

Lösning till tentamen: Språkteknologi för språkvetare (SV2122)

14 mars 2014, 9:00 – 12:00

Kursansvarig Richard Johansson, Språkbanken, institutionen för svenska språket

Hjälpmedel Inga

Betygsgränser Väl godkänt: 24p, Godkänt: 15p, Max: 30p

Observera:

- Skriv läsbart: oläsligt räknas som felaktigt.
- Numrera de papper du lämnar in.
- Om du bara hinner slutföra en uppgift delvis, lämna in din lösning i alla fall.
- Om någon fråga är oklar, passa på att fråga när den kursansvarige kommer till tentamenslokalen.

Observera: Jag har skrivit ganska utförliga svar här, men du kan få full poäng även om du uttryckt dig mer kortfattat. Jag rättar baserat på hur mycket jag tycker att svaret uttrycker de viktigaste idéerna i vad jag efterfrågar.

Uppgift 1 av 6: Relationer mellan ord (5 poäng)

Din uppgift: Hitta minst fem ordpar i ordlistan nedan, samt ange vilken lexikal-semantic relation det råder mellan orden i varje par. De relationer som är tänkbara i det här fallet är *antonymi*, *hyponymi*, *meronymi*, *polysemi* och *synonymi*. Observera att samma ord kan ingå i flera olika relationer.

Ordlistan: *vän*, *sträng*, *fisk*, *ankomst*, *gitarr*, *glas*, *avfärd*, *kompis*, *lax*

Lösning:

- *vän* är **synonym** till *kompis*
- *lax* är **hyponym** till *fisk*
- *sträng* är **meronym** till *gitarr*
- *ankomst* är **antonym** till *avfärd*
- Exempel på polysemi: *sträng* är en tråd eller en bit text i t.ex. Python; *glas* är ett material, ett dryckeskärl eller en volymenhet; *lax* är en fisk eller en sedel.

Uppgift 2 av 6: Fornisländska substantiv (5 poäng)

Fornisländska hade liksom svenska flera olika sätt att böja substantiv. Till exempel *hestr* ('häst') och *hamarr* ('hammare') tillhörde vad man kallar *a*-typen, och *vinr* ('vän') och *staðr* ('plats') hörde till *i*-typen. Här visas böjningsmönstren i obestämd form för dessa fyra exempel:

	Singular	Plural		Singular	Plural
Nominativ	hestr	hestar	Nominativ	vinr	vinir
Akusativ	hest	hesta	Akusativ	vin	vini
Dativ	hesti	hestum	Dativ	vin	vinum
Genitiv	hests	hesta	Genitiv	vinar	vina
	Singular	Plural		Singular	Plural
Nominativ	hamarr	hamrar	Nominativ	staðr	staðir
Akusativ	hamar	hamra	Akusativ	stað	staði
Dativ	hamri	hømrum	Dativ	stað	stoðum
Genitiv	hamars	hamra	Genitiv	staðar	staða

Här kan vi göra en del observationer. Till att börja med påverkas ett *a* av ett efterföljande *u* så att det uttalas längre bak, och skrivs då *ø*. Denna process kallas *u-omljud*. Den andra saken att lägga märke till är att *hamarr* fungerar som en hel del moderna svenska substantiv: *synkop* gör att det andra *a*:et faller bort om böjningsändelsen innehåller en vokal.

Din uppgift: Beskriv i stora drag hur man skulle konstruera ett program som hanterar fornisländsk substantivböjning.

Lösning: Substantivets böjningsformer beskrivs med hjälp av böjningstabeller; t.ex. för a-typen gör vi en tabell där vi har *-r* för nominativ singular, *-ar* för nominativ plural, etc, och för i-typen en annan böjningstabell där vi t.ex. har *-ir* för nominativ plural. I lexikonet anger vi då att rötterna *hest* och *hamar* kopplas till *a*-tabellen och *vin* och *stað* till *i*-tabellen.

Vi beskriver de fonologiska processerna (omljud och synkop) med hjälp av omskrivningsregler. *u*-omljudet motsvaras då av en regel som skriver om ett *a* till *o* om det har ett *u* efter sig. Synkopen blir kanske lite mer komplicerad, men en början kan vara att om en rot innehåller två vokaler, och böjningen innehåller någon vokal, så avlägsnas den andra av rotens vokaler.

Precis enligt dessa principer konstruerar vi ett morfologiprogram med t.ex. Xerox-verktygen. Ett sådant program kommer då att både kunna *generera*, dvs producera rätt böjningsform, och *analysera*, dvs avgöra grundform och morfologisk kategori för en given form.

Uppgift 3 av 6: Oförskämtheter (5 poäng)

På en tidning beslutar man sig för att tillåta sina läsare att lämna kommentarer till de artiklar som publiceras på tidningens websida. Enligt svensk lag är tidningen då juridiskt ansvarig för det som skrivs i kommentarerna: tidningens ansvarige utgivare kan dömas om en läsare skriver något som är brottsligt, t.ex. om det kan räknas som förtal eller uppvigling. Man vill dessutom hålla en god stämning i kommentarsfältet så att diskussionerna inte urartar.

Man bestämmer sig därför för att använda ett automatiskt program som kontrollerar läsarnas kommentarer och skickar dem till en granskare om de verkar olämpliga, t.ex. om de verkar formulera sig på ett brottsligt sätt eller om de innehåller oförskämtheter.

Din uppgift: Föreslå något tillvägagångssätt vi kan tänka oss att använda för att konstruera ett program av denna typ.

Lösning: Detta är ett dokumentkategoriseringsproblem, ungefär som de program för spamfiltrering och åsiktskategorisering som vi har sett i kursen. Det kan förmodligen också lösas på liknande sätt, genom att använda en tabell med informativa ord som ger en signal om vilken kategori varje kommentar tillhör. Tabellen, och "styrkevärdena" för orden i tabellen, kan tillverkas antingen manuellt eller genom korpusbaserade metoder. För den sistnämnda varianten behöver vi dock en annoterad korpus att "träna upp" programmet på, dvs en samling av kommentarer där en människa kategoriserat varje enskild kommentar som antingen lämplig eller olämplig.

Uppgift 4 av 6: Ordbetydelser och informationssökning (5 poäng)

Vissa ord har ett flertal betydelser, t.ex. *cykel*. Å andra sidan finns det en del ord som betyder samma sak, t.ex. *mat* och *föda*.

Din uppgift: Förklara varför dessa två fenomen försvårar informationssökning. Vilken effekt har respektive fenomen på precisionen och täckningen (*recall*)? Ge gärna exempel.

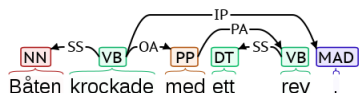
Lösning: Om ett ord har flera betydelser så kan vi få sökträffar som inte innehåller det vi var ute efter, t.ex. om jag söker på *cykel* (och var ute efter fordon) så kommer också att få träffar på den andra betydelsen av *cykel* (dvs ungefär "omgång"), vilket alltså innebär att vi får lägre precision i dessa fall (precision = antalet relevanta förslag dividerat med antalet förslag). Det vore tänkbart att försöka lösa detta problem genom att utföra betydelsedisambiguering (*word sense disambiguation*), men det är oftast svårt att göra i praktiken.

Å andra sidan, om jag t.ex. söker efter dokument som beskriver vilken *mat* som vesslor äter så kan det ju vara så att det finns informativa dokument som i stället använder ordet *föda*. Eftersom vi missar relevanta dokument av detta skäl så påverkar vi täckningen negativt (täckning =

antalet relevanta förslag dividerat med totala antalet relevanta i samlingen). Effekten av detta problem kan lindras genom *sökfrågeexpansion* (*query expansion*): man tillåter inte bara sökordet, utan också dess synonymer. Risken är förstås att man då minskar sökningens precision i stället.

Uppgift 5 av 6: Grammatisk analys (5 poäng)

Vi analyserar meningen *Båten krockade med ett rev.* med en ordklassmärkare och en syntaktisk analysator (*parser*), t.ex. med hjälp av Språkbankens annoteringslaboratorium. Resultatet ser ut så här:



ord	msd	lemma
Båten	NN. UTR. SIN. DEF. NOM	båt
krockade	VB. PRT. AKT	krocka
med	PP	med
ett	DT. NEU. SIN. IND	en
rev	VB. PRT. AKT	riva
.	MAD	

Din uppgift: Vad finns det för olika problem med ovanstående analys? Vad tror du problemen kan bero på och hur hänger de ihop?

Tips: Om du är osäker på någon av koderna (NN, SS, VB, etc) så fråga!

Lösning: De problem som finns är 1) att *rev* har analyserats som ett verb (VB) med grundformen *riva* i stället för substantiv (NN), och 2) att *ett* har blivit subjekt (SS) till *rev*, när det borde ha varit ett bestämningsattribut.

Problem 2 är en direkt följd av problem 1: den syntaktiska analysen (*dependensparsern*) gör fel på grund av den felaktiga ordklassanalysen. Det är då svårt för parsern att göra rätt, eftersom ett verb brukar ta subjekt men inte bestämningsattribut.

Problem 1 orsakas förmodligen av det faktum att *rev* är extremt mycket vanligare i korpusar som verb än som substantiv. Det kan till och med vara så att det inte förekommer över huvud taget som substantiv i den annoterade korpus (i detta fall Stockholm/Umeå-korpusen) som användes för att utveckla ordklassanalysprogrammet.

Uppgift 6 av 6: Översättning (5 poäng)

Den bästa översättningen av engelska *a difficult case* till svenska är förmodligen *ett svårt fall*.

Din uppgift: Diskutera olika tänkbara sätt att konstruera ett automatiskt översättningsprogram som kan producera texten *ett svårt fall* snarare än någon sämre möjlighet t.ex. *en svår fall*, *ett svårt kasus*, *ett besvärligt hölje* etc. Nämn minst två möjligheter att lösa detta problem.

Lösning: Problemet kan tänkas lösas med en "djup" (betydelsebaserad) eller en "ytlig" (korpusbaserad) metod.

Med en djup metod kan vi tänka oss följande. Vi gör först en ordklassanalys och en syntaktisk analys. Därefter använder vi betydelsedisambiguering (*word sense disambiguation*) för att avgöra att det är *fall*-betydelsen av *case* som avses, och inte ett lingvistiskt kasus eller ett

hölje eller en burk. Därefter representeras betydelsen för hela uttrycket, t.ex. genom att använda FrameNet. Vi har då kommit till en abstrakt mellanrepresentation som kan omvandlas åt andra hållet (*generering*) för att skapa texten på målspråket. Vid genereringen får vi förstås se till att den skapade texten tar hänsyn till svenskans genusregler så att vi inte får *en svår fall*. Observera att denna metod är svår att använda i praktiken men kan fungera bra i begränsade sammanhang, t.ex. sport, väder, manualer, etc.

Korpusbaserade metoder baseras på att man samlar in statistik från korpusar. Den klassiska IBM-metoden använder två olika sannolikheter: *översättning av ord* och *kombination av ord*. Översättningssannolikheterna samlas in genom att observera i *parallella korpusar* hur ord brukar översättas; ordföljdssannolikheterna kan samlas in i vilken stor enspråkig korpus som helst. Översättningssannolikheterna kommer då att visa att *fall* är en vanligare översättning av *case* än *kasus*, och att *svår* och *svårt* är vanligare från *difficult* än *besvärligt*. Ordföljdssannolikheterna hjälper oss att välja grammatiska följder (*svårt fall*) snarare än ogrammatiska följder (*svår fall*), och visar dessutom att av de möjliga grammatiska följderna så är *svårt fall* vanligare än t.ex. *besvärligt hölje*.

Det finns också mer avancerade korpusbaserade översättningsmetoder. I frasbaserade system (t.ex. Googles översättning) används översättningssannolikheter för *fraser* i stället för bara enstaka ord.

Du får full poäng på denna uppgift om du ger minst två (förnuftiga) förslag. Det kan t.ex. vara en djup och en ytlig metod, eller två ytliga t.ex. ordbaserad eller frasbaserad översättning.