



# Koreferenz mit BART

Spezifikationsvortrag zum Softwareprojekt im Sommersemester  
2014

Julian Baumann, Xenia Kühling, Sebastian Ruder

27. Mai 2014

# Inhalt

- 1 Einführung
- 2 BART
- 3 Stanford Sieves
- 4 Module und Aufgaben
  - Module
  - Aufgaben
- 5 Zeitplan
- 6 Softwarespezifikation
- 7 Quellen

# Problematik: Koreferenz

*John Simon, Chief Financial Officer of Prime Corp since 1986 saw his pay jump 20 percent, to 1.3 million dollar, as the 37-year-old also became the financial service company's president.<sup>1</sup>*

- Unterschiedliche Beschreibungen beziehen sich auf gleiche Entitäten
  - John Simon
  - he
  - the 37-year-old

---

<sup>1</sup>Beispiele von Yannick Versley

# Anwendungen: Information Extraction

*Towards the end of the war, under extreme pressure from the Nazi Party, **Furtwängler** fled to Switzerland. [...] **He** died in 1954 in Ebersteinburg close to Baden-Baden.*

**Q: Wann starb Furtwängler?**

→ Wie kann man Koreferenz auflösen?



# BART

- Beautiful Anaphora Resolution Toolkit
- Entstanden im Projekt  
*Exploiting Lexical and Encyclopedic Resources For Entity Disambiguation* im John Hopkins Summer Workshop 2007
- System für automatische Koreferenzresolution
- Weiterentwicklungen im Rahmen von shared tasks, für verschiedene Sprachen (Italienisch, Chinesisch)



# Wie funktioniert BART?

- Modularer Aufbau:
- Vorverarbeitungsphase
- Extraktion NP- Kandidaten, NP- Merkmale, Kandidatenpaare



# Wie funktioniert BART?

- Resolution mit Soon Algorithmus
- Kandidatenpaare werden paarweise anhand ihrer Merkmale verglichen
- Ergebnisse

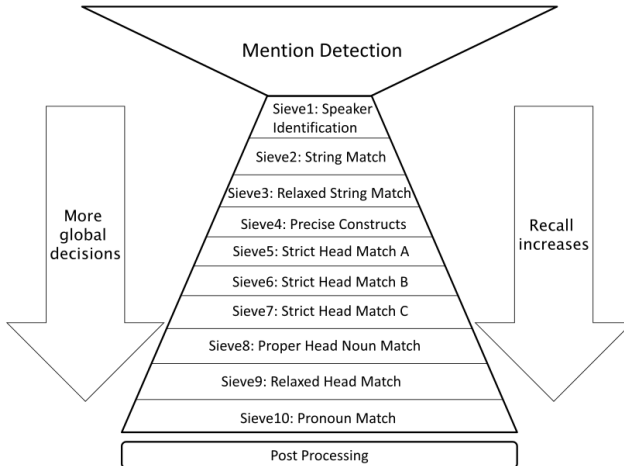


# Aufgabe

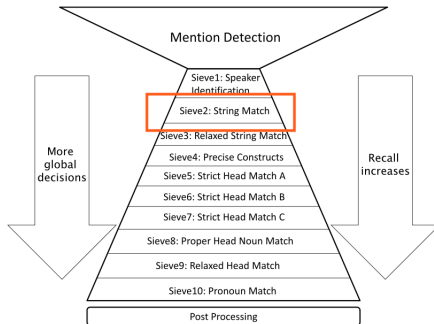
- Koreferenzresolution in BART mit neuem Ansatz:
  - Vorwiegend regelbasiertes System der Stanford-NLP-Gruppe
  - Bestes Ergebnis bei CoNLL-2011 shared task



# Aufbau des Stanford Systems

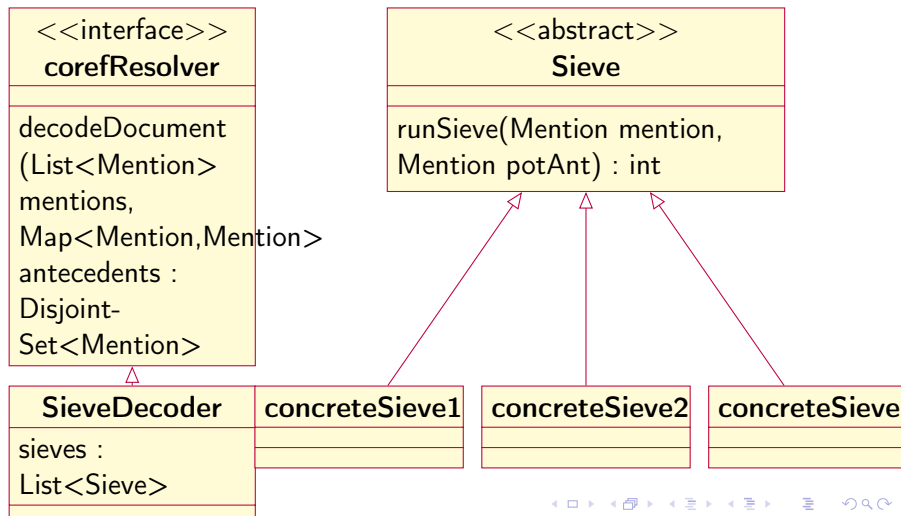


# StringMatch Sieve



[John]<sub>1</sub><sup>1</sup> is [a musician]<sub>2</sub><sup>2</sup>. [He]<sub>3</sub><sup>3</sup> played [a new song]<sub>4</sub><sup>4</sup>.  
 [A girl]<sub>5</sub><sup>5</sup> was listening to [the song]<sub>6</sub><sup>6</sup>.  
 "[It]<sub>7</sub><sup>7</sup> is [[my]<sub>9</sub><sup>1</sup> favorite]<sub>8</sub><sup>8</sup>," [John]<sub>10</sub><sup>1</sup> said to [her]<sub>11</sub><sup>11</sup>.

# Generelle Architektur



# Discourse Entity

## DiscourseEntity

mentions : set<Mention>

nextID : int

discourseID : ID

genders : set<Gender>

numbers : set<G

words : set<String>

heads : set<String>

firstMention : Mention

representativeMention : Mention

DiscourseEntity(Mention m)

mergeEntities(Mention m) : void

getMostRepresentativeMention() : void

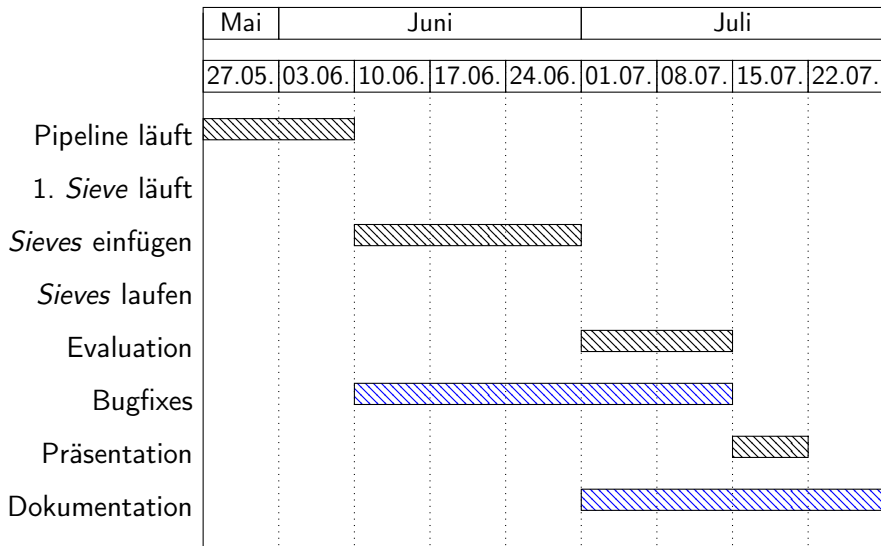
# Aufgabenverteilung

bis 10.06.: Aufteilung der Pipeline

- *DiscourseEntity*: Julian Baumann
- *Sieve & StringMatchSieve*: Xenia Kühling
- *SieveDecoder*: Sebastian Ruder

ab 10.06.: Aufteilung der Sieves

- *RelaxedStringMatchSieve*, *PreciseConstructsSieve*,  
(*SpeakerIdentificationSieve*)
- *StrictHeadMatch[ABC]Sieve*, *RelaxedHeadMatch*
- *ProperHeadNounMatch*, *PronounMatch*





# Softwarespezifikation

- Datenformate
- Interfaces
- Datenstrukturen
- Ressourcen
  - Korpus
- Programmierumgebung
  - Eclipse 4.3.2 mit IvyDE (*dependency management* und EGit und GitHub zur Versionskontrolle



# GitHub, ivyde egit