Datorn behöver statistik och grammatik

ANNA SÅGVALL HEIN

Det är lätt att skoja med en del resultat av maskinöversättning: "Vad kan vi lära av det" blir "What can we faith of it". Då gör man sig dum. I ett språkpolitiskt perspektiv kan maskinöversättning förstärka demokrati och medborgarinflytande, och den tekniska utvecklingen går fort. Anna Sågvall Hein, professor i datorlingvistik, ger här en grundkurs. Vilken översättningskvalitet är önskvärd? Vilka olika metoder finns för maskinöversättning? Vad står på tur för forskningen?

Maskinöversättning handlar om att få datorer att översätta från ett språk till ett annat. Erfarenheten visar att det med varierande resultat är möjligt. Man kan förse datorer med språkkunskap och programmera dem till att använda kunskapen för att översätta. Den stora frågan är hur bra det kan bli. Rent allmänt kan man säga, att ju mer språkkunskap översättningsprogrammet har till sitt förfogande, desto bättre blir översättningen. Maskinöversättning inriktas som regel mot fack- och brukstext av olika slag. Skönlitterär översättning är en konst i sig, där datorn främst kan erbjuda språkteknologiska hjälpmedel i form av lexikon, texter och språkprov jämte språkligt anpassade sökmetoder.

Kvalitet

När det gäller mänsklig översättning är kraven på kvalitet höga. Man förväntar sig vanligen samma språkliga kvalitet på den översatta texten (målspråkstexten) som på originaltexten (källtexten), och att målspråkstexten förmedlar innehållet i källtexten på ett adekvat vis. Vid maskinöversättning är bilden mera varierad. Här handlar det inte enbart om att få fram en översättning som håller samma, eller i det närmaste samma, kvalitet som en mänskligt översatt text. Det finns också andra,

mindre ambitiösa syften. Ett kan vara att snabbt få fram en grovöversättning som ger en ungefärlig uppfattning om textens innehåll. Då kan man överse med vissa språkliga brister, som t.ex. kongruensfel, otillfredsställande artikelbruk, felaktig ordföljd etc. Man kan då tala om *begriplighetskvalitet*. Ett annat syfte kan vara att skapa en bas för informationssökning på olika språk. Sådan sökning inriktas som regel mot texternas innehållsord, och i det sammanhanget spelar grammatiska brister i de översatta texterna mindre roll.

Begriplighetskvalitet är ett mycket ospecifikt begrepp. Om de fel som finns i översättningen inte är fler och mer allvarliga än att det lönar sig att redigera översättningen, så kan man tala om *redigeringskvalitet*. Det är också ett vagt uttryck, och olika översättare gör olika bedömningar. I kommersiella sammanhang har man jämfört den tid det tar att maskinöversätta och redigera en maskinöversatt text med den det tar att översätta manuellt från grunden. Vissa uppgifter pekar på att man gör en tidsvinst på mellan 50 och 70 procent med maskinöversättning.

Mycket forsknings- och utvecklingsarbete läggs ned på att få fram metoder för att utvärdera översättningskvaliteten. Det handlar både om mänskliga bedömningar och metoder för att mäta kvaliteten automatiskt. Vid mänsklig utvärdering brukar man ta fasta på allmänna kriterier som gäller begriplighet och trohet mot källtexten. Sådan utvärdering är tidskrävande och dyr och kan inte tillämpas storskaligt. Olika bedömare ger dessutom ofta olika svar, som måste vägas samman. Det behövs alltså kompletterande metoder. De automatiska utvärderingsmetoderna går ut på att låta datorn jämföra den maskinöversatta texten med mänskligt översatt text. Jämförelsen resulterar i likhetsmått, som visat sig stämma relativt väl med mänsklig utvärdering. En förutsättning är dock att referenstexten håller hög kvalitet. Allra bäst faller måtten ut om man har tillgång till flera olika referensöversättningar.

Språkligt avancerade översättningssystem kan anpassas till avgränsade ämnesområden och texttyper, så kallade domäner. Då kan man uppnå en redigeringskvalitet, som ligger mycket nära vad som krävs för publicering. Exempel på domäner som visat sig lämpade för maskinöversättning är väderleksrapporter, datormanualer, litteratur för underhåll av bilar och kursplaner.

Domänanpassning behövs både för översättningssystemets lexikondel och grammatikdel. I en text om bilunderhåll kan man räkna med att finna tekniska termer av ett slag som inte finns med i ett allmänspråkligt lexikon, t.ex. *axialspel*. Ordet *åker* som verbform finns troligen med, men inte i betydelsen *odlingsyta*. Att få datorn att skilja ut olika betydelser är svårt, och ju färre alternativ översättningssystemets lexikon upptar, desto bättre förutsättningar för rätt val i översättningen. Ämnesanpassningen innebär alltså att man kompletterar lexikonet med ämnesspecifik terminologi men också gärna reducerar antalet betydelser knutna till flertydiga ord. Vanligen bygger man det domänanpassade lexikonet utifrån en samling översatta texter av aktuellt slag, en så kallad *översättningskorpus*. Domänanpassning av grammatiken innebär att man kompletterar med texttypiska konstruktioner och stildrag, t.ex. användning av bestämd och obestämd form.

Vid maskinöversättning av allmänspråkliga texter kommer man inte lika

långt i kvalitet som med domänavgränsad text. Framför allt beror det på att lexikonet måste ta upp alla tänkbara betydelser av likalydande ord, och att datorn sedan måste försöka skilja ut den rätta i textsammanhanget. När man översätter allmänspråklig text, får man som regel nöja sig med begriplighetskvalitet.

Översättningsmetoder

Det finns olika metoder för maskinöversättning. Grovt sett kan de grupperas i direktöversättning, transferöversättning och statistisk översättning.

Direktöversättning är den enklaste och äldsta metoden. Översättningen sker ord för ord eller fras för fras med hjälp av ett översättningslexikon. Språkskillnader som yttrar sig som översättningsproblem hanteras med hjälp av specifika regler, vilka vanligtvis formuleras för enskilda ord. Olika direktöversättningssystem uppvisar mycket olika resultat; det beror på omfång och kvalitet på lexikonet samt på hur väl översättningsreglerna täcker de aktuella översättningsproblemen. De kan ha att göra med ordföljd, böjning, prepositionsbruk och tempus. Ett avancerat direktöversättningssystem har också viss grammatisk kompetens. Den omfattar vanligen ordklass- och böjningsinformation samt vissa ordföljdfrågor. Men någon fullständig meningsanalys görs inte. Därför kan datorn få svårt att skilja mellan till exempel subjekt och objekt.

Ett system för **transferöversättning** grundar sig på en fullständig analys av hela meningen; det är en avgörande skillnad gentemot ett direktöversättningssystem. Endast genom en fullständig analys av hela meningen kan man med säkerhet känna igen subjekt, objekt och andra grammatiska funktioner.

Översättningen i ett transfersystem sker i tre huvudsteg: 1) analys av källspråksmeningen med en satslösningsstruktur som resultat, 2) översättning av källspråkets satslösningsstruktur till en satslösningsstruktur på målspråket samt 3) genering av en målspråksmening utifrån målspråkets satslösningsstruktur. Själva översättningssteget benämns *transfer*, d.v.s. överföring. Det är också där som översättningsproblemen tas om hand med hjälp av särskilda transferregler.

Man vinner mellan 50 och 70 procent av tiden med maskinöversättning.

Satslösningsstrukturen innehåller uppgift om grammatiska funktioner. Transferreglerna kan därigenom göras mer generella än översättningsreglerna i ett direktöversättningssystem, vilka vanligtvis tar sikte på enskilda ord. Det innebär att antalet översättningsregler kan göras mindre, och att det är lättare att få regelsystemet heltäckande än i ett direktöversättningssystem. Avgörande för hur bra transfersystemet översätter är kvalitet och täckning på grammatiken.

Grammatiska skillnader som måste kunna hanteras av ett översättningsprogram för kvalitetsöversättning kan yttra sig på många vis. I ett svensk-engelskt sammanhang är t.ex. ordföljden mellan subjekt och predikat ett generellt problem. Där svenskan har omvänd ordföljd, alltså predikatet före subjektet (*då kom han*) till följd av inledande adverbial har

engelskan rak (*then he came*). För att systemet ska kunna ge en korrekt engelsk översättning av en svensk mening med omvänd ordföljd måste det alltså hitta subjektet. Det kräver i sin tur att satsinledande adverbial känns igen, oberoende av om de utgörs av enskilda ord eller hela satser. För uppgifter av detta slag är ett transfersystem mer lämpat än ett direktöversättningssystem.

Med hjälp av transferregler kan man också uttrycka lexikala val som kan göras utifrån den grammatiska kontexten. Om man t.ex. vill översätta verbet *anta* till engelska, så finns det åtminstone två alternativ, *suppose* eller *admit*. För valet mellan de båda kan man utgå från objektet. Om objektet är en att-sats, bör översättningen bli *suppose*, medan den bör bli *admit* om objektet är en person. Exempel av liknande slag är *lämna information → provide information* (inte *leave*), *ägare till en bil → owner of a car* (inte *to*). Även många andra problematiska konstruktionsskillnader mellan språken kan uttryckas mer generellt och bättre i ett transfersystem än i ett avancerat direktöversättningssystem.

Väderleksrapporter, kursplaner och litteratur för bilunderhåll passar maskinöversättning.

Så kallad **statistisk maskinöversättning** är en översättningsstrategi som kommit på bred front under senare år. Det är en form av direktöversättning utan särskilda översättningsregler. Det specifika med metoden är att översättningslexikonet skapas automatiskt genom datamaskinell bearbetning av stora mängder översatt text. Käll- och måltext länkas ihop meningsvis, frasvis och ordvis, huvudsakligen med hjälp av statistiska metoder. Utmärkande för detta så kallade länklexikon är att det innehåller många olika alternativ men ingen lingvistisk information. Förutom länklexikonet innehåller de statistiska översättningssystemen en enkel ordföljdsmodell för målspråket, som också den är automatgenererad på statistisk grund. Därtill kommer metoder för att välja ut den mest sannolika översättningen bland de många alternativ som skapas.

Mycken forskning ägnas i dag åt statistisk maskinöversättning och resultaten är förvånansvärt bra. Men det är svårt att komma fram till tillförlitliga utvärderingsmodeller, som kan tillämpas storskaligt.

Forskning pågår också om hur man ska kunna komplettera de statistiska modellerna med lingvistisk kunskap för att förbättra kvaliteten. Den största fördelen med statistisk maskinöversättning är att det kan gå på ett par veckor att skapa ett system för ett nytt språkpar, förutsatt att man har stora översättningsmängder av god kvalitet. Det handlar om hundratusentals meningspar. De första försöken med statistisk maskinöversättning gällde engelska och hindi respektive arabiska.

Systran

Det mest använda översättningssystemet i dag heter *Systran*, som enligt uppgift från företaget utför sju miljoner översättningar per dag. Man hävdar vidare att 500 årsverken har investerats i forskning och utveckling av systemet. Det kan närmast

betraktas som ett avancerat direktöversättningssystem. Systran står för *Sys*tem *trans*lation. Det kom i en första version för översättning mellan ryska och engelska redan 1969. I dag omfattar systemet mer än 36 språkpar. Det finns i en kommersiell version, som i huvudsak är inriktad på översättning av allmänspråk, men till vilken man kan köpa särskilda ämneslexikon. Man kan prova den allmänna versionen på <http://babelfish.altavista.com>. Vidare finns det en version som är speciellt anpassad till den europeiska unionens behov, EC Systran. Med hjälp av denna version utfördes år 2000 närmare 100 000 översättningsuppdrag. Lexikonet omfattar mer än 1 600 000 lexikonenheter fördelade på 20 ämneslexikon. Kvaliteten på översättningen mellan de olika språkparen varierar beroende på hur mycket de olika delsystemen utvecklats och ”tränats”. Längst har man kommit med engelska-franska.

Systranmoduler för översättning från svenska och danska till engelska utvecklades under 2003 i samarbete mellan EU, det franska företaget Systran och, för svenskans del, Institutionen för lingvistik vid Uppsala universitet. Från forskningsassistenterna Ebba Gustavii och Eva Pettersson lånar jag fyra exempel på översättningsproblem, som dök upp i arbetet med den svensk-engelska modulen.

Ex. 1: *Enskilda företagare som inte bildat bolag klassificeras hit.* → *Individual entrepreneurs that have not formed companies are classified here.*

Systemet känner igen *bildat* som uttryck för perfektum, trots att hjälpverbet är utelämnat, och översätter korrekt *have formed* med negationen *not* på rätt plats. Passivformen *klassificeras* översätts också korrekt i rätt tempus.

Ex. 2: *När byarna kontaktades hade de inte ens utsatts för influensa.* → *When the villages were contacted had they not even been exposed to flu.*

Systemet hittar inte subjekt och predikat och ger därför fel ordföljd. Svårigheten med att finna subjektet ligger i att meningen inleds med en bisats, och att systemet inte gör någon fullständig satsanalys.

Ex. 3: *Vad kan vi lära av Arrawetestammen?* → *What can we faith of the Arawete?*

På grund av frågesatskonstruktionen hittar systemet inte sambandet mellan *kan* och *lära* och ser därför inte heller att *lära* är ett verb.

Ex. 4: *Extrapoleringen går till så här.* → *The extrapolation goes to so here.*

Systemet känner inte till partikelverbet *gå till* och översätter därför felaktigt ord för ord med olyckligt resultat.

Den svensk-engelska modulen utvecklades under åtta månader. Målsättningen var att översättningarna skulle uppvisa begriplighetskvalitet. När projektet var slut återstod fortfarande en del problem, till exempel frågor, sammansatta konjunktioner, ordföljd i underordnade satser och vissa uttryck som *65-åring* (översätts ej), *klockan 12* (*o'clock 12*). Ändå visade det sig att den svensk-engelska modulen hävdade sig väl i jämförelse med andra moduler som utvecklats under väsentligt längre tid, t.ex. den grekisk-engelska. Det svensk-

engelska systemet, jämte ytterligare 34 språkpar, finns att prova på <www.systransoft.com>. Den svensk-engelska modulen tillsammans med ytterligare 14 språkpar finns numera också tillgänglig kommersiellt.

Multra och Mats

Ett exempel på ett transfersystem är *Multra*, som utvecklats vid Uppsala universitet. Multra översätter från svenska till engelska men kan vidareutvecklas till att omfatta andra språkpar och språkriktningar. Målsättningen med Multra är hög översättningskvalitet inom begränsade domäner. Den första domän som utforskades i arbetet med Multra var servicelitteratur för buss- och lastbilsunderhåll från Scania CV AB; här följer ett exempel på översättning med Multra:

> Ex. 5. *Eftersom denna beskrivning behandlar flera liknande modeller av växellådor, kan utseendet på komponenter skilja sig något från det som visas på bilderna.* → *As this description deals with several similar models of gearboxes, the appearance of components can differ slightly from what is shown on the illustrations.*

Översättningen bygger på en fullständig analys av den svenska meningen. Meningen har omvänd ordföljd beroende på ett inledande adverbial, en bisats. Bisatsen har analyserats, och analysatorn har återfunnit det böjda verbet *kan* i positionen därefter och subjektet *utseendet på komponenter* i positionen närmast efter verbet. Subjektet och övriga delar av meningen har översatts, och den engelska meningen har kunnat genereras med rak ordföljd i enlighet med den engelska grammatiken.

Som andra transfersystem ställer Multra stora krav på språkkunskap och det är svårt att få fram grammatik som täcker all meningsvariation. För meningar som inte täcks helt av grammatiken produceras ingen sammanhängande satslösningsstruktur, enbart delanalyser. En översättning, som bygger på delanalyser är mindre tillförlitlig än en som grundar sig på en fullständig analys. I sin ursprungliga form översatte Multra därför enbart sådana meningar där det fanns en fullständig analys; övriga meningar lämnades oöversatta. Men erfarenheten visade att det är bättre med en bristfällig översättning än ingen översättning alls. Multra har därför kompletterats med strategier för översättning av delanalyser i de fall grammatiken är otillräcklig.

Detta har skett i *Mats*, som är en utvidgad version av Multra. Mats levererar alltid en översättning; det är med andra ord robust. De delar av den översatta texten som inte bygger på en fullständig satslösningsstruktur markeras med en särskild färg för att underlätta granskning och eventuell redigering. Även andra problem som har att göra med ofullständighet i språkbeskrivningen färgmarkeras, t.ex. att ett ord saknas i lexikonet. Matssystemet är därigenom inte bara robust utan också transparent, d.v.s. man kan bland annat se vilken analys som gjorts och vilka regler som tillämpats.

Den som vill se prov på en översättning med Mats kan gå till <www.lingfil.uu.se/MATS/demo.html>. Det är en jordbrukstext från EU. Exemplet upptar också en jämförelse med en mänskligt utförd översättning, d.v.s. en referensöversättning, och två likhetsmått, som visar den maskinöversatta textens överensstämmelse med referensöversättningen. På

samma adress finner man också den färgmarkerade översättningen. Man kan också följa de olika stegen i översättningen av en exempelmening samt titta in i lexikonet. Fortsatt forskning och utveckling av Mats inriktas bland annat mot att effektivisera analysen, vidareutveckla grammatiken samt förbättra metoderna för robust översättning. Man arbetar också med att utveckla effektiva metoder för att anpassa lexikon och grammatik till nya domäner. Att inkludera nya språkpar är också angeläget.

Återanvändning

Tidigare översättningar utgör en viktig kunskapskälla i arbetet med maskinöversättning. Man brukar i det här sammanhanget tala om *återanvändning* av översättningar. Ett användningsområde är statistisk maskinöversättning, som nämnts ovan. Det finns nu datorprogram som automatiskt kopplar samman käll- och målspråksmeningar parvis med en precision på närmare 100 procent även om texterna skiljer sig åt i fråga om meningsindelning. Av sådana meningspar kan man bygga så kallade *översättningsminnen*, som kan användas både för manuell och automatisk översättning. Man kan också gå vidare med sådana meningslänkade texter och låta datorn söka ut översättningar av ord och sammanhängande fraser. Även dessa så kallade ordlänkningsprogram kan lyckas riktigt bra och fånga upp närmare 80 procent av översättningarna. De material som skapas av ett sådant program kan gå direkt in i ett statistiskt översättningssystem som dess översättningslexikon. De kan också vidareutvecklas och tillföras lingvistisk information för användning i andra typer av översättningssystem.

Hellre en vägbeskrivning på dålig svenska än en obegriplig på god kinesiska.

Utvecklingen av lexikonet för svenska Systran byggde på ordlänkningstekniker. Först togs en rå version av ett svensk-engelskt översättningslexikon fram. Därefter vidareutvecklades och förfinades en del av lexikonet för ett ämnesområde (jordbrukstexter). Den totala översättningskorpusen, ur vilken lexikonmaterialet extraherades, bestod av 118 svenska EU-dokument med engelska översättningar, som täckte 20 ämnesområden. Totalt handlade det om 773 551 löpande ord; jordbruksdelen bestod av 103 589 ord. Det resulterande lexikonet omfattar totalt 27 363 uppslagsord och jordbruksdelen 6 114 samt 300 flerordsenheter inklusive 127 partikelverb.

Framtiden

Allt oftare kommer den enskilde medborgaren att träffa på maskinöversatt text. Det är nu t.ex. möjligt att med en enkel knapptryckning få fram en översättning av en webbsida. Alltfler kommer att utnyttja den möjligheten. Särskilt intressant blir det givetvis att snabbt kunna få fram en grovöversättning från språk som man själv inte behärskar. Genom utveckling av den mobila IT-tekniken kommer man i ökande utsträckning att få tillgång till maskinöversättning via webben. Det är av stor betydelse, när man är ute på resor i länder där engelska inte är var mans egendom. I en praktisk situation kan det vara bättre att få en vägbeskrivning på dålig

svenska än en obegriplig på god kinesiska. Samtidigt finns i det i dessa sammanhang ingen kvalitetsgaranti, och man löper risk att vänja sig vid ett dåligt språk och påverkas av det i sin egen språkutövning. Det är alltså av stor vikt att kvalitetssäkra och höja kvaliteten på maskinöversättning.

Potentialen för fortsatt utveckling och forskning som leder till högre översättningskvalitet är stor. Det handlar inte främst om att finna nya metoder utan att vidareutveckla och kombinera dem som finns.

Även om man har ett bra översättningssystem så får man räkna med många månaders arbete för att anpassa det till nya domäner och språk. Det handlar om att bygga upp språkliga resurser samt träna och utvärdera systemen till önskad kvalitet. En viktig forskningsuppgift i det sammanhanget är att vidareutveckla de metoder som redan nu finns för att automatisera uppbyggandet av de språkliga resurserna. Det viktigaste råmaterialet för detta arbete är stora mängder redan översatt text. Att samla in sådana datamängder och lingvistiskt analysera och kategorisera dem är av strategisk betydelse för den fortsatta utvecklingen av maskinöversättningen.

Utveckling av maskinöversättning från och till svenska är viktigt för att stärka svenskans ställning. Med maskinöversättning från svenska till främmande språk kan svenska i ökande utsträckning användas som källspråk i yrkeslivet och risken för domänförluster minska.

Maskinöversättning från främmande språk till svenska kan göra den globala textproduktionen storskaligt tillgänglig, inte minst den som sker på Internet. Härigenom ökar möjligheterna för den enskilde medborgaren att förstå och ta del av den internationella utvecklingen, politiskt såväl som ekonomiskt. I det sammanhanget kan konstateras att engelskans andel av textmängden på Internet kontinuerligt minskar.

Uppenbart är också att det finns särskilda behov av maskinöversättning från och till invandrarspråken och de svenska minoritetsspråken för att stärka deras och språkbrukarnas ställning i det svenska samhället. Det är många och språkteknologiskt delvis outforskade språk det handlar om. Här framstår den vidare utvecklingen av de statistiskt baserade maskinöversättningssystemen som särskilt lovande. Man kan snabbt och med små personella resurser åstadkomma begriplig maskinöversättning, förutsatt att man har tillgång till stora mängder översatt text.

Att utveckla maskinöversättning för svenska är alltså en uppgift som bör placeras högt på den språkpolitiska dagordningen, inte minst av demokratiska och näringspolitiska skäl. Låt mig sluta med att än en gång nämna några av de mest angelägna uppgifterna: uppbyggandet av språkliga resurser för nya domäner och språk för existerande översättningssystem, storskalig insamling och analys av översatt text för olika slags återanvändning, förbättring av statistisk maskinöversättning med lingvistiska metoder samt vidareutveckling av metoder för automatisk kvalitetsbedömning. ■

Anna Sågvall Hein är professor i datorlingvistik vid Uppsala universitet.