Einführung in die Pragmatik und Diskurs: Anapherresolution

I. Kruijff-Korbayova/A. Horbach

Universität des Saarlandes

Sommersemester 2011

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

Referenzresolution

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?

Übersicht

Vorlesungsplan

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

Kernlektüre

- Jurafsky & Martin (2009), Kapitel 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
- Referenzausdrücke, die auf dieselbe Entität verweisen, sind **koreferent**
- **Referent** ist die Entität, auf die sich ein Sprecher mit einem Referenzausdruck bezieht.

Deiktische vs. Anaphorische Referenzen

Deixis und Anapher

- **Deiktische Referenz (Deixis):** Referanz auf eine Entität im situativen Kontext des Äußerung z.B.
  - auf den Sprecher (*ich/wir, mein/unser*)
  - Hörer (*du/Sie, ihr/sie, Frau X /Herr X, Herr Professor, meine Damen und Herren*)
  - Ort (*hier, da, dort*),
  - Zeit (*jetzt, heute, morgen, dieses Jahr*);
- **Anaphorische Referenz (Anapher):** Referenz auf eine Entität durch den Verweis auf eine vorher erwähnte Entität, realisiert durch ein Antezedens
  - Koreferenz (identity of reference): 'alte' Diskursentität
  - Bridging/Assoziation (Inferenz): 'neue' Diskursentität

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
- **Koreferenz vs. Anaphorik**
  - cross-document coreference (=nicht anaphorisch)
  - Anaphern, die nicht koreferent sind (*Everybody has his own destiny.*)

Koreferenzresolution vs. Anapherresolution

Koreferenzresolution: finde die Koreferenzketten in einem Text.

Anapherresolution: finde das Antezedens einer Anapher.

Beispiel: Koreferenzresolution

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}

Beispiel: Anapherresolution

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* $\Rightarrow$ *Sophia Loren*
- *the actress* $\Rightarrow$ *Sophia Loren*
- *the U2 singer* $\Rightarrow$ *Bono*

Koreferenzresolution

Schwierigkeiten:

- unterschiedliche Form $\not\Rightarrow$ verschiedene Referenten
  (*Sophia Loren* vs. *the actress* vs. *she*)
- gleiche Form $\not\Rightarrow$ gleiche Referenten
  (*die Katze*, *Michael Jackson* der Sänger vs. *Michael Jackson* der britische General)

Ambiguität und desambiguierende Faktoren

Jane told Peter he was in danger.
⇒ Kongruenz (Genus, Numerus etc.): *he* = *Peter*

Peter told John that he is running the business for himself.
⇒ *syntactic constraints: himself* = *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*

Ambiguität und desambiguierende Faktoren

Tony Blair met President Yeltsin. The old man had just recovered from a heart attack.
$\Rightarrow$ Weltwissen: *the old man* = *Yeltsin*

Jan hat ein Kaninchen. Hans hat ein Pferd. Maria füttert es.
$\Rightarrow$ *Recency (Distanz): es* = *Pferd*

Jan kaufte Hans ein Computerspiel. Maria kaufte ihm ein Buch.
$\Rightarrow$ Parallelismus: *ihm* = *Hans*

Ambiguität und desambiguierende Faktoren

Probleme

- Selektionsbeschränkungen können verletzt werden:

  (1) Hans hat ein Computerspiel gekauft. Es läuft nicht.

- Es ist unklar, inwiefern sich diese Präferenzen auf andere anaphorische Ausdrücke (nicht Personalpronomen) anwenden lassen:

  (2) Jan besuchte die Messe mit Hans. Er kaufte ein Computerspiel.

  (3) Jan besuchte die Messe mit Hans. Der kaufte ein Computerspiel.

Motivation

- Theoretisches Interesse: Erklärung, wie Refenzen 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: Ananphernresolution notwendig zur korrekten Interpretation, Generierung von anaphorischen Ausdrücken zur Effizienz und Natürlichkeit

Referenzresolution

- Aufgabe: Berechne die Bedeutung eines Referenzausdrucks:
- Wesentlich zur Konstruktion eines Diskursmodels:
  - evoke (introduce) "new" discourse referents
  - access "old" discourse referents
- 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 Antezendenten
  3. finde passendes Antezedens für jede Anapher

Ansätze zur Referenzresolution

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)

- Pronomenresolution basierend auf syntaktischer Represäntation, 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

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.
3. Traverse all branches to the left below X in left-to-right breadth-first order. Propose as antecedent any compatible NP node which has an NP or S node between it and X.
4. If X is the highest 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. 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. Propose any compatible NP node as the 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. Propose any compatible NP node as the antecedent.
9. Go to step 4.

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

Examples:

(4) The castle in camelot remained the residence of the king until 536, when he moved it to London.

(5) a driver in his truck

(6) a driver of his truck (illustrates when rule 6 blocks a candidate)

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 potientiellen Link mit der folgenden Äußerung U_{n+1}.
- Die partielle Ordnung der C_f wird u.A. durch die grammatische Role 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.

Brennan et al. 1987

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

Brennan et al. 1987

Anapherresolution im Rahmen der Centering Theory:

1. generiere C_b, C_f für alle möglichen Referenzzuweisungen
2. Filter (Selektionspräferenzen, Regel 1, . . .)
3. ordne bei Regel 2 (d.h. löse Anapher so auf, daß die vorausgesagte Kohärenz des Textes maximiert wird)

Lappin & Leass (1994)

- kein explizites semantisches oder pragmatisches Wissen
- mehrere Filter (syntaktisch, morphologisch etc.)
- saliency-based discourse model (wie stark ist ein potentieller Antzendens aktiviert?)
- corpus-basiertes Parametertuning

Saliency Faktoren (unterschiedlich gewichtet)

- wann erwähnt? (sentence recency) $\Rightarrow$ 100
- Subjektposition? $\Rightarrow$ 80
- Existenzkonstruktion? $\Rightarrow$ 70
- direktes Objekt? $\Rightarrow$ 50
- indirektes Objekt? $\Rightarrow$ 40
- head noun? $\Rightarrow$ 80
- non-PP? $\Rightarrow$ 50

Lappin & Leass (1994)

Ablauf (vereinfacht):

1. identifiziere mögliche Antezendenten
2. wende Filter an, um Antezendenten auszuschließen
3. ermittle saliency Werte für verbleibende Antezedenten
4. wähle Antezedens

Evaluation:

- 86% Accuracy

Koreferenresolution - Ergebnisse

Aktuelle Ergebnisse (basierend auf verschiedenen Autoren):

NP form	Precision	Recall	F-measure
All	83%	53%	65%
Definite NPs	66%	21%	33%
Proper names	94%	62%	75%
Demonstrative NPs	23%	23%	23%
Personal pronouns	85%	85%	85%
Possessive pronouns	80%	85%	82%

Herausforderungen

- Koreferenzen und Bridging involvieren lexikalische/konzeptuelle Relationen
- Anaphern mit geteilten/mehreren Antezedenten
- Vage Anaphern
- Pleonastische Pronomen (nicht anaphorisch)
- Nicht-nominale Antezedenten
- Nicht-nominale Anaphern
  - z.B, Adverben, Diskurskonnektoren
  - Verb-Substitutionen end Ellipsen
  - Tense and mood
- Was über reinen Text hinausgeht:
  - Dialog
  - Multimodaler Kontext
  - gesprochene Sprache

Koreferenzen

(7) Peter bought a Ferrari. It is red.

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

(9) Peter bought a car. The vehicle is red. Synonymie

(10) Peter bought a Ferrari. The car is red. Hyperonymie

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

(12) We have two customers.

a. The bold head came an hour ago. Meronymie

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

(13) Peter bought a Ferrari. The beast is red. Metapher

Bridging

(14) Peter bought a Ferrari.
A door has a dent and the engine stops. (part-whole)

(15) Papers were reviewed by a committee.
The chair was female. (set-member)

(16) Mix the flour, butter, eggs and milk.

a. Kneed the dough until smooth and shiny.

b. Spread the paste in the baking form.

c. Stir the batter until all lumps are gone.

d. Let it rest. (result/outcome of a process)

(17) Peter crashed against a wall.
The noise woke up the neighbours. (cause-effect)

(18) Peter is reading. The book is exciting. (pred-arg)

Pluralanaphern mit mehreren Antezedenten

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

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

$\Rightarrow$ Sets of entities evoked by discontinuous expressions in text.
Note also that sometimes there is an ambiguity between the "set-reading" ("collective") and a "distributed" reading.

Long Distance Anaphors

(21) 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.

Vage Anaphern

(22) A. I mean, the baby is like seventeen months and she just screems.

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 Antezendens. Möglicherweise referieren sie eher auf das allgemeine "Diskurstopic", aber ihre tatsächliche Bedeutung ist oft unklar.

Pleonastic/Expletive Pronouns

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

(24) Es regnet.

(25) Es gibt zwei Möglichkeiten.

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

Anaphern mit nicht-nominalen Antezedenten

z.B. Diskursdeixis:

(26) A. . . . we never know what they are thinking.

B. that's right. I don't trust them, maybe I guess it's because of what happened over there with their own people, how they threw them out of power . . .

(sw3241)

(27) Now why she didn't take him over there with her? No, she didn't do that.

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

Nicht-nominale Anaphern

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

- zeitliche und räumliche Anaphern ('dort', 'dann')
- Tempus/Modus

  (28) Peter entered the room. He turned on the light.

  (29) A wolf might come. He would eat you.

- Diskurskonnektoren, Diskursmarker

  (30) If there is a red light, stop.

  a. Otherwise, you can go on.

  b. Otherwise, you might get a fine.

Zusammenfassung

- 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 eione 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. (Prezision 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.