Coreference Resolution

Maschinelles Lernverfahren zur Koreferenz-Auflösung

Nikolina Koleva

UdS

12. Mai 2011

Womit befasst sich die Koreferenz-Auflösung?

Womit befasst sich die Koreferenz-Auflösung?

Referieren zwei sprachliche Ausdrücke auf das selbe Individuum?

Womit befasst sich die Koreferenz-Auflösung?

Referieren zwei sprachliche Ausdrücke auf das selbe Individuum?

- Bezugswort: worauf referiert wird
- Anapher: der referierende Ausdruck

Womit befasst sich die Koreferenz-Auflösung?

Referieren zwei sprachliche Ausdrücke auf das selbe Individuum?

- Bezugswort: worauf referiert wird
- Anapher: der referierende Ausdruck

"Alice spielt draußen. Sie genießt das schöne Wetter."

Womit befasst sich die Koreferenz-Auflösung?

Referieren zwei sprachliche Ausdrücke auf das selbe Individuum?

- Bezugswort: worauf referiert wird
- Anapher: der referierende Ausdruck

"Alice spielt draußen. Sie genießt das schöne Wetter."

Koreferenzkette: **Alice-Sie**

Motivation

Warum ist Koreferenz-Auflösung wichtig?

Motivation

Warum ist Koreferenz-Auflösung wichtig?

1. zur Diskursanalyse

Motivation

Warum ist Koreferenz-Auflösung wichtig?

1. zur Diskursanalyse

2. allgemein zum Sprachverstehen

Motivation

Warum ist Koreferenz-Auflösung wichtig?

1. zur Diskursanalyse
2. allgemein zum Sprachverstehen
3. wichtige Teilaufgabe in NLP-Systeme

   $\rightarrow$ insbesondere in Informationsextraktion(IE) Systeme

Der Ansatz von Soon et al.

Der Ansatz von Soon et al.

1. Bestimmung der Markables (potenzielle Bezugswörter oder Anapher)

Der Ansatz von Soon et al.

1. Bestimmung der Markables (potenzielle Bezugswörter oder Anapher)
2. Festlegung ihrer Attributvektoren

Der Ansatz von Soon et al.

1. Bestimmung der Markables (potenzielle Bezugswörter oder Anapher)
2. Festlegung ihrer Attributvektoren
3. Generieren der Trainingsinstanzen

Der Ansatz von Soon et al.

1. Bestimmung der Markables (potenzielle Bezugswörter oder Anapher)
2. Festlegung ihrer Attributvektoren
3. Generieren der Trainingsinstanzen
4. Trainieren des Modells

Der Ansatz von Soon et al.

1. Bestimmung der Markables (potenzielle Bezugswörter oder Anapher)
2. Festlegung ihrer Attributvektoren
3. Generieren der Trainingsinstanzen
4. Trainieren des Modells
5. Bilden der Koreferenzketten für das Testdokument

Markable

- sprachlicher Ausdruck, auf den sich andere sprachliche Ausdrücke beziehen können

Markable

- sprachlicher Ausdruck, auf den sich andere sprachliche Ausdrücke beziehen können
- oft: Markable Grenzen $\neq$ Token Grenzen

Markable

- sprachlicher Ausdruck, auf den sich andere sprachliche Ausdrücke beziehen können
- oft: Markable Grenzen $\neq$ Token Grenzen
-

Separately, Clinton transition officials said that Frank Newman, 50, vice chairman and chief financial officer of BankAmerica Corp. , is expected to be nominated as assistant Treasury secretary for domestic finance.

Bestimmung der Markables

Bestimmung der Markables

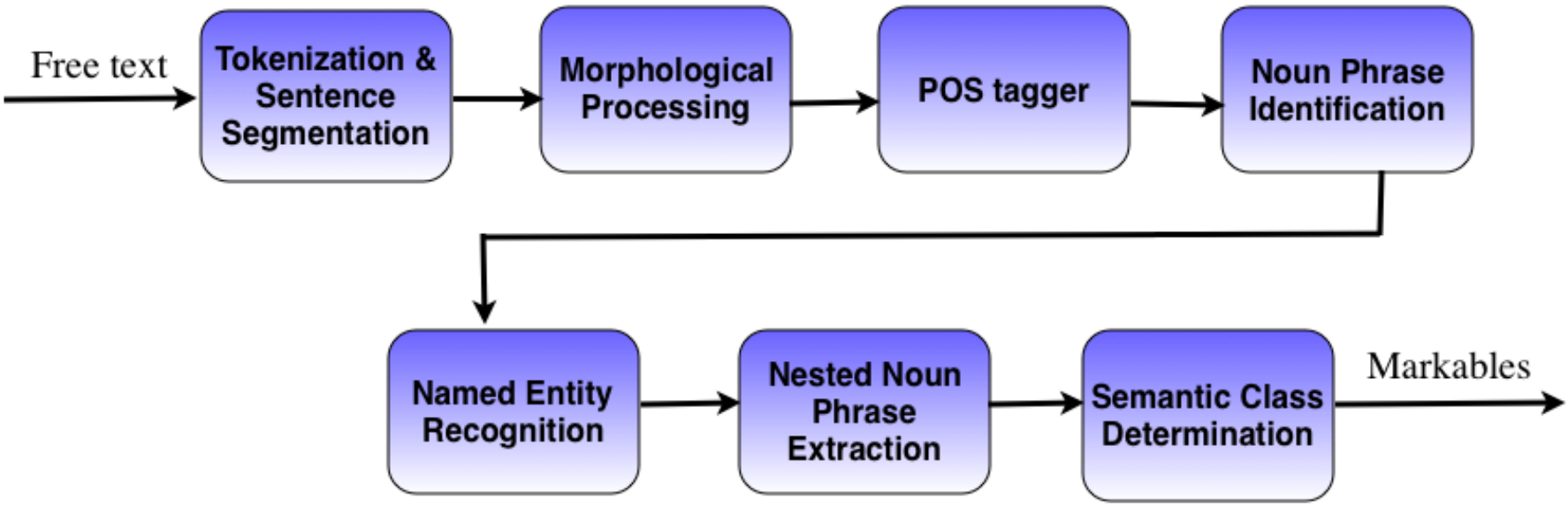
- mittels einer Pipeline von NLP module

Bestimmung der Markables

- mittels einer Pipeline von NLP module

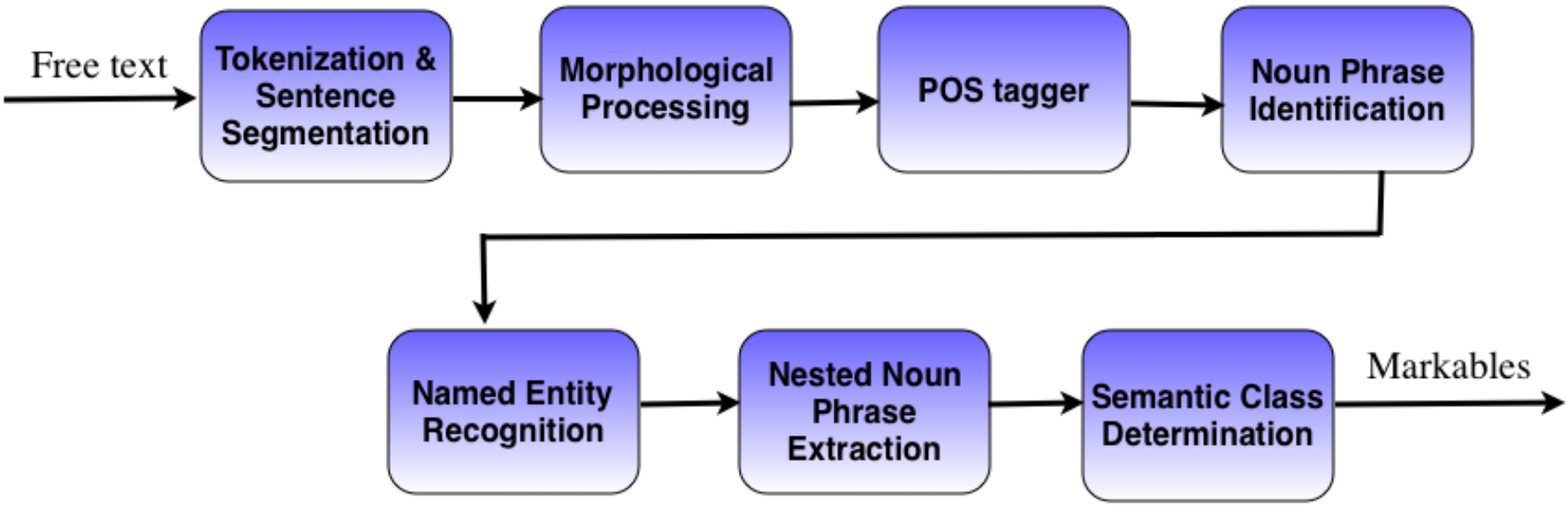
Bestimmung der Markables

- mittels einer Pipeline von NLP module



Bestimmung der Markables

- mittels einer Pipeline von NLP module



- getestet gegen 100 Trainingsdokumente $\rightarrow$ 85 % Übereinstimmung

12 Attribute der Markables

12 Attribute der Markables

i ist das potenzielle Bezugswort; **j** ist die Anapher

12 Attribute der Markables

i ist das potenzielle Bezugswort; **j** ist die Anapher

- *DIST*: Distanz zwischen i und j (0,1,2,3...)

12 Attribute der Markables

i ist das potenzielle Bezugswort; **j** ist die Anapher

- *DIST*: Distanz zwischen i und j (0,1,2,3...)
- *I_PRONOUN*: ist i ein Pronomen (true, false)

12 Attribute der Markables

i ist das potenzielle Bezugswort; **j** ist die Anapher

- *DIST*: Distanz zwischen i und j (0,1,2,3...)
- *I_PRONOUN*: ist i ein Pronomen (true, false)
- *J_PRONOUN*: ist j ein Pronomen (true, false)

12 Attribute der Markables

i ist das potenzielle Bezugswort; **j** ist die Anapher

- *DIST*: Distanz zwischen i und j (0,1,2,3...)
- *I_PRONOUN*: ist i ein Pronomen (true, false)
- *J_PRONOUN*: ist j ein Pronomen (true, false)
- *STR_MATCH*: stimmen die Strings i und j überein(true, false)
  $\rightarrow$ *a boy* und *this boy*

12 Attribute der Markables

i ist das potenzielle Bezugswort; **j** ist die Anapher

- *DIST*: Distanz zwischen i und j (0,1,2,3...)
- *I_PRONOUN*: ist i ein Pronomen (true, false)
- *J_PRONOUN*: ist j ein Pronomen (true, false)
- *STR_MATCH*: stimmen die Strings i und j überein(true, false)
  $\rightarrow$ *a boy* und *this boy*
- *DEF_NP*: ist j definite Nominalphrase (true, false)
  $\rightarrow$ *the car/girl* etc.

12 Attribute der Markables

i ist das potenzielle Bezugswort; **j** ist die Anapher

- *DIST*: Distanz zwischen i und j (0,1,2,3...)
- *I_PRONOUN*: ist i ein Pronomen (true, false)
- *J_PRONOUN*: ist j ein Pronomen (true, false)
- *STR_MATCH*: stimmen die Strings i und j überein(true, false)
  $\rightarrow$ *a boy* und *this boy*
- *DEF_NP*: ist j definite Nominalphrase (true, false)
  $\rightarrow$ *the car/girl* etc.
- *DEM_NP*: ist j demonstrative Nominalphrase (true, false)
  $\rightarrow$ *this car/girl* etc.

12 Attribute der Markables

- *NUMBER*: kongruieren i und j bezüglich Numerus (true, false)

12 Attribute der Markables

- *NUMBER*: kongruieren i und j bezüglich Numerus (true, false)
- *SEMCLASS*: is-a Hierarchy

12 Attribute der Markables

- *NUMBER*: kongruieren i und j bezüglich Numerus (true, false)
- *SEMCLASS*: is-a Hierarchy
  - wenn *SEMCLASS*(**i** bzw. **j**) $\subset$ *SEMCLASS*(**j** bzw. **i**)
    oder *SEMCLASS*(i) = *SEMCLASS*(j) $\rightarrow$ true

    The top model, Mrs. Claudia Schiffer has three children.
    top model $\in$ "person"; Mrs. Claudia Schiffer $\in$ "female"
    "person" $\subset$ "female"

12 Attribute der Markables

- *NUMBER*: kongruieren i und j bezüglich Numerus (true, false)
- *SEMCLASS*: is-a Hierarchy
  - wenn *SEMCLASS*(**i** bzw. **j**) $\subset$ *SEMCLASS*(**j** bzw. **i**)
    oder *SEMCLASS*(i) = *SEMCLASS*(j) $\rightarrow$ true

    The top model, Mrs. Claudia Schiffer has three children.
    top model $\in$ "person"; Mrs. Claudia Schiffer $\in$ "female"
    "person" $\subset$ "female"
  - *SEMCLASS*(i) != *SEMCLASS*(j) $\rightarrow$ false

12 Attribute der Markables

- *NUMBER*: kongruieren i und j bezüglich Numerus (true, false)
- *SEMCLASS*: is-a Hierarchy
  - wenn *SEMCLASS*(**i** bzw. **j**) $\subset$ *SEMCLASS*(**j** bzw. **i**)
    oder *SEMCLASS*(i) = *SEMCLASS*(j) $\rightarrow$ true

    The top model, Mrs. Claudia Schiffer has three children.
    top model $\in$ "person"; Mrs. Claudia Schiffer $\in$ "female"
    "person" $\subset$ "female"
  - *SEMCLASS*(i) != *SEMCLASS*(j) $\rightarrow$ false
  - *SEMCLASS*(i) = unknown, *SEMCLASS*(j) = unknown
    und *STR_MATCH* = false $\rightarrow$ unknown

12 Attribute der Markables

- *GENDER*: kongruieren i und j bezüglich Genus (true, false, unknown)

12 Attribute der Markables

- *GENDER*: kongruieren i und j bezüglich Genus (true, false, unknown)
  - *SEMCLASS*(i und j)=unknown $\rightarrow$ *GENDER*(i und j)=unknown

12 Attribute der Markables

- *GENDER*: kongruieren i und j bezüglich Genus (true, false, unknown)
  - *SEMCLASS*(i und j)=unknown $\rightarrow$ *GENDER*(i und j)=unknown
- *PROPER_NAME*: sind i und j Eigennamen(true, false)

12 Attribute der Markables

- *GENDER*: kongruieren i und j bezüglich Genus (true, false, unknown)
  - *SEMCLASS*(i und j)=unknown $\rightarrow$ *GENDER*(i und j)=unknown
- *PROPER_NAME*: sind i und j Eigennamen(true, false)
- *ALIAS*: sind i und j **Named Entities**, die auf das selbe Entität referieren (true, false)

12 Attribute der Markables

- *GENDER*: kongruieren i und j bezüglich Genus (true, false, unknown)
  - *SEMCLASS*(i und j)=unknown $\rightarrow$ *GENDER*(i und j)=unknown
- *PROPER_NAME*: sind i und j Eigennamen(true, false)
- *ALIAS*: sind i und j **Named Entities**, die auf das selbe Entität referieren (true, false)
  - **Person** *Mr. Simpson* und *Bent Simpson*

12 Attribute der Markables

- *GENDER*: kongruieren i und j bezüglich Genus (true, false, unknown)
  - *SEMCLASS*(i und j)=unknown $\rightarrow$ *GENDER*(i und j)=unknown
- *PROPER_NAME*: sind i und j Eigennamen(true, false)
- *ALIAS*: sind i und j **Named Entities**, die auf das selbe Entität referieren (true, false)
  - **Person** *Mr. Simpson* und *Bent Simpson*
  - **Datum** *01-08* und *Jan. 8*

12 Attribute der Markables

- *GENDER*: kongruieren i und j bezüglich Genus (true, false, unknown)
  - *SEMCLASS*(i und j)=unknown $\rightarrow$ *GENDER*(i und j)=unknown
- *PROPER_NAME*: sind i und j Eigennamen(true, false)
- *ALIAS*: sind i und j **Named Entities**, die auf das selbe Entität referieren (true, false)
  - **Person** *Mr. Simpson* und *Bent Simpson*
  - **Datum** *01-08* und *Jan. 8*
  - **Organization** *IBM* und *International Business Machines Corp.*

12 Attribute der Markables

- *GENDER*: kongruieren i und j bezüglich Genus (true, false, unknown)
  - *SEMCLASS*(i und j)=unknown → *GENDER*(i und j)=unknown
- *PROPER_NAME*: sind i und j Eigennamen(true, false)
- *ALIAS*: sind i und j **Named Entities**, die auf das selbe Entität referieren (true, false)
  - **Person** *Mr. Simpson* und *Bent Simpson*
  - **Datum** *01-08* und *Jan. 8*
  - **Organization** *IBM* und *International Business Machines Corp.*
- *APPOSITIVE*: ist j Apposition von i (true, false)

Beispiel für einen Attributvektor

Separately, Clinton transition officials said that Frank Newman, 50, vice chairman and chief financial officer of BankAmerica Corp. , is expected to be nominated as assistant Treasury secretary for domestic finance.

Beispiel für einen Attributvektor

Attribut	Attributwert	Kommentar
DIST	0	**i** und **j** sind im selben Satz
I_PRONOUN	false	**i** ist kein Pronomen
J_PRONOUN	false	**j** ist kein Pronomen
STR_MATCH	false	**i** und **j** stimmen nicht überein
DEF_NP	false	**j** ist keine definite NP
DEM_NP	false	**j** ist keine demonstrative NP
NUMBER	true	**i** und **j** sind beide im Singular
SEMCLASS	true	**i** und **j** sind beide Personen
GENDER	true	**i** und **j** sind beide maskulin
PROPER_NAME	false	nur **i** ist Eigenname
ALIAS	false	**j** ist keinen Alias von **i**
APPOSITIVE	true	**j** ist Apposition von **i**

Table: Attributvektor für das Markables-Paar (**i**=Frank Newmann, **j**=vice chairman)

Trainingsinstanzen

Trainingsinstanzen

... ((*Eastern Airlines*)$^{5}_{a2}$ *executives*)^{6} notified ((*union*)$^{7}_{e1}$ *leaders*)^{8} that (*the carrier*)^{9} wishes to discuss (*selective* ((*wage*)$^{10}_{c2}$ *reductions*)$_{d2}$)^{11} on (*Feb.* 3)$^{12}_{b2}$. (((*Union*)$^{13}_{e2}$ *representatives*)^{14} *who could be reached*)$_{f1}$ said (*they*)$^{15}_{f2}$ hadn't decided whether (*they*)$^{16}_{f3}$ would respond. By proposing (*a meeting date*)$^{17}_{b3}$, (*Eastern*)$^{18}_{a3}$ moved (*one step*)^{19} closer toward reopening (*current high – cost contrast agreements*)^{20} with ((*its*)$^{21}_{a4}$ *unions*)$^{22}_{e3}$.

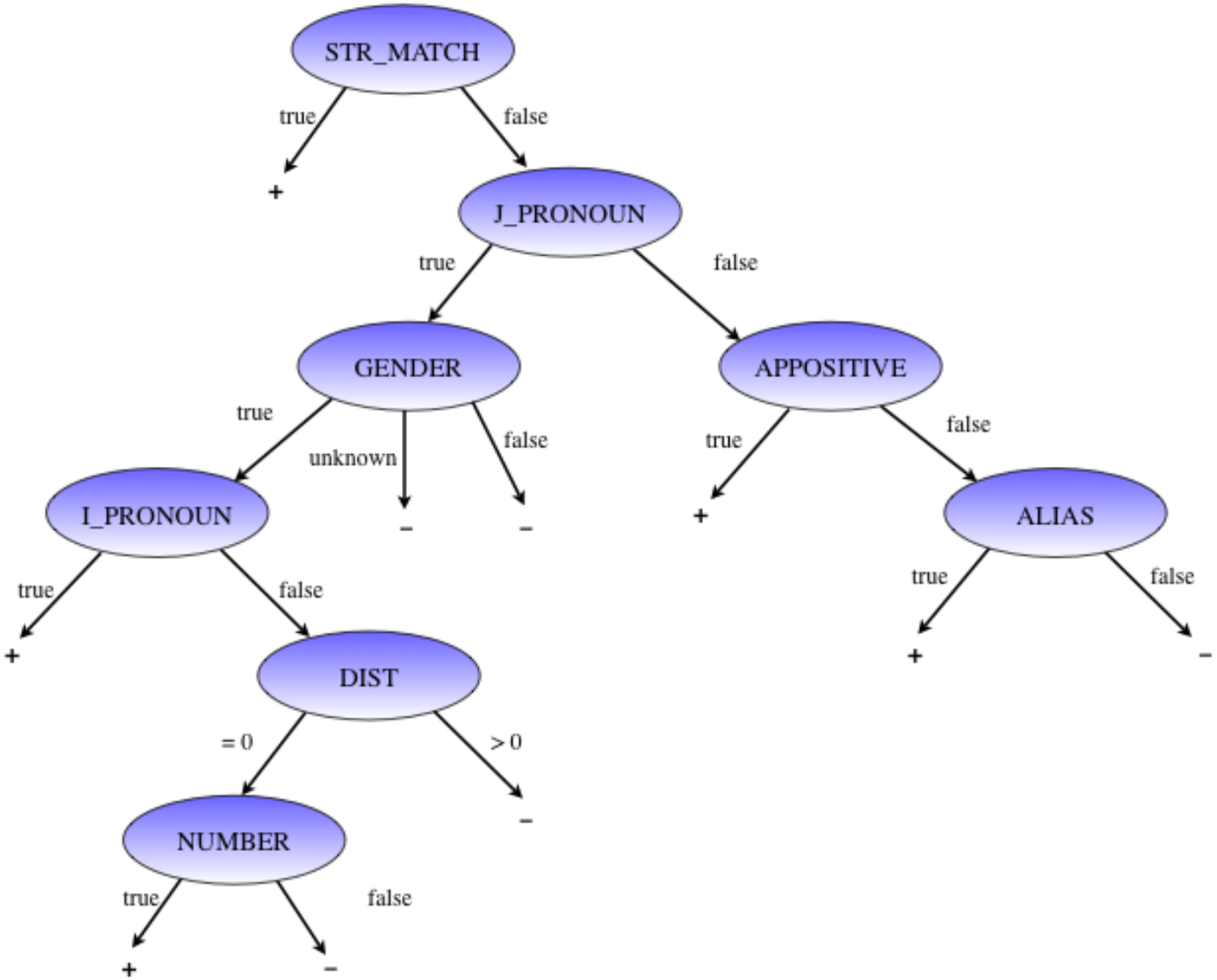
Trainingsinstanzen

... ((*Eastern Airlines*)$^{5}_{a2}$ *executives*)^{6} notified ((*union*)$^{7}_{e1}$ *leaders*)^{8} that (*the carrier*)^{9} wishes to discuss (*selective* ((*wage*)$^{10}_{c2}$ *reductions*)$_{d2}$)^{11} on (*Feb.* 3)$^{12}_{b2}$. (((*Union*)$^{13}_{e2}$ *representatives*)^{14} *who could be reached*)$_{f1}$ said (*they*)$^{15}_{f2}$ hadn't decided whether (*they*)$^{16}_{f3}$ would respond. By proposing (*a meeting date*)$^{17}_{b3}$, (*Eastern*)$^{18}_{a3}$ moved (*one step*)^{19} closer toward reopening (*current high – cost contrast agreements*)^{20} with ((*its*)$^{21}_{a4}$ *unions*)$^{22}_{e3}$.

- positive Beispiele
  - ((*union*)^{7}, (*union*)^{13})
  - ((*union*)^{13}, (*its unions*)^{22})
- negative Beispiele
  - ((*the carrier*)^{9}, (*union*)^{13})
  - ((*wage*)^{10}, (*union*)^{13})
  - ((*selective wage reductions*)^{11}, (*union*)^{13})
  - ((*Feb.* 3)^{12}, (*union*)^{13})

MUC-6 Klassifikator erstellt mit C5

Message Understanding Conference (MUC)



Koreferenzketten für ein Testdokument

...(**Ms. Washington**)[73] 's candidacy is being championed by (*several powerful lawmakers*)[74] including ((**her**)[76] *boss*)[75], (*Chairman John Dingell*)[77](Dr.(*Mich.*)[78]) of (*the House Energy and Commerce Committe*)[79]. (**She**)[80] currently is (*a counsel*)[81] to (*the committee*)[82].(**Ms. Washington**)[83] and (*Mr. Dingell*)[84] have been considered (*allies*)[85] of (*the* (*securities*)[87] *exchanges*)[86], while (*banks*)[88] and ((*futures*)[90] *exchanges*)[89] have often fought with (*them*)[91].

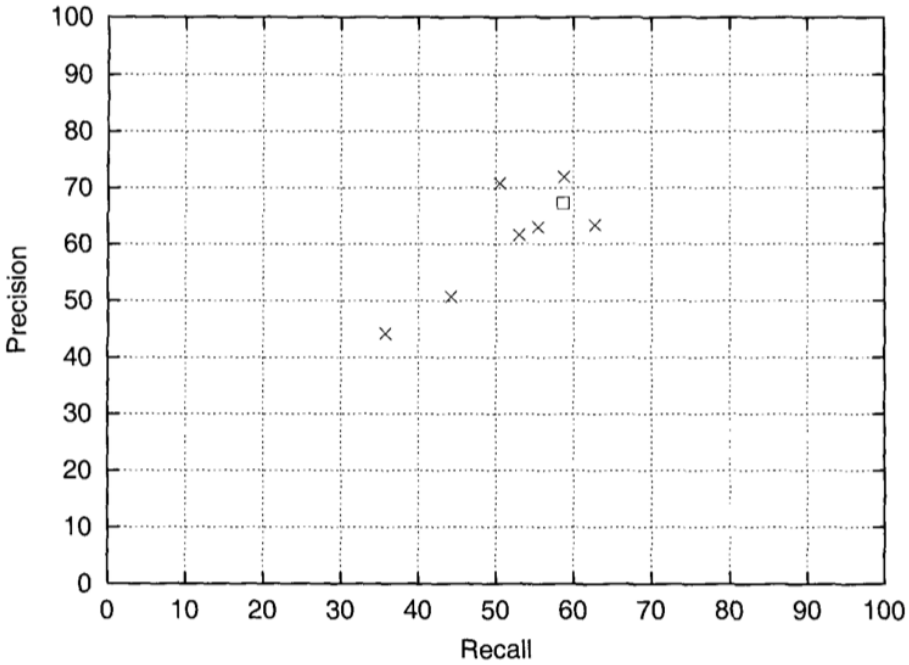
getestete Markables-Paare zur Koreferenzkettenbildung

Bezugswort	Anapher	Attributvektor*	Koreferent?
(*several powerful lawmakers*)[74]	(**her**)[76]	0, t, f, u, f, f, t, f, f, f, f, f	Nein
(**Ms. Washington**)[73]	(**her**)[76]	0, t, t, t, f, f, t, f, f, f, f, f	Ja
(*the House Energy and Commerce Committe*)[79]	(**She**)[80]	1, f, t, f, f, f, t, f, f, f, f, f	Nein
(*Mich.*)[78]	(**She**)[80]	2, f, t, f, f, f, t, f, f, f, f, f	Nein
(*Chairman John Dingell*)[77]	(**She**)[80]	3, t, t, f, f, f, t, f, f, f, f, f	Nein
(**her**)[76]	(**She**)[80]	3, t, t, t, f, f, t, f, f, f, f, t	Ja
(*the committee*)[82]	(**Ms. Washington**)[83]	1, f, t, f, f, f, f, f, f, f, f, f	Nein
(*a counsel*)[81]	(**Ms. Washington**)[83]	1, f, t, u, f, f, f, f, f, f, f, f	Nein
(**She**)[80]	(**Ms. Washington**)[83]	1, t, t, t, f, f, f, f, f, f, f, t	Nein
(*the House Energy and Commerce Committe*)[79]	(**Ms. Washington**)[83]	2, f, t, f, t, f, f, f, f, f, f, f	Nein
(*Mich.*)[78]	(**Ms. Washington**)[83]	3, f, t, f, t, f, f, f, f, f, f, f	Nein
(*Chairman John Dingell*)[77]	(**Ms. Washington**)[83]	4, t, t, f, t, f, f, f, f, f, f, f	Nein
(**her**)[76]	(**Ms. Washington**)[83]	4, t, t, t, f, f, f, f, f, f, f, t	Nein
(*her boss*)[75]	(**Ms. Washington**)[83]	4, t, f, f, f, f, f, f, f, f, f, f	Nein
(*several powerful lawmakers*)[74]	(**Ms. Washington**)[83]	4, t, f, u, f, f, f, f, f, f, f, f	Nein
(**Ms. Washington**)[73]	(**Ms. Washington**)[83]	4, t, t, t, t, t, f, f, f, t, f, f	Ja

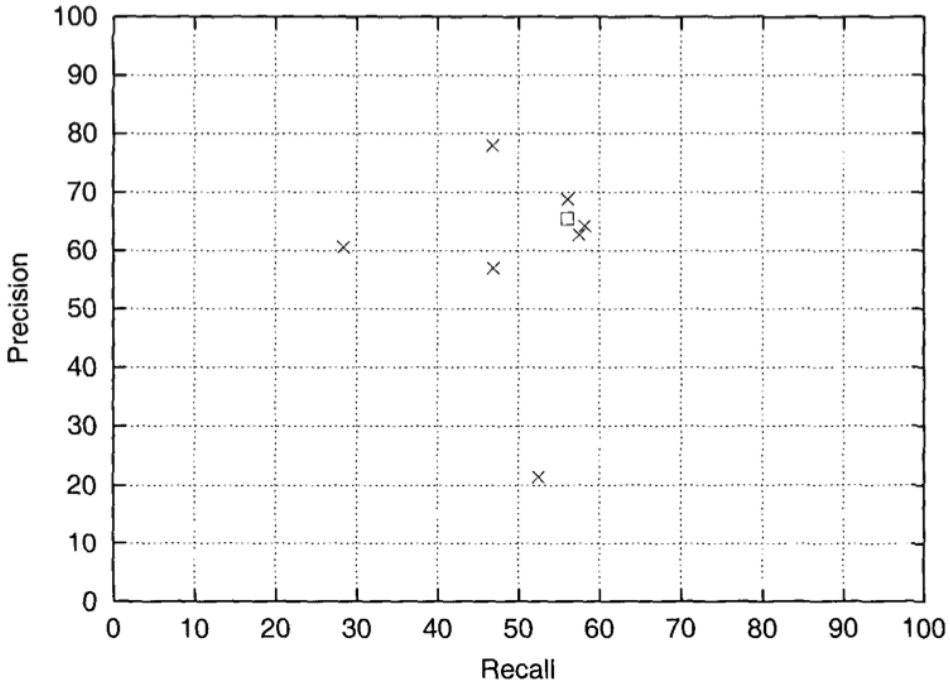
* dist, semclass, number, gender, proper-name, alias, j-pronoun, def-np, dem-np, str-match, appositive, i-pronoun

Evaluation: Precision und Recall

MUC - 6

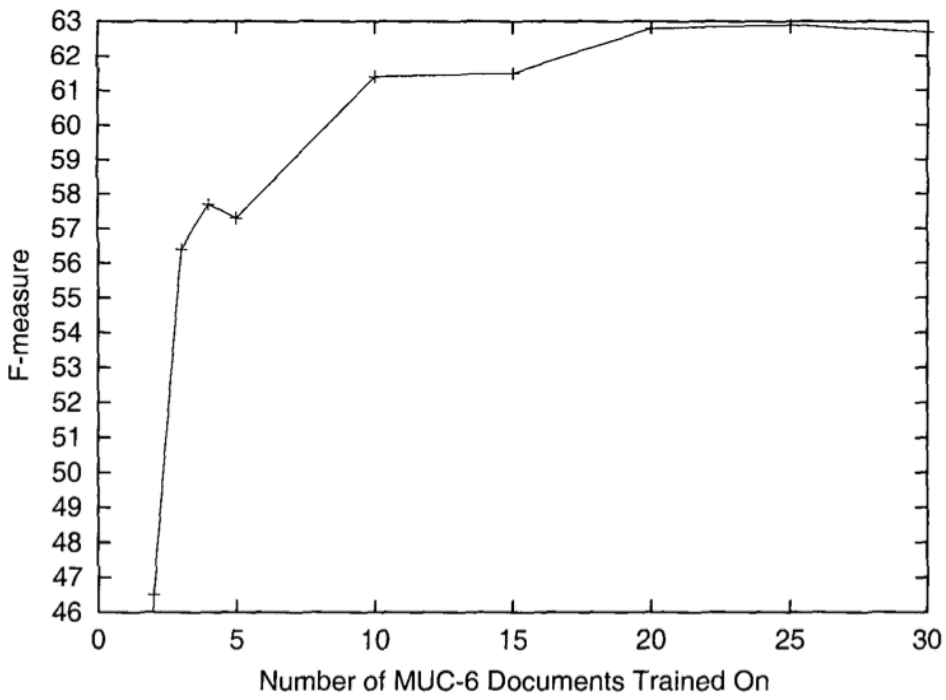


MUC - 7

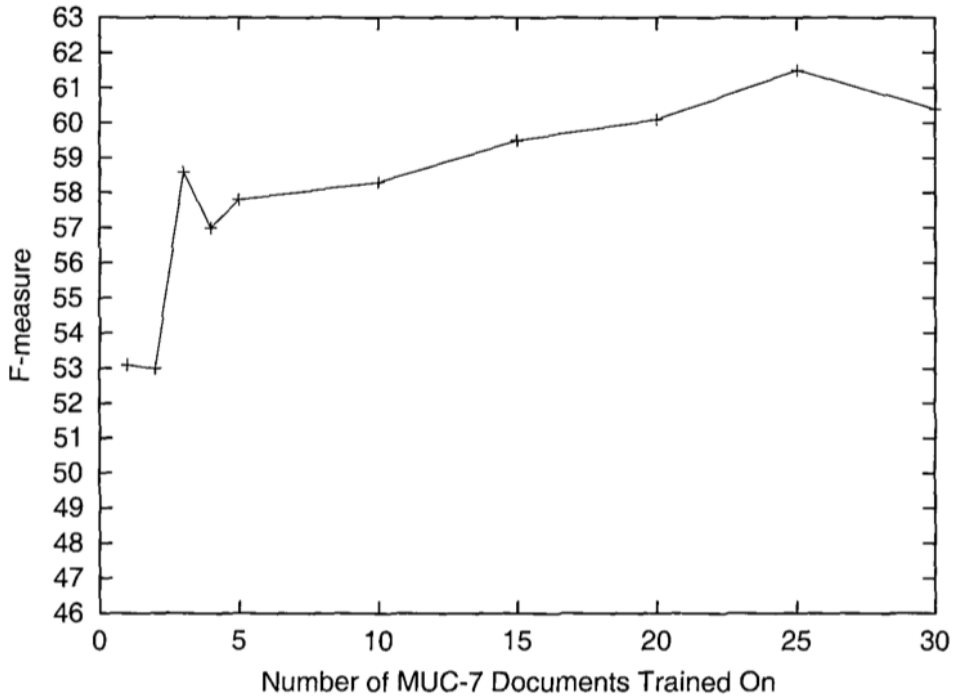


Evaluation: F-Score

MUC - 6



MUC - 7



Zusammenfassung

- inkrementelle Verarbeitung des Eingabetextes

Zusammenfassung

- inkrementelle Verarbeitung des Eingabetextes
- hinreichende Attribute der Markables

Zusammenfassung

- inkrementelle Verarbeitung des Eingabetextes
- hinreichende Attribute der Markables
- Umsetzung auf erhältliche Datensätze

Zusammenfassung

- inkrementelle Verarbeitung des Eingabetextes
- hinreichende Attribute der Markables
- Umsetzung auf erhältliche Datensätze
- echtes Testen und umfassende Evaluation

Referenzen

[1] Wee Meng Soon, Daniel Chung Yong Lim and Hwee Tou Ng, 2001, "A Machine Learning Approach to Coreference Resolution of Noun Phrases" , *Computational Linguistics*, 27:4, S. 521-544

Danke für Ihre Aufmerksamkeit!