Taal- en Informatietechnologie

Emiel Krahmer &
Antal van den Bosch

{E.J.Krahmer,antalb}@uvt.nl

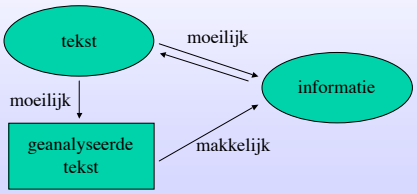
eerste semester 2003-2004

Taal- en Informatietechnologie

- Informatietechnologie:
  - informatie halen uit tekst
  - samenvatten, classificeren
  - *slimme* search engines
- Taaltechnologie:
  - zinnen analyseren
  - dialogen voeren
  - informatie omzetten in tekst

Informatie en tekst

- Informatie zit verpakt in taal



Over de cursus

- BDM/TKI
- 2001/2002 "Automatische bewerking en verwerking van informatie"
- colleges en open-boek tentamen
- twee bonustaken (1/2 punt per stuk)
  - informatietechnologie
  - taaltechnologie
- Blackboard
- http://ilk.uvt.nl/~antalb/tint/

Over de cursus (2)

- dinsdagen 14.45 - 16.30 AZ01
- laatste college 2 december

- Emiel Krahmer, R104
  - E.J.Krahmer@uvt.nl
- Antal van den Bosch, R116
  - Antal.vdnBosch@uvt.nl

Over de cursus: opbouw

- week 1: introductie
- week 2: basic NLP
- week 3: document retrieval
- week 4: information extraction
- week 5: NLP voor IE I
- week 6: NLP voor IE II

Over de cursus: opbouw (2)

- week 7: text categorization
- week 8: text & web mining
- week 9: named entities & coreference
- week 10: question answering
- week 11: spraaktechnologie
- week 12: de praktijk: Textkernel

Scope van de cursus

- Inzicht in state of the art in taal- en informatietechnologie
  - wetenschappelijk onderzoek
  - toepassingen
- Niet:
  - spraaktechnologie (zijdelings)
  - visuele informatietechnologie
  - bibliotheekwetenschappen

Jackson & Moulinier

- "Natural language processing for online applications"
- NLP: analyse / synthese van gesproken en geschreven taal
- I.t.t. programmeertalen is NL ambigu
  - *Visiting aunts can be a nuisance*
  - *She boarded the airplane with two engines*

NLP

- Informatieniveaus waarop ambiguiteit opgelost kan worden:
  - Syntax en semantiek
  - Pragmatiek en context (van gebruik)
- Twee visies:
  - Good old-fashioned AI, logica, formele/mathematische taalkunde
  - Statistische NLP en lerende systemen

Rode draad

- Om accurate NLP te doen is begrip nodig
- Keuze:
  - Diep modelleren: is lastig (AI is erop kapot gelopen)
  - Oppervlakkig modelleren, korte bochten, trucs, heuristieken: werkt redelijk tot verrassend goed

"Bluff your way into..."

- "Technologie" ~ toepassingen
- Informatietechnologie:
  - IR = information retrieval
  - IE = information extraction
  - a.k.a. text mining, web mining (analogie met data mining)
  - question answering
  - document classification

Bluff your way into IR

- Relevante documenten vinden
  - zonder rommel ertussen (precision)
  - zoveel mogelijk (recall)
- Web search: grootschalig prutsen met booleaans zoeken
  - gebruikers accepteren lage p & r
  - "wonder" Google gebruikt *hubs & authorities* theorie
- Weinig taaltechnologie

Bluff your way into IE

- Informatie halen uit tekst
  - zelfde precision & recall doel
- Text mining:
  - www.flipdog.com
  - Medisch voorbeeld (Swanson, 1991)

IE: medisch voorbeeld

- Kennis uit titels medische artikelen
  - Stress is associated with migraines
  - Stress can lead to loss of magnesium
  - Calcium channel blockers prevent some migraines
  - Magnesium is a natural channel blocker
  - Spreading cortical depression (SCD) is implicated in some migraines
  - High levels of magnesium inhibit SCD
  - Migraine patients have high platelet aggregability
  - Magnesium can suppress platelet aggregability
- Magnesium deficiency ~ migraines!

Bluff your way into web mining

- Analoog aan data mining
  - bier en luiers
- Tekst is informatie, maar slecht doorzoekbaar: omzetten naar DB
- Kennis opbouwen over domein
  - ...docent van cursus ...
  - ...ik volg de cursussen ...
  - ...lid van vakgroep ...
  - ...onderdelen van BAMA...

Bier en luiers

Bluff your way into the future

- Question answering
  - wie vond de telefoon uit?
  - waarom is John F. Kennedy vermoord?
- Automatisch samenvatten
- Gepersonaliseerde informatie-filtering

Bluff your way into LT

- Diepst doordrongen: MS Word
  - Maar begin er niet over!
- Spelling- en stijlchecking
- Automatisch vertalen: babelfish (Systran)
- Mens-machine dialoogsystemen
- Verborgen technologie in informatietechnologie (AI included)

Bluff your way into dialogue

- Turing test (1950)
- Elisa (Weizenbaum, 1966)
- Loebner prize (1990s-present)
- maar:
  - doel is converseren, niet begrijpen
  - "echte" dialoog veronderstelt begrip
- Reisinformatie per sprekende computer

Elisa

- P: men are all alike.
- E: in what way?
- P: they are always bugging us about something or other.
- E: can you think of a specific example?
- P: well, my boyfriend made me come here.
- E: go on.
- P: perhaps I could learn to get along with my mother.
- E: tell me more about your family.
- ...

Bluff your way into 2001

- (Kubrick, 1969)
- HAL (=IBM) 9000 wordt opgeleverd op 12 januari 1997
  - bestuurt ruimteschip
  - prettige stem
  - speelt schaak
  - maakt nooit fouten

En na 2001

- Speech-to-speech translation (VERBMOBIL)
- Sprekende helpdesks
- Ambient intelligence
- Communicatieprotocollen huishoudelijke apparaten