Orientatie Alfa-informatica
Computer-taalkunde

Gosse Bouma

www.let.rug.nl/~gosse/orient/

Overzicht

- Taaltechnologie
  - Toepassingen,
  - *Computer-taalkunde,*
  - Woorden en reguliere expressies
  - Wat je ook doet, de semantiek gooit roet…
- Corpustaalkunde,
  - Zoeken en tellen in teksten
  - Corpus Internet

Taaltechnologie?

- ICT-toepassingen waar kennis van taal een rol speelt:
  - spellingcorrectie
  - tekst naar spraak (demo Fluent Dutch)
  - automatisch vertalen (demo Alta Vista)
  - dialoogsystemen (intelligente *voice response*) (NS reisinformatie)
  - rapporten genereren (weerbericht, beursnieuws)

Meer toepassingen

- Spraakherkenning (Philips FreeSpeech, Lernhout & Hauspie,...),
- Intelligente Information Retrieval (concepten, morfologie, multilinguaal, multimediaal),
- Document (email) classificatie,
- Samenvatten.

Wat is computer-taalkunde?

- Taalkundig onderzoek met behulp van de computer:
  - *taaltechnologie,*
  - *testen* van taalkundige theorieën,
  - *automatisch leren* van taalkundige kennis.

Spellingcorrectie

- Fouten vinden is tamelijk eenvoudig
- Correcties voorstellen is lastiger:
  - onmiddelijk → onmiddellijk
  - pselling → spelling
  - pijnzen → pijnzin, peinzen
  - slaolm → slalom, slaolie, slakom
  - kompjoeter → computer
  - N.B. MS Office accepteert *pijnzen* en *slaolm*!

Woorden

- (Bijna) iedere toepassing maakt gebruik van een woordenboek
- Sommige toepassingen bestaan vrijwel alleen uit een woordenboek:
  - spellingcorrectie
  - afbreken
  - tekst naar spraak
  - spraakherkenning
  - vertaalhulp

Hoeveel woorden zijn er?

- Groene Boekje : 125K
- Words-L (public domain woordenlijst voor spellingcorrectie en afbreken): 250K
- Celex (lexicale database) : 325K
- Van Dale:…

Geen woordenlijst is volledig.

- Voorbeeld:
  - *Eindhoven corpus*
  - 1 mln woorden.
  - 40% hiervan ontbreekt in Celex
- Toepassingen die alleen gebruik maken van een woordenlijst gaan dus vaak de fout in:
  - false alarms (spellingcorrectie)
  - afbreekfouten

Afbreken

- Op basis van lettergreepstructuur:
  - *af-bre-ken*, niet *afbr-eke-n*
- Maak het begin van de lettergreep zo lang mogelijk:
  - *ha-mer*, niet *ham-er*
  - *al-fa-bet*, niet *alf-ab-et*
- Met inachtneming van morfeemgrenzen:
  - *lamp-licht*, niet *lam-plicht*
  - *fietslamp-je* vs. *slagboom-pje*

Afbreekalgoritme:

- Verdeel een woord in samenstellende delen (morfemen),
- Verdeel de delen in lettergrepen,
- Zorg ervoor dat het begin van de lettergreep zo lang mogelijk is.

‘Stemming’

- herleiden van een woord tot een stam
  - fietsen, fietste, gefietst --> fiets,
  - lopen, gelopen, liep --> loop
  - varken --> varken
- nuttig voor veel toepassingen
  - information retrieval,
  - zinsontleden, ….
- Demo : Xerox

Woordsoorten

- Benoemen van woorden op woordsoort (zelfst nw, ww, bijv nw, vz, lidw, …)
- herleiden van een woord tot een stam
  - fietsen --> fiets,
  - leven --> leef
  - varken --> varken
- nuttig voor veel toepassingen
  - zinsontleden,
  - automatisch vertalen,
  - information retrieval

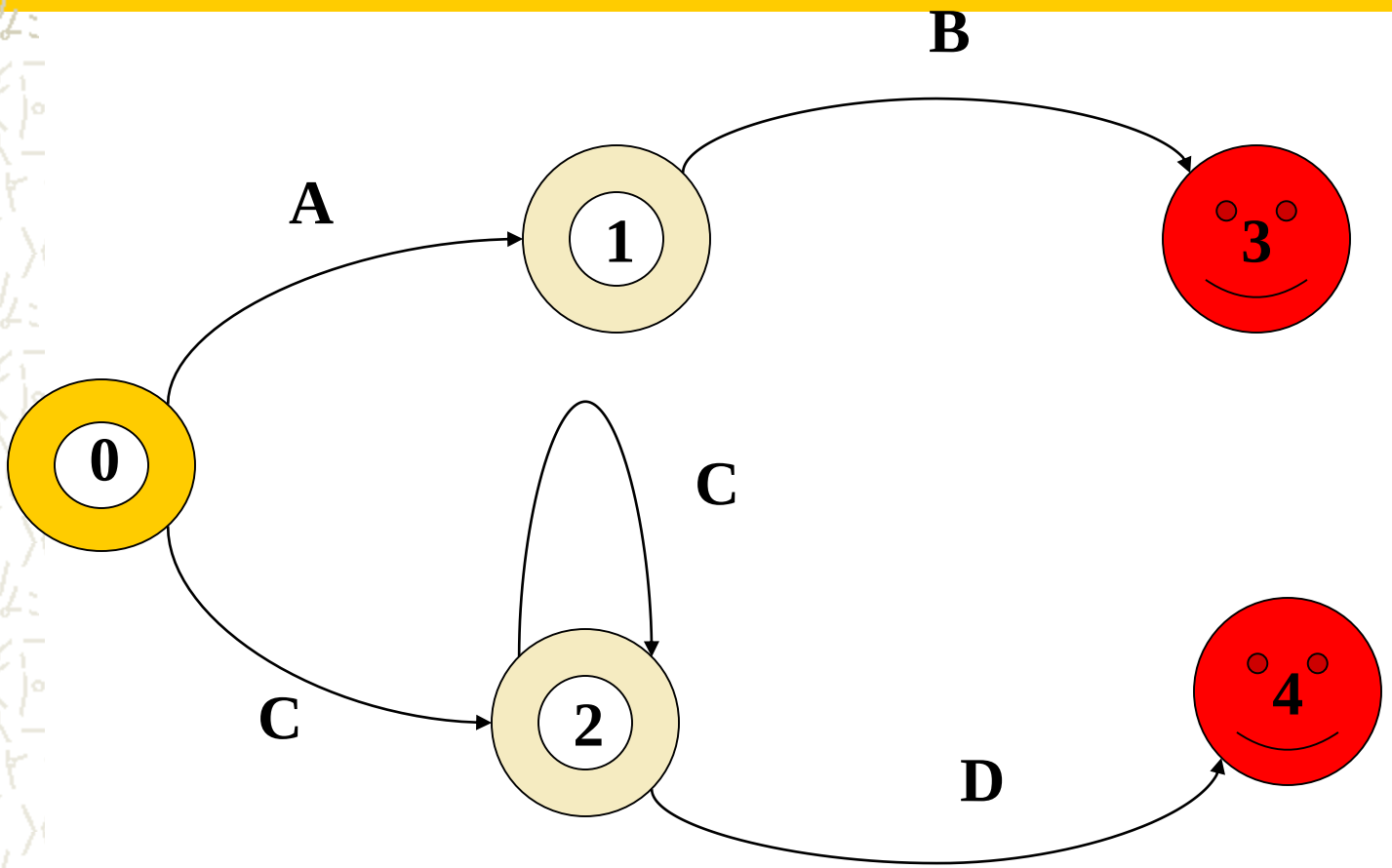
Part-of-Speech tagging

- fiets -> zelfst. nw? werkwoord?
- fietsen -> zelfst. nw? werkwoord (infinitief, ott mv?)
- De fietsen staan in de schuur.
- We fietsen naar school.
- Maak gebruik van de woorden in de context om de juiste categorie te bepalen.
- Demo: ilk.kub.nl

Eindige Automaten

- De eenvoudigste machines om taal (reeksen symbolen) te verwerken zin eindige automaten.
- Een automaat bestaat uit
  - een aantal toestanden
  - transities
  - een begintoestand
  - één of meer eindtoestanden

Eindige Automaten



Reguliere expressies

- Handige manier om automaten te definiëren.
- A* = nul of meer A's
- A+ = één of meer A's
- [A, B] = een A gevolgd door een B
- {A, B} = een A of een B
- [A, B^] = een A optioneel gevolgd door een B
- etc….

Reg Ex voor woordsoorten

- Bijvoeglijke naamwoorden:
- o.a. woorden die *eindigen op -e*
- [? *, e]
- *maar zelfst. nw eindigen vaak op -je!*
- [?*, e] & ~ [? *, j, e]
- [? * , e] - [?*, j, e]
- [?*, ? - j, e]

Opdracht 1

- www.let.rug.nl/~gosse/orient/
- Probeer de reguliere expressie demo op het web
- Bedenk een reg ex voor bijv. nw'en,
- Test op een willekeurige verz. Voorbeelden

Wat je ook doet, de semantiek gooit roet

- Makkelijk ?
  - Spellingcontrole,
  - Afbreken,
  - OCR,
  - Tekst naar spraak,
  - Information Retrieval,
  - Voice Response,
  - Part of Speech tagging,
  - Samenvatten,
  - Rapporten genereren.
- Moeilijk?
  - Grammaticale controle (d/t fouten)
  - Dicteersystemen (grote woordenschat)
  - Volledige syntactische en semantische analyse
  - Automatisch vertalen

Corpustaalkunde

- Corpora (verzamelingen tekst) bevatten veel nuttige informatie over het gebruik van taal,
- Tekst is elektronisch beschikbaar op CD of via Internet,
- Corpustaalkunde probeert taalkundige kennis te vinden in corpora,
- Bijna alle taaltechnologie maakt gebruik van corpora.

Onderzoeksmethode 1: “Literatuuronderzoek”

- Verzin een vraag,
- Lees boeken en artikelen over het onderwerp,
- (doe wat denkwerk,)
- Doe verslag van je bevindingen

Onderzoeksmethode 2: “Data-gestuurd, Experimenteel”

- Verzin een vraag,
- **verzamel data, bedenk een experiment**
- **tel, experimenteer, vergelijk,**
- (doe wat denkwerk),
- Doe verslag van je bevindingen

- Corpusonderzoek is een manier om deze methode binnen de letteren te hanteren.

Corpora

- Veel vragen over taal kun je alleen/beter beantwoorden door te kijken naar echt taalgebruik.
- Corpus: een verzameling tekst of gesproken taal
- B.v. British National Corpus:
  - 100 mln woorden,
  - Allerlei tekstsoorten, stijlen, auteurs
  - Voorzien van woordsoort

Voorbeeld: “X laat zich”

From: gj@cogsci…

To: gosse@let.rug.nl, …

Subject: Vraag

Gegeven voorbeelden als

“de deur laat zich openen met een sleutel”

“de auto laat zich starten door contact te maken”

Heeft een van jullie dan het gevoel “vrije wil” (bewust of onbewust) aan de sleutel/ de auto toe te kennen?

“Laat zich” in Eindhoven corpus

het laat zich verstaan dat het afzoeken van….

Dat het gevaarvolle avontuur slaagt , laat zich voorspellen

Het bedenken van een dergelijke naieve gewapende overval laat zich moeilijk verenigen met…..

de cassette laat zich net zo gemakkelijk inbrengen en uitnemen

De combinatie van schone stad en industriestad laat zich moeilijk rijmen

Uit de aantekeningen laat zich reconstrueren hoe onze schrijfster…

Hij slaat zich verwoed op een knie

Ongeveer 25 resultaten (2 false positives), met persoon/abstractum/dat-zin/apparaat als onderwerp

Zoeken op het Web

- Het web is niet zonder meer geschikt voor corpusonderzoek,
- Maar bevat wel veel data (ook voor minder courante talen)
- WebCorp: www.webcorp.org.uk
- Netkwic: www.let.rug.nl/vannoord_bin/netkwic
- Search-engine die tekstfragmenten als resultaat oplevert.

Corpus Internet

- Nederlands Corpus Internet is naar schatting meer dan 100 mln woorden groot. (Oostendorp & VdWouden, ts. Ned. Taalkunde, 1998)
- Is alweer een Nederlands woord, en weeral Vlaams?
- Is “best wel” gebonden aan bepaalde registers?
- (nieuwe woorden, tussenklanken,…)

Opdracht 2: Corpus internet

- Bekijk de vragen over spelling en betekenis bij de Taaladviesdienst (Onze Taal)
- Zoek voor een probleemwoord op het Web naar de verhouding tussen goede en foute spelling,
- Zoek op het Web naar voorbeelden van woorden met een moeilijke betekenis.