Koreferenzresolution mit BART

Spezifikationsvortrag zum Softwareprojekt im Sommersemester 2014

Julian Baumann, Xenia Kühling, Sebastian Ruder

27. Mai 2014

Inhalt

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

[1] 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

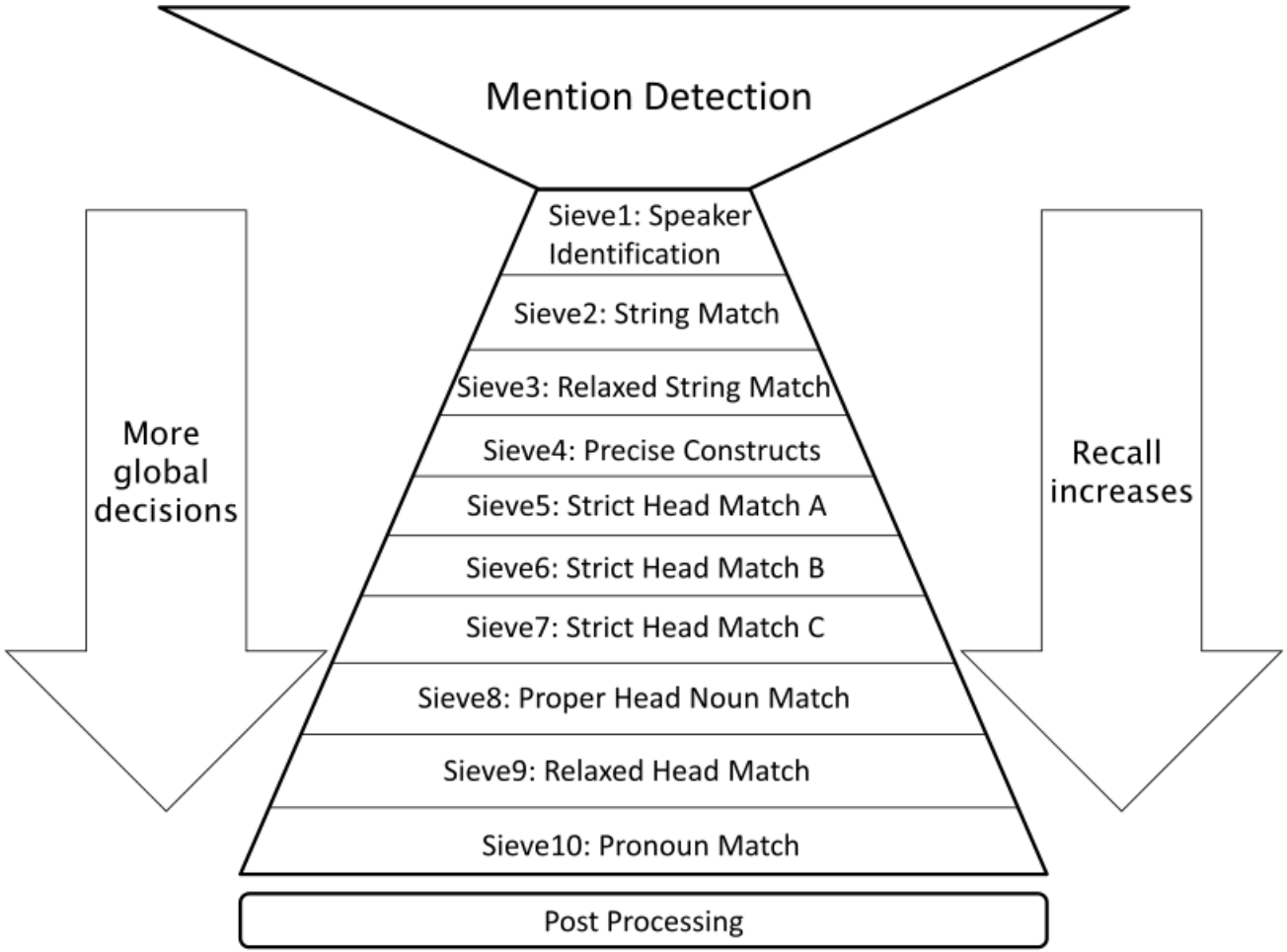
Wie funktioniert BART?

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

Problemstellung

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

Aufbau des Stanford Systems



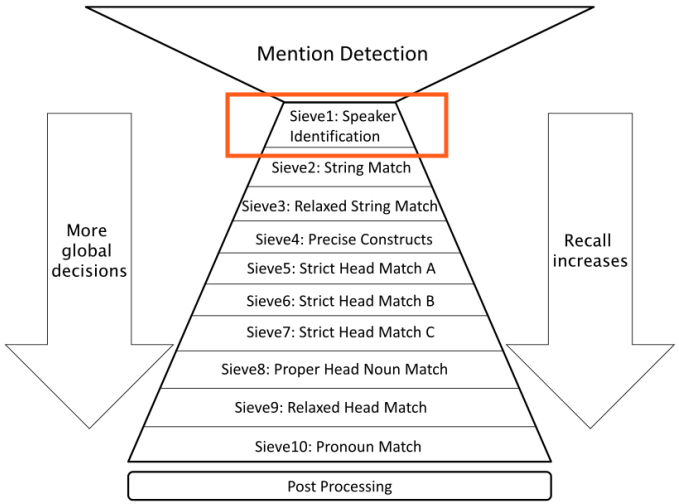
Input + Mention Detection

Beispielsatz:

> *John is a musician. He played a new song. A girl was listening to the song. "It is my favorite," John said to her.*

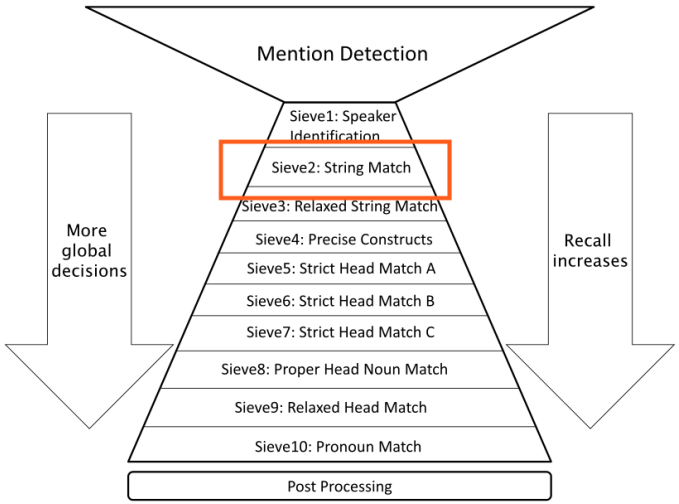
$[\text{John}]_1^1$ is $[\text{a musician}]_2^2$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$.
$[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^6$.
"$[\text{It}]_7^7$ is $[[\text{my}]_9^9$ favorite$]_8^8$," $[\text{John}]_{10}^{10}$ said to $[\text{her}]_{11}^{11}$.

Speaker Identification



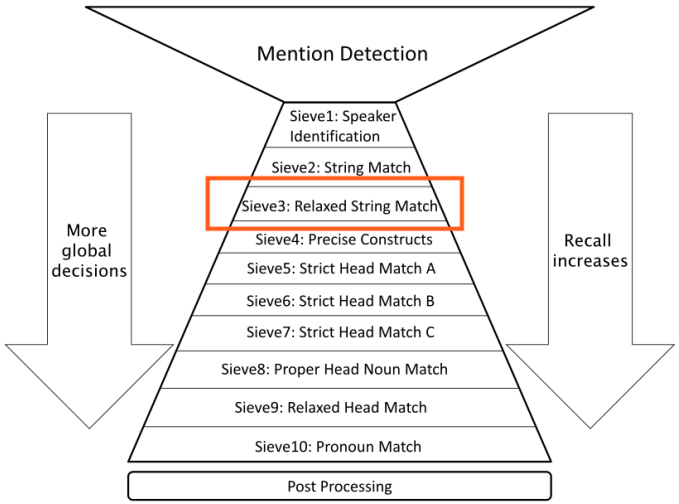
$[\text{John}]_1^1$ is $[\text{a musician}]_2^2$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$.
$[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^6$.
"$[\text{It}]_7^7$ is $[[\textbf{my}]_9^9$ favorite$]_8^8$," $\textbf{[John]}_{10}^9$ said to $[\text{her}]_{11}^{11}$.

String Match



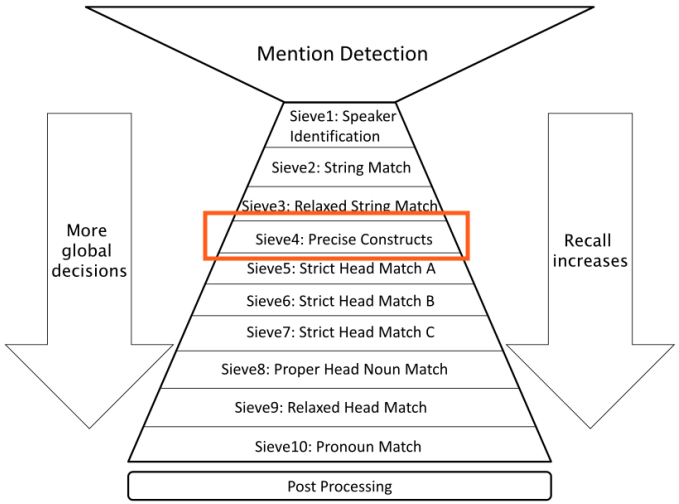
[John]$^{1}_{1}$ is [a musician]$^{2}_{2}$. [He]$^{3}_{3}$ played [a new song]$^{4}_{4}$.
[A girl]$^{5}_{5}$ was listening to [the song]$^{6}_{6}$.
"[It]$^{7}_{7}$ is [[my]$^{1}_{9}$ favorite]$^{8}_{8}$," **[John]**$^{1}_{10}$ said to [her]$^{11}_{11}$.

Relaxed String Match



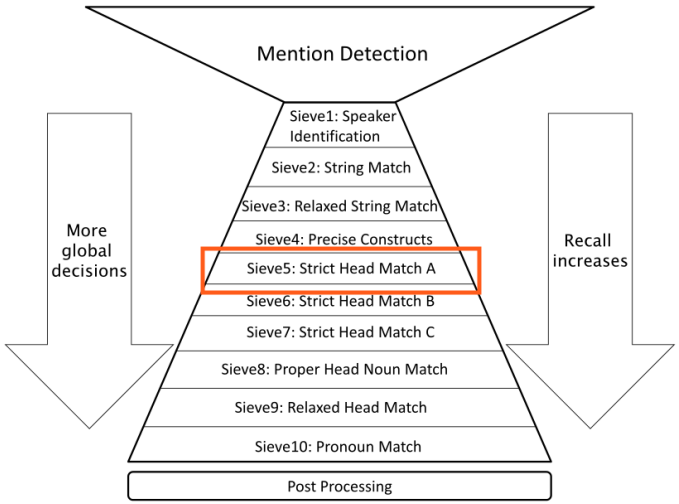
$[\text{John}]_1^1$ is $[\text{a musician}]_2^2$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$.
$[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^6$.
"$[\text{It}]_7^7$ is $[[\text{my}]_9^1$ favorite$]_8^8$," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^{11}$.

Precise Constructs



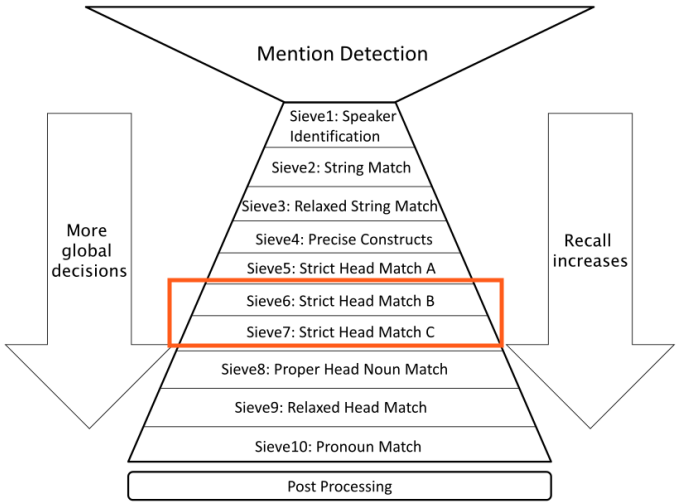
[John]$^1_1$ is **[a musician]**$^1_2$. [He]$^3_3$ played [a new song]$^4_4$.
[A girl]$^5_5$ was listening to [the song]$^6_6$.
"**[It]**$^7_7$ is **[[my]**$^1_9$ **favorite]**$^7_8$," [John]$^1_{10}$ said to [her]$^{11}_{11}$.

Strict Head Match A



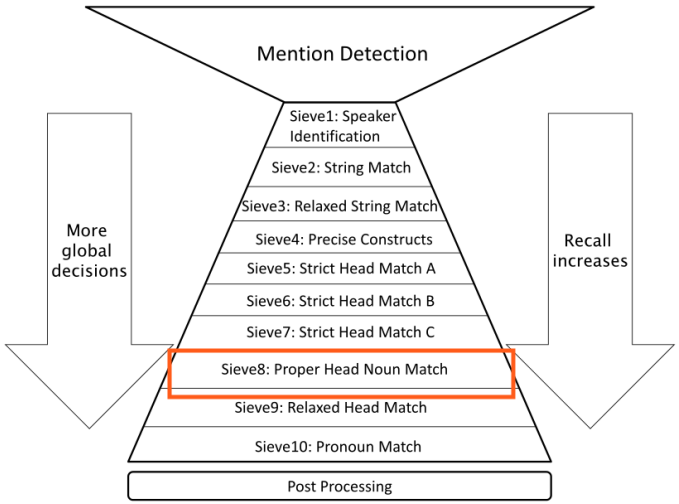
$[\text{John}]_1^1$ is $[\text{a musician}]_2^1$. $[\text{He}]_3^3$ played **[a new song]**$_4^4$.
$[\text{A girl}]_5^5$ was listening to **[the song]**$_6^4$.
"$[\text{It}]_7^7$ is $[[\text{my}]_9^1$ favorite$]_8^7$," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^{11}$.

Strict Head Match B C



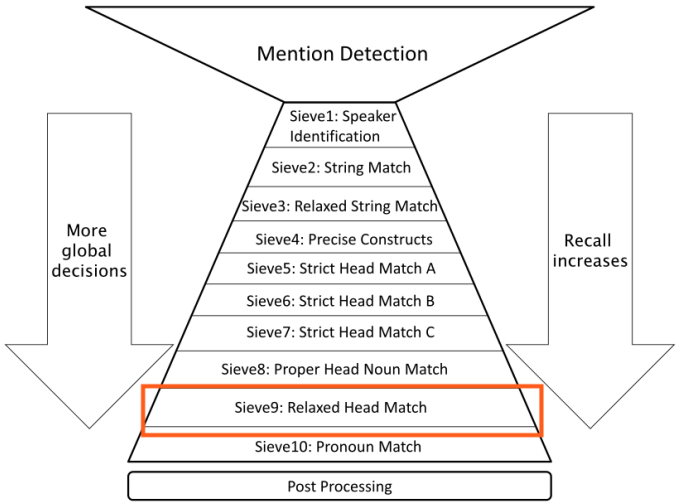
$[\text{John}]_1^1$ is $[\text{a musician}]_2^1$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$.
$[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^4$.
"$[\text{It}]_7^7$ is $[[\text{my}]_9^1$ favorite$]_8^7$," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^{11}$.

Proper Head Noun Match



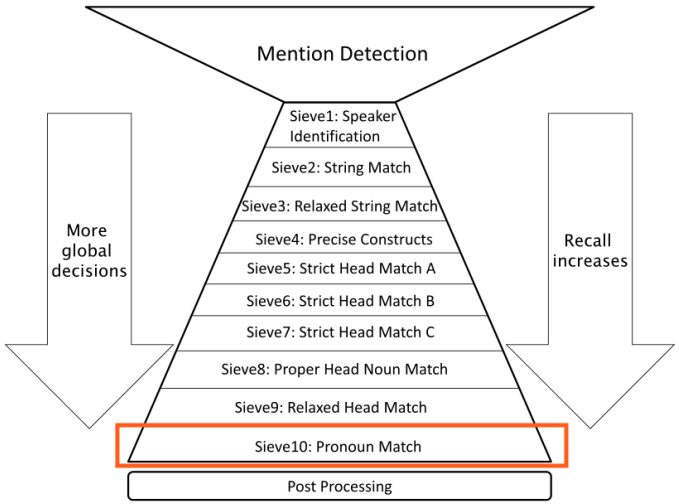
$[\text{John}]_1^1$ is $[\text{a musician}]_2^1$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$.
$[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^4$.
"$[\text{It}]_7^7$ is $[[\text{my}]_9^1$ favorite$]_8^7$," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^{11}$.

Relaxed Head Match



$[\text{John}]_1^1$ is $[\text{a musician}]_2^1$. $[\text{He}]_3^3$ played $[\text{a new song}]_4^4$.
$[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^4$.
"$[\text{It}]_7^7$ is $[[\text{my}]_9^1$ favorite$]_8^7$," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^{11}$.

Pronoun Match



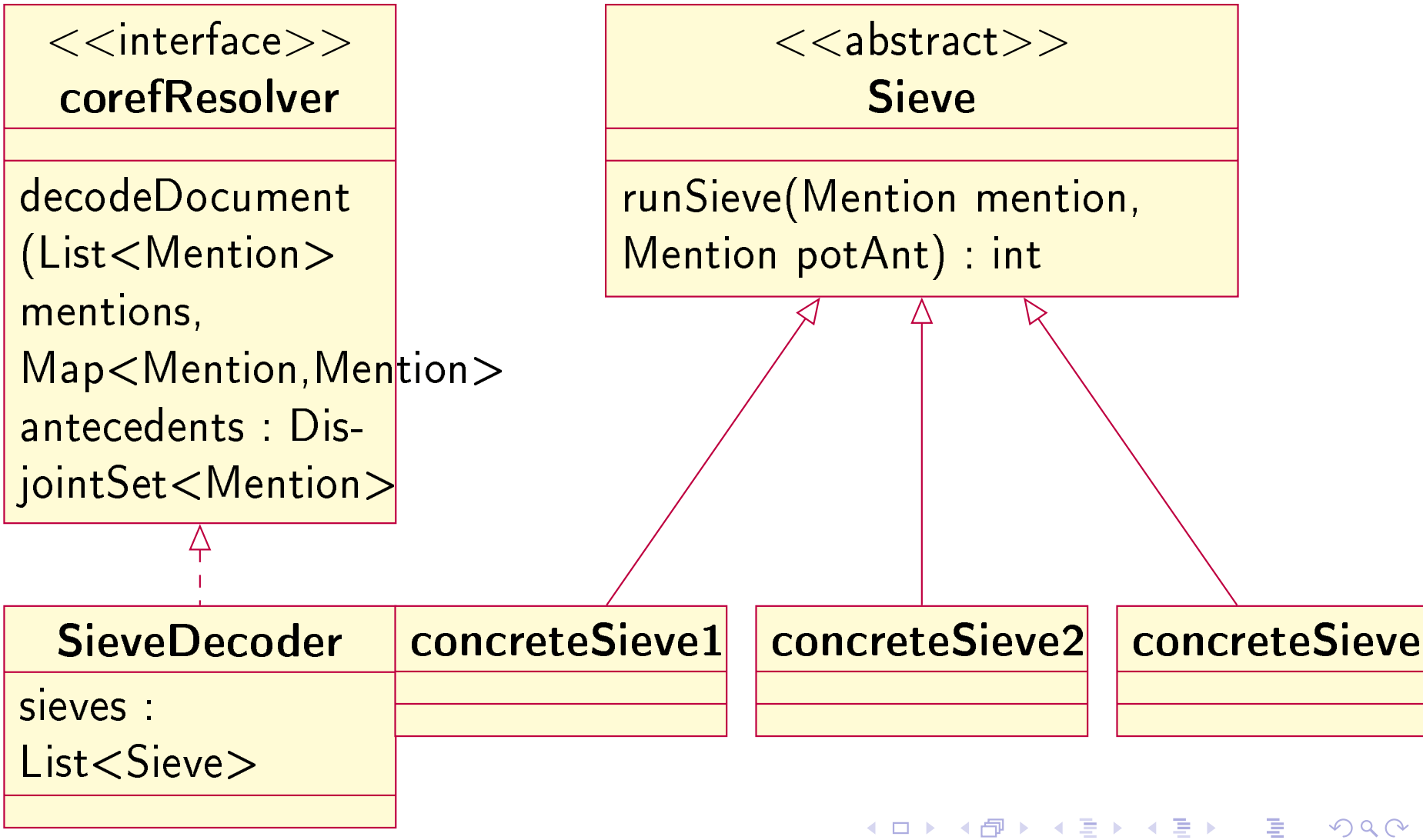
[John]$^1_1$ is [a musician]$^1_2$. **[He]**$^1_3$ played [a new song]$^4_4$.
[A girl]$^5_5$ was listening to **[the song]**$^4_6$.
"**[It]**$^4_7$ is [[my]$^1_9$ favorite]$^4_8$," [John]$^1_{10}$ said to **[her]**$^5_{11}$.

Post Processing + Final Output

$[\text{John}]_1^1$ is **a musician**. $[\text{He}]_3^1$ played $[\text{a new song}]_4^4$.
$[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^4$.
"$[\text{It}]_7^4$ is **[my]$_9^1$ favorite**," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^5$.

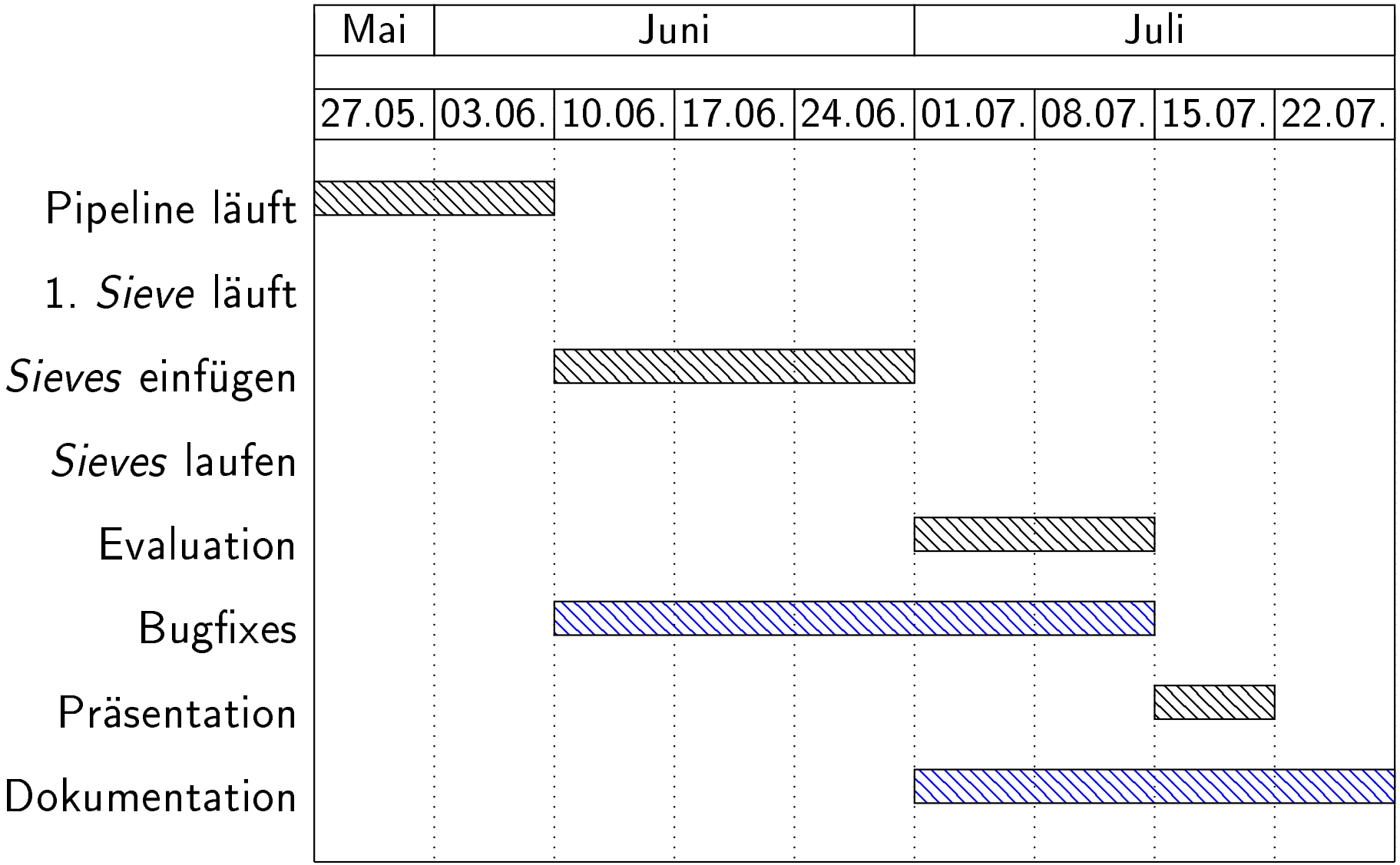
$[\text{John}]_1^1$ is a musician. $[\text{He}]_3^1$ played $[\text{a new song}]_4^4$.
$[\text{A girl}]_5^5$ was listening to $[\text{the song}]_6^4$.
"$[\text{It}]_7^4$ is $[\text{my}]_9^1$ favorite," $[\text{John}]_{10}^1$ said to $[\text{her}]_{11}^5$.

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


Mai
Juni
Juli
27.05.
03.06.
10.06.
17.06.
24.06.
01.07.
08.07.
15.07.
22.07.
Pipeline läuft
1. *Sieve* läuft
Sieves einfügen
Sieves laufen
Evaluation
Bugfixes
Präsentation
Dokumentation

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

Quellen


S. Broscheit, M. Poesio, S. P. Ponzetto, K. J. Rodriguez, L. Romano, O. Uryupina, Y. Versley, and R. Zanoli.
Bart: A multilingual anaphora resolution system.
In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 104–107, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.


H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky.
Deterministic coreference resolution based on entity-centric, precision-ranked rules.
Comput. Linguist., 39(4):885–916, Dec. 2013.


Y. Versley, M. Poesio, and K. Rodriguez.
Bart: A coreference framework, dgfs fall school folien, 2009.