Anaphora Resolution

Vašek Němčík
(xnemcik@fi.muni.cz)

CZPJ, FI MU Brno

4. prosinec 2009

Plán

- Stručný úvod (co a proč)
- Nejzajímavější systémy a algoritmy
- Saara

Co je to “anafora”?

- “The thousand injuries of Fortunato$_i$ I had borne as I best could; but when **he$_i$** ventured upon insult, I woved revenge. ... It was about dusk, one evening during the supreme madness of the carnival season, that I encountered **my friend$_i$**. **He$_i$** accosted me with excessive warmth, for **he$_i$** had been drinking too much. **The man$_i$** wore motley. **He$_i$** had on a tight, parti-striped dress ...”
- teoreticky velmi složité
- diskurs, referující výrazy, reference
- anafora, anafora, antecedent, anaphora resolution

Defince AR jako úkolu

- nalézt v textu anaforické výrazy
- určit ke každému z nich antecedent
- určit typ vztahu
  - koreference (stejná entita)
  - bridging (asociativní/nepřímá anafora)
    obecný sémantický vztah mezi entitami

Nejzajímavější:

- pronominální koreference
- textová anafora

AR Systémy

knowledge-poor

- “kacířství motivované praktickými potřebami”
- RAP, Kennedy&Boguraev, CogNIAC, MARS
- jednoduchá (snadno spočítatelná) data (morfologické značky a povrchová syntax)
- salience-based algoritmy

strojové učení a statistika

- korpusy anotované na koreferenci jako tréninková data
- AR ale není klasifikační problém → třeba “překódovat”
- McCarthy and Lehnert, 1995: *Antecedent Anaphor Y/N*

Saara

- (System for Automatic Anaphora Resolution and Analysis)
- možnost pracovat s různými zdroji dat
  - PDT 2 (jediná česká anotovaná data)
  - Synt (⇢ analýza volného textu)
  - *BNC (⇢ Sketch Engine)*
  - ...
- přesto ale používání stejných algoritmů a vyhodnocování

Saara

- 2 roviny abstrakce:
  - markable rovina
  - rovina konkrétní technické struktury věty
- algoritmy pracují pouze s markables (+jejich atributy) a vybranými metodamy, které zprostředkovávají informace o struktuře věty
- Metody:
  - seřaď podle gramatické role
  - markable je subkonstituent/závislý
  - shoda v morfologii apd.

Saara

Struktury

- text – věta – klause (+konkrétní realisace věty)
- markable (množina tokenů)
- vztahy mezi markables
- množiny vztahů (positivní/negativní; ekvivalence)

3. rovina abstrakce (“supervisor layer”, Byron & Tetreault)

- krátký program, který pouze definuje, které moduly se jak použijí
- “skoro deklarativní”

Saara

AR

- načtení dokumentu (z konkrétního formátu)
- pre-processing
  - rozdělení do klausí
  - detekce nevyjádřených subjektů
  - model diskursu
    detekce markables (referujících výrazů)
- AR: model diskursu ⇢ ekvivalenční třídy nad markables
- výstup: MMAX2 XML

Vyhodnocení

- reimplementované algoritmy

MUC-6	Precision	Recall	pers.	dem.
Recency	41,8	37,3	45,9	13,4
Hajičová 87	41,3	36,8	46,7	14,4
HHS 95	41,3	36,8	48,1	16,7
Hobbs	38,9	33,9	30,6	9,2
Centering BFP	52,2	39,2	45,9	16,1
Lappin&Leass RAP	49,9	46,3	53,8	27,2

Výhled ...

- experimentování
- BNC vertikály – Sketch Engine
- valence
- strojové učení
- ...

Co je anafora?

(bum – bác – @#$%&#%)*

A: Co je to?!

B: To je anafora.

A: Co je anafora?

B: Ne, není.

A: To není anafora?

B: Ne ne, to je anafora!

A (s povzdechem): OK, takže co není anafora?

B: Přesně tak!