Question Answering en Informatie Extractie

Gosse Bouma

Information Science
University of Groningen

Natuurlijke Taalverwerking II, 2006/2007

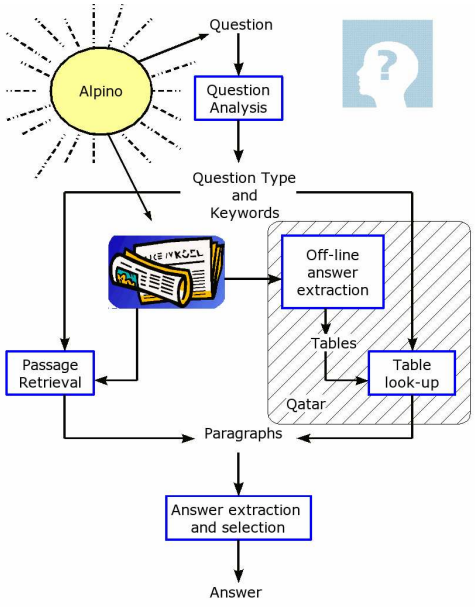
Overzicht

- Question Answering Technieken
  - Question Analysis
  - Answer Extraction
  - Answer Ranking
- Informatie Extractie
  - Evaluatie
  - Automatisch leren van patronen

Question Answering

- Gegeven een vraag (een zin) van een gebruiker
- Vind het antwoord op de vraag in verzameling tekst
- Vereist
  - Question Classification
    - wie, wanneer, wat is, ...
  - Information Retrieval
    - zoek naar relevante tekstfragmenten
  - Answer Extraction
    - vind het antwoord in de tekst
  - Answer Ranking)
    - geef het beste antwoord eerst

Joost: a QA system for Dutch

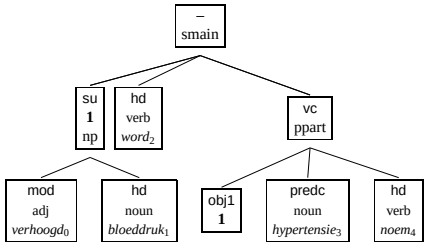


Alpino

- Lexicale Analyse
  - Part-of-Speech Tagging
  - Named Entity Tagging
  - Regels voor onbekende woorden
- Syntactische Analyse voor het Nederlands
  - 500+ regels, 50.000+ woorden
  - Regels voor ongrammaticale en/of moeilijke zinnen
  - Disambiguatie
- CLEF-corpus volledig geparsed

Dependency Relations

- Verhoogde bloeddruk wordt hypertensie genoemd
- *High blood pressure is called hypertension*



⟨blood-pressure, mod, high⟩
⟨is, su, blood-pressure⟩
⟨is, vc, call⟩
⟨call, predc, hypertension⟩
⟨call, obj, blood-pressure⟩

Question Analysis

Category	Example
Location	Uit welk land komt het konikpaard?
Date	In welk jaar werd Suriname onafhankelijk?
Measure	Hoeveel rugvinnen hebben reuzenhaaien?
Abbreviation	Waar staat de afkorting WWW voor?
Capital	Wat is de hoofdstad van Ethiopië?
How	Hoe drinken Engelsen hun bier het liefst?

Question Analysis

Category	Example
Person	Welke paus was de voorganger van paus Paulus VI?
Function	Wie was de mannelijke leider van de Rote Armee Fraktion?
Organization	Welk bedrijf werd opgericht door de uitvinder George Eastman ?
Name	Wie leidde de Kon-Tiki-expeditie?
Which	Welke giftige stof zit in bittere amandelen?

Answer Extraction

- event_date(Event): zoek een datum die een syntactische relatie heeft met Event
- capital(Country): zoek in tabel
- which(Class): zoek een naam die lid is van Class
- ...

Answer Selection

- Type-score: score van de regel die is gebruikt om het antwoord te vinden
- Namen: voorkeur voor zinnen met dezelfde namen als in de vraag
- Syntaxis: voorkeur voor antwoord-zinnen die lijken op de vraag
- IR-score: voorkeur voor zinnen die volgens Information Retrieval relevant zijn
- Frequentie: voorkeur voor antwoorden die vaak gevonden worden

QACLEF

- Cross Language Evaluation Forum
- QACLEF:
  - 200 vragen en vervolgvragen
  - Antwoord vinden in Wikipedia (50 mln woorden), Algemeen Dagblad 94/95, NRC 94/95 (80 mln woorden)
- Hoeveel goede antwoorden vind systeem?

Resultaten van Joost

Q:	Wie is de moordenaar van John Lennon ?
QA:	`function(moordenaar,John Lennon)`
A:	10 mei: 1955 - Mark David Chapman , moordenaar van John Lennon
Q:	Waar werd hij vermoord ?
QA	: `location(John Lennon,nil)`
A:	John Lennon: John Lennon op All Music Guide
Opm1.	All Music Guide is als locatie gemarkeerd
Opm2.	*in de avond werd Lennon voor zijn huis in het Dakota gebouw aan 72nd Street vermoord*
Q:	Hoe vaak werd Lennon geraakt ?
QA:	frequency
A:	Mark David Chapman: Lennon werd vier keer geraakt en overleed rond 23:15 .

Question Analysis

Q:	Wat is single malt whiskey?
QA:	`which(malt whiskey)`
A:	Lagavulin Single Malt
Q:	Welk type auto zat hij in tijdens de aanslag ?
QA:	`which(type)`
A:	Voorbeeld
Q:	Wanneer is het volgens de Joodse kalender Jom Ha-atsmaoet ?
QA:	`event_date(het)`
A:	Op 14 mei 1948 (op de joodse kalender 5 ijar 5708)
Q:	Waar komt de naam vandaan
QA:	`location(naam,nil)`
A:	voor een versie van GNU/Linux

No Question Type

Hoeveel broers en zussen had deze man ?
Wat werd Karel I na de moord op zijn oom
Waarvan was Panaji de hoofdstad voor 1987 ?
Wanneer begon men met de bouw van de Dom ?
Wat zijn de belangrijkste vormen van openbaar vervoer ?
Waarvan is CLEF de Europese tegenhanger ?
Het hoeveelste Eurovisiesongfestival was de 1958 editie ?
Wat is mede ?
Wanneer werd hij gebruikt in het Koninkrijk Joegoslavie ?
Wat werd zijn functie daarna ?

Answer Extraction

Q:	Wanneer is het volgens de Joodse kalender Jom Ha-atsmaoet ?
A:	Op 14 mei 1948 (op de joodse kalender 5 ijar 5708)
Q:	Wat was het hoofddoel van de Hanze ?
A:	Het hoofddoel van deze samenwerking was om een handelsmonopolie te veroveren...
Q:	Naar welke neuroloog is het syndroom Gilles de la Tourette vernoemd ?
A:	Het syndroom is vernoemd naar de Franse neuroloog Georges Gilles de la Tourette

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?

Afkortingen

- Vind de betekenis van afkortingen
- `Hij bouwde de Automatic Computing Engine ( ACE ) .`
- `Op 1 juli 2006 telde de gemeente Amsterdam 741.623 inwoners ( bron : CBS )`
- Schrijf een programma dat voor iedere zin bepaalt
  - Of het een afkorting + volledige term bevat
  - Schrijf gevonden afkortingen + volledige term naar een bestand
- Test op (40K) zinnen met daarin 2 hoofdletter achter elkaar uit Wikipedia

Afkortingen

BCA	Badminton Combinatie Amersfoort
KMSKA	Koninklijk Museum voor Schone Kunsten
MUHKA	Museum van Hedendaagse Kunst Antwerpen
PTA	Passagiers Terminal Amsterdam
GVB	Gemeentevervoerbedrijf Amsterdam
AE	astronomische eenheid
UCL	Université catholique de Louvain
NMBS	Nationale Maatschappij der Belgische Spoorwegen
LSFB	Frans-Belgische Gebarentaal
ISBN	September 26

Precision en Recall

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

Precision en Recall

- Precision bepalen :
  - tel in de eerste N resultaten van het systeem hoeveel volledige termen goed zijn
- Recall (1):
  - Maak een bestand met alleen maar zinnen waarin een afkorting en een volledige term staan,
  - Test voor hoeveel van deze zinnen je systeem een resultaat geeft.
- Recall (2):
  - Kies N afkortingen
  - Tel hoe vaak ze met volledige betekenis in de tekst staan
  - Tel hoe vaak je systeem de betekenis gevonden heeft

Automatisch patronen leren

- Zelf regels bedenken, implementeren, testen, is tijdrovend
- Kan het met minder inspanning?
- Automatisch leren van relaties
  - Land-hoofdstad
  - Land/Stad-inwoneraantal
  - Land-munteenheid
  - Persoon-geboortedatum
  - Film-Regisseur
  - ...

Automatisch patronen leren

- Seed-list: lijst met voorbeelden van de relatie
  - Frankrijk-Parijs
  - Griekenland-Athene
  - Japan-Tokyo
  - Togo-Lomé
  - ...
- Verzamel zinnen met beide elementen van de relatie
  - *Volgens een woordvoerder bij de VN-missie in Soechoemi , de hoofdstad van Abchazië*
  - *Volgens een bekendmaking in de Abchazische hoofdstad Soechoemi*
  - *autoriteiten in de Abchazische hoofdstad Soechoemi*
  - *in de Afghaanse hoofdstad Kaboel*
- Vind de patronen die Land en Hoofdstad verbinden
  - CAPITAL , de hoofdstad van COUNTRY
  - de COUNTRY-ADJ hoofdstad CAPITAL
- Test patronen op precision/recall