Textfiltrering, sammanfattning indexering, sökning

Hercules Dalianis
email:hercules@dsv.su.se
tel 08-790 91 05

Innehåll

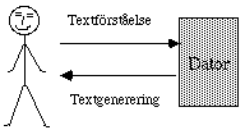
- Bakgrund
- Filtrering
- Indexering
- Stemming/Ordböjning
- Sökning
- Sammanfattning/extrahering

Bakgrund

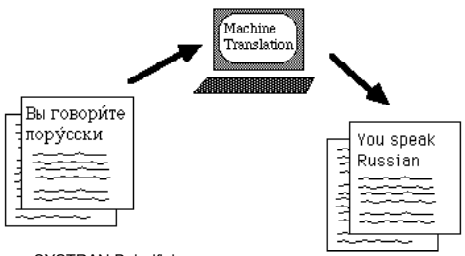
- Lingvistik
- Datorlingvistik
- Forskning sedan 1950 talet
- Maskinöversättning
- Naturligt språkgränssnitt
- Taligenkänning
- Talsyntes

Datorlingvistik

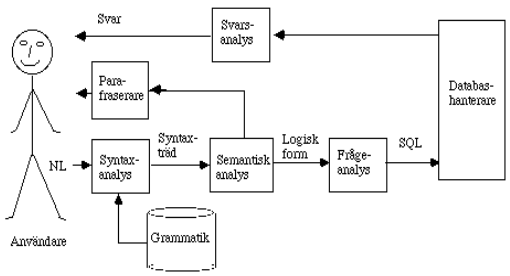
- Datorer



- Kommunikation Människa - Dator
  - —Parsning (F r st e lse)
  - —Generering (Skapande)



- SYSTRAN-Babelfish
- Metal
- + 60 till



- NL-gränssnitt á la 80-tal

NL-gränssnitt

HSQL-(Help systen for structured query language) prototypen är implementerad mot en databas som innehållerinformation om olika sjukhus och deras verksamheter.
Exempel på generering av SQL från naturligt språk samt parafrasering från SQL tillbaka till naturligt språk

Fråga: vilken diagnos har Amster ?
SQL:
SELECT DISTINCT T2.other_info, T1.name, T2.reg_no
FROM PATIENT T1, DIAGNOSIS T2
WHERE (T1.name = `Amster K.`) AND
(T2.reg_no=T1.reg_no)
PARAFRAS AV SQL I NATURLIGT SPRÅK
vilka diagnoser har en patient som heter Amster K.

Fråga: vem behandlar Adler.
SQL:
SELECT DISTINCT T3.name, T1.name,T2.empl_no, T3.reg_no
FROM DOCTOR T1, ATTDOCTOR T2, PATIENT T3
WHERE (T1.name = `Adler J.`) AND
(T2.empl_no = T1.empl_no) AND
(T3.reg_no=T2.reg_no)
PARAFRAS AV SQL I NATURLIGT SPRÅK:
Vilka patienter behandlar doktor Adler J.

Fråga: vem behandlar Hansson.
SQL:
SELECT DISTINCT T1.name,T2.empl_no, T3.reg_no, T3.name
FROM DOCTOR T1, ATTDOCTOR T2, PATIENT T3
WHERE (T2.empl_no = T1.empl_no) AND
(T3.reg_no = T2.reg_no) AND
(T3.name=T2.`Hansson A.`)
PARAFRAS AV SQL I NATURLIGT SPRÅK:
Vilka doktorer behandlar patienten Hansson A.

Textgenerering

- Generering av naturligt språk
- Generering av väder- eller börsrapporter direkt från rådata
- På flera språk
- Generering av bruksanvisningar på flera språk på en gång (Scania)
- Generering från formella språk tex UML

Indata till GOSSIP:

```
ee(martin,ttyp0,login,[],8:20:03,_,_,0).
ee(martin,ttyp0,editor,[f1],8:30:00,9:10:32,0:40:32,240).
ee(martin,ttyp1,editor,[f2],8:42:21,9:13:14,0:30:53,183).
ee(martin,ttyp0,logout,[],9:21:05,_,1:01:02,0).
ee(martin,ttyp0,login,[],10:17:32,_,_,0).
ee(martin,ttyp0,editor,[f1],10:20:58,12:15:27,1:54:29,1200).
ee(martin,ttyp1,editor,[f2],11:00:39,11:32:48,0:32:09,185).
ee(jessie,ttyd0,login,[],11:03:46,_,_,0).
ee(jessie,ttyd0,editor,[f5],11:12:45,12:48:22,1:35:37,573).
ee(jessie,ttyd1,compiler,[f4],11:23:32,11:31:01,0:07:29,300).
ee(jessie,ttyd1,editor,[f3],11:32:25,11:45:56,0:13:31,70).
ee(jessie,ttyd1,editor,[f4],11:47:34,11:59:09,0:11:35,65).
ee(jessie,ttyd1,compiler,[f4],12:04:47,12:08:32,0:03:45,186).
ee(jessie,ttyd1,editor,[f3],12:09:57,12:16:34,0:06:37,15).
ee(jessie,ttyd1,editor,[f4],12:18:43,12:39:24,0:20:41,154).
ee(martin,ttyp0,logout,[],12:20:21,_,2:02:49,0).
ee(jessie,ttyd1,editor,[f7],12:56:01,13:15:02,0:19:01,143).
ee(jessie,ttyd0,editor,[f6],12:59:56,13:20:43,0:20:47,187).
```


Utdata från GOSSIP:

The system was used for 7 hours 32 minutes and 12 seconds. The users of the system ran editors and compilers during this time. Compilers were run six times (the cpu-time equal to 46% of the total cpu-time). Editors were run twelwe times (the cpu-time equal to 53% of the total cpu-time). Two users, Martin and Jessie, logged on to the system. Jessie used the system for 63 % of the time in use.

IR+lingvistik

- IR-Information Retrieval
- Informationssökning/Åtkomst
- (Textmining)
- Taggning/Parsning
- Statistik+Lingvistik

Filtrering

- Om vi betraktar en grupp dokument på svenska
- Vi ser att alla texter har en grupp gemensamma ord, sk stoppord,
- Stopporden är så kallade vanligt förekommande icke-betydelse bärande ord tex *och, eller, på, under, genom etc*

- Vi filtrerar bort stopporden med en stoppordlista eller stopplista
- Kvar får vi de betydelsebärande orden som skiljer dokumenten åt.
- Ett dokument innehåller ord som tex *stamceller, leukemi, blodsjukdomar, sjukhus, forskare, etc*
- Ett annat dokument innehåller ord som *Arafat, Israel, sprängning, FN, demonstrationer, etc*

Index

- De betydelsebärande orden är oftast substantiv och adjektiv samt en del verb
- Vi kan kalla dem nyckelord eller indexord. Ord som skulle kunna finnas i ett index till dokumentet.
- Jmfr sökmotorers index
- Vi kan klassificera/ indexera dokumenten

tf, idf och tf * idf

- tf = term frequency, antal unika termer per dokument.

bilverkst der = 5 ggr = tf
bilverkstad = 1 ggr = tf
bilar = 4 ggr = tf
bil = 1 ggr = tf
att = 7 ggr = tf,
och = 32 ggr = tf
eller 25 ggr = tf

- idf = inverse document frequency,

idf = N/di dr N totalt antal dokument och di r antal dokument d r termen f rek ommer.

- Högt idf för en term ger en hög signifikans, dvs en ovanlig term

Ex 100 dokument totalt/ 4 dokument med termen *bilverkst der* = 20 = idf

Om alla dokumenten inneh ller termen = idf = 100/100 = 1

tf • idf ger ett mått på signifikans i ett dokument. Termer med höga tf • idf är signifikanta i ett specifikt dokument.

tf • idf = 5 * 20 = 100 för bilverkstäder i vårt dokument

tf • idf = 1 * 20 = 20 för ett annat dokument med tf = 1

- Det finns många varianter på tf * idf
- Normalisering pga olika dokumentlängd kan behövas.

Sökmotorer

- tf * idf brukar inte användas av Internet-sökmotorer, svårt att beräkna
- Man kan estimera tf * idf
- Precision/Recall(Täckning) mått på hur bra en sökmotor är

Precision/Recall

- Precision =
  funna relevanta dokument/
  totalt antal funna dokument
- Recall/Täckning =
  funna relevanta dokument /
  totalt antal relevanta dokument

Informationsprecision

- Precisionen ökar om man finner det man söker efter.
- Sök efter *bilverkstad* på Internet
- Man hittar bara *bilverkstad* och inte *bilverkstäder.*
- => Alltså bra att kunna böja orden
- För att även kunna hitta *bilverkstäder*

Lemma

- Morfologi - lemma
- Varje ord har en grundform/normalform eller lemma
- Det är värdefullt om vi i ett språk som har mycket morfologi kan ta fram grundformen. Tex svenska, finska etc
- *bilverkstad => bilverkstäder, bilverkstäderna, bilverkstädernas,...*

Stemming-Ordböjning

- Ta bort ordböjningen ta fram stammen
- (jmfr trunkering bilverkstad*)
- Enklare form av lemma / grundform
- Bättre täckning vid sökning
- Men hur går det med precisionen?

Stemming

- Stemming på svenska
- Utvärdering
- På mindre textmängder ≈ 50 000 texter
- 15 % bättre precision
- 18 % bättre relativ täckning

Nyckelordlista

- En nyckelordlista eller ett morfologiskt betydelsebärande lexikon
- Innehåller alla viktiga ord sk nyckelord inom en domän samt deras böjningar.
- Domän kan vara nyhetstexter, (inrikes, utrikes, ekonomi, sport) motorcykeltexter, eller ngt annat man är intresserad av.

- Denna nyckelordlista kan du använda för att indexera ett dokument på ett språkkänsligt vis.
- Flera olika nyckelordlistor kan användas för att kategorisera dokument i grupper/domäner.
- Har man stoppordlistor kan man kategorisera dokument i olika språk
- Man kan även köpa dessa verktyg/lexikon
- Vi har en automatisk språkigenkännare för 40 språk.

Indexering

Tre sätt

1) Ta bort stopporden räkna resten av orden.
   —De flesta s kmotorer g r s idag.
2) tf * idf och cut-off
3) Stemming och tf * idf
   Normalisera dokumentet med stemmern
   Normalisera fr gan med stemmern
   Eller
   Expandera fr gan med stemmern. Jmfr bilverkst*

4) Använd en morfologiskt betydelsebärande lexikon på vad som tillåts i dokumentet.
   —Du f r normalisering av texten och nyckelord
   —Eventuellt kan du behandla alla vanligt f re kommande ok nda ord som r l ngre n 8 tecken som betydelseb rande ord