Named Entity Disambiguation

Software-Projekt WS 2011/2012

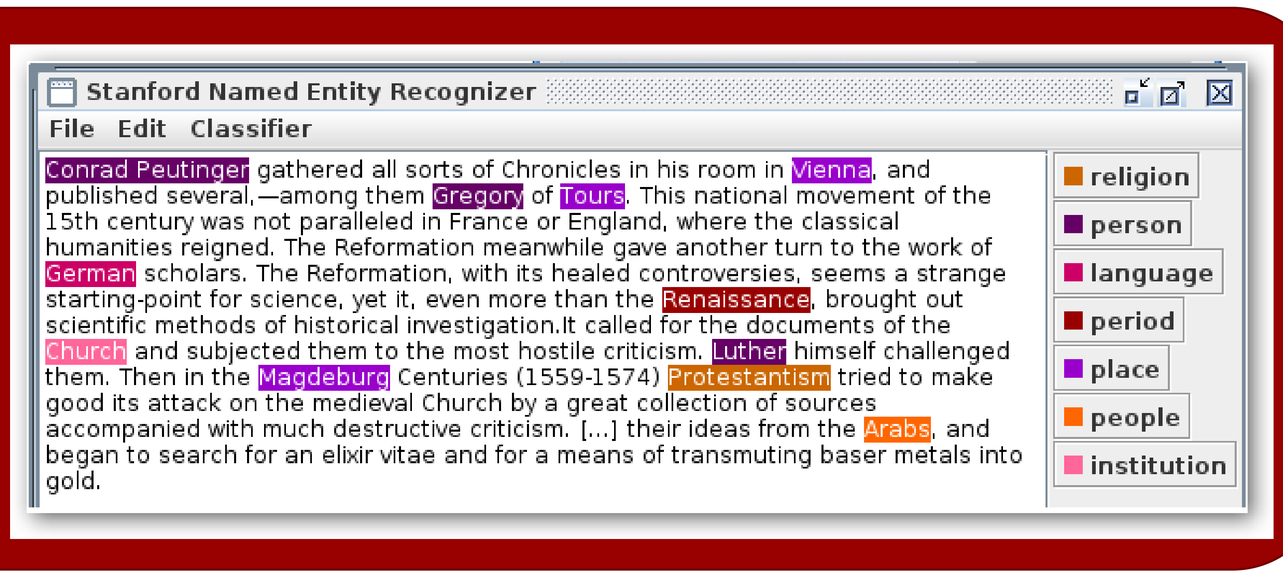
Betreuer: Kathrin Spreyer

Vanessa Reubold
vanessa.reubold@googlemail.com

Department of Computational Linguistics
Heidelberg University, Germany

Ladislav Paterko
linux.ch@seznam.cz

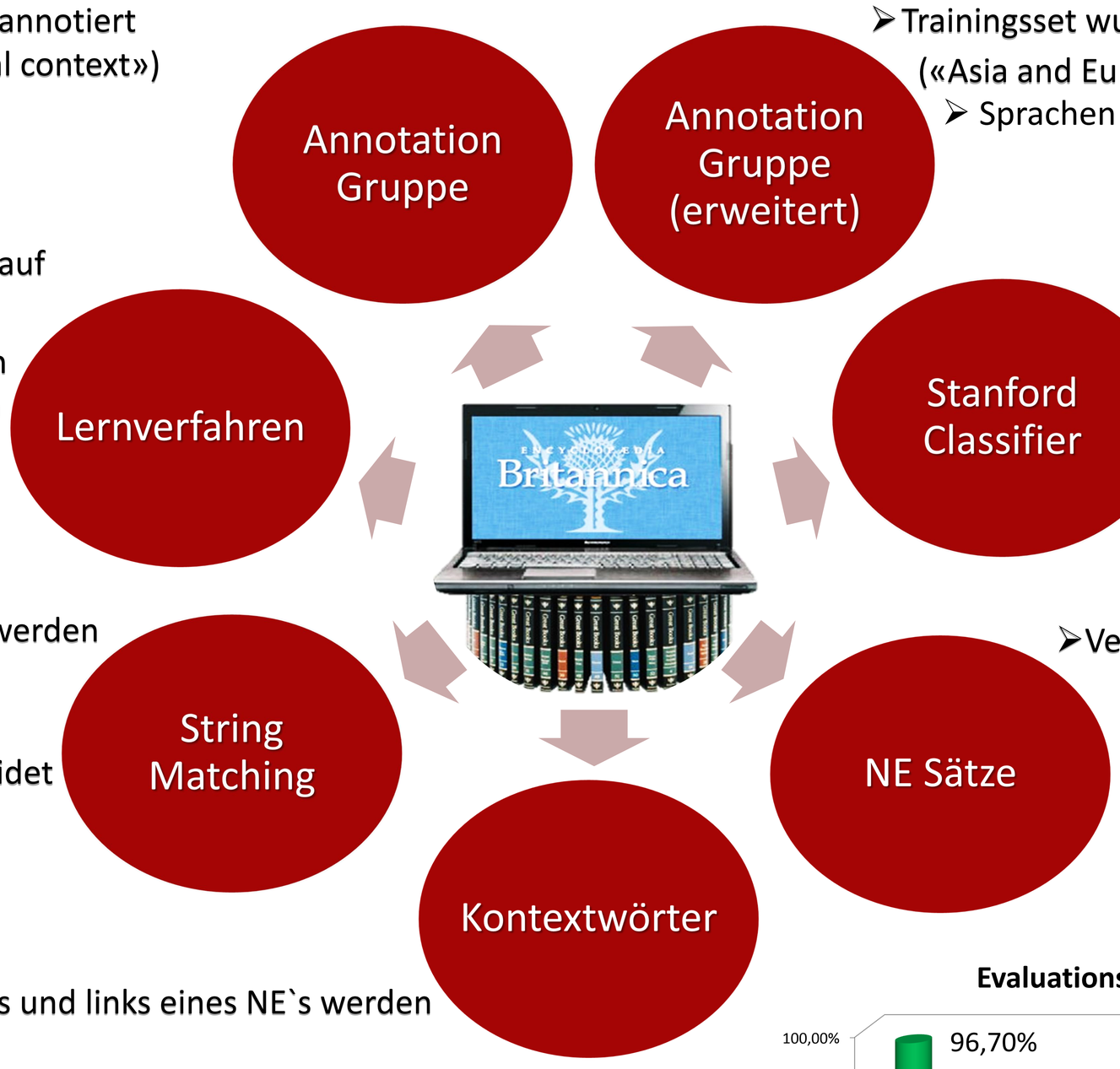
- englischsprachige Texte verarbeiten
- Named Entities extrahieren und einer Kategorie zuordnen
- NE's mit passendem Link aus Wikipedia versehen



- Aufteilung des Textes mit insgesamt 4040 ¦ 5284 NE's
  - Trainingset 2825 ¦ 3699
  - Developmentset 802 ¦ 1057
  - Testset 413 ¦ 528
- Trainingsset bearbeiten
- Classifier trainieren
- testen und evaluieren

Probleme

- Nicht-englischsprachige Wörter
- Sprachgebrauch 19. Jahrhundert
- Fehlerhafte Interpunktion
- "leere" Tags
- Fehler bei manueller Annotation



Annotation Gruppe

- Trainingsset wurde manuell annotiert («Asia and Europe in a global context»)

F-Score: 52,83 %

Annotation Gruppe (erweitert)

- Trainingsset wurde manuell annotiert («Asia and Europe in a global context»)
  - Sprachen und Religionen hinzugefügt

F-Score: 61,70 %

Stanford Classifier

- Verwendung des mitgelieferten Classifiers von Stanford

conll.4.class.di.stsim.crf.ser.gz

F-Score: 35,75 %

NE Sätze

- Verkleinerung des Trainingssets
  → Sentencesplitter von Paul Clough
  → Sätze mit NE`s bleiben erhalten

F-Score: 64,77 %

Kontextwörter

- Die 5 Wörter rechts und links eines NE`s werden extrahiert

Lernverfahren

- Bester Classifier wiederholt auf Trainingset angewendet
- Neu gefundene NE`s werden extrahiert und manuell dem Trainingsset hinzugefügt

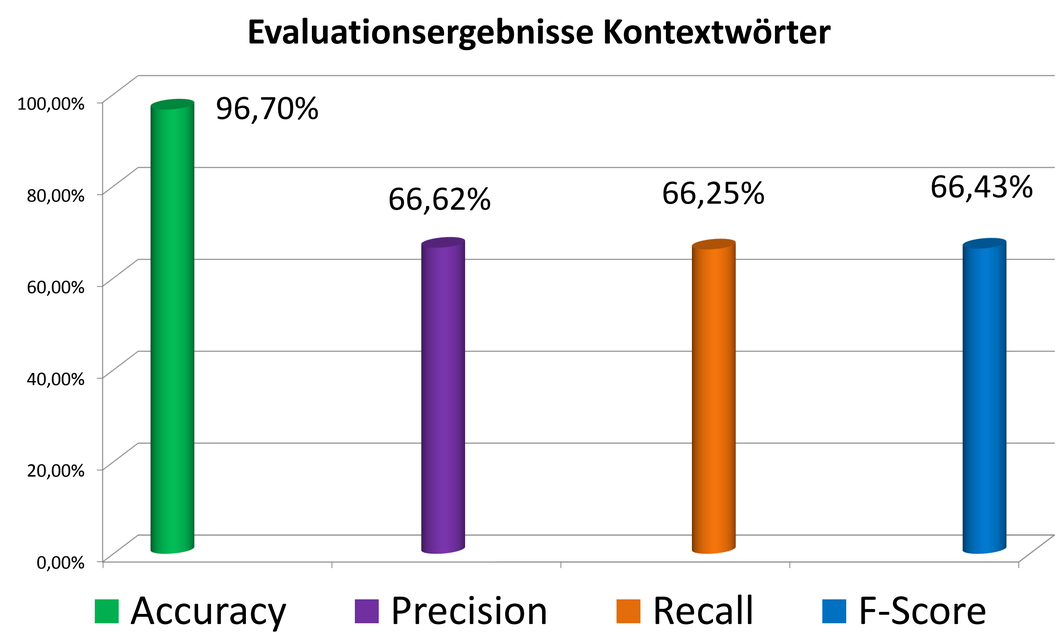
String Matching

- Alle Vorkommen eines NE`s werden als solches annotiert
- Bei unterschiedlichen Annotationsklassen, entscheidet die Häufigkeit

F-Score: 54,51 %

Fehlerbehebung

- Zeichen wie „'s" | „ , " | „ . " | „ ' " | „(" | „)" | „the" werden als „O" annotiert
- Umwandlung der Annotation von Token wie z.B.: „from", „to", „century", welche als „person", „people" oder ähnliches Annotiert wurden in „O" um
- Entfernung von leeren Tags („<name type=„person" />")



Referenzen:
http://nlp.stanford.edu/ner/index.shtml http://ir.shef.ac.uk/doughie/software.html http://www.cs.umass.edu/~mccallum/papers/crf-tutorial.pdf http://www.aclweb.org/anthology-new/P/P05/P05-1045.pdf