Automatisk översättning

En analys av Google Translate

HÅKAN ANDERSSON
och ELIN JOHANSSON

KTH Datavetenskap
och kommunikation

Examensarbete
Stockholm, Sverige 2010

Automatisk översättning

En analys av Google Translate

H Å K A N A N D E R S S O N
o c h E L I N J O H A N S S O N

Examensarbete i datalogi om 15 högskolepoäng
vid Programmet för datateknik
Kungliga Tekniska Högskolan år 2010
Handledare på CSC var Johan Boye
Examinator var Mads Dam

URL: www.csc.kth.se/utbildning/kandidatexjobb/datateknik/2010/andersson_hakan_OCH_johansson_elin_K10038.pdf

Kungliga tekniska högskolan
Skolan för datavetenskap och kommunikation

KTH CSC
100 44 Stockholm

URL: www.kth.se/csc

Innehållsförteckning

Förord 3
Abstract 4
Inledning 5
Sammanhang 5
Bakgrund 5
Syfte 5
State of the Art 5
Utförande 6
Översättningsalgoritmer 6
Regelbaserade 6
Statistiska 7
Hybrida 7
Svenska till Engelska 8
Generella meningar 8
Felaktig meningsbyggnad 9
Enstaka ord 10
Engelska till Svenska 11
Generella meningar 11
Felaktig meningsbyggnad 11
Enstaka ord 12
Översättning på redan översatt material 13
Generella meningar 13
Felaktig meningsbyggnad 14
Enstaka ord 14
Resultat 15
Analys av data 15
Analys av algoritmen 15
Diskussion 16
Felkällor 16
Slutsatser 16
Referenser 18

Förord

Då vi är två stycken som har jobbat på denna rapport så har vi valt att göra vissa uppdelningar. Kursen kräver i detta fall att vi inför ett förord där vi delger läsaren med vem som har skrivit vilka delar i rapporten. Överlag kan det sägas att uppdelningen har varit sådan att Elin Johansson har behandlat det som har med språket eller lingvistiken att göra medan Håkan Andersson har arbetat med de mer tekniska delarna som bland annat behandlar algoritmen och maskinöversättningen. Denna uppdelning gäller både delarna benämnda Utförande och Resultat.

Vi har tillsammans jobbat med att utföra tester och samla ihop information från källor. Vi har även samarbetat i analyserna genom att diskutera fram rimliga slutsatser med varandra, även om dessa legat utanför vårt tilldelade område. Inledning och Diskussion har vi skrivit tillsammans.

Slutligen har vi läst igenom varandras texter och kommit med konstruktiv kritik samt gjort ändringar där sådana har krävts.

Abstract

Google Translate is a service that offers you instant translation between multiple different languages on websites as well as on text files. But how well does it really work and what are its strengths and weaknesses? This thesis aims to discover how well Google Translate works as a substitute for manual translation and also to analyze the algorithms behind this service.

Inledning

Sammanhang

Så länge olika språk har funnits så har språkbarriären inneburit ett problem för människan. Länge har det enda alternativet varit att ha en mänsklig tolk som, i tal eller skrift, fått agera som en bro mellan de olika språken. Under nittonhundratalet utvecklades dock en helt ny möjlighet för människor att kommunicera med varandra med hjälp av en mekanisk översättning (Hutchins *et al* 1992), men tekniken var långt ifrån felfri.

Mycket har hänt och nu under början av tjugohundratalet finns tekniken för mekanisk översättning tillgänglig för allmänheten genom fler olika tjänster. Vi har valt att fokusera på en av dessa, nämligen företaget Google's internetbaserade översättningstjänst Google Translate.

Bakgrund

Google startade som en söktjänst på internet 1998 med målet att "*organisera världens information och göra den universellt tillgänglig och användbar*". Under åren har flera andra tjänster tagits fram av Google för att främja detta mål, och Google Translate är en av dem. I februari 2009 utökades Google Translate med ytterligare språk och tjänsten täcker nu direkt översättning mellan 41 olika språk, eller 98% av de språk som läses på internet (Google 2010).

Sedan juni 2009 finns funktionaliteten för användaren att mata in en bättre översättning ifall denne inte skulle vara nöjd med Google Translate's översättning. På så sätt kan alltså användarna agera lärare åt den automatiska översättningen. Svenska har funnits tillgänglig i Google Translate sedan maj 2008 och det är mellan svenska och engelska som vi kommer att fokusera vår undersökning på.

Syfte

I följande rapport ämnar vi utföra och dokumentera en utvärdering i hur väl Google Translate fungerar på översättning mellan svenska och engelska. Vi har även för avsikt att diskutera dels kring pålitligheten hos denna tjänst och dels om automatisk översättning kan ses som ett substitut för den manuella varianten. Genom att, med hjälp av Google Translate, översätta ett antal texter mellan svenska och engelska så kan en utvärdering av hur bra Google Translate fungerar presenteras. Även en analys av Google Translate's starka respektive svaga sidor skall upprättas.

Den andra delen i projektet är att försöka analysera resultaten av våra tester ur ett algoritmperspektiv. Det vore intressant att veta om vi kan dra några slutsatser kring hur Google Translate's algoritm fungerar eller vilken typ av maskinöversättning den är baserad på.

Syftet med rapporten är att utvärdera pålitligheten hos automatisk översättning med hjälp av Google Translate. Samt att uppvisa en teori i hur algoritmerna bakom Google Translate kan tänkas fungera.

State of the Art

Det är svårt att definiera en *state of the art* i området för automatisk översättning då vi inte har tillgång till några säkra metoder för att mäta hur lyckad en översättning är. Dock är Google Translate en välkänd tjänst som används av människor världen över och med företagets konstanta arbete att utveckla och förbättra den så får vi anta att de strävar efter att hela tiden hålla sig konkurrenskraftiga och uppdaterade.

Utförande

Vi har valt att utmana Google Translate genom att (1) bedöma översättningen på generella meningar med korrekt meningsbyggnad och korrekt grammatik, (2) bedöma översättningen på meningar med medvetet felaktig meningsbyggnad samt (3) bedöma översättningen på enstaka ord. Dessa testfall är gemensamma både för översättning från svenska till engelska och från engelska till svenska. Avslutningsvis har vi valt att notera Google Translate's arbete på material som tjänsten redan har översatt en gång.

Varje stycke kommer att inledas med en eller flera fraser samt deras översättning enligt Google Translate för att sedan följas av en kort diskussion kring resultaten av detta test. Vidare diskussioner kring resultat eller funktionsdugligheten hos Google Translate's algoritm kommer att skötas i en separat del av rapporten.

Vi inleder med en förklaring av några olika tillvägagångssätt för att automatiskt översätta texter då det kan vara bra med en överblick i hur det fungerar. Värt att poängtera är att Google har valt att skriva sin egen algoritm för att sköta översättningen hos Google Translate och denna är givetvis en företagshemlighet (Google 2010).

Översättningsalgoritmer

Maskinöversättning kan delas upp i ett flertal grupper, bland annat regelbaserade, statistiska och hybrida. Regelbaserade kan i sin tur delas upp i transfer-based, interlingual-based och dictionary-based.

Regelbaserade

Transfer-based maskinöversättning är vanligast av de regelbaserade varianterna. Men för att en översättning skall kunna konstrueras korrekt behövs det en tolkning av källan som fångar betydelsen av meningarna. Det finns flera varianter av transfer-based men de flesta följer ett och samma mönster. Kortfattat kan man säga att de applicerar ett flertal fördefinierade språkregler till texten som skall översättas (Hutchins *et al* 1992). Nackdelen med Transfer-based maskinöversättning är att regler mellan samtliga språk måste finnas, detta är praktiskt taget omöjligt att uppnå. Algoritmen illustreras i bild 1 nedan.

Interlingual-based maskinöversättning fungerar så att källspråket översätts till en abstrakt språkoberoende formulering med en mycket enkel grammatik som i sin tur översätts till målspråket. Fördelen med en sådan maskinöversättning är att regler mellan varje språkpar inte behövs. En nackdel är dock att det skulle vara mycket komplicerat att skapa regler mellan det interlinguala språket och samtliga språk som talas idag (Hutchins *et al* 1992). Algoritmen illustreras i bild 1 nedan.

Dictionary-based maskinöversättning använder sig av lexikon och översätter ord för ord i en text. Detta sätt är inte pålitligt för meningar eller texter över lag då text på ett språk som översätts ordagrant till ett annat språk sällan har likvärd betydelse. Den här formen av maskinöversättning bör endast användas till att översätta exempelvis listor med ord (Hutchins *et al* 1992).

Statistiska

Statistical-based maskinöversättning angriper översättningsproblemet som ett sök-problem. Genom att låta en mening av källspråket agera indata så maximerar algoritmen produkten av en serie sannolikhetsmodeller som ett översatt utdata (Goutte 2009).

Hybrida

Hybrid maskinöversättning kan ses som en kombination av de regelbaserade algoritmerna och de statistiska. På så sätt kan den hybrida varianten utnyttja styrkorna från båda sorterna samt försöka minska ner de svagheter som kan tänkas finnas (Hutchins *et al* 1992).

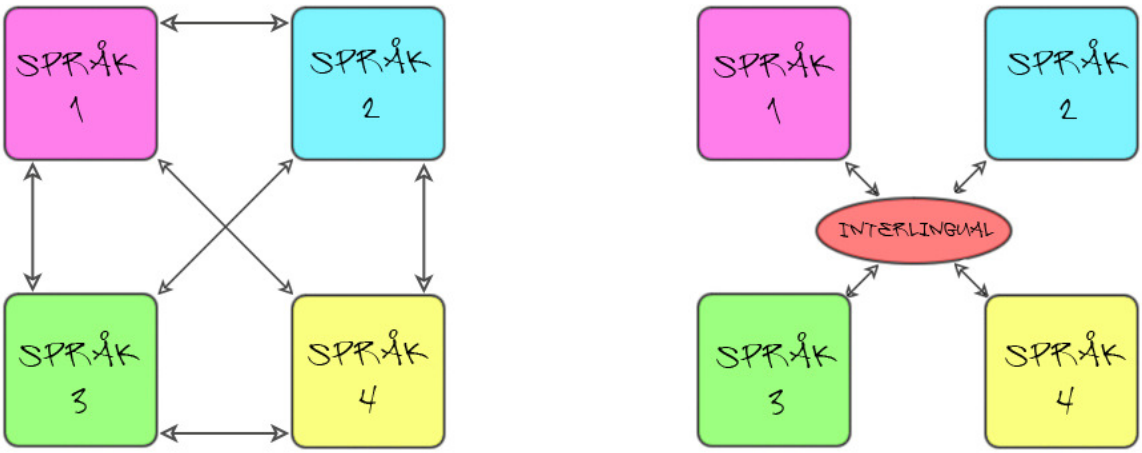

Bild 1 illustrerar till vänster en transfer-based maskinöversättning och till höger en interlingual-based

Svenska till Engelska

Nedan redovisas meningar översatta från svenska till engelska. De ord som översatts inkorrekt eller som av någon annan anledning är värda att notera kommer att markeras med rött.

Generella meningar

Svenska	Engelska
Blinka lilla stjärna där, hur jag undrar vart du är Bää bää vita lamm, har du någon ull?	Twinkle, twinkle little star, how I wonder where you are Baa Baa Black Sheep, have you any wool?
Denna handbok är avsedd för både lärare och studenter. De flesta kapitel behandlar ämnen som förmodligen är av större intresse för lärarna, men varje kapitel har också ett avsnitt som föreslår hur studenterna kan ha nytta av att läsa texten. (KTH 2010)	This manual is intended for both teachers and students. Most chapters deal substances that are probably of more interest to teachers, but each chapter has also a section that suggests how students can benefit from reading the text.

Tabell 1.1 Svenska till Engelska, generella meningar

Det övre exemplet i tabell 1.1 visar tydligt att Google Translate hanterar dessa välkända fraser utan några problem alls, trots att orden inte stämmer överens mellan språken. Vi kan se att *Blinka* översätts till *Twinkle, twinkle* och att svenskans *vita lamm* blir till *black sheep*, översättningar som inte alls är logiska om man inte känner till låttexterna hos de båda språken.

När det kommer till det undre exemplet har merparten av översättningen gett en korrekt engelsk text, men på två ställen hittar vi brister i översättningen. Att *behandlar ämnen* har översatts till *deal substances* är en ordagrann översättning från svenskan som inte håller i den engelska texten och som onekligen skulle förvirra en läsare. En mindre förvirrande, men fortfarande felaktig, översättning är *has also* som istället borde ha blivit *also has*.

Om vi funderar mer på en datateknisk nivå, vad för sorts algoritmer skulle på ett logiskt sätt kunna översätta på det här viset? Första tanken är då att någon form av statistisk översättning har tillämpats, att *blinka* blir *twinkle, twinkle* visar egentligen att en interlingual eller dictionary baserad översättning inte har tillämpats i det här fallet.

Det andra exemplet är lite mer komplicerat, här skulle det mycket väl kunna vara antingen interlingual eller statistisk översättning. Det som talar emot interlingual översättning är att den översatta texten är väldigt grammatiskt korrekt, en interlingual översättning har få och väldigt enkla grammatiska regler och kan troligtvis inte generera en sådan bra översättning. Däremot skulle det kunna vara någon form av statistisk översättning, men då blir det istället oklart hur det kan uppstå en översättning som *has also* eller *deal substances.* En förklaring till detta är i sådana fall att det finns så pass mycket exempel på just dessa ordföljder att de skulle vara de statistiskt bästa valen.

En annan idé är att den statistiska översättningen utförs först och efteråt så korrigeras de ordföljder som är för statistiskt svaga genom att helt enkelt göra en ordagrann översättning av dessa. Detta skulle förklara *has also* samt *deal substances* som en direkt översättning av *har också* samt *behandlar ämnen.*

Felaktig meningsbyggnad

Svenska	Engelska
Min mamma är en sjuk sköterska	My mom is a nurse sick
Jag gillar kyckling lever	I like chicken liver
problem att spela musiken i en sådan följd så att de som dansar på dansgolvet dansar så länge som möjligt.	problem playing music in such a sequence so that the dancing on the dance floor dancing as long as possible.

Tabell 1.2 Svenska till Engelska, felaktig meningsbyggnad

Särskrivning är ett vanligt problem och Google Translate behandlar det med varierande framgång. Det första försöket med *en sjuk sköterska* blir inte så lyckat på engelska medan *en kyckling lever* översätts till den korrekta bedömningen. Här skall dock noteras att om man översätter *kyckling lever* ordagrant så får man korrekt engelska medan motsvarande översättning med *sjuk sköterska* ger ett felaktigt svar. Engelskan har generellt särskrivningar på många ställen där vi i svenskan skriver ihop två ord, så en felaktig särskrivning på svenska kan i många fall utan korrektion ge en korrekt översatt mening på engelska.

Nästa mening är oklar redan på svenska med flera frågetecken i formuleringen och översatt till engelska blir den inte mycket bättre. Värt att reflektera över här är att det känns som att det saknas ord runt översättningen *dancing*. En mer korrekt översättning hade varit *the ones dancing* respektive *keep on dancing*.

Vidare ur en algoritmsynpunkt finns klara tecken här på att Google Translate i vissa fall tar hänsyn till särskrivningar och i vissa fall inte. I det andra exemplet om kycklinglevern finns tecken på att en statistisk översättning har misslyckats och en dictionary baserad översättning har tagit över. Det skulle även kunna vara så att Google Translate har tagit hänsyn till särskrivning och tagit fram den korrekta översättningen med hjälp av en statistisk översättning.

Enstaka ord

Svenska	Engelska	
	Substantiv	Verb
Rulla	Roll	Roll
	Muster-roll	Reel
	List	Scroll
	Register	Move
		Wheel
		Bowl
		Trundle
		Toss
		Welter
		Heave
		Labor
		Labour
Direktöversättning		
Rulla	Scroll	
rulla	Roll	

Tabell 1.3 Översättning av enstaka ord

Det framgår här att samma ord får olika översättningar beroende på om första bokstaven är en gemen eller versal. Detta kan verka konstigt men överlag har vi konstaterat att Google Translate är väldigt känslig för radbrytningar, interpunktion samt bruk av gemener och versaler. I det här fallet får ordet *rulla* en annan betydelse med stor begynnelsebokstav förmodligen på grund utav att när ordet används i början av en mening så har det oftast en annan betydelse än om det används mitt i en mening.

Första tanken är att ifrågasätta varför en statistisk modell skulle användas för enstaka ord. En tanke är att det borde vara effektivare att använda en dictionary baserad översättningsalgoritm till dessa. Det som skulle tala emot en dictionary baserad översättning är att Google Translate lär sig hela tiden, det vill säga att någon form av statistik bör finnas med i översättningen. En dictionary baserad översättning skulle troligtvis inte heller ta hänsyn till stora och små bokstäver utan det finns en möjlighet att någon form av sannolikhetstabell används. När *rulla* används i mitten av en mening är det mest sannolikt att det är *roll* man menar och i början av en mening är det förmodligen *scroll* man menar.

Engelska till Svenska

Nedan redovisas meningar översatta från engelska till svenska. De ord som översatts inkorrekt eller som av någon annan anledning är värda att notera kommer att markeras med rött.

Generella meningar

Engelska	Svenska
The itsy-bitsy spider, climbed up the water spout	Imse Vimse spindel, klättrade upp för trån
This guide is for teachers and students. Most of the chapters discuss issues that are probably more interesting to teachers but we have also added a section in each chapter suggesting how students would benefit from reading it. (KTH 2010)	Denna guide är för lärare och elever. De flesta av de kapitel som diskuterar frågor som förmodligen är mer intressant att lärare men vi har också lagt till ett avsnitt i varje kapitel vilket tyder på hur elever skulle ha nytta av att läsa den.

Tabell 2.1 Engelska till Svenska, generella meningar

I den övre texten har vi i likhet med tabell 1.1 en välkänd text som får en korrekt översättning trots att denna ordagrant inte stämmer. Sedan kan det diskuteras huruvida *trån* är en korrekt översättning då det är en förkortning i talspråk för ordet *tråden*.

Nästa steg var att översätta en längre text i likhet med den i tabell 1.1. Detta gav ett liknande resultat som översättningen från svenska till engelska, alltså merparten av den översatta texten är korrekt men ett fåtal fel har smugit sig in. Här skulle man kunna tänka sig att ta bort *de* och *som* från de första rödmarkerade orden i den svenska texten. När det gäller ordföljden *intressant att lärare* vore det mer korrekt med *intressant för lärare* och till sist istället för *vilket tyder på* borde det ha stått *som föreslår* eller motsvarande.

Felaktig meningsbyggnad

Engelska	Svenska
I has a car and me friend have a bike	Jag har en bil och min underbara kompis har en cykel
I have two foots I have two gooses	Jag har två Bottensatser Jag har två gooses

Tabell 2.2 Engelska till Svenska, felaktig meningsbyggnad

I de allra flesta fall spelar det ingen roll för översättningen om man i engelskan skriver fel på *have* och *has*, detta kan bero på att vi i svenskan har samma böjning på verben oberoende av vem det är som gör någonting. Exempelvis *jag har* och *han har* får samma böjning på verbet i svenskan medan *I have* och *he has* får olika böjningar.

En annan intressant del av meningen är att användandet av brittiska slangordet *me* istället för *my* resulterar i att vi får ordet *underbara* inlagt i översättningen, något som inte fanns i originalmeningen.

Ett liknande fenomen inträffar i nästa rad i tabellen där ordet *foots,* en felaktig benämning på fötter i plural, översätts till det helt nya ordet *Bottensatser*. Värt att nämna är att ordet *foots* har direktöversättningen *Bottensatser* i Google Translate så följaktligen var detta det mest logiska alternativet som översättaren hittade. Sedan kan det diskuteras hur korrekt denna direktöversättning är.

Ytterligare ett misslyckat försök att sätta ett ord i plural resulterade i att Google Translate inte ens hittade ett ord för *gooses* och skickade således tillbaka ordet på originalspråket.

Enstaka ord

Engelska	Svenska	
	Substantiv	Verb
Place	Plats	Placera
	Ort	Lägga
	Ställe	Sätta
	Rum	Placera sig
	Loka	Inplacera
	Uppgift	Sätta in
	Hem	Anbringa
	Utrymme	Erinra sig
	Öppen plats	Lokalisera
	Anställning	Skaffa plats
	Ställning	
Direktöversättning		
Place	Placera	
Place	plats	

Tabell 2.3 Översättning enstaka ord

Här, i likhet med tabell 1.3, så varierar översättningen beroende på om gemener eller versaler har använts på begynnelsebokstaven. Samma diskussion om varför detta sker kan föras här, men för att undvika risk för upprepningar hänvisas till stycket tillhörande tabell 1.3.

Översättning på redan översatt material

Nedan redovisas meningar som översatts en gång för att sedan översättas tillbaka till originalspråket. Meningen med testet är att se hur väl Google Translate fungerar på material som tjänsten redan har översatt. De ord som översatts inkorrekt eller som av någon annan anledning är värda att notera kommer att markeras med rött.

Generella meningar

Icke översatt	Översatt en gång	Översatt två gånger
This guide is for teachers and students. Most of the chapters discuss issues that are probably more interesting to teachers but we have also added a section in each chapter suggesting how students would benefit from reading it. (KTH 2010)	Denna guide är för lärare och elever. De flesta av de kapitel som diskuterar frågor som förmodligen är mer intressant att lärare men vi har också lagt till ett avsnitt i varje kapitel vilket tyder på hur elever skulle ha nytta av att läsa den.	This guide is for teachers and students. Most of the chapters discuss issues that are probably more interesting to teachers, but we've also added a section in each chapter, suggesting how students would benefit from reading it.
Denna handbok är avsedd för både lärare och studenter. De flesta kapitel behandlar ämnen som förmodligen är av större intresse för lärarna, men varje kapitel har också ett avsnitt som föreslår hur studenterna kan ha nytta av att läsa texten. (KTH 2010)	This manual is intended for both teachers and students. Most chapters deal substances that are probably of more interest to teachers, but each chapter has also a section that suggests how students can benefit from reading the text.	Denna handbok är avsedd för både lärare och elever. De flesta kapitlen behandlar ämnen som förmodligen är av större intresse för lärare, men varje kapitel har också ett avsnitt som visar hur elever kan dra nytta av att läsa texten.

Tabell 3.1 Generella meningar

Den första översättningen här är engelska till svenska och sedan tillbaka till engelska. Om vi tittar på första och sista kolumnen hos denna ser vi att det enda som skiljer texterna åt är två kommatecken och hopskrivningen av *we have* till *we've*. Notera att de två första kolumnerna här motsvarar de i tabell 2.1 (för första översättningen) och 1.1 (för andra översättningen), och att brister finns i båda mittenkolumner även om dessa inte är utmärkta med röd text. Trots detta är det en i det närmaste felfri översättning tillbaka till engelska!

I den andra översättningen i tabellen ovan skiljer sig texten på ett flertal ställen, men det är inte alltid en felaktig grammatik. Vissa ord har bytts ut mot synonymer och på några ställen har ordets böjning ändrats. Överlag är de båda texterna i kolumn tre mer grammatisk korrekta än texterna i kolumn två, något som är värt att notera.

I de här exemplen kan man fundera huruvida Google Translate använder sig av en mycket välstrukturerad interlingual översättningsalgoritm eller en statistisk sådan. I en interlingual översättning finns ett abstrakt basspråk som texten först översätts till, detta översätts i sin tur till målspråket och man har en färdig text.

I det första exemplet skiljer sig inte ursprungstexten speciellt mycket ifrån måltexten efter andra översättningen så det skulle mycket väl kunna vara en interlingual översättning. Om vi då kollar på det andra exemplet, svenska till engelska till svenska så börjar man se lite fler skillnader.

Om det skulle vara en statistisk översättning så betyder det att den statistiskt bästa lösningen till översättningen från svenska till engelska inte alls är lika som den statistiskt bästa översättningen från engelska till svenska. Detta kan tyckas vara väldigt konstigt.

Felaktig meningsbyggnad

Icke översatt	Översatt en gång	Översatt två gånger
I has a car and me friend have a bike	Jag har en bil och min underbara kompis har en cykel	I have a car and me friend have a bicycle
Min mamma är en sjuk sköterska	My mom is a nurse sick	Min mamma är sjuksköterska sjuk

Tabell 3.2 Felaktig meningsbyggnad

Den första översättningen här är tagen från tabell 2.2 och översätts från en ursprungligen inkorrekt engelsk mening till en korrekt svensk mening och sedan tillbaka till inkorrekt engelska. *I has* från ursprungsmeningen rättas visserligen till och blir *I have* i sista kolumnen, men den felaktiga *me friend have* kvarstår. Även här översätts slangfrasen *me friend* till *min underbara kompis* och vice versa.

I nästa översättning går vi från inkorrekt svenska till inkorrekt engelska och tillbaka till inkorrekt svenska, som är ännu svårare att tyda.

Enstaka ord

Icke översatt	Översatt en gång	Översatt två gånger
Rulla	Scroll	Bläddra
Rulla	roll	rulle
Place	Placera	Place
Place	plats	location

Tabell 3.3 Enstaka ord

Det intressanta i denna tabell är att orden i kolumn ett och tre inte alltid är samma, alltså bara för att A på svenska blir till B på engelska betyder det inte att B på engelska blir till A på svenska. Detta kan vara värt att tänka på, speciellt då översättningarna på redan översatt material i tabell 3.1 blev så pass snarlika originaltexten. Återigen är det skillnad på om vi använder gemener eller versaler på begynnelsebokstäverna, och även här får vi ibland annorlunda svar.

Dessa översättningar tyder på att en interlingual översättning inte används. För i en interlingual översättning av ett enstaka ord skulle resultatet bli detsamma om man översätter tillbaka till källspråket. Om man istället tänker sig den statistiska översättningen så verkar det mer logsikt. Exempelvis så skulle det då vara en högre sannolikhet att när *Scroll* i början av en mening betyder det *Bläddra* istället för *Rulla*.

Resultat

Analys av data

Resultatet från översättningen varierar mycket beroende på vilka slags texter eller meningar som skall översättas. Det är mycket viktigt att radbrytningar, interpunktion och versaler sköts på ett korrekt sätt. Dessa saker kan, enligt våra undersökningar, försämra ett resultat betydligt mer än böjningsfel, särskrivningar eller liknande språkfel. Det är kanske inte helt överraskande att en automatisk översättare är känsligare för *human error* än vad en person som översätter ett dokument manuellt skulle vara, och trots korrekt grammatik så behöver de flesta av vår testdata viss efterbehandling för att den översatta texten skall vara korrekt.

Däremot finns många fördelar med en automatiserad översättning, inte minst i att den inte kräver någon väntetid. En helt automatiserad översättare som Google Translate ger en omedelbar och kostnadsfri översättning dygnet runt och oavsett var i världen du befinner dig, förutsatt att du har en internetuppkoppling och jobbar med några av de språk som tjänsten stödjer.

Översättningen på redan översatt material bjöd på en hel del positiva överraskningar, i vissa fall blev texter som översatts fram och tillbaka nästan identiska med ursprungstexterna trots att mellansteget uppvisade vissa brister. I de tester vi har gjort så har dubbel översättning (alltså översättning tillbaka till originalspråket) ofta gett ett bättre resultat än en enkel översättning mellan språken. Man kan alltså säga att en delvis inkorrekt mellanstegstext resulterar i en bättre översättning än en korrekt text som inte översatts via tjänsten. Denna förbryllande upptäckt är värd att fundera vidare över.

Analys av algoritmen

Utifrån de resultat och tankar kring de översättningar vi gjort så är det troligt att Google Translate använder sig av en något modifierad statistisk översättningsalgoritm. Det kan vara så att när en översättning utförs så görs statistiska beräkningar på hur stor sannolikheten är att ett ord efterföljs av ett annat ord och sedan prioriteras de med högst sannolikhet i svaret.

Efter våra undersökningar och tester har vi även kommit fram till att det troligtvis finns en undre gräns för hur osäker sannolikheten får vara och när detta inträffar översätts ordet eller ordföljden ordagrant för att undvika svar som är för osäkra.

Det som talar emot en interlingual översättningsalgoritm är de fall där sångtexter översätts till något som inte alls har samma språkliga betydelse, men som i låtarna är korrekta. Det finns även fraser som *me friend* som översätts till *min underbara kompis* vilket inte alls är en korrekt språklig översättning.

En transfer baserad översättningsalgoritm är i stort sätt omöjlig att skapa för många språk då regler måste finnas mellan samtliga. Vi kan då anta att en sådan algoritm inte finns i Google Translate.

Det är väldigt svårt om inte omöjligt att hitta detaljer i hur algoritmerna fungerar med hjälp av att översätta texter, troligtvis är algoritmen oerhört komplicerad och består av väldigt många delar och finesser som för användaren är omöjliga att upptäcka.

Diskussion

Felkällor

Då vår rapport baserats på testdata som vi inte har metoder att analysera på ett korrekt och objektivt sätt så har vi behövt göra en hel del antaganden i våra analyser. Detta gäller i synnerhet för algoritmanalysen då denna är i det närmaste omöjlig att sköta på ett vetenskapligt sätt med de resurser vi har haft till vårt förfogande. Det framgår i rapportens syfte att vi ämnar uppvisa en teori kring hur denna algoritm kan fungera och vi vill göra läsaren uppmärksam på att denna teori bygger på antaganden och analyser från vår sida som inte skall ses som slutgiltiga.

Det är även svårt att mäta hur pass väl en översättare fungerar jämfört med en annan då vi inte har en mall för detta. De slutsatser vi har dragit kring huruvida Google Translate fungerar som ett substitut till manuell översättning eller inte bygger på vår subjektiva uppfattning om vad som krävs av en sådan tjänst.

Vi vill poängtera att trots att vi har gjort vårt bästa för att vara så objektiva och noggranna som möjligt så kan vi inte garantera att våra slutsatser är korrekta på grund utav att felkällorna i denna rapport kan vara så pass omfattande.

Slutsatser

Vi frågade oss i syftet om en helt automatiserad översättning såsom Google Translate kan användas som ett substitut för den manuella varianten. Efter tester och analyser av resultatet på dessa kan vi dra slutsatsen att trots att dessa automatiserade översättare blivit bättre och bättre så är resultaten fortfarande för svaga för att helt ersätta en manuell översättning. Däremot kan vi båda se fördelar med att använda dessa tjänster i kombination med det manuella arbetet, exempelvis som ett sätt att snabbt och kostnadseffektivt grovöversätta en text för att sedan korrekturläsa och rätta till de felaktiga översättningarna manuellt.

Ett annat användningsområde för tjänster som Google Translate är i områden där det är viktigare att förstå helheten i texten än att ha en grammatiskt korrekt översättning. Exempel på detta är informell kommunikation över språkbarriärer, översättning av hemsidor på internet som inte själva har lagt upp sidan på andra språk eller som ett lexikon för att slå upp ord eller fraser snabbt och lägesoberoende.

Den andra delen i vårt syfte var att försöka dra slutsatser om hur algoritmen bakom Google Translate fungerar genom att analysera våra testdata. Detta är något som vi redan från början visste skulle bli väldigt svårt då denna algoritm är en väl bevarad företagshemlighet. Vi har dock kunnat dra vissa slutsatser om vissa delar av algoritmen och detta får vi vara nöjda med.

Vi kan nu i slutskedet av vår undersökning konstatera att algoritmen förmodligen använder sig av en kombination av flera olika maskinöversättningar men med en viss betoning på den statistiska delen. Det finns ett flertal exempel i våra översättningar där en statistisk översättningsalgoritm är den enda förklaringen.

Sammanfattningsvis har vi genom arbetet med denna rapport lärt oss en hel del om automatisk översättning och att denna typ av tjänst, trots att den inte är pålitlig nog att fungera som ett substitut för manuell översättning, helt klart borde kunna användas som ett komplement till detta. Självfallet kommer vi inte att kunna återskapa Google Translate's algoritm med hjälp av denna undersökning, men syftet med att dra vissa slutsatser och uppvisa en teori kring hur algoritmen bakom Google Translate fungerar konstaterar vi som uppfyllt.

Referenser

Cyril Goutte (2009), *Learning machine translation*, Cambridge, Massachusetts USA, Neural information processing series.

Sandra Guy (1999). Say what? High-tech language translation: the next generation. *The Rotarian*, volume 175 No 2, s. 12-13.

W John Hutchins, Harold L Somers (1992). *An introduction to machine translation*. London : Academic Press INC. 362 s. ISBN: 0-12-362830-X.

Google (2010). Google's hemsida.
Hämtat från <http://www.google.com> under perioden 2010-01 till 2010-05.

Google Translate (2010). *Google Translate* översättningstjänst.
Hämtat från <http://translate.google.com/> under perioden 2010-01 till 2010-05.

KTH (2010). *Hjälp studenterna att undvika plagiering*.
Hämtat från <http://www.kth.se/vil/learninglab/plagiat> under perioden 2010-01 till 2010-05.