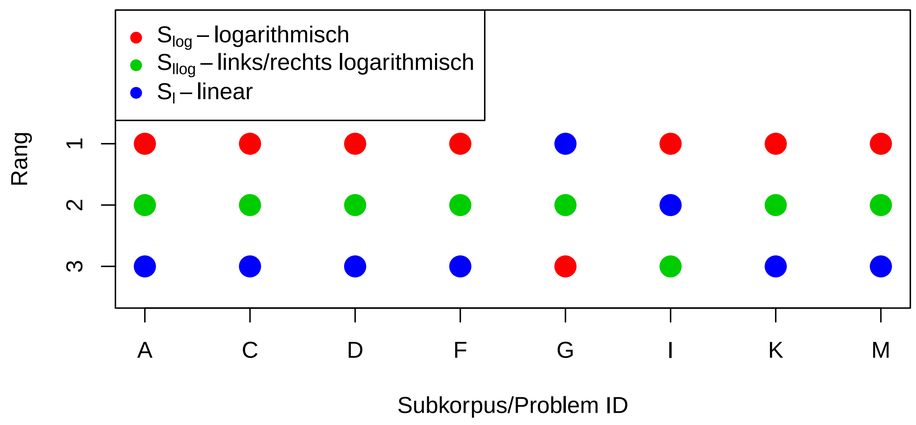

Abbildung 1: Wie schneiden die verschiedenen Maße ab? Korpus aus Juola 2004.
x-Achse: 8 Aufgaben zur Autorenbestimmung.
y-Achse: Relativer Rang der definierten Maße (1 am besten).

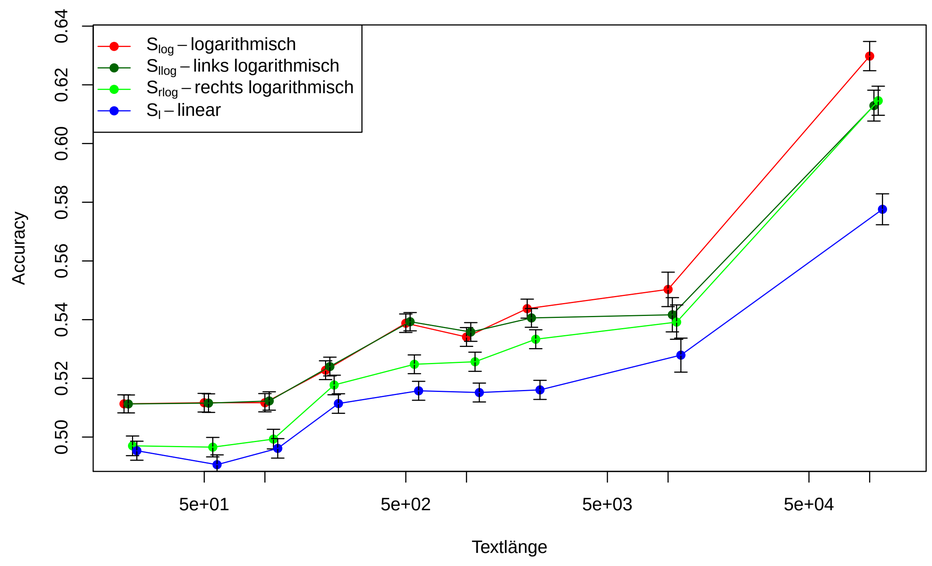

Abbildung 2: Wie hängt die Performanz der verschiedenen Maße von der Trainingstextlänge ab?
Die Länge von A wächst von links nach rechts.
Die Länge von B bleibt konstant.
Korpus aus Baroni und Bernardini 2006

Bottom line:

Je mehr Logarithmus, desto besser.

Literatur

Baroni, Marco und Silvia Bernardini (2006). "A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text". In: *Literary and Linguistic Computing* 21.3 (Sep. 2006), S. 259–274. ISSN: 0268-1145. DOI: 10.1093/llc/fqi039.

Clement, Ross und David Sharp (2003). "Ngram and Bayesian classification of documents for topic and authorship". In: *Literary and Linguistic Computing* 18.4, S. 423–447.

Juola, Patrick (2004). "Ad-hoc Authorship Attribution Competition". In: *Proceedings 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*. Göteborg, Sweden, S. 175–176.