

# Entitätserkennung auf einem Web-Korpus

23. September 2014

Manuel Ruder

Albert-Ludwigs-Universität Freiburg



**UNI  
FREIBURG**



- Einleitung
- Fehleranalyse
- Verbesserung der Entitätserkennung
- Evaluation



- Fortsetzung „Semantische Suche auf einem WEB-Korpus“ von Phillip Bausch (2014)
- Dabei noch vorhandene Probleme
  - Extrahieren von Texten aus HTML
  - Erkennung / Verlinkung von Entitäten

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

**False positive**  
Adam\_McKay

Chase McKay, 21, had been drinking and arguing with his common law wife.

**False positive**  
Chase

- 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...”*

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

*oder*

- Stanford Named Entity Recognizer

# Verbesserung der Entitätserkennung



Stanford Named Entity Recognizer

File Edit Classifier

The Philadelphia School District signed a two-and-a-half year civil rights agreement with the U.S. Justice Department to address anti-Asian immigrant violence at a Philadelphia high school.

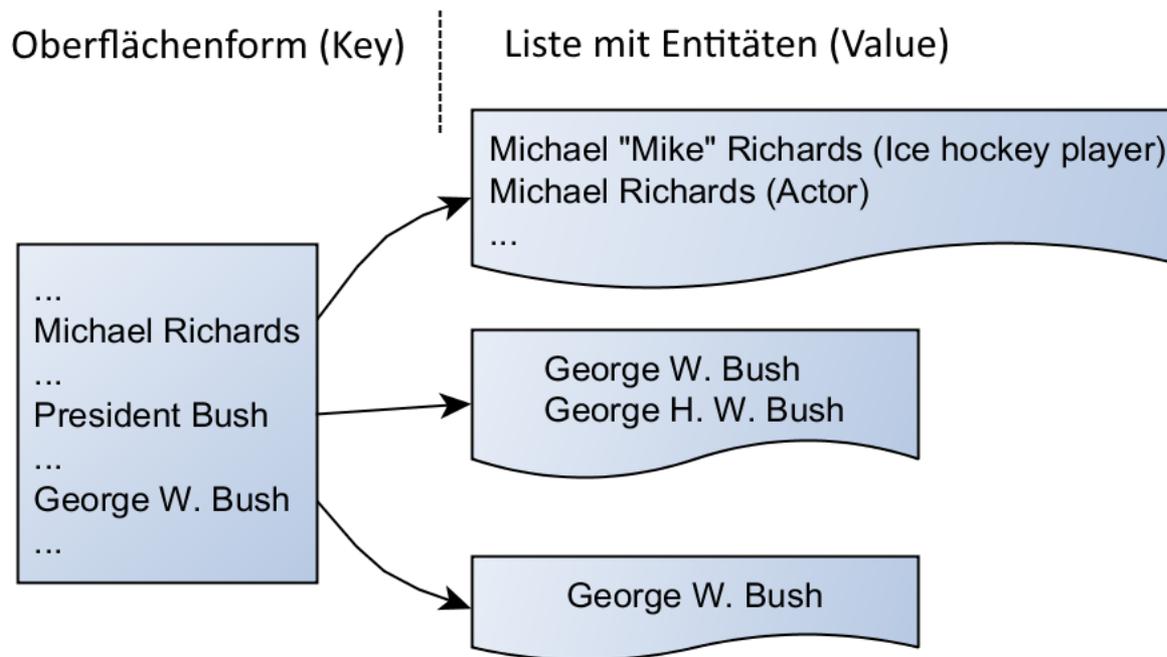
"Schools have an obligation to ensure a safe learning environment for everyone. We will continue to use all of the tools in our law enforcement arsenal to ensure that all students can go to school without fearing harassment," Thomas E. Perez, Assistant Attorney General for the Civil Rights Division said in a written statement.

The complaints were triggered by events on December 3, 2009, during which large numbers of Asian immigrant students from South Philadelphia High School were assaulted in and around the school throughout the day.

LOCATION  
PERSON  
ORGANIZATION  
MISC

Run NER

- Mehrere Entitäten pro Name



➔ Disambiguation notwendig

- Dismabiguation von Mehrdeutigkeiten
  - Popularitätswert
  - Kontext

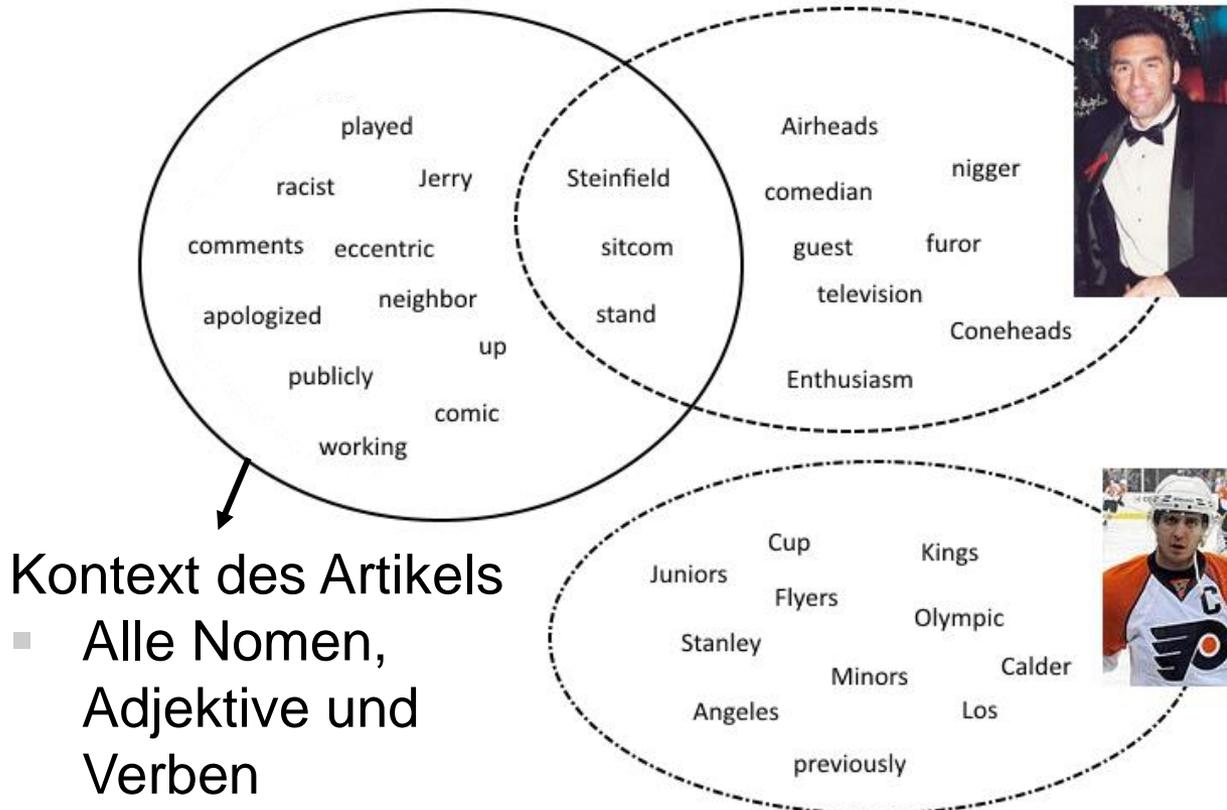
*Beispiel:*

***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 Entitäten:

- Ice-Hockey Spieler Michael „Mike“ Richards (19 451)
- Stand-up Comedian Michael Richards (10 438)

## ■ Dismabiguation mit Kontext-Vektoren



Kontext des Artikels

- Alle Nomen, Adjektive und Verben

Kontext der Entitäten

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

- Ähnlichkeit zweier Vektoren durch Skalarprodukt

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

- Produkt aus Ähnlichkeit und Popularität

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

- Maximierungsproblem

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

- Testdatensätze

## Eigene Auswahl

- **16** Dokumente
- **670** Ground Truth Annotationen
- **200** Koreferenzen annotiert

## ERD-50

- **50** Dokumente
- **1100** Ground Truth Annotationen
- Keine Koreferenzen annotiert

- Evaluationsmethode

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

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

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



## Evaluations-Demo

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

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

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

**Tabelle:** Vergleich Disambiguation

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

**Tabelle:** Vergleich Disambiguation bei idealem Finden von Eigennamen



## Broccoli-Demo