Natuurlijke-Taalverwerking I

Gosse Bouma en Geert Kloosterman (pract)

2e semester 2005/2006

Overzicht

- Week1 :
  - Inleiding, Context-vrije grammatica.
- Week 2-3 : Definite Clause Grammar
  - Regels, gebruik van variabelen, parse-bomen, betekenis, ..
- Week 4-5 : Automatisch Ontleden
  - Top-down vs bottom-up, shift-reduce en chart parsing, ...
- Week 6-7 : Unificatie-grammatica
  - Feature-structuren en unificatie, macro's, vraag-zinnen, ...

Studiehandleiding

- Zie Nestor of `www.let.rug.nl/~gosse/ntv1`
- Links naar de syllabus, aanvullende literatuur, college-aantekeningen, practicumopdrachten
- Practicum start volgende week
- Beoordeling
  - Practicum (4 opdrachten, 50%)
  - Tentamen (50%)
  - Beide onderdelen moeten voldoende zijn

Wat is natuurlijke-taalverwerking?

- Het ontwikkelen van programma's en toepassingen waarbij kennis van de structuur en de betekenis van natuurlijke taal een rol speelt.

Voorbeelden

- Grammatica-correctie (identificeren en corrigeren van grammaticale fouten in tekst),
- Automatisch vertalen
- Automatisch e-mail beantwoorden,
- Informatie Extractie
  - ⋆ Google `define citroenzuur`
    `Citroenzuur is een zwak organisch zuur. Het komt in citrusvruchten voor en is een natuurlijk conserveermiddel en antioxidant. Daarnaast wordt het gebruikt om een zure smaak aan voedsel te geven. ..`

Automatisch vertalen....

Term is de erkenning een uitdaging in aan het vakgebied verbonden informatieherwinning geweest. De terugwinning van informatie baseert zich zwaar op de identificatie van de relevante concepten, welke door termijnen worden vertegenwoordigd.

...is moeilijk

Term recognition has been a challenge in domain-specific information retrieval. The discovery of knowledge relies heavily on the identification of relevant concepts, which are represented by *terms*.

Term is de erkenning een uitdaging in aan het vakgebied verbonden informatieherwinning geweest. De ontdekking van kennis baseert zich zwaar op de identificatie van relevante concepten, die door termijnen worden vertegenwoordigd. `babelfish.altavista.com, worldlingo`

...is moeilijk

De termijn erkenning is een uitdaging in domijn-specifiek informatie inwinnen geweest. De ontdekking van kennis steunt op zwaar op de identificatie van relevante begrip, die door termijnen vertegenwoordigd worden. `ets.freetranslation.com`

Term is de erkenning een uitdaging in aan het vakgebied verbonden informatieherwinning geweest. De ontdekking van kennis baseert zich zwaar op de identificatie van relevante concepten, die door termijnen worden vertegenwoordigd.

Makkelijk en Moeilijk

Makkelijk	Moeilijk
Spellingcontrole	Grammaticale controle
Voice Response	Volledige spraak-herkenning
Rapporten genereren uit tabellen	Samenvatten van artikelen
Vertaalhulp	Automatisch vertalen
Domein-specifieke dialoogsystemen	Turing-test
Web-search	Automatisch vragen beantwoorden

Automatisch vragen beantwoorden (2 voor 12)

- `Wie is de voorzitter van het Europese Parlement?`
- Klaus Hänsch , voorzitter van het Europese Parlement , drukte het iets sterker uit...
- Ook de voorzitter van de CDA-delegatie in het Europese Parlement , oud-minister Maij ....

Automatisch vragen beantwoorden

- `Wanneer vond de Duitse hereniging plaats?`
- Sinds de Duitse hereniging in oktober 1990 is de sterfte in Oost-Duitsland sterk toegenomen.
- Al in 1962 voorspelde hij de Duitse hereniging en het uiteenvallen van de Sovjet-Unie.

Zelfs makkelijke toepassingen zijn moeilijk

- Spellingcorrectie:
  - ⋆ Lijkt gemakkelijk: markeer alle woorden die niet in het woordenboek staan,
  - ⋆ Maar is moeilijk: geen woordenboek is volledig, iedere dag worden nieuwe woorden geïntroduceerd.

Omvang van een Woordenboek

- 125K (*Groene Boekje*)
- 500K+ (*van Dale*).
- Soms ontbreekt 40% van de woordtypes in een tekst in het woordenboek.
  - ⋆ Tokens: aantal woorden in een tekst,
  - ⋆ Types: aantal verschillende woorden in een tekst.

Meer cijfers

- Kun je een goede woordenlijst afleiden uit een corpus (verzameling tekst):

Woorden	Corpus	OOV
20K	110M	6.6%
40K	145M	4.5%
60K	125M	3.6%

- OOV = *out of vocabulary rate*, aantal woordtokens dat niet in het woordenboek staat.

Spellingcorrectie $\neq$ opzoeken

- Deze jongen vind(t) je aardig.
  - ⋆ (Daarom wil hij een date.)
  - ⋆ (Daarom wil jij een date.)
- Wel/geen spelfout hangt af van betekenis.

Brandt Corstius

- De derde wet van de computer-taalkunde:
  - ⋆ Na een bepaalde tijd, bv 1 jaar, werken, krijg je 80% goede resultaten. Elke halvering van de *gap* tussen 80 en 100% betekent een vermenigvuldiging van de aanvankelijk bestede tijd met een vaste factor die groter is dan 1.
- Wat je ook doet, de semantiek gooit roet.

Rol van de Taalkunde

- Kennis van taal en spraak:
  - de structuur van woorden (morfologie),
  - uitspraak (fonologie),
  - zinsbouw (syntaxis),
  - betekenis (semantiek).

Computationele Taalkunde

- Het gebruik van de computer voor taalkundig onderzoek:
  - Komt de woordvolgorde *heeft geslapen* vaker voor dan *geslapen heeft*?
  - Welke regels voor het toekennen van klemtoon aan Nederlandse woorden werken het beste?
  - Kun je met Machine Learning betere regels vinden?

Rol van de Informatica

- Technieken en algoritmen:
  - eindige (*finite state*) automaten (voor snel analyseren en herkennen van strings),
  - parsers voor context-vrije grammatica, ...

Rol van de Kunstmatige Intelligentie

- Taalverwerking is een aspect van menselijke intelligentie,
- Natuurlijke-taalverwerking modelleert een aspect van menselijke intelligentie,
- Technieken: *Machine Learning*, Ontologische kennis en netwerken

Rol van de Informatiekunde

- Slim coderen (m.n. in XML) van informatie (woordenboeken, grammatica's, tekst- en spraakcorpora, *treebanks*, etc.)
- Toepassingen: Samenvatten van web-content, Informatie extractie, on-line vertalen, multilinguale zoekmachine's, etc.

Grammatica

- Veel toepassingen vereisen een kennis van de structuur van zinnen (zinsbouw, grammatica):
  - ⋆ Grammatica-correctie (*jan word ziek*),
  - ⋆ Automatisch vertalen,
  - ⋆ ....

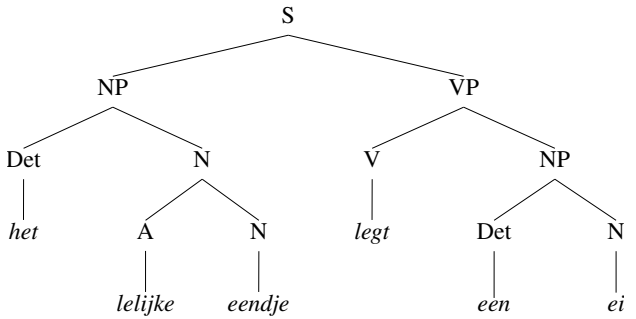
Grammatica

- Een taal is een (oneindige) verzameling zinnen,
- Zinnen zijn reeksen woorden,
- Niet alle reeksen woorden zijn zinnen,
- Een grammatica beschrijft
  - ⋆ welke reeksen woorden goede zinnen vormen,
  - ⋆ en wat de structuur van die reeksen is

Context-vrije Grammatica

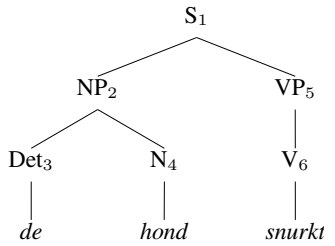
S	→	NP VP	Det	→	*een*
NP	→	Det N	Det	→	*het*
N	→	A N	N	→	*eendje*
VP	→	V	N	→	*ei*
VP	→	V NP	V	→	*legt*
			A	→	*lelijke*

Boomstructuren



Herschrijf-grammatica

S
NP VP
Det N VP
de N VP
de hond VP
de hond V
de hond snurkt



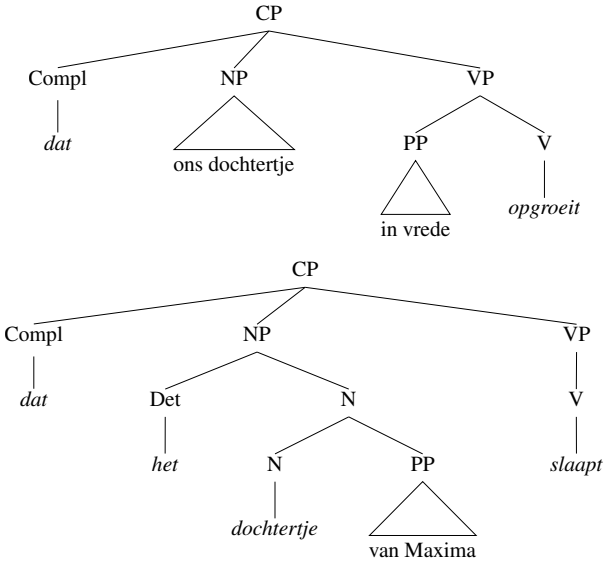
Taal en Grammatica

- Een reeks woorden W wordt herkend door grammatica G, wanneer je, door S te herschrijven, W kunt genereren.
- Alle reeksen die door G worden herkend, vormen de taal van G.

Ambiguïteit

- Wij willen dat ons dochtertje in vrede opgroeit
- Wij hopen dat het dochtertje van Maxima slaapt
- VP $\rightarrow$ VP PP
- N $\rightarrow$ N PP

Ambiguïteit



Ambiguïteit groeit exponentieel

- Wanneer deel 1 van een zin 5 mogelijke analyses heeft, en deel 2 3, heeft de hele zin 3×5 analyses
- Grammatica's die duizenden analyses aan een zin van 20 woorden toekennen zijn niet ongewoon.

All-and-only principe

- All: Een grammatica moet alle zinnen van een taal kunnen herkennen,
- Only: Een grammatica mag geen ongrammaticale zinnen herkennen.
- Bijna alle grammatica's voldoen niet aan All,
- Veel grammatica's voldoen niet aan Only.

Hoe nuttig is CFG?

- Een groot deel van het Nederlands kun je met CFG beschrijven, maar
  - ⋆ Duizenden regels nodig,
  - ⋆ Sommige aspecten van de taal zijn niet context-vrij,
- Definite Clause Grammar lijkt op CFG, maar levert
  - ⋆ Compactere grammatica's,
  - ⋆ Meer expressieve kracht.