

Koreferenzresolution 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
- 5 Aufgaben
- 6 Zeitplan
- 7 Softwarespezifikation

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.¹

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

¹Beispiele 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

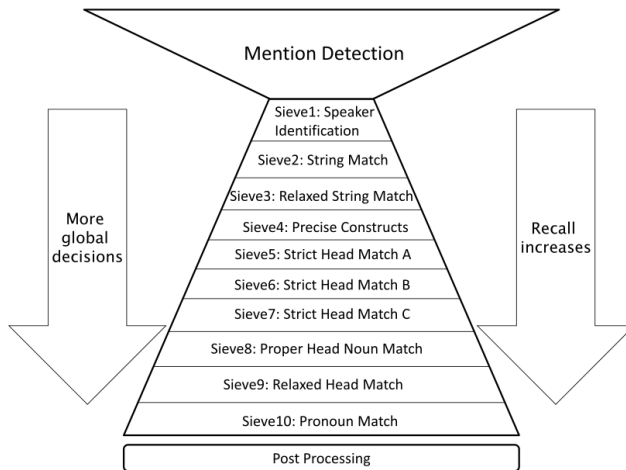
Wie funktioniert BART?

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

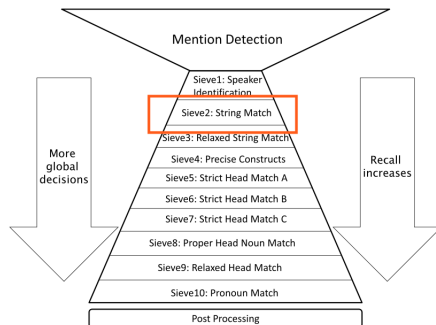
Problemstellung

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

Aufbau des Stanford Systems

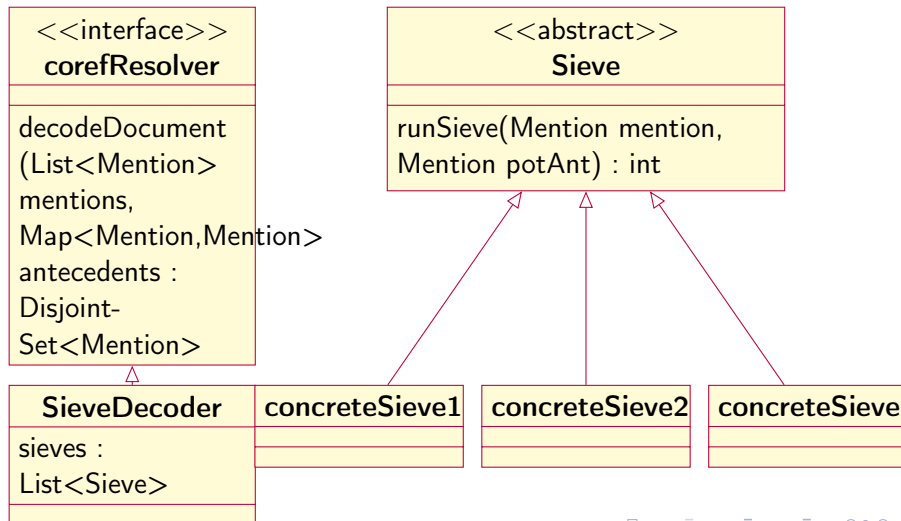


StringMatch Sieve



[John]₁¹ is [a musician]₂². [He]₃³ played [a new song]₄⁴.
 [A girl]₅⁵ was listening to [the song]₆⁶.
 "[It]₇⁷ is [[my]₉¹ favorite]₈⁸," [John]₁₀¹ said to [her]₁₁¹¹.

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

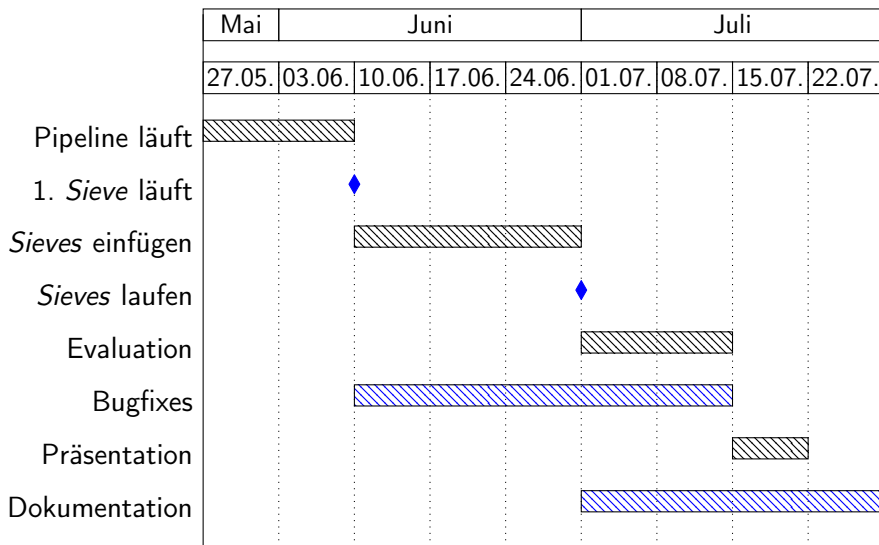
Aufgabenverteilung

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
 - MMAX2, Java, .config
- BART-Version: Klon von Yannicks bitbucket *repository* (<https://bitbucket.org/yannick/bart>); Stand 05.05.14
- Korpora
 - TüBA-D/Z 2008 MMAX2 (Deutsch)
 - Penn Treebank (Englisch)
 - Turin University Treebank/ISST (Italienisch)
- Programmierumgebung
 - Eclipse 4.3.2 mit IvyDE (*dependency management*) sowie EGit und GitHub zur Versionskontrolle