Rickard Domeij, Nada, KTH

Lösningar till tenta i Översiktskurs i språkteknologi på Språkkonsultutbildningen vid Stockholms universitet, kurskod NS 8020. 2002-03-04.

Fråga 1 a-b: Informationssökning och textsammanfattning (4 p)

a) Stemming är när ett datorprogram tar bort prefix eller suffix på ord så att bara stammen blir kvar. Det kan användas inom informationssökning för att hitta varianter på ett eftersökt ord, t.ex. böjningsformer.

b) Ett nyckelord är ett ord som väl beskriver innehållet i en text. Dessa ord kan ses som en delbeskrivning av en text, en minisammanfattning. Ett textsammanfattningssystem kan automatiskt ta fram nyckelord genom att bl.a. titta på ords frekvens i texter och utifrån det bedöma vad som är mer eller mindre viktigt.

Fråga 2 a-c: Rättstavning (5 p)

a) Suffixregler gör att Stava kan godkänna böjningsformer som inte finns med i ordlistan.

b) När ett ord som slutar på "arna" ska stavningskontrolleras så tar Stava fram stammen X (det som står före arna) och kontrollerar sedan om X, Xen och Xar finns med i ordlistan. Om alla tre orden finns med godkänns Xarna.

c) stolarna

Fråga 3 a-b: Statistik (4 p)

a) Rättstavning. Rangordning av rättelseförslag (rangordna vanliga ord högre än ovanliga).

b) Informationssökning på webben. Söktjänsten ska presentera de webbsidor som innehåller sökordet. Webbsidor med många förekomster av ordet ska rangordnas högt.

Fråga 4: Morfologi (4 p)

Analys innebär att orden analyseras, ofta med hjälp av ett lexikon och böjningsregler. Orden tilldelas ordklass, grundform och böjningsinformation. Analysen kan vara flertydig och ofta får ett ord flera tolkningar. Ordet *bilar* kan analyseras både som substantiv i utrum obestämd form pluralis med lemmat *bil* och som ett verb i presens med lemmat *bila*..

Generering innebär att programmet skall böja ett ord utifrån angivna krav. Programmet skall alltså skapa nya ordformer som tillhör språket. Om programmet får in lemmat *bil* och kraven är att det skall böjas i utrum obestämd form pluralis skall ordformen *bilar* komma ut från programmet.

Fråga 5: Syntax och parsning (5 p)

Viktiga steg i analysen är till exempelföljande:

1. Förbearbetning med ordgränsigenkänning.

2. Morfologisk analys där ordformerna tilldelas information om ordklass, grundform och böjningsegenskaper.
3. Morfologisk disambiguering där endast en ordklass väljs ut beroende på kontext.
4. Frasanalys t.ex. igenkänning av nominalfraser.
5. Satsanalys t.ex. igenkänning av satsgränser och hela satsers ingående struktur.

Äv. semantisk och pragmatisk analys kan göras.

Fråga 6 a-c: Datorstödd språkgranskning (9 p)

a) Vid mönsterigenkänning görs enbart en analys av vissa kombinationer av ordklasser t.ex. två på varandra följande supinumformer för att hitta dubbel supinum. Vid full parsning görs en fullständig analys av strukturen i hela satser. Mönsterigenkänning är en effektivare och mer robust metod, medan parsning är noggrannare och ger därmed upphov till färre falska alarm.

b) Granskning av problem i tecken, grammatik och stil.
Tecken: tex (felaktig förkortning)
Grammatik: ett stor hus (inkongruens)
Stil: medans (talspråkligt)

c) En fördel kan vara att det är lättare att sprida och få gehör för t.ex. Svenska skrivregler. Det är också bra för små språk att ha språkteknologiskt stöd för att hävda sig i konkurrensen med större språk. En nackdel kan vara att programmen indirekt, genom att sätta fokus på korrekthet i språket, drar bort uppmärksamheten från viktigare saker som läsare och organisation. Kanske kan en ovan skribent tro att allt är bra om programmet inte klagar på något.

Fråga 7 a-c: Semantik och pragmatik (4 p)

a) Satsfunktion signaleras med grammatiska medel, t.ex. visar med ordföljd och interpunktionstecken att det rör sig om en deklarativ sats, imperativsats eller frågesats. Talhandlingen är det som talaren avser med satsen när han yttrar den i en given situation. Ofta finns en tydlig koppling mellan satsfunktion och talhandling (man använder en frågesats för att fråga ngt), men inte alltid. I exemplet Kan du skicka saltet? används en frågesats för att be/uppmana någon att skicka saltet.

b) b) Enligt Grice principer frågar man inte om något som man redan vet svaret på. Det vet mottagaren och därför tolkar denne inte frågan i a) som en fråga. I stället försöker mottagaren luska ut vad den frågande egentligen kan ha menat, något som är relevant i kontexten och även stämmer med övriga samarbetsprinciper. På så sätt kommer mottagaren fram till att den frågande troligen vill få sig saltet tillskickat. Den tolkningen stämmer med samarbetsprinciperna.

Fråga 8 a-d: Talteknologi och dialogsystem (7 p)

a) Ett taligenkänningssystem kan analysera talat språk på ett begränsat sätt. Ett talsyntessystem kan läsa upp text.

b) Automatisk nummerupplysning är ett exempel på ett dialogsystem som besvarar frågor om folks telefonnummer med telefonkatalogen som databas. Med taligenkänning analyseras talet och görs om till text i formen av en sökfråga. Efter databasuppslagning läses svaret upp med talsyntes.

c) En domän är ett begränsat område för en tillämpning, t.ex. nummerupplysning. Om domänen är tydligt avgränsad och språket som används inom domänen har starkt begränsad vokabulär, entydig semantik och enkel syntax är förutsättningarna för att göra ett användbart system goda.

d) Allt som ett system säger kan tolkas i enlighet med Grice principer vilket gör att om systemet säger nåt som verkar helt irrelevant omtolkas det av människan till nåt som är relevant. Det gör att människor lätt tillskriver dumma datorsystem mer intelligens än de besitter. Eliza är ett exempel på ett system som bygger på väldigt enkla principer där systemet bara eftersöker huvudordet i en yttrad mening och använder det i sin replik, ändå luras många att tro att det är intelligent.

Fråga 9 a-d. Datorstödd översättning (8 p)

a) Ett maskinöversättningssystem översätter mer eller mindre helautomatiskt medan ett översättningsminne fungerar som stöd för mänskliga översättare i deras arbete.

b) Ett maskinöversättningssystem lämpar sig framför allt till grovöversättningar av text när kravet på kvalitet är lågt, t.ex. när man vill skumma igenom texter för att se om de verkar intressanta. Ett översättningsminne gör effektiviserar arbetet med översättning av tekniska manualer genom att spara tidigare gjorda översättningar för återanvändning. Innehåller även specialiserade terminologidatabaser.

c) Flertydiga ord (ty. *meine*, eng. *mine*/*mean*), och skillnader i ordföljd (t.ex. placeringen av *inte* i svenska resp. engelska bisatser)

d) Ett kontrollerat språk är ett konstruerat språk som påtvingats en förenklad syntax, en entydig semantik och en begränsad vokabulär. Om man tränar översättare att skriva i det kontrollerade språket kan översättningar till andra språk göras helautomatiskt med bibehållen kvalitet.