Lösningsförslag för tentamen i Språkteknologi, DH2418 2008-10-22

1. Informationssökning

PageRank är ett försök att modellera sidors trovärdighet. Grundidén är att en sida som många länkar till är trovärdig.

En sida får en ranking som beror av rankingen de sidor som länkar till den har. För en sida som länkar till många sidor fördelas dess ranking jämnt mellan dessa.

En grupp av sidor som bara länkar till varandra utgör ett problem. De blir en ranksänka. Man löser detta problem genom att införa en dämpningsfaktor.

2. Språkgenerering

Språkgenerering är svårt eftersom att det ofta är svårt att bedöma i vilken kontext som det som genereras hamnar, och därmed om det generade verkligen har någon betydelse för en människa. I fallet med genererade rättningsförslag från en stavningskontroll kan det lätt bli så att vid rättstavning av t.ex. sammansättningar så kan ord genereras som är svåra att förstå vad de betyder i det aktuella sammanhanget eller överhuvudtaget. En felstavning som ”haklbanan” (halkbanan) kan generera förslag som hallbanan och halbanan från en svensk rättstavningskontroll.

I ett maskinöversättningsprogram kan man i många fall få till syntaxen korrekt, men det är osäkert vad den genererade meningen egentligen betyder (semantik) i just i det språk och den genre som översätts. Företeelsen finns kanske inte ens på målspråket. Om den lexikala databasen framförallt är baserad på myndighetsspråk kan flera lustiga översättningar uppstå för att för få betydelser av orden finns tillgängliga för programmet. En sporttext som innehåller ord som ”mål” och ”pass” översätts kanske till engelska som ”objective” repsektive ”passport”.

3. Ordklasstaggning

a) Det är en förenklad modell av naturligt språk, i detta fall en modell för vilka ordklasser ord har i en mening.

b) Antagandet är: sannolikhetsfördelningen för vilken ordklass ett ord har ges av vilka ordklasser ordet kan ha och vilken ordklass föregående ord har.

c) Tillstånd: ordklasser, signaler: ord.

4. Morfologi

Den mest tydliga mekanismen för bildning av nya ord i svenskan är sammansättningar av redan existerande ord. Man kan sätta samman framförallt substantiv, adjektiv, verb och egennamn för att bilda nya ord när behov uppstår. Sammansättningsmekanismen tillåter att flera led sätts samman, men tvåledssammansättningar dominerar, som t.ex. kantarellhund (en hund som är bra på att lukta sig fram till kantareller i skogen). De slutna ordklasserna berörs inte av denna mekanism, t.ex. är det svårt att se hur man skapar sammansättningar av prepositioner, konjunktioner, determinerare, eller subjunktioner.

5. Flertydighet

Ordklasstaggningsproblemet måste ses som ett mer eller mindre löst fall av flertydighetsproblemet, dvs. att bestämma rätt ordklass för ett ord i en given kontext. Visserligen varierar resultaten mellan olika språk bland annat beroende på hur mycket av syntaxen som markeras genom morfologin, men i cirka 96 % av fallen gör en standardtaggare av idag rätt (lågt räknat). Dagens ordklasstaggare är tillräckligt bra för att vara mycket användbara i en rad tillämpningar. Givetvis spelar de sista procenten i korrekthet roll, och feltaggningar får vissa spridningseffekter, men för engelska och svenska kan man lugnt ta en ordklasstaggare, och få mycket nytta av den.

Ordens betydelser i olika kontexter har en bra bit kvar till någon tillförlitlig lösning för löpande text. Om vi vill disambiguera semantiskt flertydiga ord, och om vi vill ha en generell lösning på problemet så finns den ännu inte.

6. Mänskliga resurser

Om vi förutsätter att det finns en bra ordklasstaggare så kan vi ta den som utgångspunkt för att identifiera olika namn. Eftersom det är webbtexter som skall taggas bör vi bestämma med kunden om det endast är svenska namn som skall ingå, eller om t.ex. även engelska namn skall ingå. Vi bör också bestämma om det är vissa genrer som är viktigare än andra. Det kan tänkas att kunden mest intresserar sig för texter om ekonomi skrivna på svenska.

Om vi har en ordklasstaggare, har vi också någon form av lexikon där en hel del namn finns med. Vi kan plocka ut dessa, och låta lingvisten märka upp dem med lämplig tagg för t.ex. förnamn, efternamn, företagsnamn, ortsnamn, länder. Givetvis finns det flera andra viktiga typer men på grund av projektets längd koncentrerar vi oss på dessa. När detta är gjort har vi ett ganska bra startlexikon som dock är långt ifrån är fullständigt. Detta startlexikon kan användas som en initialgissning för en maskininlärningsalgoritm, och tillsammans med en ordklasstaggare som en bas för lingvistens annoteringsarbete, se nedan.

En stor del av lingvistens tid bör spenderas på uppmärkning av olika namntyper i kontext i den eller de genre(r) som kunden är intresserad av. Vi behöver exempel på hur olika namn används, och vilka ledtrådar som finns om dem i kontexten. När detta görs kommer lingvisten att upptäcka att det finns flertydighetsproblem att hantera, t.ex. att Volvo Personvagnar ibland endast benämns endast Volvo. Lingvisten kommer också att få

hantera att en del förnamn används som efternamn (t.ex. Ingvar). En del andra namn sammanfaller med vanliga ord som Tapper och Stolt, och kan ställa till det för namnigenkänningen. Det finns sammanfattningsvis en rad problem med namnigenkänning som bäst upptäcks genom inspektion av data, och efterföljande uppmärkning. Med vårt uppmärkta data kan vi träna och utvärdera en maskininlärningsalgoritm för namnigenkänning. Arbetsuppgifterna för lingvisten blir alltså att först bygga upp ett baslexikon, och sedan annotera olika namn i av kunden givna genrer. Genom att successivt märka upp alltmer material maskinellt kommer lingvistens arbete alltmer att handla om att rätta maskinella taggningar än att analysera namn för namn. Ett datorstöd som håller reda på lingvistens tidigare taggningar av samma typ gör att lingvistens arbete förenklas.

7. Utvärdering

Jag skulle förklara för min uppdragsgivare att jag kommer att använda mig av K-fold crossvalidation, eller dess extremare variant – leave-one-out – för att iterativt kunna träna på större delen av mängden och utvärdera på en mindre tills dess att man slutligen tränat och utvärderat hela på hela mängden exempel. På så vis kan vi utnyttja lingvisternas arbete maximalt.

För att säkerställa resultaten måste vi dock försöka ställa upp någon form av minimikrav, och helst också ett mål att sträva mot. Minimikravet brukar kallas för baseline och i det här fallet kan man t.ex. tänka sig två mycket enkla sådana. Den ena är att tagga varje nytt ord med en framslumpad ordklasstagg, och den andra är att tilldela varje nytt ord den vanligaste taggen satt på något ord av lingvisterna. Kan vi inte slå dessa naiva metoder så kan vi lika gärna använda dem som att använda den ordklasstaggaren vi håller på att utvärdera.

För att veta hur långt taggaren har kvar för att nå upp till lingvisternas taggningskvalitet så måste vi först ta reda på vilken kvalitet lingvisterna egentligen håller. Detta kan vi t.ex. göra med en s.k. Human Ceiling, dvs att vi låter två eller fler lingvister tagga upp samma mängd text och så mäter vi samstämmigheten dem emellan. Därefter kan vi mäta vår taggares prestanda mot dels vår baseline, dels mot vår Human Ceiling, och förhoppningsvis hamnar den närmare den senare än den förra.

8. Informationssökning och maskininlärning

Extrahera träningsmängd ur det historiska materialet genom att söka på alla förekomster av börsnoterade företag. Markera företag som efter den aktuella textmassan steg eller sjönk som positiva respektive negativa exempel, övriga som ointressanta.

Ett företag kan representeras i en term-dokument-matris som sammanslagningen av alla texter i vilket namnet förekommer. Det är antagligen bra att tillämpa mycket filtrering – att ha en stor stopplista. Kanske kan man extrahera en representation som försöker avspegla om texterna innehåller positiva eller negativa omdömen. För att göra det kan ytterligare en maskinlärningsalgoritm användas.

Tillämpa sedan en maskininlärningsalgoritm på träningsdata, så att den lär sig skillnaden

mellan företag vars aktier kommer att stiga, sjunka eller vara oförändrade. Den enklaste tänkbara är en centroid-baserad metod; textrepresentationen för företag som stiger, sjunker och är oförändrade sammanfattas med varsin centroid.

När ny aktuell text blir tillgänglig konstruerar man företagsrepresentationer på samma sätt för dessa texter. De nya företagsrepresentationerna jämförs med centroiderna. De företag som är mest lika ”köp”-centroiden signaleras som värda att köpa, osv.

9. Grammatikkontroll

Om vi väljer regelbaserade metoder skulle vi behöva två regelsamlingar. Vi behöver först en regelsamling som hittar så många fel som möjligt för hög täckning, och sedan en regelsamling som är mer precis för att uppnå hög precision. Utan att ha genomfört en rejäl utvärdering är det svårt att säga hur pass naiva vi kan vara för att hitta så många felaktiga ”dom” som möjligt i text. Men en bra början skulle vara att leta efter ”dom” som följs av optionellt adjektiv, och ett substantiv i ett ordklasstaggat indata. Det vill säga ett mönster där ”dom” används som determinerare. Det andra mönstret att basera en regel på skulle vara preposition följt av ”dom”. För att uppnå riktigt hög täckning måste vi våga leta efter andra förekomster av ”dom”, t.ex. när ”dom” inte föregås av determinerare eller adjektiv för att undvika att få träffar på ”en fällande dom” osv. Dessa mönster skulle säkerligen räcka för att hitta många fel, och kan därmed utgöra den första regelsamlingen.

Förmodligen får vi ett stort antal falska alarm på grund av att ”dom” är semantiskt flertydigt (juridisk term och kyrka); att två nominalfraser kan stå i rad; att skribenter inte alltid följer förutsägbara normer när de skriver samt att ordklasstaggaren ibland gör fel. Genom att testköra våra regler på stora mängder text, och genom att strama åt våra regler utifrån de falska alarm som uppstår kan vi få fram en regelsamling som ger hög precision. I användargränssnittet låter vi användarna välja mellan hög täckning eller precision, och beroende på användarens val så väljs respektive regelsamling.

10. Namnmatchning

Tillämpning 1: Här vill man att namn som uttalas lika ska matcha.
Använd en rättstavningsalgoritm med uttalsmetrik, alltså där avståndet mellan två ord (namn) bestäms av hur stor skillnad det är på uttalen, till exempel hur många ljud som skiljer. Om inget namn skiljer på mindre än två ljud så finns det nog ingen bra matchning.

Tillämpning 2: Här har forskaren själv skrivit sitt namn i artikeln, så den bör vara rättstavad, frånsett sånt som att ringar och prickar över bokstäverna kan ha fallit bort, att bara förnamnsinitialer används och att förnamn och mellannamn kan ha fallit bort. Jämför därför namnen med en metrik som tar hänsyn till det. Kolla också att författaren har rätt universitetstillhörighet, dvs. kolla om KTH, Kungl. tekniska högskolan, Royal institute of technology eller liknande varianter nämns efter författarnamnet.