Lösningsförslag för tentamen i Språkteknologi, DH2418, 2007-10-26

1. Textsammanfattning (4 poäng)

En lingvistisk metod som ordklassdisambiguering är mycket användbar i en textsammanfattare för att skilja ut innehållsord från funktionsord. En statistisk metod kan göra en frekvensberäkning av textens innehållsord. En heuristisk metod skulle kunna vikta upp innehållsord som finns i rubriker. Funktionsord är i de flesta textmängder långt mycket vanligare än innehållsord. För att hitta de vanligaste innehållsorden måste vi alltså först identifiera och sålla bort funktionsorden.

2. Ordklasstaggning (4 poäng)

En generell metod som passar på alla okända ord är att gissa på alla taggar som finns (men helst inte slutna ordklasser) och sedan låta disambigueraren välja.

Ordklass (och särdrag) för ett sammansatt ord ärvs från efterledet. En annan metod är alltså att gissa hur ordet är sammansatt (vilket kan göras med t. ex. Stavas metod) och sedan använda lexikoninformationen för efterledet. Om det finns flera möjliga sammansättningsuppdelningar med olika efterled så får man lägga till alla möjliga efterleds taggar och låta disambigueraren välja.

En tredje metod är att titta på ordets sista bokstäver och gissa på taggar som är vanliga för ord som slutar så.

3. Informationssökning (4 poäng)

a) Varje rad motsvarar ett ord/en term. Varje kolumn ett dokument/en text. Värdet i varje matriselement representerar hur viktigt man tycker ordet är för att beskriva texten. Värdena är tf*idf för ordet och texten.

b) Två texters likhet beräknas genom att jämföra deras två kolumner i matrisen. Jämförelsen kan göras genom att beräkna cosinus av vinkeln mellan dem.

4. Morfologi (4 poäng)

Stemming för ner orden till en paradigmoberoende stam, t. ex. “cykel”, “cyklar” och “cykling” stemmas till den gemensamma stammen “cykl”. En lemmatiserare håller sig inom böjningspradigmet, och skulle ge följande lemman cykel, cykla/cykel och cykling av exemplen ovan.

Stemming kan till skillnad från lemmatisering, användas för att överbrygga böjningsparadigmer och på så sätt öka täckningen i ett informationssökningssystem. Dock måste man se upp så att stemmingreglerna inte övergenererar, dvs. producerar allt för generella stammar. Detta leder ofta till en sänkning av precisionen (t.ex. genom att “tomte” och “tomat” båda stemmas ner till “tom”). Lemmatisering i sin tur är mer lingvistiskt exakt, och kan separera frukten banan från ordet bana i bestämd form (banan), vilket ger högre precision.

5. Språkstatistik (4 poäng)

Det är framförallt funktionsord, men också extremt vanliga verb, substantiv och namn som kommer i översta toppen av en frekvenslista. Korta ord som “är”, “och” och “i” hamnar i toppen. Toppen av frekvenslistan kan användas för flera olika saker, till exempel som en bra bas för en stoppordlista i en sökmotor.

6. Statistiska metoder (6 poäng)

Samla nya och gamla korpusar av texttyper som är relevanta för Språkrådet. Förmodligen är tidningstext mest relevant. Tokenisera texterna.

Det gäller att från den nya korpusen plocka bort så många ord som möjligt som inte är nyord. Börja med att plocka bort alla namn och förkortningar, antingen genom att använda en named-entity-taggare eller bara titta på versalinledda ord. Plocka vidare bort alla ord som inte ser ut som svenska ord, lämpligen genom att kolla alla bokstavsfyrgram som Stava gör. Kör resten av orden genom en rättstavningskontroll och plocka bort alla kända ord. Eftersom många nyord är nya sammansättningar så kan man inte sålla bort alla sammansatta ord. Plocka sedan bort alla ord som förekommer i gamla korpusen, och helst också böjningsformer av dessa, till exempel genom att bygga ett stavningskontrollprogram med orden i gamla korpusen.

Återstående ord är kandidater till att vara nyord. För att ett ord ska vara ett intressant nyord ska det ha förekommit ett par gånger i flera olika texter. Ta därför bort alla ord som inte uppfyller det. Om det är alltför många utländska ord (förmodligen mest engelska)
ord kvar kan man köra en engelsk stavningskontroll på resultatet och
kasta bort dom ord som godkänns.

Skicka sen resterande ord till Språkrådet.

7. Utvärdering (6 poäng)

a) Om man inte utvärderar delkomponenterna var för sig vet man inte vilken inverkan de har på systemets prestanda, och därmed inte vilken komponent det ger mest att arbeta mer med.

b) För att utvärdera taggaren kan man tvinga den att ge varje ord i en text exakt en tagg och jämföra den mot vad en människa satt för tagg på motsvarande ord i samma text. Detta handgjorda "facit" kallas ofta för en guldstandard. En baseline kan vara att sätta en slumpvis vald tagg på varje ord, en annan kan vara att alltid låta taggaren sätta den tagg vi tror är vanligast (t. ex. substantiv). Slår inte taggaren båda dessa mycket naiva metoder är det ingen ide att vi använder den. För att veta hur mycket mer vi ska jobba med taggaren innan vi kan nöjda med dess inverkan på systemets prestanda kan vi ta fram en s.k. human ceiling. För att göra detta kan vi låta flera människor tagga samma text och beräkna hur ofta de sätter samma tagg på samma ord. Då vet vi hur svår uppgiften är, och hur mycket till det lönar sig att arbeta med taggaren.

8. Textkategorisering/klustring (8 poäng)

Textmängder: Använd största delen av de taggade texterna som träningsmängd, resten som testmängd. Applicera sedan på de äldre texterna. Alternativt tagga en del av de äldre texterna manuellt.

Representation: Till exempel term-dokumentmatrisen. Stilistiska mått är antagligen också användbara.

Algoritm: Vilken etablerad övervakad maskininlärningsalgoritm som helst bör kunna ge ett bra resultat. Enklast - en centroid per texttyp, kolla likhet mot dessa. Naive Bayes-klassificerare är mycket bra på att lösa den här typen av uppgifter. Det är dock bra att kontrollera/utvärdera flera maskininlärningsmetoder!

Utvärdering: Se textmängder ovan för testmängd. Utvärdera med t. ex. precision. Observera att man bör ta hänsyn till att det kanske fungerar olika bra för olika texttyper. Metoden kanske behöver ta hjälp av människor när den är osäker, vilket kan behöva vägas in i utvärderingen.

9. Maskininlärning och skrivverktyg (5 poäng)
Det finns givetvis flera olika lämpliga metoder men Random Indexing är en lämplig metod eftersom den bland annat är resursnål och enkel att anpassa för en specifik texttyp.

Random Indexing (RI) är en metod att utvinna relaterade ord ur fritext. Kortfattat associeras varje unikt ord med en *Random Label*. Metoden skapar sedan en *kontextvektor* för varje ord genom att gå igenom en textmängd ord för ord (*indexering*). Kring varje ord upprätthåller metoden ett *kontextfönster* på t. ex. fyra ord före och fyra ord efter. Två ord som förekommer i liknande kontexter, dvs tillsammans med liknande ord, kommer härigenom att få liknande kontextvektorer. Man avgör deras likhet genom att jämföra kontextvektorerna med ett *avstånds- eller likhetsmått*, t. ex. det kartesiska avståndet mellan dem.

Man kan använda ett "random index" för varje texttyp för att ge synonymer för just den texttypen. För varje texttyp använder man de texter som finns för den (för att fånga upp ordanvändningen som är specifik för texttypen) plus någon mer generell textmängd (för att fånga upp allmänt språkbruk, det som en ovan skribent använder). Som generell textmängd kan man använda alla texter eller texter någon annanstans ifrån.

När en användare skriver ett mer allmänt ord i en specifik texttyp kan den texttypens "random index" föreslå relaterade specialiserade ord. För att göra det måste man alltså beräkna likheten mellan det allmänna ordet och alla specialiserade. De mest lika presenteras som förslag. Att beräkna alla likheter är väldigt tidskrävande och beräknas med fördel i förväg.

10. Grammatikkontroll (5 poäng)
En större långsiktig satsning skulle ge en möjlighet att utveckla gedigna generella resurser, som en stor ordklasstaggad balanserad textkorpus, en stor felkorpus samt en morfologisk analysator och generator. Utifrån felfrekvenser i felkorpusen skulle man bestämma vilka fel som främst skall attackeras. Det är också viktigt att spara undan en ganska stor del av felkorpusen som en testkorpus. Grammatikkontrollen i sig kunde utvecklas med en bas i flera olika metoder. En regelbaserad metod skulle finnas med – de har visat sig ge bra och pålitliga resultat. Denna skulle kräva omfattande utvecklingsarbete, och för att spara på resurser skulle en färdig regelmotor köpas in. Statistiska metoder skulle kunna bidra genom att upptäcka andra fel än den regelbaserade. Ytterligare en metod baserad på någon annan maskininlärningsalgoritm skulle vara bra, om man vill ha en grammatikkontroll som agerar utifrån olika metoders gemensamma beslut genom ett röstningsförfarande.

En begränsad satsning skulle kräva mer försiktighet, men en lämplig början skulle vara att få fram en manuellt ordklasstaggad korpus, på i storleksordningen 50 000 ord. Denna del skulle kräva mycket resurser eftersom korpusinsamling och uppmärkning i sig kan vara nog så krävande för mindre språk. Vidare skulle man kunna ta en fritt tillgänglig ordklasstaggare (open source) och

träna denna på det taggade materialet. Det skulle kunna ge en hyfsad ordklasstaggare som skulle användas för att märka upp mer text. En tränad lingvist med stora kunskaper i sydsamiska skulle kunna hyras in för att snabbt kunna upptäcka de mest fatala taggningsfelen, och korrigera dessa för bättre framtid träning av en betydligt större korpus.

Att bygga ett regelbaserat system skulle kräva stora resurser och omfattande manuella insatser därför är det bättre att satsa på statistiska metoder och maskininlärning. En felkorpus skulle kunna skapas genom enkla reguljära uttryck som matchar den felkonstruktion man intresserad av i korpusen samt plocka fram en annan ordform med samma lemma från korpusen, och användas som utbytesord. Som ett led i denna process märks felen upp, och kan användas som träningsmaterial. Vilka fel som skall attackeras får ske i samråd t. ex. med några lärare i sydsamiska.

Eftersom resurserna är begränsade skulle man bli tvungen att skära ner på korrektionsdelen av grammatikkontrollen. Andra prioriteringar kan givetvis göras men denna del är kanske mest uppenbar eftersom den kostsam och svår. Kvaliteten på hela grammatikkontrollen blir totalt sett väsentlig lägre än för den grammatikkontroll som utvecklas i den större satsningen.