Diplom-Verteidigung

Sprachübergreifendes Retrieval von ähnlichen Dokumenten aus großen Textkollektionen

Katja Schöllner

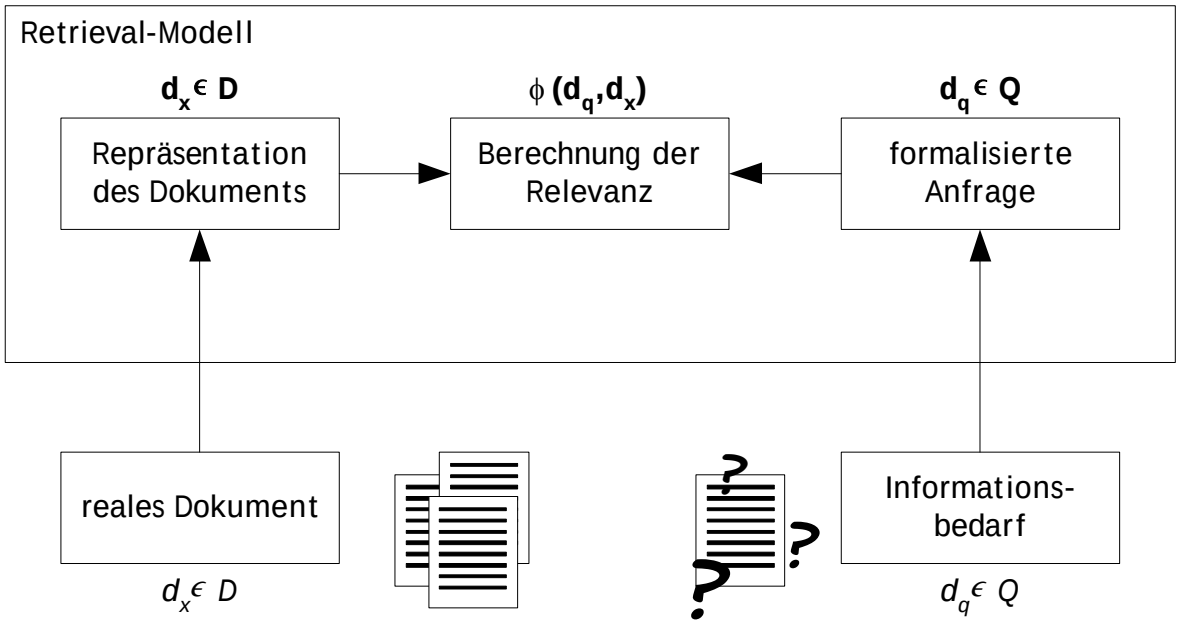
Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Mediensysteme
Web Technology & Information Systems

11.12.2008

- sprachübergreifendes Retrieval ähnlicher Dokumente:
  - gegeben ist ein Dokument d_q in der Sprache L
  - gesucht sind Dokumente d_x in L'
  - Ähnlichkeit zwischen d_q und d_x soll möglichst groß sein
- Anwendungsfall: Plagiatanalyse
  - verdächtiges Dokument d_q: Teile fremder Dokumente übersetzt und ohne Quellenangabe eingefügt
  - gesucht sind Originaldokumente d_x
- in dieser Arbeit: Untersuchung von 3 möglichen Retrieval-Ansätzen

Inhalt

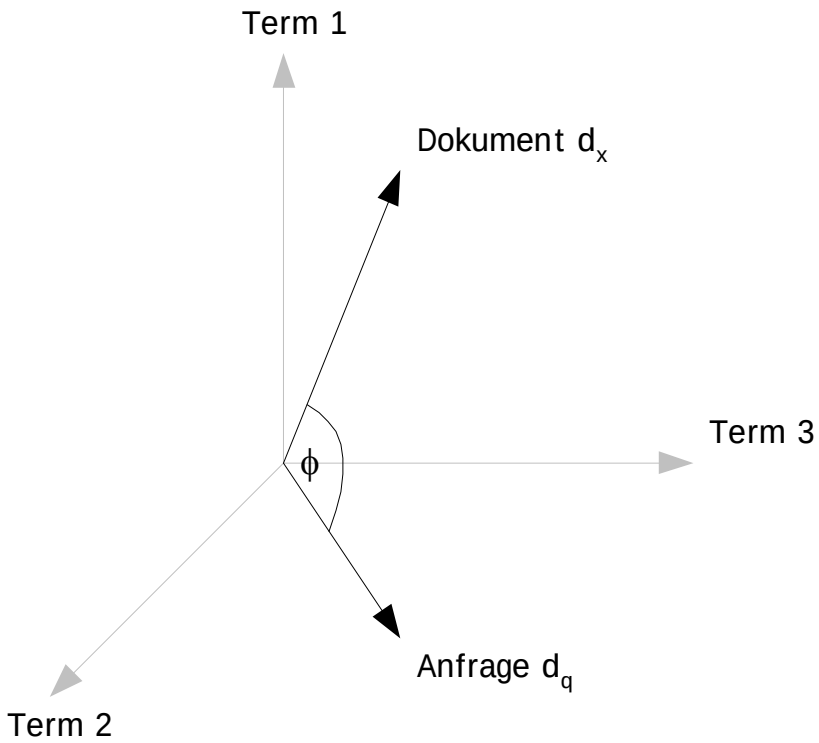
Information Retrieval: Ähnlichkeitssuche



Retrieval-Modelle

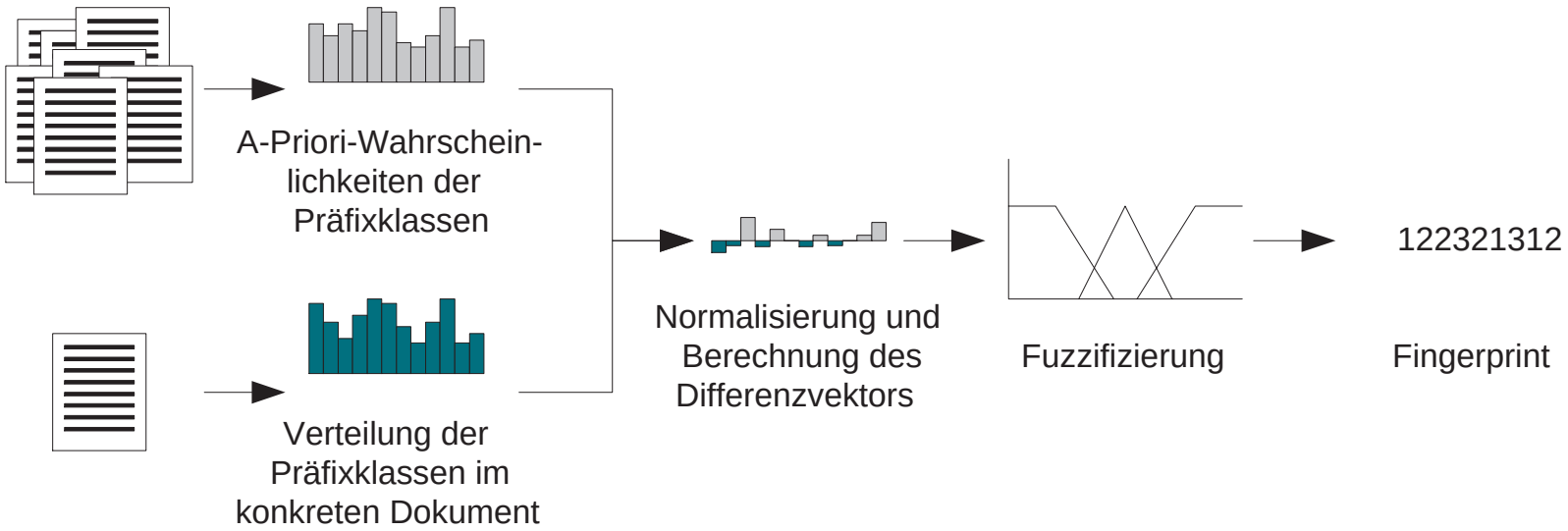
- Vektorraummodell
- Fuzzy-Fingerprinting

Vektorraummodell



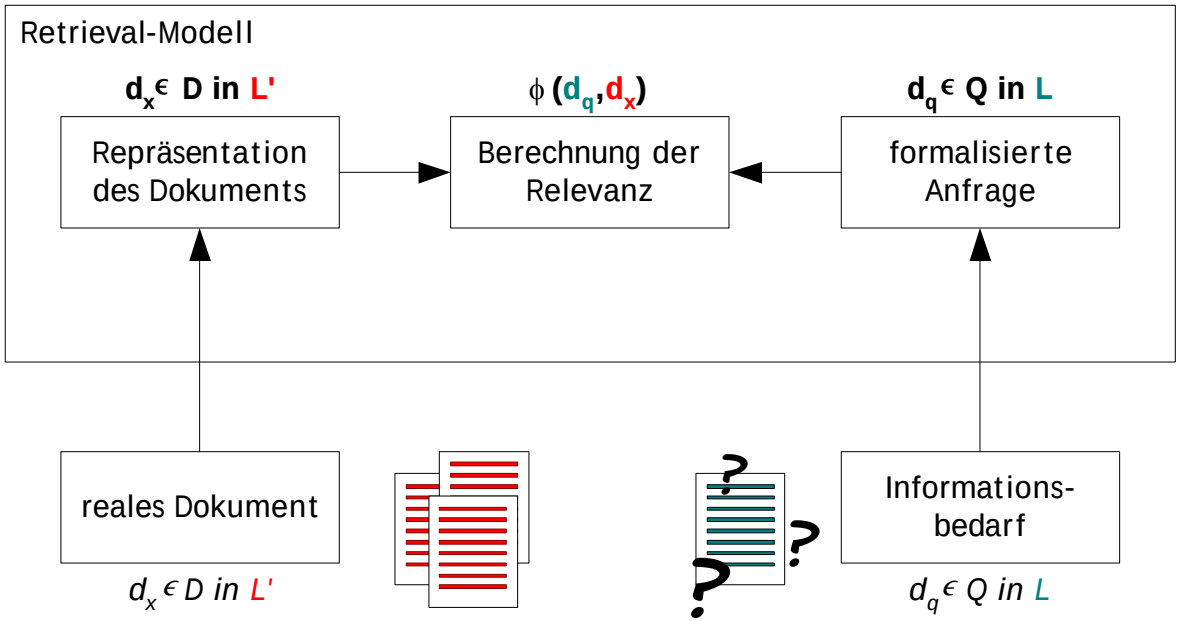
- Kosinus-Ähnlichkeit: $\varphi_{cos}(\mathbf{d_q}, \mathbf{d_x}) = \frac{<\mathbf{d_q}, \mathbf{d_x}>}{\|\mathbf{d_q}\| \|\mathbf{d_x}\|}$

Fuzzy-Fingerprinting



- ähnlichkeitssensitive Hashfunktion: bildet ähnliche Dokumente mit hoher Wahrscheinlichkeit auf denselben Hashwert ab

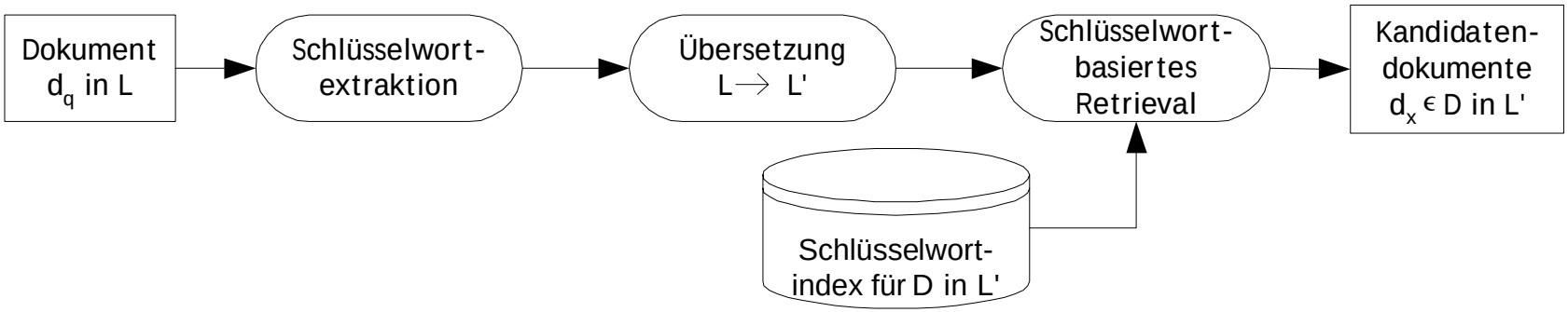
Cross Language Information Retrieval



Retrieval-Modelle

- Schlüsselwortübersetzung $L \longrightarrow L'$ + Vektorraummodell
- Textübersetzung $d_q \longrightarrow d'_q$ + Vektorraummodell
- Textübersetzung $d_q \longrightarrow d'_q$ + Fuzzy-Fingerprinting

Schlüsselwortübersetzung (SÜ)



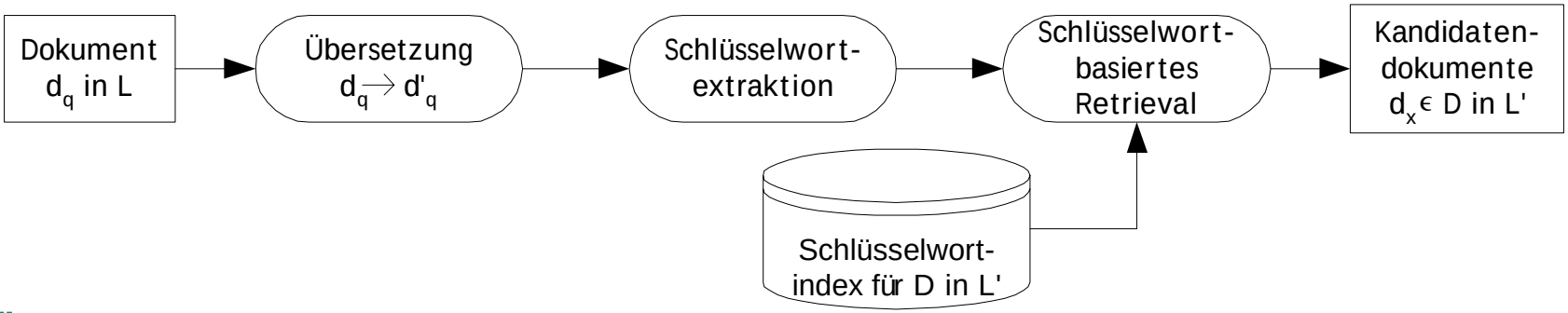
Übersetzung

- wörterbuchbasiertes Verfahren → Ambiguitätsproblem

Schlüsselwortbasiertes Retrieval: Anfragestrukturierung

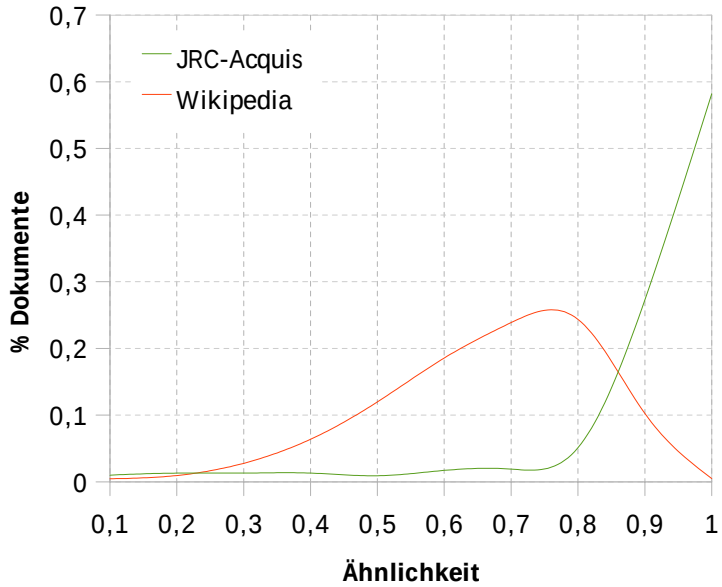
- Verfahren “Close-End-Query”
- nicht eine einzige Anfrage, sondern mehrere Teilanfragen
- Vereinigung der Ergebnismengen

Textübersetzung (TÜ)



Übersetzung

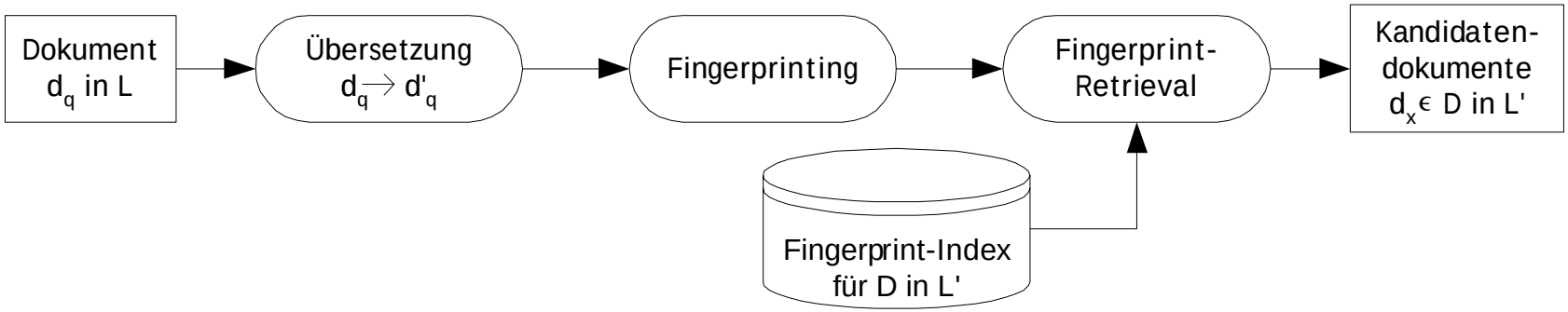
- Verwendung des Übersetzungsservice von Google
- statistisches Lernverfahren, korpusbasiert



Schlüsselwortbasiertes Retrieval: Anfragestrukturierung

- Verfahren “Close-End-Query”

Fuzzy-Fingerprinting (FFP)



Übersetzung

- Verwendung des Übersetzungsservice von Google

Fragestellungen

- Retrieval-Eigenschaften
- schlüsselwortbasierte Verfahren
  - geeignete Anzahl Terme
  - Wortgruppen oder einzelne Wörter
  - Einfluss der Art der Übersetzung
- Eignung von FFP für maschinell übersetzte Texte
- Einfluss der Textlänge
- Laufzeitverhalten

Testkollektionen

- zwei bilinguale Korpora:

Wikipedia

- 138.324 deutsche und englische Artikel, die Sprachpaare bilden
- Artikel behandeln dasselbe Thema
- keine 1-zu-1-Übersetzungen

JRC-Acquis

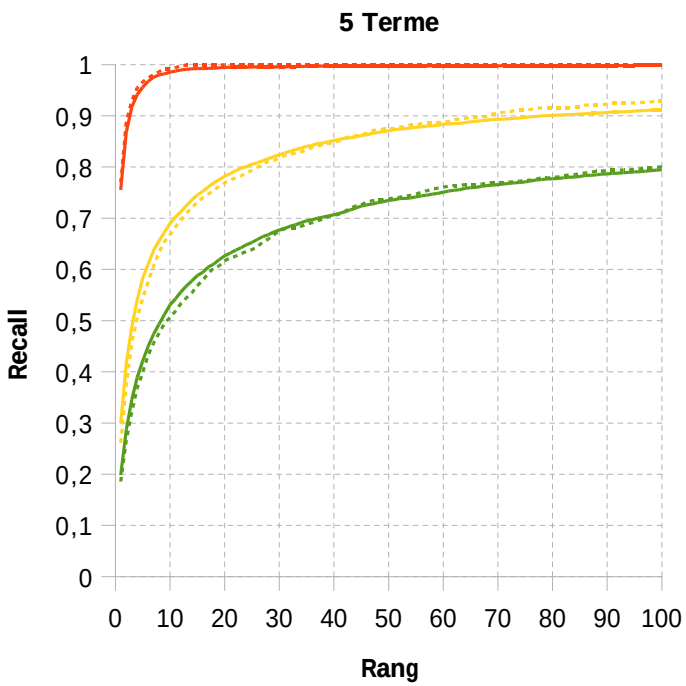
- 20.593 deutsche und englische Dokumente des “Acquis Communautaire” (dt.: gemeinschaftlicher Besitzstand) der EU
- direkte Übersetzungen
- auf Teilsatzebene einander zugeordnet

Ablauf der Experimente

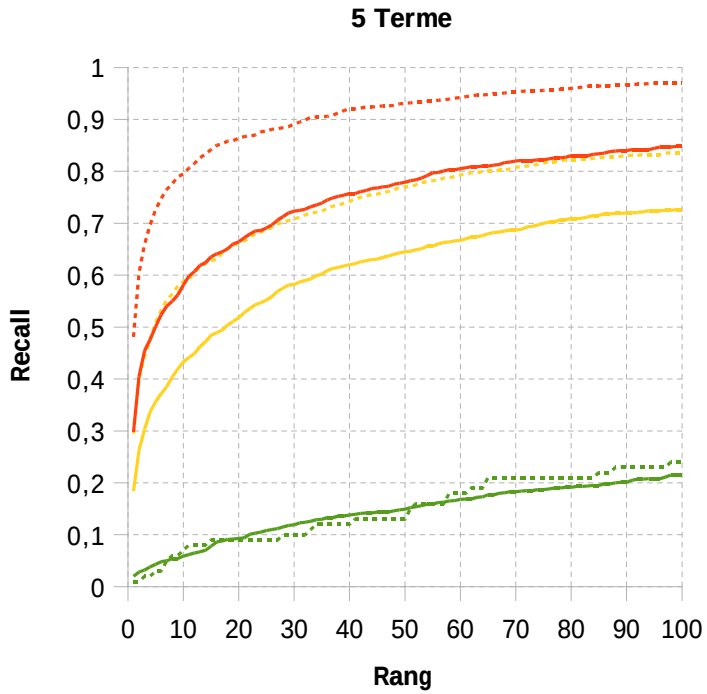
- Auswahl von jeweils 1000 zufälligen deutschen Dokumenten aus Wikipedia- und JRC-Acquis-Kollektion
- SÜ, TÜ und FFP für jedes Dokument
- Bestimmung der Retrieval-Qualität
  - Recall: Anteil der gefundenen relevanten Dokumente an der Menge der relevanten Dokumente
  - Precision: Anteil der relevanten Dokumente an der Menge der gefundenen Dokumente
  - Rang, F-Measure, Mean-Average-Precision

Schlüsselwortübersetzung vs. Textübersetzung

Extraktion von 2, 5 und 10 einzelnen Wörtern bzw. Wortgruppen



Wikipedia

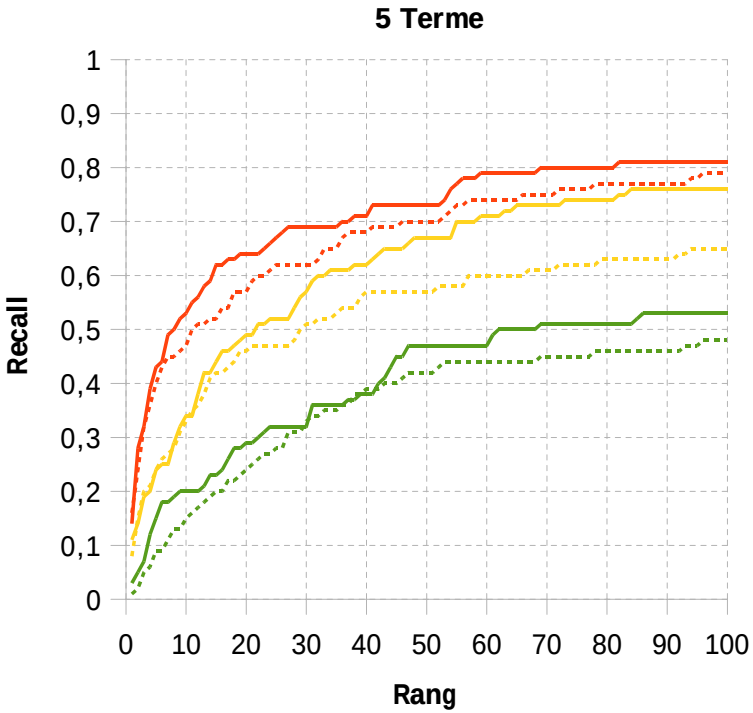


JRC-Acquis

monolingual, 1-Term
monolingual, n-Term
TÜ, 1-Term
TÜ, n-Term
SÜ, 1-Term
SÜ, n-Term

Schlüsselwortübersetzung vs. Textübersetzung

Wikipedia-Webexperiment



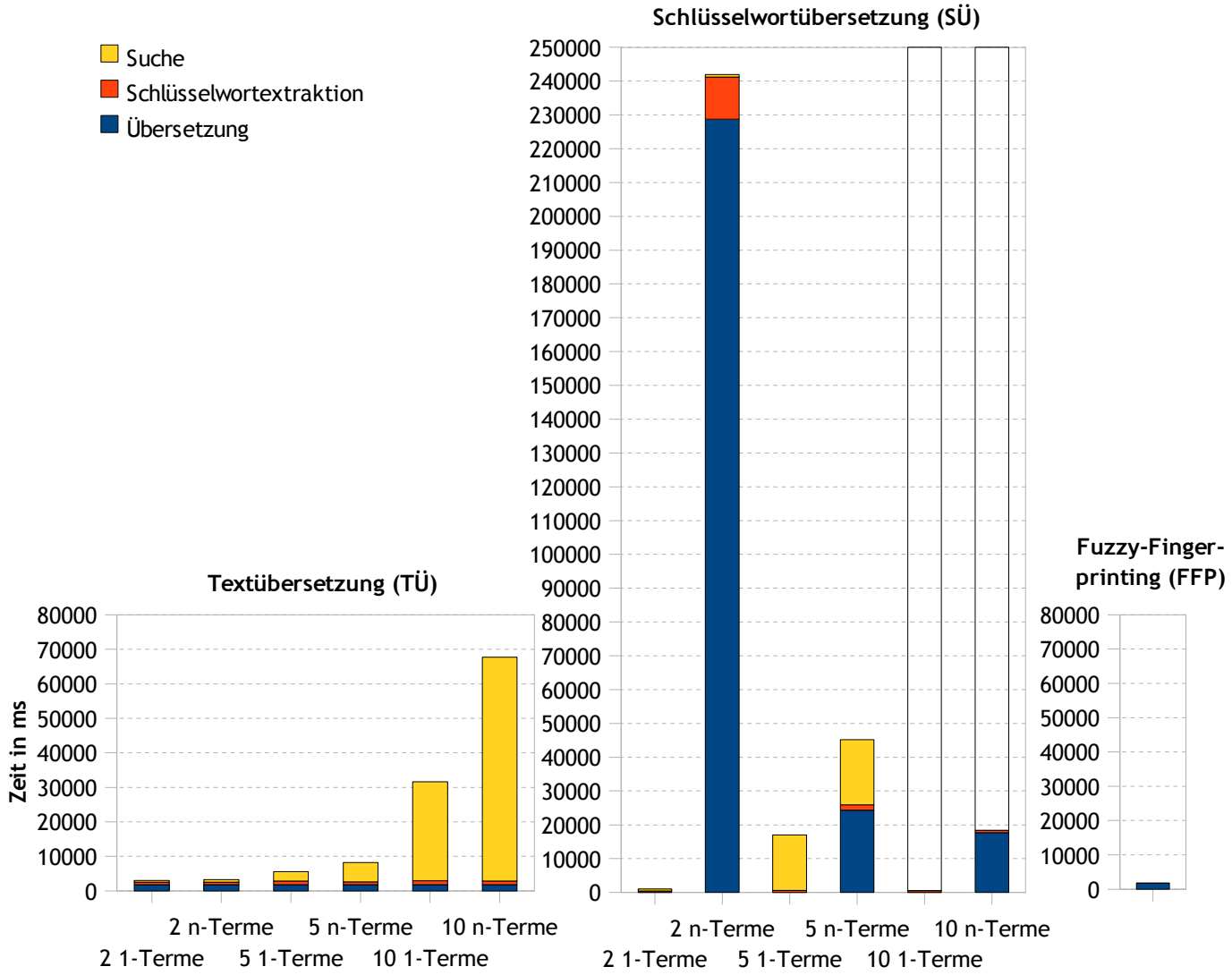
Fuzzy-Fingerprinting

mit verschiedenen Parameterkombinationen:

- Recall: 0.04..0.16 (Wiki), 0.07..0.24 (JRC-Acquis)
- Precision: 0.0004..0.0043 (Wiki), 0.01..0.11 (JRC-Acquis)

Abhängigkeit von der Textlänge

Laufzeitverhalten



Gesamtübersicht

Kollektion		FFP	SÜ	TÜ
Wikipedia	Rang	-	14.81	11.69
	Recall	0.1028	0.8004	0.9285
	Precision	0.0015	0.0080	0.0092
	MAP	-	0.2891	0.3982
	F-Measure	0.0030	0.0158	0.0184
JRC-Acquis	Rang	-	39	12.98
	Recall	0.1710	0.2400	0.8360
	Precision	0.0523	0.0024	0.0084
	MAP	-	0.0249	0.3970
	F-Measure	0.0801	0.0048	0.0166
gesamt	Laufzeit	1770 ms	45202 ms	8157 ms

- SÜ und TÜ: 5 n-Terme
- FFP: zwei Hashfunktionen (1.25 (65) und 1.55 (65))

Fazit

- FFP
  - schnellstes Verfahren
  - geringer Recall, geringe Precision
  - abhängig von Länge des zu untersuchenden Textes
- SÜ
  - scheitert an Ambiguität
  - abhängig von Qualität verschiedener Ressourcen wie Wörterbuch
  - geringerer Recall als TÜ
- TÜ
  - beste Kombination aus Laufzeit und Retrieval-Qualität
  - hoher Recall, Precision über Parameter steuerbar
  - steigende Übersetzungsqualität durch statistisches Lernverfahren

Vielen Dank!