Tokenisierung in Suchmaschinen

::: MASTERSEMINAR SUCHMASCHINEN, CIS, SOMMERSEMESTER 2023 :::

Überblick

- Tokenisierung allgemein
- Sprachen ohne Leerzeichen zwischen Wörtern
- Trainings- und Evaluierungsdaten
- Algorithmen
- Evaluation
- Weitere Anwendungsgebiete

Tokenisierung

Aufteilen eines Textes in indizierbare Token

Recht trivial für westliche und viele andere Sprachen

schwierig für Chinesisch, Japanisch, Thai und einige andere

المطبخ العربي هو أحد المطابخ التي تغطي الوطن العربي من شرقه إلى غربه، وقد تأثر المطبخ العربي بمطابخ بلاد الشام، وتركيا، والأمازيغ، والأقباط وغيرهم من الشعوب في تلك المناطق

教育部即將首次辦國小英語教師英語能力檢核測驗

ยทไาษาภงยสีเห์ะารคเงสัมรกแรปโ
มรกแรปโ#ห์ะารคเงส์#งยสีเ#ยทไาษาภร

Tokenisierung am Beispiel Deutsch

Unter dem Begriff deutsche Küche bzw. deutsche Cuisine fasst man verschiedene regionale Kochstile und kulinarische Spezialitäten in Deutschland zusammen. … Als „typisch deutschen“ Fleischlieferanten betrachtet man das Schwein … ebenso das (hart oder weich gekochte) Frühstücksei. In Deutschland gibt es viele Brot- und Brötchensorten, traditionell vor allem Grau- und Schwarzbrotsorten (u. a. Pumpernickel, Mischbrot, Vollkornbrot usw.). Eine reiche Palette an Mehlspeisen und Knödelgerichten wie beispielsweise Dampfnudel, Germ-, Zwetschgen-, Semmel- und Leberknödel … z. B. Birnen mit Kloß („Birn’ un’ Klütje“).

Tokenisierung allgemein (nur Text): Herausforderungen

- Satzzeichen und andere Sonderzeichen (Zeilenumbruch …)

- Bindestriche *Tor-Chancen Torchancen*

- Trennstriche

- Apostrophe: *Prud'hon/ Prudhon ; we‘ll / we will*

- Akronyme *USA U.S.A. , OK (auch für Oklahoma)*

- Abkürzungen

- Zahlen (38 000; 38000; 38 Tausend; 2.0, 2,0 2,00)

- Datumsangaben

- Maßangaben DIN A4 DinA4….

Tokenisierung und Morphologie

In einigen Sprachen:

- Angehängte Pronomina (*arrivederci) (Klitika),* Artikel, Konjunktionen und andere Wortarten
- Komposita

Tokenisierung in Elasticsearch

Tokenisierung in Elasticsearch (Beispiele)

Analyzer	Character filter	Tokenizer	(token) filter
standard	html_strip	standard	word_delimiter_graph
simple	mapping: replaces characters	whitespace	lowercase
custom		letter: breaks at all non-letters	uppercase
			NGram
		thai	
smartcn (Chinese)		smartcn (Chinese)	

Eine zufällige Liste??

- Chinesisch
- Japanisch
- Thai
- Khmer
- Klassisches Griechisch
- Spätklassisches Latein

คนไทยบริโภคข้าวเป็นอาหารหลัก

發端於春秋戰國時的齊國和魯國

Tokenisierungsalgorithmen am Beispiel eines Thai-Tokenizers

Thai

คนไทยบริโภคข้าวเป็นอาหารหลัก โดยนิยมกัน 2 ชนิดคือ ข้าวเหนียวและข้าวเจ้า คนไทยภาคอีสานและภาคเหนือนิยมกินข้าวเหนียวเป็นหลัก ส่วนคนไทยภาคกลางและภาคใต้นิยมกินข้าวเจ้าเป็นหลัก ประเทศไทยที่ผูกพันกับสายน้ำเป็นหลัก ทำให้อาหารประจำครัวไทยประกอบด้วยปลาเป็นหลัก ทั้ง ปลาย่าง ปลาปิ้ง จิ้มน้ำพริก กินกับผักสดที่หาได้ตามหนองน้ำ ชายป่า หากกินปลาไม่หมดก็สามารถนำมาแปรรูปให้เก็บไว้ได้นาน ๆ ไม่ว่าจะเป็นปลาแห้ง ปลาเค็ม ปลาร้า ปลาเจ่า

Vague indication of content (Google translate):

Thailand's rice is the staple food consumed by two popular types of rice and rice flour. The eastern and northern Thailand is mainly eaten rice. The central and southern part of Thailand is mainly eaten rice. Thailand's bond with the main stream. Make food kitchen Thailand consists of fish, mainly fish, grilled fish, roasted chili eating fresh vegetables at any swamp forest if you eat fish, then he can be processed to be stored for a long time whether the fish dry. salted pickled condiments

Englisch ohne Leerzeichen

thaicuisineismoreaccuratelydescribedasfourregionalcuisinescorrespondingtothefourmain regionsofthecountry:northern,northeastern(orisan),central,andsouthern,eachcuisineshari ngsimilarfoodsorfoodsderivedfromthoseofneighboringcountriesandregions: burmatothenorthwest,thechineseprovinceofyunnanandlaostothenorth,vietnamandcambo diato theeast,indonesiaandmalaysiatothesouthofthailand inadditiontothesefourregionalcuisines thereisalsothethairoyalcuisinewhichcantraceitshistorybacktothecosmopolitanpalacecuisin eoftheayutthayakingdom(1351–1767ce)

Was ist das Problem mit fehlenden Leerzeichen?

Suchmaschinen: Indizieren. Teilstring-Suche ist nicht wirklich eine Alternative (Performanz, Präzision u.a.)

Maschinelle Übersetzung: Arbeitet mit Wörterbüchern

Enitätenextraktion

Definition der Aufgabe (visuell)

… in Worten

thaicuisineismoreaccuratelydescribedasfourregionalcuisinescorresponding tothefourmainregionsofthecountry:northern,northeastern(orisan),central,andsouthern,eachcuisinesharingsimilarfoodsorfoodsderivedfromthoseofneighboringcountriesandregions

→Identifiziere Tokens (Wörter) in einer zusammenhängenden Zeichenkette

Anforderungen:

- Minimiere die Zahl von falschen Token
- Performanz: Das Modul sollte kein Flaschenhals in der Pipeline sein.

Daten für Training und Evaluierung

Repräsentatives Korpus ausreichender Größe

Korrekt segmentierte Text um:

→ Eine Liste korrekter Token zu erhalten

→ Tokenfrequenz

→ Tokensequenzen mit Frequenz/Wahrscheinlichkeit

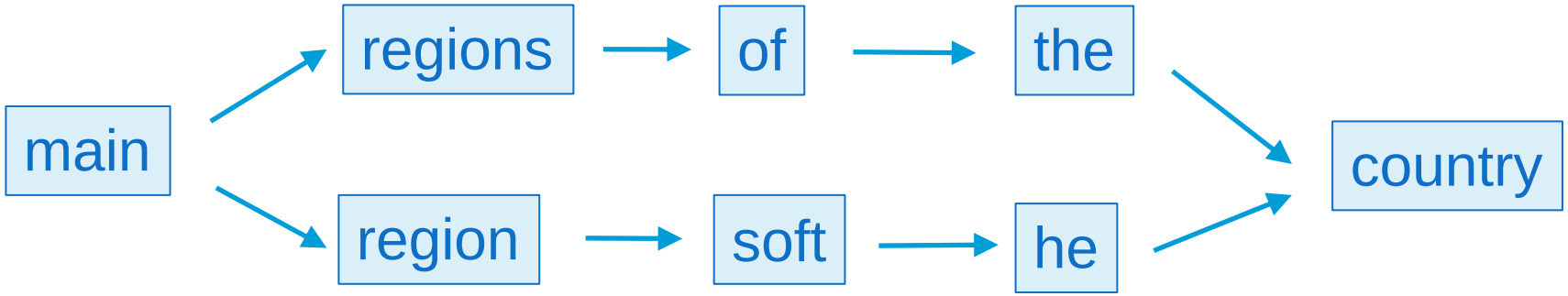
Segmentierungsalgorithmus I

Identifiziere Tokens durch Nachschlagen im Lexikon

+ einfache Statistik: Tokenfrequenz/Wahrscheinlichkeit
+ Sprachmodelle – Wahrscheinlichkeit von Tokensequenzen
  - Bigramme
  - Trigramme

$$P(NGram|L) = \frac{f\big(NGram(L)\big)}{N}$$

Segmentationsalgorithmus II



mainregionsofthecountry

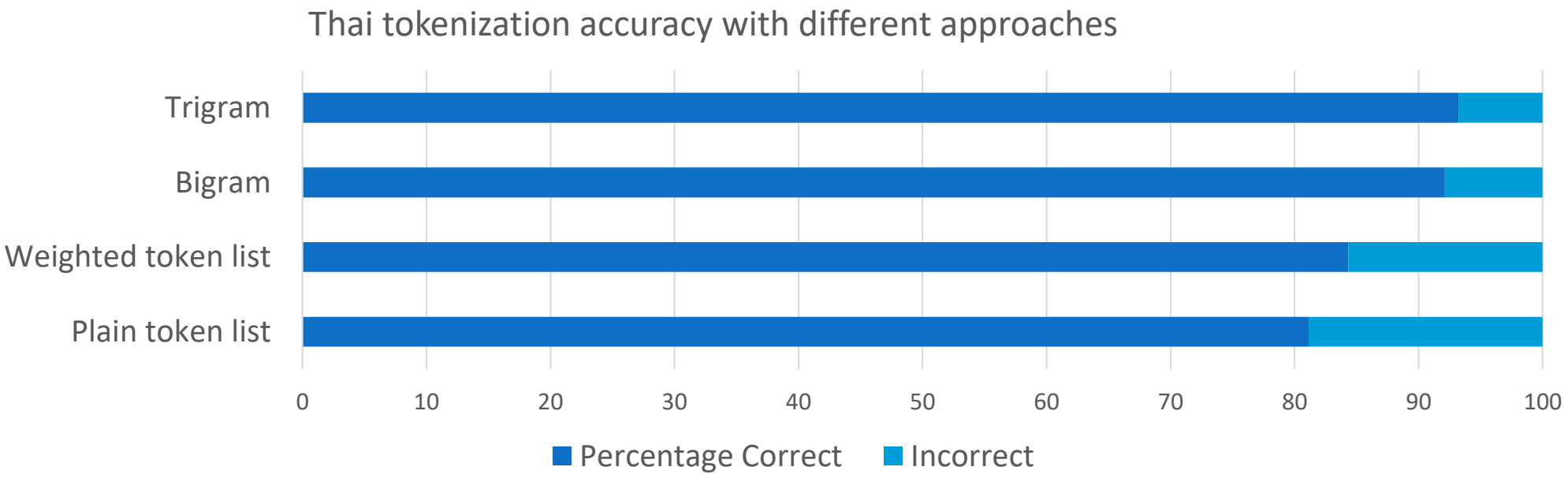
Unbekannte Token

Auch Token, die nicht im Trainingskorpus vorkommen, können valide Token sein. Die Wahrscheinlichkeit von unbekannten Token sollte geringer sein als die von bekannten Token.

Lidstone-Glättung (Beispiel für Einzelwörter, Ähnlich für N-Gramme):

$$P(W|L) = \frac{f\big(W(L)\big) + \boldsymbol{\lambda}}{N + \boldsymbol{\lambda}\, B}$$

Validation


Thai tokenization accuracy with different approaches
Trigram
Bigram
Weighted token list
Plain token list
0
10
20
30
40
50
60
70
80
90
100
Percentage Correct
Incorrect

Performanz

Ausreichende Performanz durch

- Effizientent Lookup (Automaten, Hashes)
- Kompressionstechnologien für Wörterbücher
- Intelligentes Abschneiden des Suchbaumes

Andere Algorithmen

Statistisch

- Mehr Kontext
- Textsortenspezifische Segmentierung
- HMMs, neuronale Netze

Regelbasiert:

- Lokale Grammatiken

Zusätzliche Tokenslisten und Benutzerwörterbuch

Tokenlisten

- Produktnamen
- Eigennamen
- Orte

Benutzerwörterbücher

Token vs. korrekte Segmentierung