Recall Error Analysis for Coreference Resolution

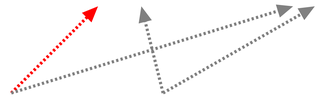
Sebastian Martschat and **Michael Strube**
Heidelberg Institute for Theoretical Studies gGmbH

Heidelberg Institute for Theoretical Studies
HITS

Aim and Motivation

Coreference resolution is a complex, difficult task. How can we **assess errors** made by a coreference resolution system? How can we analyze errors to **identify challenges** in coreference resolution?

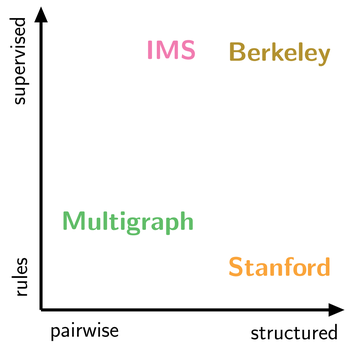
After the discussion, **Obama** confirmed **he** will return.



Then **the president** and **his** bodyguards left.

Experimental Settings

- **CoNLL-2012 English development data**: 343 documents among seven genres
- focus on an analysis of **recall errors involving only names and nouns**: one of the main performance bottlenecks
- analyze systems following different paradigms, reflecting major coreference approaches



Toolkit available at

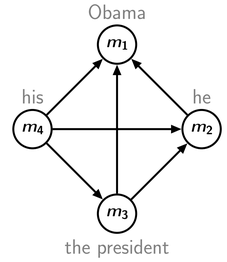
`http://smartschat.de/software`

Method

- represent **entities as graphs**
- recall errors: mismatches in **spanning trees** of reference entities compared to system output (switch roles for precision errors)

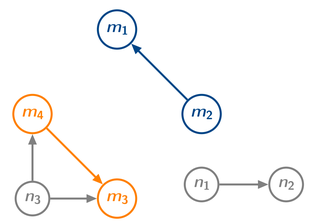
Construct graph for reference entity

- build a complete graph respecting mention ordering



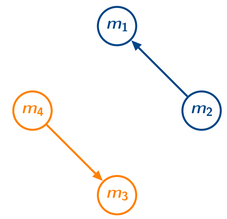
Construct graph for system output

- same principle as for reference entities



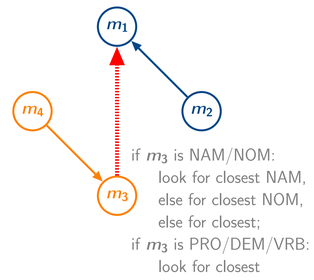
Compute partition

- retain edges in reference entity that are also in system output



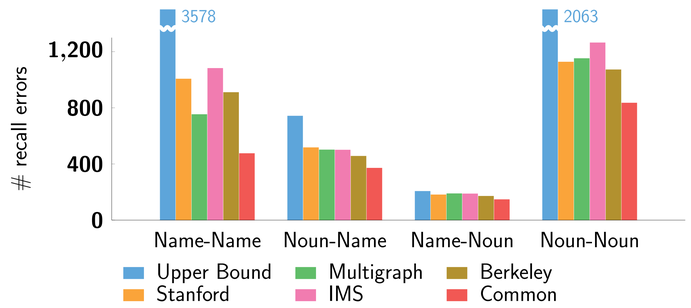
Construct spanning tree

- first compute arbitrary spanning trees for connected components
- to connect: choose edges motivated by Ariel's accessibility theory
- **edges** of tree not in partition are errors



Analysis

We are interested in **common errors**: analyze challenges in coreference resolution.



Proper name pairs: 475 common errors

- especially difficult classes: ORG and DATE
- 32%: complete string match (*China* and *China's*)
- 46%: token overlap (*the Cole* and *the USS Cole*)
- 22%: no overlap, mostly acronyms, alias, or dates (*Florida* and *the Sunshine State* or *1989* and *last year*)

Anaphor noun, antecedent name: 371 common errors

- especially difficult: PERSON
- most: hyponymy (*the prime minister* and *Mr. Papandreou*)
- 18%: token overlap (*the entire park* and *the Ocean Park*)
- 204 different heads: 142 only once, top ten heads 24% of errors; top heads *company*, *group*, *government*, *country* and *nation*

Common noun pairs: 835 common errors

- 21%: anaphor indefinite/bare plural (*a young airman* and *the man*)
- 41%: head match (*the same colors* and *the colors*)
- 59%: no head match – mostly hyponymy or synonymy, but also many other phenomena such as metaphor (*the products* and *the Justin merchandise* but also *the public's money* and *the purse*)

Conclusions

- **graph-based method** for error extraction
- incorporate linguistic information during spanning tree computation
- analysis of nominal/name errors: core set of challenging errors
- highlighted and quantified **usefulness of world knowledge**