Verfahren zur Wortsegmentierung nicht-alphabetischer Schriften

Timo Homburg
Angewandte Computerlinguistik (ACoLi)
Johann Wolfgang Goethe-Universität, Frankfurt am Main

DH Summit 2015
Berlin, Deutschland
3.3.2015

Dieses Poster bietet eine Übersicht über die Ergebnisse und Vorgehensweisen der am Lehrstuhl für Angewandte Computerlinguistik (ACoLi) der Goethe Universität Frankfurt eingereichten Masterthesis "Verfahren zur Wortsegmentierung nicht-alphabetischer Schriften" von Timo Homburg.

Problemstellung: Wortsegmentierung bzw. Worterkennung ist in der Computerlinguistik eine elementare Aufgabe für die maschinelle Analyse eines Textes. Ausgehend von einer Zerlegung eines Textes in bekannte Einheiten, sogenannte Tokens, ist es möglich eine umfassende linguistische Analyse anzustreben. In der Masterthesis werden Algorithmen vorgestellt, die für die Worttrennung im Chinesischen entwickelt wurden und auf der akkadischen Keilschrift ausgetestet. Auf dem Poster werden die einzelnen Typen von Segmentierungsalgorithmen sowie der Evaluationsprozess, die Visualisierung und die Ergebnisse dargestellt.

Der Lehrstuhl für Angewandte Computerlinguistik (ACoLi) an der Goethe Universität Frankfurt am Main wurde im Januar 2013 gegründet um die Aktivitäten der Digital Humanities im Bereich Natural Language Processing (NLP) zu untersützen. Aufbauend auf vorheriger Forschung von Juniorprofessor Dr. Christian Chiarcos entwickelt ACoLi technische Infrastrukturen zur Analyse, zum Aufbau und zur automatischen Verarbeitung von linguistischen Daten und unterhält Infrastrukturen zur Speicherung, Abfrage und Visualisierung von linguistischen Experimenten. Ein Interessensgebiet der Forschung ist die Analyse von außereuropäischen und historischen Sprachen, welche mit diesem Poster aufgegriffen werden.

Segmentierung:

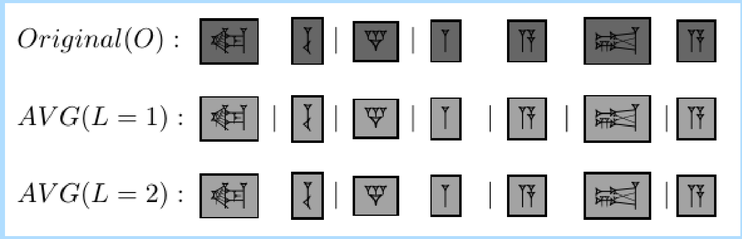
Satz: 这是一张有趣的海报

Worte: 这是 / 一张 / 有趣 / 的 / 海报

(Zhèshì yìzhāng yǒuqù de hǎibào)
(Das ist ein interessantes Poster)

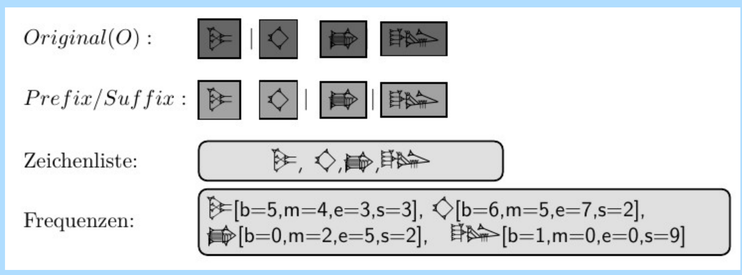
Regelbasierte Verfahren und Baseline

Baseline Verfahren:
- Verfahren mit eingeschränkter Funktionalität
- Referenzpunkt zur Evaluation fortgeschrittener Methoden



Beispiel Mittlere Wortlänge: Mittlere Wortlänge des Korpus als Baselinemethode der Klassifikation

Regelbasierte Verfahren:
- Händisch entwickelte Regeln werden verwendet
- Scoringsysteme auf Zeichenbasis werden verwendet
- Scoringsysteme gewichten Zeichen(folgen) auf Segmentierbarkeit



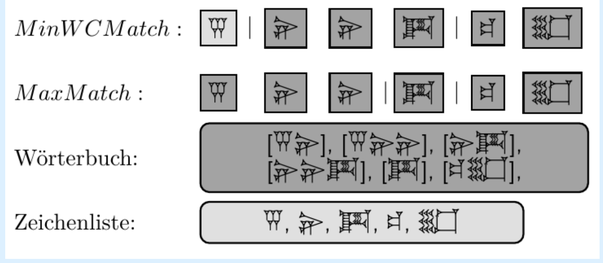
Prefix/Suffix Segmentierung: Klassifiziere mittels Frequenzen auf Zeichenebene Beginn/Mittel/End und Einzelzeichens

Wörterbuchbasierte Verfahren

Idee:
- Leite von bekannten Worten eine Segmentierung her
- Arbeite auf Grundlage von ggf. aus Korpora generierten linguistisch angereicherten Wörterbüchern
- Linguistische Anreicherungen: Wortfrequenzen und Zeichenfrequenzen pro Position, Vorhergehende/Nachfolgende Worte, Zeichenvarianz und Wortkategorien

.

Beispiele:
- **MaxMatch:** Versucht von Beginn zu Ende des Textes jeweils das längste Wort abzutrennen
- **MinWCMatch:** Minimiere die Anzahl der Worte im gegebenen Text
- **MaxProbMatch:** Maximiere die Gesamtwahrscheinlichkeit aller Worte im jeweiligen Text



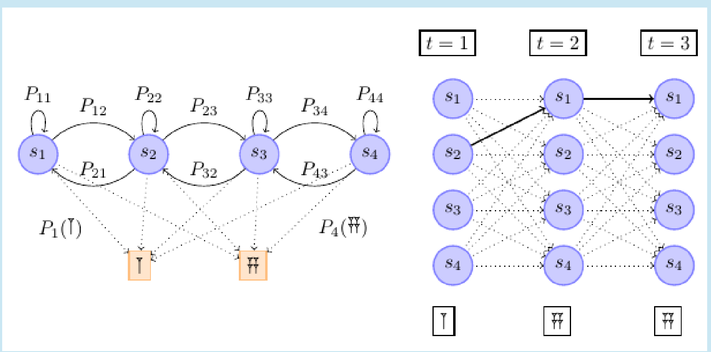
Überwachtes maschinelles Lernen

Idee:
- Annotiere eine Klasse zu jedem Zeichen des Textes
- Klassifiziere auf Grundlage von FeatureSets aus dem Chinesischen
- Verwende im Chinesischen erfolgreiche Verfahren des maschinellen Lernens

Verwendete Verfahren des maschinellen Lernens:
- Clusteringverfahren (kNN,kMeans)
- Entscheidungsbaumverfahren (C45)
- Statistische Verfahren (NaiveBayes,MaxEnt)
- Sequenzlabeling (HMM,CRF)
- Perzeptronenlernen

我 妈 妈 是 中 国 人
S B E S B I E

Beispielannotation der Klassen:
Begin, Intermediate, End, Single Character

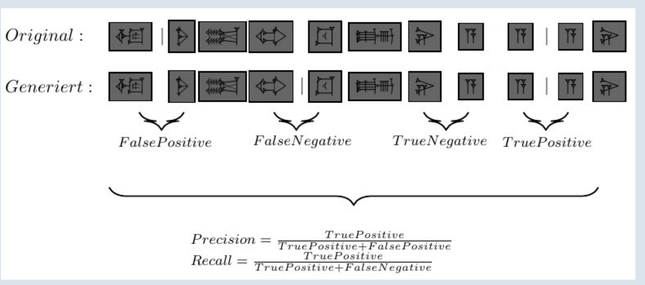


Beispielklassifikation mit HMM*:
Abbildung eines Hidden Markov Models für die Klassifikation mit den Observationen zweier Keilschriftzeichen
*Grafik modifiziert von https://gist.github.com/mblondel/472540

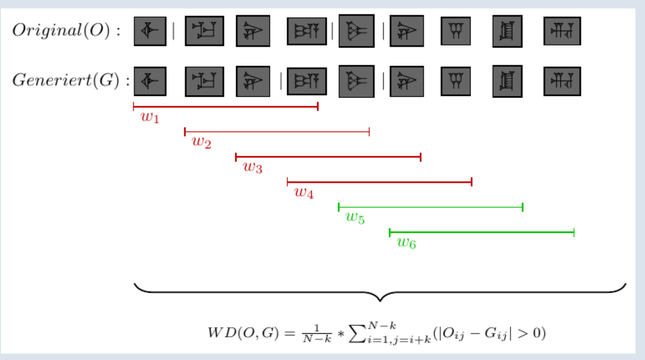
Evaluation

Metriken:
- **Characterbasierte Metriken** Analysiere die Richtigkeit der jeweiligen Entscheidung pro Zeichen
- **Wortbasierte Metriken** Wie viele Worte wurden korrekt segmentiert?
- **Wortfenstermetriken** Wie viele Wortfenster wurden richtig bzw. Falsch segmentiert?
- **Editierdistanzmetriken** Wie groß ist die Editierdistanz der generierten Wortgrenzen zu den richtigen Wortgrenzen?

Vergleichsmaße:
- **Precision:** Wie viel Prozent der vom Algorithmus gesetzten Grenzen waren korrekt?
- **Recall:** Wie viel Prozent der vom Algorithmus gesetzten Grenzen entsprechen der erwarteten Segmentierung?
- **F-Score:** Gewichtetes harmonisches Mittel aus Precision und Recall



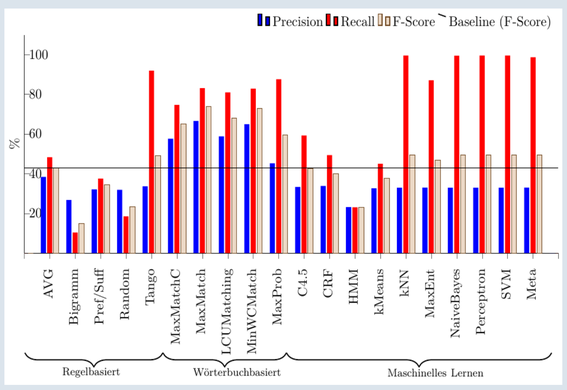
Binäre Evaluation: Entscheide für jeden Character ob eine Wortgrenze danach richtig oder falsch gesetzt bzw. nicht gesetzt wurde.



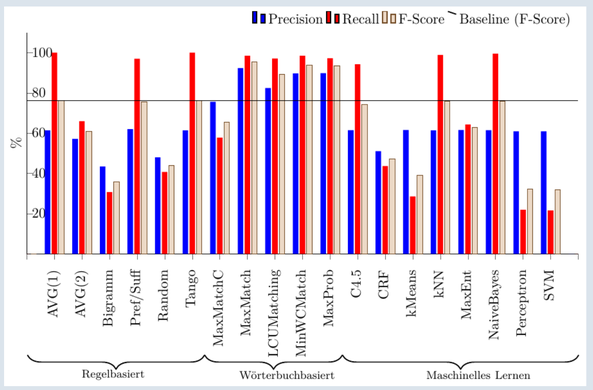
Wortfensterevaluation: Wende ein Scoringsystem für Wortfenster an um zu einer Aussage über die Güte der Segmentierung zu kommen.

Experimente und Ergebnisse

Aufbau der Experimente:
- 80% eines mittelbabylonischen Korpus als Trainingsmenge und 20% des Korpus als Testmenge
- 80% eines chinesischen Korpus als Trainingsmenge und 20% des Korpus als Testmenge



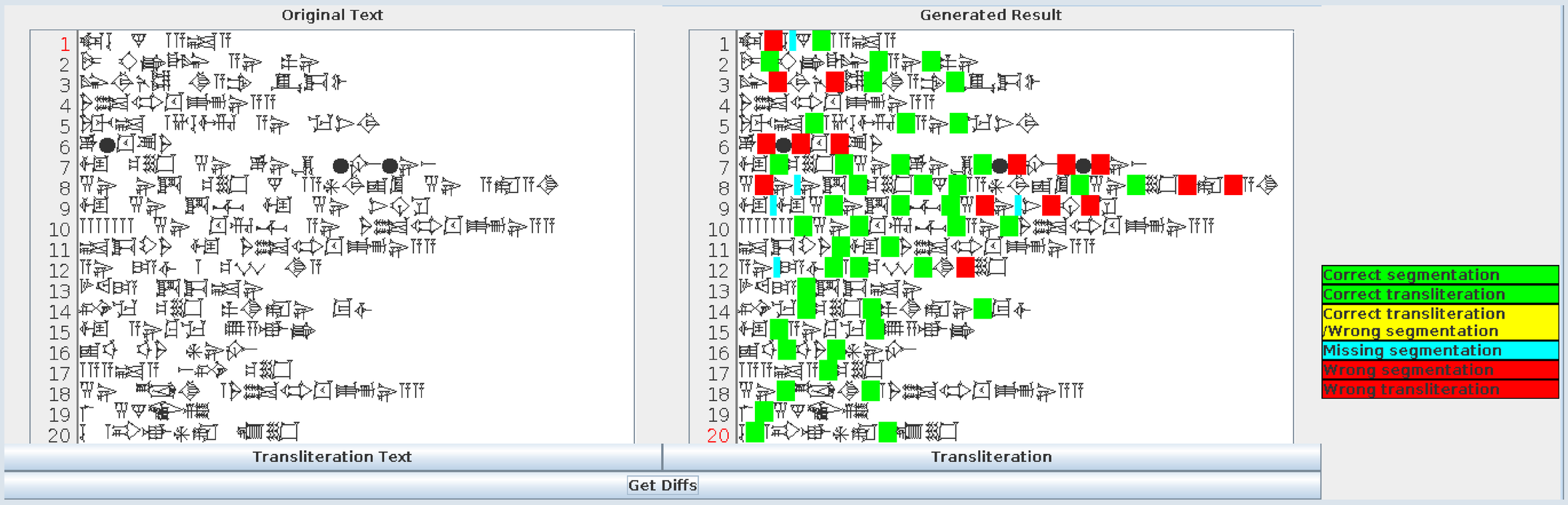
Mittelbabylonisch(Train) auf Mittelbabylonisch(Test)



Chinesisch(Train) auf Chinesisch(Test)

Ergebnisse:
- Baseline Verfahren erreichen im Chinesischen einen F-Score von >70% und im Akkadischen von ca. 45%
- Regelbasierte Verfahren können im Allgemeinen die Baseline nicht überbieten
- Wörterbuchbasierte Verfahren können in der Keilschrift mit einem F-Score von ca. 70% überzeugen
- Überwachtes Maschinelles Lernen erreicht meist mindestens die Baseline, scheitert jedoch an nichtoptimierten FeatureSets und zu geringen Trainingsmengen für die Keilschrift

Visualisierung



Ziele der Visualisierung:
- Ein einfach zu bedienendes Werkzeug für Interessenten aller Fachbereiche
- Klare Strukturierung der Ergebnisse → Anzeige des Ergebnisses mit Signalfarben
- Rot: Falsche Segmentierung
- Grün: Richtige Segmentierung
- Blau: Fehlende Segmentierung

Timo Homburg, B.Sc. (timo.homburg@gmx.de), Goethe Universität Frankfurt Institut für Informatik
Betreuer der Arbeit: Prof. Dr. Christian Chiarcos (chiarcos@em.uni-frankfurt.de) PD Dr. habil. Thomas Richter (richter@em.uni-frankfurt.de)