Informatie Extractie met Reguliere Expressies

Gosse Bouma

Information Science
University of Groningen

Natuurlijke Taalverwerking II, 2006/2007

Overzicht

- Web Search, Information Extraction, Question Answering,
- Reguliere Expressies en Informatie Extracite,
- Question Answering als Informatie Extractie.

Web Search (Information Retrieval)

- Zoek relevante documenten over een bepaald onderwerp.
- **Vraag:** keywords (en boolese connectieven),
- Armstrong AND Cycling AND NOT Tour
- Zoeken in (indexen van) Documenten,
- **Antwoord:** lijst van relevante documenten.

Information Retrieval, Web Search

- Robuust (Gigabytes aan documenten, 1000-en queries),
- Technieken goed onderzocht,
- Nadeel:
  - Niet *direct* antwoord op een vraag,
  - Documenten bevatten relevante en irrelevante informatie.

Information Extraction

- Extraheer relevante informatie uit ongestructureerde tekst,
- Vul hiermee een **database**.
- **Vraag:** database query,
- **Antwoord:** matching entries uit de database.
- **Nadeel:** Systeem bepaalt wat relevant is.

Een vacature-site

- Haalt vacatures van het web, uit de krant,
- Ongestructureerde informatie,
- Wordt doorzocht op relevante velden,
- b.v. functie, opleiding, bedrijf, plaats, salaris,...
- Informatie wordt opgeslagen in database.

Academic Transfer

```
Twee Promovendi Waarnemen en Bewegen (100% elk)
Groningen, 38 uur per week
Faculteit der Medische Wetenschappen
Functiebeschrijving
.....
Functie-eisen Opleidingsniveau/vaardigheden:
Universitair - afgeronde universitaire opleiding
bewegingswetenschappen, psychologie, (bio)fysica
....
Aanbod De RUG biedt een salaris van minimaal ? 1.702,-
bruto per maand in het eerste jaar...

Contractduur: 4 jaar Maximaal aantal uur per week: 38
```

citeseer.nj.nec.com

- Haalt papers van het web (ps, pdf),
- Wordt doorzocht op auteur, titel, abstract, verwijzingen,..
- Database, met bovendien:
  - Active bibliography,
  - Similar documents,
  - Users who viewed this document also viewed,
  - Citatie-index,

Question Answering

Q:	Wanneer vond de Duitse hereniging plaats?
A:	in 1989
Q:	Wie is bestuursvoorzitter van Fiat?
A:	Giovanni Agnelli
Q:	Wie stelde een embargo in tegen Irak?
A:	de Veiligheidsraad van de Verenigde Naties
Q:	Waar staat GATT voor?
A:	General Agreement on Tariffs and Trade
A:	Algemene Overeenkomst inzake Tarieven en Handel
Q:	Hoe lang duurt het om een Duitse herder te trainen?
A:	10 tot 15 dagen
Q:	Waaruit wordt tolueen gewonnen?
A:	olie

Question Answering (2 voor 12)

- Gegeven een vraag in natuurlijke taal,
- Vind het exacte antwoord in een verzameling documenten,

Question Answering

- When was the Ebola virus first encountered?
  - `244 persons died of the Ebola virus, that was first found in Zaire in 1976`
- How did Jimi Hendrix die?
  - `...and when on September 18, 1970, Jimi Hendrix died of an overdose, her reaction...`
- What is the capital of Russia?
  - `The riders had a tour of Moscow this morning. Tomorrow morning they are leaving the Russian capital..`

Question Answering

- What is leukopenia?
  - `Leukopenia is a shortage of white blood cells (less than 4×109 per liter)...`
- What causes RSI?
  - `A potential explanation for RSI caused by light movements is that the same muscles are constantly involved.`
- Can one be allergic for soy?
  - `Food which relatively often induces allergic reactions are: cow milk, soy, peanuts, fish, ...`

Resources

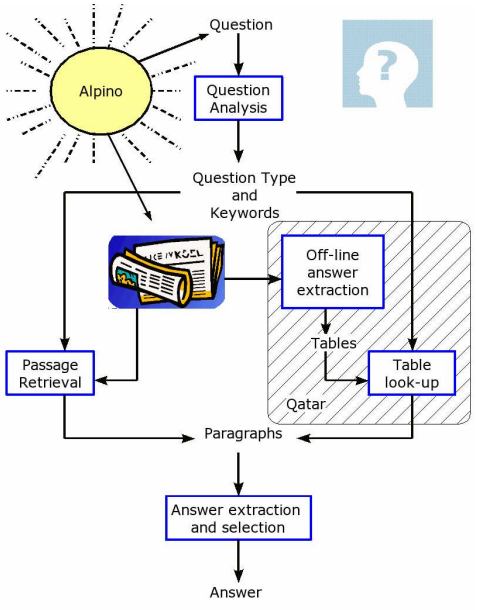
- Open Domain: CLEF evaluation exercise
  - Find answers for factoid questions
  - in Algemeen Dagblad, NRC 1994, 1995
  - 80 mln words
  - CLEF 2007: use `wikipedia.nl` as well (50 mln words)
- Closed Domain: Medical QA
  - Find answers to medical questions
  - in encyclopedia and reference material, web documents
  - 3 mln words

Summary

High-quality syntactic information is useful for QA

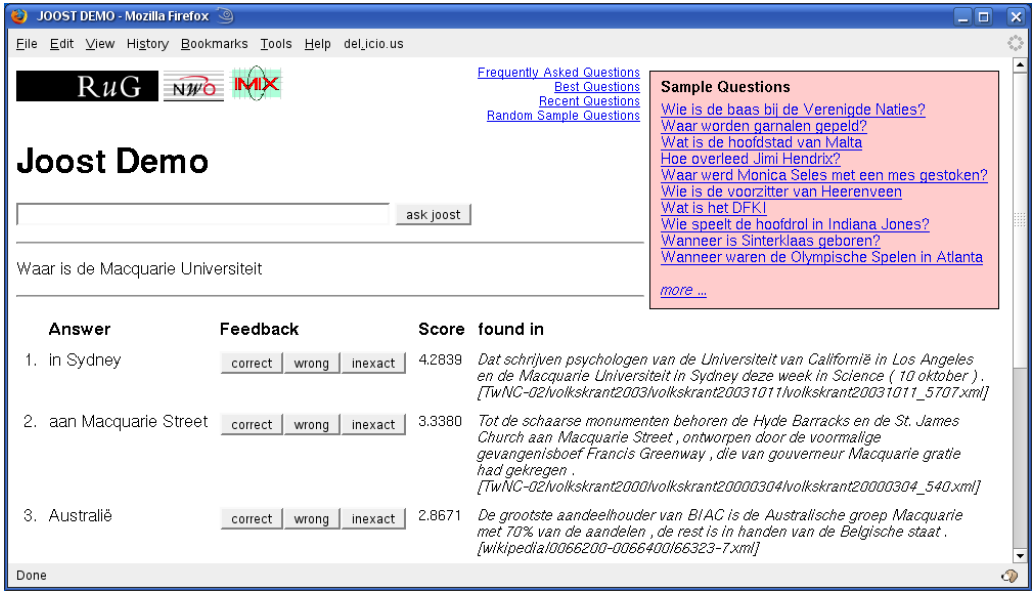
- question analysis
- answer extraction
  - on-line and off-line
  - acquiring ontological information
- answer ranking
- information retrieval

Joost: a QA system for Dutch



Question Answering

Question Answering



Frequently Asked Question Types

- For frequenty asked question types, answers are searched off-line
  - How many inhabitants does Location have?
  - When was Person born?
  - Who won the Nobelprize for literature in 1990?
  - What does the abbreviation ADHD mean?
  - What causes Frei syndrome?
  - What are the symptoms of poisoning by mushrooms?

Information Extraction

- Extraheer de geboortedata van personen
- Extraheer de geboorteplaats (of het geboorteland) van personen

Information Extraction Strategie

1. Zoek naar voorbeelden,
2. Ontdek patronen,
3. Test precisie van deze patronen,
4. (Pas patronen aan, en itereer naar 1)
5. Doe extractie.

Geboortejaren

```
> zgrep ' 19[0-9][0-9] ' *.gz
Eind 1994 werd een ' aanvaardbare ' plaats gevonden
Van Zon ( 1950 ) is na zijn studie fysische
geografie gepromoveerd
Op 24 december 1991 las ik in NRC Handelsblad
Sinds 1991 probeert een clubje ambtenaren het
```

Geboortejaren II

```
> zgrep ' ( 19[0-9][0-9] ) ' *.gz
Van Zon ( 1950 ) is na zijn studie fysische
geografie gepromoveerd
het Celloconcert ( 1966 ) van György Ligeti
de innovatieve Deense maker van Europa ( 1991 )
het WK in Frankrijk ( 1998 )
vluchtelingenstromen in Thailand ( 1979 ) , Somalië
( 1980 ) en Sudan ( 1985 ) .
```

Geboortejaren III

```
> zgrep ' ( 19[0-9][0-9]-' *.gz
de oorlog tussen Irak en Iran ( 1980-'88 )
Moord in Rotterdam ( 1905-1967 )
de komende Kunstenplanperiode ( 1997-2000 )
de biografie van de Charlotte Salomon ( 1917-1943 )
de schilder Pyke Koch ( 1901-1991 )
kunstverzamelaar Abram Tsjoetnovsky ( 1910-1985 ) .
een ode aan de Turkse onafhankelijkheidsoorlog (
1919-23 )
tijdens de eerste jaren ( 1965-1969 ) van het
bewind
```

Geboortejaren IV

```
> zgrep '[A-Z][a-z]* [A-Z][a-z]* ( 19[0-9][0-9]-'
*.gz
de biografie van de Charlotte Salomon ( 1917-1943 )
de schilder Pyke Koch ( 1901-1991 )
kunstverzamelaar Abram Tsjoetnovsky ( 1910-1985 ) .
het kabinet-Den Uyl ( 1973-1977 )
Boris Vian ( 1920-1959 )
```

Eigenlijke Extractie

- Vind de relevante delen:
- Naam + geboortedatum,
- Print de naam en de geboortedatum

Andere patronen?

```
> zgrep 'geboren in 19[0-9][0-9]-' *.gz
Mac Rebennack , geboren in 1941
Ze is geboren in 1957 in Revnice
Carl Barks ( geboren in 1901
Oppenheim ( Amerikaan , geboren in 1938 )
Kanan werd geboren in 1949
De Amerikaan John Baldessari bijvoorbeeld ( geboren
in 1931 )
```

Precision en Recall

- Precision : $\frac{\text{Goede geëxtraheerde relaties}}{\text{Geëxtraheerde relaties}}$
- Recall : $\frac{\text{Goede geëxtraheerde relaties}}{\text{Relaties in de tekst}}$

Informatie Extractie Methodologie

- Handmatig
  - Moeilijk (m.n. bedenken van nieuwe patronen),
  - Tijdrovend.
- Machine Learning
  - Aantal voorbeelden handmatig coderen,
  - Programma ontdekt patronen om vergelijkbare gevallen te vinden