

Einführung in die Pragmatik und Diskurs: Anapherresolution

I. Kruijff-Korbayova/A. Horbach

Universität des Saarlandes

Sommersemester 2013

basierend auf Folien von C. Sporleder und I. Kruijff-Korbayova

Der Briefträger streichelte den Hund. Plötzlich biss er zu.
Wer beißt hier wen?

Tony Blair met President Yeltsin. The old man had just recovered from a heart attack.
Wer hatte einen Herzinfarkt?

Vorlesungsplan

- Arten von Referenzen
- Constraints und Präferenzen
- Anaphern: Problem und Bedeutung
- 3 Algorithmen zur Anapherresolution
- (Interessante/Schwierige) Phänomene

Kernlektüre

- Jurafsky & Martin (2000), Kapitel 18 (2009, Kapt 21)

Zusätzliche Lektüre

- Hobbs 1978
- Brennan et. al. 1987
- (Lappin & Leass 1994)

Wiederholung

- **Referenzausdrücke** (*die Queen, der Bus, eine Katze, er ...*) referieren auf reale Entitäten
- **Referent** ist die Entität, auf die sich ein Sprecher mit einem Referenzausdruck bezieht.
- Referenzausdrücke, die auf dieselbe Entität verweisen, sind **koreferent**

Koreferenz und Anaphorik

- **Koreferenzkette** (coreference chain): eine Menge von Referenzausdrücken in einem Text, die koreferent sind
- **Anaphorik** (anaphora): ein Ausdruck verweist auf einen vorangegangenen Ausdruck (Antezedens)
- **Anapher** (anaphor): der zurückweisende Ausdruck (z.B. *sie, die Katze*)
- analog: Kataphorik (cataphora) für vorausweisende Ausdrücke
- **Deiktische Referenz (Deixis)**: Referanz auf eine Entität im situativen Kontext des Äußerung. (*Du, hier, jetzt*)
- **Koreferenz vs. Anaphorik**
 - cross-document coreference (=nicht anaphorisch)
 - Anaphern, die nicht koreferent sind (*Everybody has his own destiny.*)

Koreferenzresolution: finde die Koreferenzketten in einem Text.

Anapherresolution: finde das Antezedens einer Anapher.

Sophia Loren says she will always be grateful to Bono. The actress revealed that the U2 singer helped her calm down when she became scared by a thunderstorm while travelling on a plane.

Koreferenzketten (coreference chains):

- {Sophia Loren, she, the actress, her, she}
- {Bono, the U2 singer }
- {a thunderstorm}
- {a plane}

Sophia Loren says **she** will always be grateful to Bono. **The actress** revealed that **the U2 singer** helped her calm down when she became scared by a thunderstorm while travelling on a plane.

- *she* ⇒ *Sophia Loren*
- *the actress* ⇒ *Sophia Loren*
- *the U2 singer* ⇒ *Bono*

Schwierigkeiten:

- unterschiedliche Form \nrightarrow verschiedene Referenten
(*Sophia Loren* vs. *the actress* vs. *she*)
- gleiche Form \nrightarrow gleiche Referenten
(*sie* (die Katze) vs. *sie* (Sophia Loren), *Michael Jackson* der Sanger vs. *Michael Jackson* der britische General)

Ambiguität und desambiguierende Faktoren

*Jane told Peter **he** was in danger.*

⇒ Kongruenz (Numerus, Genus): *he* = *Peter*

*The cat did not come down from the tree. **It** was scared.*

⇒ selectional preferences: *it* = *the cat*

*Jane told Mary **she** was in danger.*

⇒ salience (Subjektposition): *she* = *Jane*

*Jane told Mary **SHE** was in danger.*

⇒ Prosodie: *she* = *Mary*

*Jane warned Mary **she** was in danger.*

⇒ lexical semantics (*warned*): *she* = *Mary*

*Peter told John that he is running the business for **himself**.*

⇒ syntactic constraints: *himself* = *Peter*

*Tony Blair met President Yeltsin. **The old man** had just recovered from a heart attack.*

⇒ Weltwissen: *the old man* = *Yeltsin*

*Jan hat ein Kaninchen. Hans hat ein Pferd. Maria füttert **es**.*

⇒ Recency (Distanz): *es* = *Pferd*

*Jan kaufte Hans ein Computerspiel. Maria kaufte **ihm** ein Buch.*

⇒ Parallelismus: *ihm* = *Hans*

Die meisten Constraints können verletzt werden:

Hans hat ein Computerspiel gekauft. Es läuft nicht.

⇒ Keine klare Unterscheidung zwischen Präferenzen und Constraints.

- Theoretisches Interesse: Erklärung, wie Referenzen funktionieren, ist Teil der Erklärung wie Sprache funktioniert.
- Praktische Bedeutung in NLP-Anwendungen:
 - Information extraction, question answering, text summarization: Anapherresolution zur Verknüpfung (und Kondensierung) von Information
 - Maschinelle Übersetzung: zur richtigen Übersetzung von anaphorischen Ausdrücken
 - NL-Interfaces und Dialogsysteme: Anaphernresolution notwendig zur korrekten Interpretation, Generierung von anaphorischen Ausdrücken zur Effizienz und Natürlichkeit

- Aufgabe: Berechne die Bedeutung eines Referenzausdrucks:
- Wesentlich zur Konstruktion eines Diskursmodells:
 - Einführen neuer Diskursreferenten vs.
 - Zugriff auf alte Diskursreferenten
- Schritte zur Anapherresolution:
 - 1 identifiziere Anapher (Ist ein Ausdruck anaphorisch oder nicht?)
Schwierigkeiten: NPs, die keine Referenzausdrücke sind:
pleonastisches *es* (*Es schneit.*) etc.
 - 2 identifiziere potentielle Antezedenten
 - 3 finde passendes Antezedens für jede Anapher

Vor 1990 ...

- Referenzresolution = Pronomenresolution
- regelbasiert (manuell erstellte Regeln)
- Beispiele:
 - SHRDLU (Winograd, 1972): komplexe Heuristiken (Fokus, Obliqueness etc.)
 - Hobbs's (1976, 1978): heuristisch gelenkte Suche in Syntaxbäumen
 - Centering-basiert (Brennan et al. 1987)

Nach 1990 ...

- corpusbasiert (co-occurrence statistics, machine learning)
- auch Referenzresolution für nicht-pronominale Ausdrücke (definite NPs, bridging; z.B. Vieira & Poesio, 2000)

Syntactic-Tree Search Algorithm for Pronominal Coreference (Hobbs 1978)

- 1 Begin at the NP node immediately dominating the pronoun.
- 2 Go up the tree to the next NP or S node. Call it X and the path to reach it p.
- 3 Traverse all branches to the left of p below X in left-to-right breadth-first order. When a compatible NP node which has an NP or S node between it and X is encountered, propose it as antecedent.
- 4 If X is the highest S node in the sentence, traverse the trees of the previous sentences starting from the more recent ones. Each tree is traversed left-to-right, breadth-first. When a compatible NP node is encountered, propose it as antecedent. If X is not the highest node, go to step 5.
- 5 From X, go up the tree to the first NP or S node encountered. Call it X and the path to reach it p.
- 6 If X is an NP node and if the path p to X did not pass through the nominal node that X immediately dominates, propose X as antecedent.
- 7 Traverse all branches below X to the left of p in left-to-right, breadth-first order. When a compatible NP node is encountered, propose it as antecedent.
- 8 If X is an S node, traverse all branches of node X to the right of p in left-to-right, breadth-first order, but do not go below any NP or S node encountered. When a compatible NP node is encountered, propose it as antecedent.
- 9 Go to step 4.

Syntactic-Tree Search Algorithm for Pronominal Coreference (Hobbs 1978)

- Pronomenresolution basierend auf syntaktischer Repräsentation, implementiert als Suche durch den Baum der das Pronomen enthält und die Vorgängerbäume:
- integriert Constraints zum Syntactic Binding
- integriert Präferenz für Antezedens in Subjektposition
- Präferenz für Recency (Distanz) durch die Suchrichtung vorgegeben
- braucht keine Semantik, aber komplette syntaktische Analyse
- abhängig von der jeweiligen Syntaxtheorie
- wir nehmen an, dass wir immer auch Agreement überprüfen

Syntactic-Tree Search Algorithm for Pronominal Coreference (Hobbs 1978)

Beispiele (Tafel):

John saw a driver in his truck.

John saw a driver of his truck.

John saw him.

A friend of John saw him.

John saw a friend of him.

Wiederholung: Centering Theory

- **Backwards Looking Center**, C_b , verbindet U_n mit der vorangegangenen Äußerung U_{n-1} .
- **Forward Looking Centers**, C_f , bilden einen potentiellen Link mit der folgenden Äußerung U_{n+1} .
- Die partielle Ordnung der C_f wird u.A. durch die grammatische Rolle des Referenzausdrucks bestimmt.
- Das höchste Element im C_f einer Äußerung ist das präferierte Zentrum C_p .
- Das C_b einer Äußerung U_n ist das am höchsten gewertete Element des C_f in U_{n-1} , das in U_n realisiert ist.

Zentrumstransitionen:

	$C_b(U_{n+1}) = C_b(U_n)$ or undefined $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

Regeln:

Regel 1: wenn ein Element von $C_f(U_n)$ als Pronomen in U_{n+1} realisiert ist, dann muss auch $C_b(U_{n+1})$ pronominalisiert sein.

Regel 2: Continue > Retain > Smooth-Shift > Rough-Shift

John benimmt sich seltsam

Gestern hat er Mike angerufen

Er war über Johns Anruf verärgert

→ inkohärent, weil Regel 1 verletzt wird

John benimmt sich seltsam

Gestern hat er Mike angerufen

Mike hat für seine Führerscheinprüfung gelernt

Er war über Johns Anruf verärgert

Anapherresolution im Rahmen der Centering Theory:

- 1 generiere C_b , C_f für alle möglichen Referenzzuweisungen (, die im Agreement übereinstimmen)
- 2 Filter (Agreement, Kontraindizierung, Regel 1)
- 3 ordne nach Regel 2 (d.h. löse Anapher so auf, dass die vorausgesagte Kohärenz des Textes maximiert wird)

Achtung: mögliche Referenzzuweisungen werden **nicht** primär danach sortiert, ob sich das Pronomen auf das präferierte Element des vorherigen Satzes bezieht! Das passiert erst bei gleichen Transitionen.

Brennan drives an Alpha Romeo.

$C_b = \text{undef}$, $C_F = \{\text{Brennan, Alpha Romeo}\}$, $C_p = \text{Brennan}$

She drives too fast.

$C_b = \text{Brennan}$, $C_F = \{\text{Brennan}\}$, $C_p = \text{Brennan}$

→ Continue

Friedman races her on weekends.

$C_b = \text{Brennan}$, $C_F = \{\text{Friedman, Brennan}\}$, $C_p = \text{Friedmann}$

→ Retain

She often beats her.

$C_b = \text{Friedman}$, $C_F = \{\text{Brennan, Friedman}\}$, $C_p = \text{Brennan}$

→ Rough-Shift

$C_b = \text{Friedman}$, $C_F = \{\text{Friedman, Brennan}\}$, $C_p = \text{Friedman}$

→ **Smooth Shift**

Grundidee: Berechne, wie aktiviert potenzielle Antezedenten sind.
Wähle als Antezedens den aktiviertesten Referenten.

- kein explizites semantisches oder pragmatisches Wissen
- mehrere Filter (Agreement,...)
- corpus-basiertes Parametertuning
- Wir berechnen Salienz für Entitäten (nicht für referentielle Ausdrücke)

Saliency Faktoren (unterschiedlich gewichtet)

- sentence recency \Rightarrow 100
- Subjektposition? \Rightarrow 80
- Existenzkonstruktion? \Rightarrow 70
- direktes Objekt? \Rightarrow 50
- indirektes Objekt? \Rightarrow 40
- head noun? \Rightarrow 80
- non adverbial PP? \Rightarrow 50

Für Pronomen

- Parallelismus \Rightarrow 35
- Kataphorik \Rightarrow -175

Ablauf (vereinfacht):

Für jeden Satz

- 1 update Diskursmodell (für alle nicht-Pronomen): errechne Salienzwert für jede Entität
- 2 für alle Pronomen
 - 1 identifiziere mögliche Antezedenten
 - 2 wende Filter an, um Antezedenten auszuschließen (Agreement, Kontraindizierung)
 - 3 ermittle Salienz für verbleibende Antezedenten (Salienzwert + Parallelismus/Kataphorik)
 - 4 wähle Antezedens mit höchstem Salienzwert und update das Diskursmodell (addiere Salienzwert des Pronomens für die entsprechende Entität)
- 3 halbiere alle bisherigen Salienzwerte

Evaluation:

- 86% Accuracy

Beispiel (Tafel):

John saw an Acura Integra at the dealership.

He showed it to Bob.

He bought it.

Herausforderungen:

Koreferenz durch semantische Relationen

*Peter bought a Ferrari. **It** is red.*

*Peter bought a new Ferrari. **The Ferrari** is red.* (Partial) Identität

*Peter bought a car. **The vehicle** is red.* Synonymie

*Peter bought a Ferrari. **The car** is red.* Hyperonymie

*Peter bought a car. **The Ferrari** is red.* Hyponymie (is-a)

We have two customers.

***The bold head** came an hour ago.* Meronymie

***The ham-sandwich** came an hour ago.* Metonymie

*Peter bought a Ferrari. **The beast** is red.* Metapher

Peter bought a Ferrari.

A door has a dent and the engine stops. (part-whole)

Papers were reviewed by a committee.

The chair was female. (set-member)

Mix the flour, butter, eggs and milk.

Knead the dough until smooth. (result/outcome of a process)

Peter crashed against a wall.

The noise woke up the neighbours. (cause-effect)

Peter is reading.

The book is exciting. (pred-arg)

Herausforderungen:

Pluralanaphern mit mehreren Antezedenten

John and Peter love their cars. They drive them every day.

John has a Ferrari and Peter has a Beatle. They drive them every day.

- A.63 I think it really depends a lot on the child, because **our daughter** is, was just a lot more levelheaded about her proc-, the process.
- B.64 Luckily I still have twelve more years to worry about it.
- A.65 Yeah [laughter]
- ...
- B.96 ... the University of Virginia. How much did it en-, end up costing?
- A.97 Uh [breathing], I think, uh, on a yearly basis, I'm trying to think. I would just make it a rough figure about, uh, with, with the travel expenses and so on, although **she** didn't come home that much, uh, actually.

A. I mean, the baby is like seventeen months and she just screams.

B. Uh-uh.

A. Well even if she knows that they're fixing to get ready to go over there. The're not there yet–

B. Uh-uh.

A. –you know.

B. Yeah. **It's** hard.

⇒ Nicht alle anapherähnlichen Ausdrücke haben ein eindeutiges Antezedens. Möglicherweise referieren sie eher auf das allgemeine “Diskurstopic”, aber ihre tatsächliche Bedeutung ist oft unklar.

When it comes to trucks, though, I would probably think to go American.

Es regnet.

Es gibt zwei Möglichkeiten.

⇒ Einige Pronomen sind nicht anaphorisch sondern nur grammatische „Slotfiller“.

Herausforderungen:

Anaphern mit nicht-nominalen Antezedenten

Laut Peter hat sich Anna ein neues Auto gekauft,

Aber das war falsch.

Aber das stellte sich als Lüge heraus.

Aber das war ein seltsame Art die Sache auszudrücken.

Das bereitete Anna große finanzielle Probleme.

⇒ speech act, proposition, event/state, etc.

Auch nicht-nominale Ausdrücke wurden als anaphorisch diskutiert, weil sie sich ähnlich wie Anaphern verhalten:

- zeitliche und räumliche Anaphern ('dort', 'dann')
- Tempus/Modus
Peter entered the room. He turned on the light.
A wolf might come. He would eat you.
- Diskurskonnectoren, Diskursmarker
If there is a red light, stop.
Otherwise, you can go on.
Otherwise, you might get a fine.

- Anaphorische Ausdrücke sind in menschlicher Sprache weit verbreitet.
- Anapherresolution ist entscheidend für die Interpretation und Generierung von Diskursen.
- Die meisten Ansätze beruhen auf ähnlichen Grundannahmen. Dabei kommen verschiedenen Faktoren zusammen:
- Morphologische und syntaktische Oberflächenfeatures spielen eine wichtige Rolle.
- Weltwissen ist zwar wichtig, automatische Anapherresolution erzielt auch ohne Weltwissen brauchbare Ergebnisse.
- Pronominale Koreferenzen und einfachere Fälle von Koreferenzen mit definiten NP können robust gelöst werden. (Präzision und Recall ca 85%)
- Einige Anaphertypen sind in der Forschung noch wenig behandelt worden (z.B. Demonstrativpronomen)
- Einige Phänomene stellen immer noch eine Herausforderung dar.