Entitätserkennung auf einem Web-Korpus

23. September 2014

Manuel Ruder

Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

Gliederung

UNI FREIBURG

- Einleitung
- Fehleranalyse
- Verbesserung der Entitätserkennung
- Evaluation

Einleitung

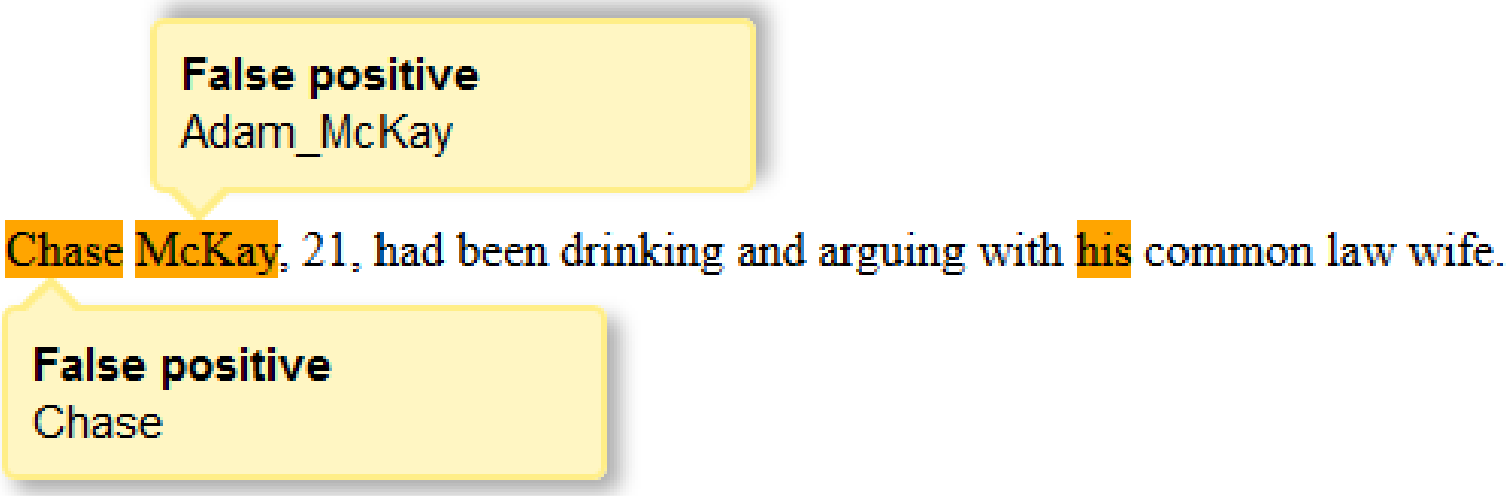
UNI FREIBURG

- Fortsetzung „Semantische Suche auf einem WEB-Korpus“ von Phillip Bausch (2014)

- Dabei noch vorhandene Probleme
  - Extrahieren von Texten aus HTML
  - Erkennung / Verlinkung von Entitäten

Fehleranalyse

- Erkennung / Verlinkung von Entitäten
  - False positives
    - Grenzen von Eigennamen oft falsch erkannt



Fehleranalyse

- Erkennung / Verlinkung von Entitäten
  - False positives
    - Grenzen von Eigennamen oft falsch erkannt
  - Mehrdeutige Namen nicht behandelt
  - Teilweise Aliase nicht erkannt

Beispiel:
*"…**Detroit** outfielder Curtis Granderson said after **the Tigers** lost 4-2…"*

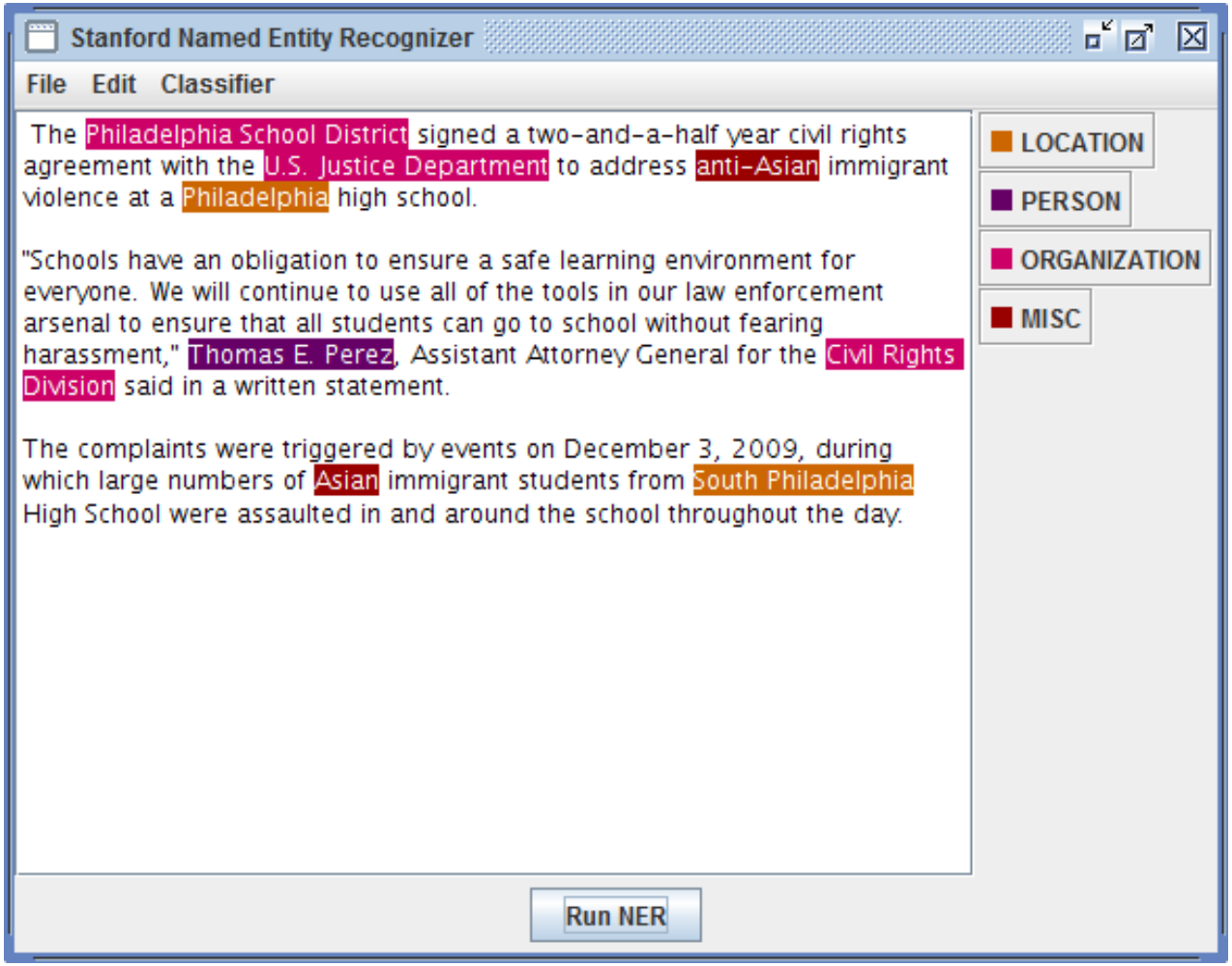
Verbesserung der Entitätserkennung

- Vermeidung von False Positives
  - Unterscheidung zwischen Eigennamen und sonstigen Nomen
    Stanford POS-Tags: NN**P**(S)- vs. NN(S)-Tag
  - Zwei Eigennamen direkt hintereinander
    → Entitäten werden verworfen

oder

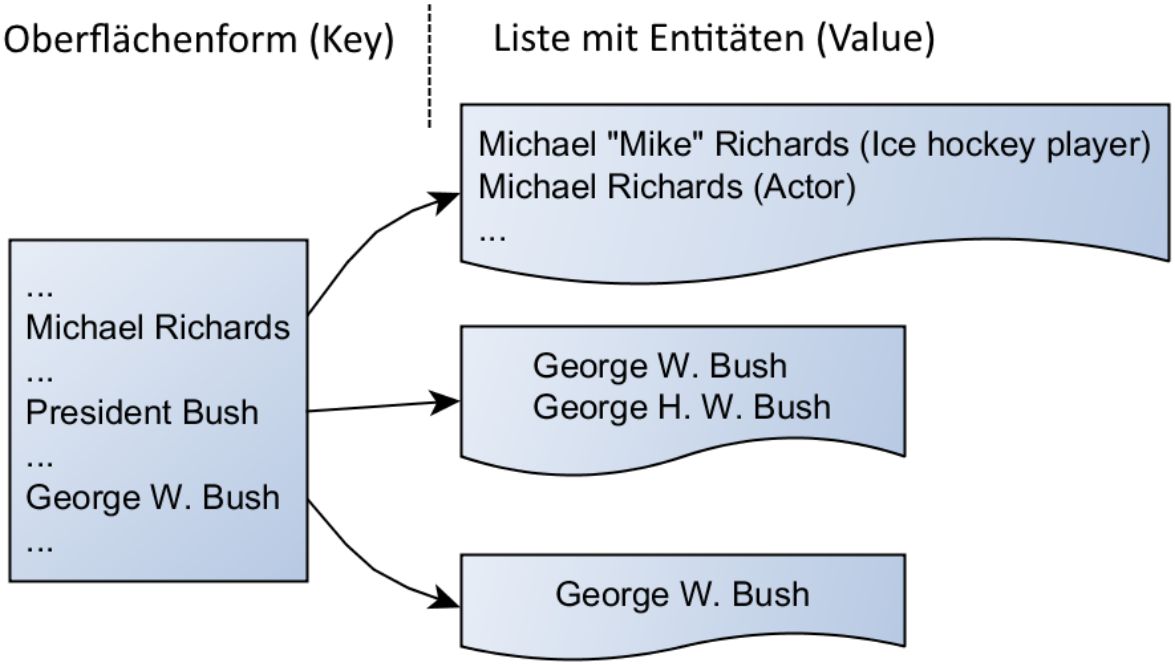
  - Stanford Named Entity Recognizer

Verbesserung der Entitätserkennung



Verbesserung der Entitätserkennung

- Mehrere Entitäten pro Name



➔ Disambiguation notwendig

Verbesserung der Entitätserkennung

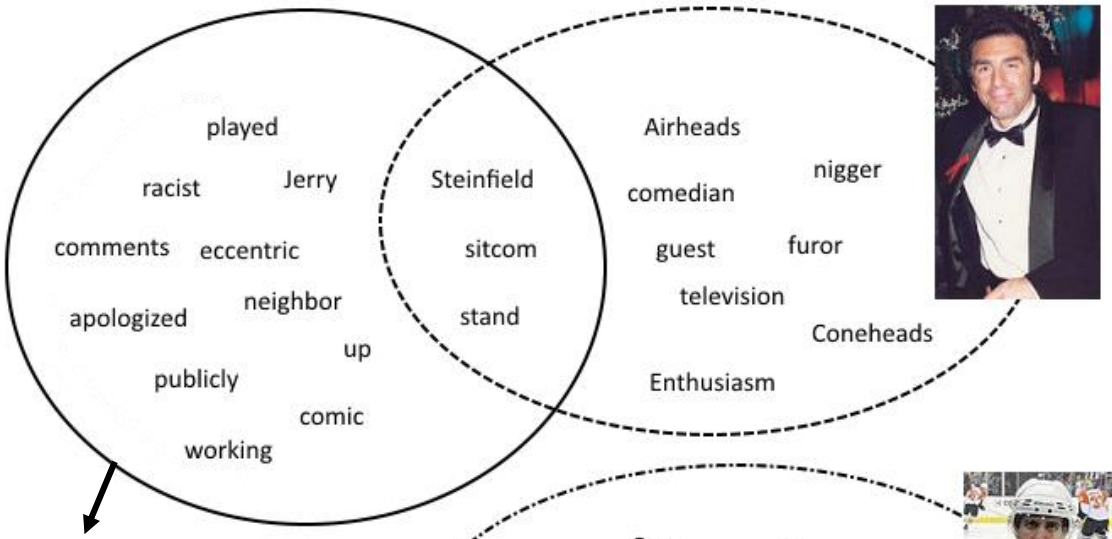
- Dismabiguation von Mehrdeutigkeiten
  - Popularitätswert
  - Kontext

Beipsiel:
***Michael Richards**, who played Jerry Seinfeld's eccentric neighbor Kramer on the sitcom "Seinfeld", publicly apologized for racist comments he made [...] he was working as a stand-up comic.*

Mögliche Enitäten:
- Ice-Hockey Spieler Michael „Mike" Richards (19 451)
- Stand-up Comedian Michael Richards (10 438)

Verbesserung der Entitätserkennung

- Dismabiguation mit Kontext-Vektoren

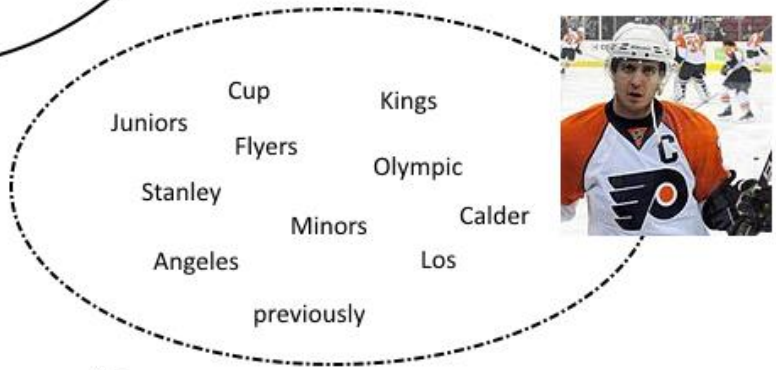


Kontext der Entitäten

- Aus Freebase-Beschreibungen
- Nach TF-IDF gewichtet
- Nur „Top-30“

Kontext des Artikels

- Alle Nomen, Adjektive und Verben



Verbesserung der Entitätserkennung

- Ähnlichkeit zweier Vektoren durch Skalarprodukt

$$\text{similariy}(c_{doc}, c_{entity}) = c_{doc} \cdot c_{entity}$$

- Produkt aus Ähnlichkeit und Popularität

$$\text{score}_{doc}(entity) = (\text{similariy}(c_{doc}, c_{entity}) + \epsilon) \cdot \text{popularity}(entity)$$

- Maximierungsproblem

$$\arg \max_{entity \in E} \text{score}_{doc}(entity)$$

Evaluation

- Testdatensätze

Eigene Auswahl	ERD-50
• **16** Dokumente • **670** Ground Truth Annotationen • **200** Koreferenzen annotiert	• **50** Dokumente • **1100** Ground Truth Annotationen • Keine Koreferenzen annotiert

Evaluation

- Evaluationsmethode

$$Precision = \frac{\#Richtig\ erkannt\ +\ \#Partiell\ richtig\ erkannt}{\#Entitäten}$$

$$Recall = \frac{\#Richtig\ erkannt\ +\ \#Partiell\ richtig\ erkannt}{\#Ground\ Truth\ Annotationen}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Evaluation

UNI FREIBURG

Evaluations-Demo

Evaluation

	Eigene Auswahl			ERD-50		
	Precision	Recall	F1	Precision	Recall	F1
Ideal	*0,814*	*0,743*	*0,777*	*0,771*	*0,637*	*0,698*
Unbeschränkt	0,583	**0,734**	0,650	0,316	**0,678**	0,431
Nil-Vorhersage	**0,643**	0,726	**0,682**	0,464	0,636	**0,536**
Stanford NER	0,640	0,663	0,651	**0,551**	0,514	0,532

Tabelle: Verschiedene Varianten zum Finden von Eigennamen / Vermeidung von False positives

Evaluation

	Eigene Auswahl			ERD-50		
	Precision	Recall	F1	Precision	Recall	F1
Ideal	*0,674*	*0,757*	*0,713*	*0,475*	*0,651*	*0,549*
Popularitätswert	0,612	0,70	0,653	**0,465**	**0,638**	**0,538**
Kontext	**0,643**	**0,726**	**0,682**	0,464	0,636	0,536

Tabelle: Vergleich Disambiguation

Evaluation

	Eigene Auswahl			ERD-50		
	Precision	Recall	F1	Precision	Recall	F1
Ideal	*0,850*	*0,776*	*0,811*	*0,794*	*0,656*	*0,718*
Popularitätswert	0,780	0,712	0,744	**0,778**	**0,643**	**0,704**
Kontext	**0,815**	**0,743**	**0,777**	0,771	0,637	0,698

Tabelle: Vergleich Disambiguation bei idealem Finden von Eigennamen

Evaluation

UNI FREIBURG

Broccoli-Demo