Språk(teknologi) är nyckeln till (artificiell) intelligens och rättvisa

Jörg Tiedemann

Vårt samhälle är fullt av ogjorda översättningsjobb. Utan teknologiska lösningar finns det helt enkelt ingen möjlighet att kunna hantera den gigantiska mängd information som borde göras tillgänglig för alla. I artikeln presenteras initiativ som bygger på öppna och säkra översättningsverktyg som beaktar användarnas integritet. Vårt samhälle måste stödja sådana, för att skapa rättvisa utan exploatering, och för att erbjuda tjänster på lika villkor även till mindre kommersiellt intressanta minoriteter.

Språk spelar en väsentlig roll i vårt mänskliga liv. Jag tvekar inte att påstå att språk är nyckeln till det vi kallar intelligens. Vårt kollektiva kunnande bygger på den omfattande möjligheten att kommunicera och dokumentera allt vi kan. Utan språk skulle vi människor inte ens komma i närheten av de färdigheter som ledde till vår dominans på jorden. Vi använder språk för att beskriva, analysera, debattera och utreda alla världens problem. Vi använder språk för att begripa, lära ut och studera nya och gamla fenomen. Det vi skrivit ner eller berättat vidare sedan generationer tillbaka används för att utveckla vårt samhälle så att vi snabbt kan anpassa oss till nya omständigheter och förändringar.

Språk fyller även en avgörande social roll. Via vårt sätt att kommunicera skapar vi gemenskap, sociala förhållanden och ett nätverk som gör oss starkare än individer var för sig. Men språk kan förstås också missbrukas, det kan manipulera våra tankar och leda till mycket missförstånd. Språk används också för att avgränsa oss från varandra och kan leda till stor irritation. Språk är med oss hela tiden och bestämmer vår vardag hela livet ut.

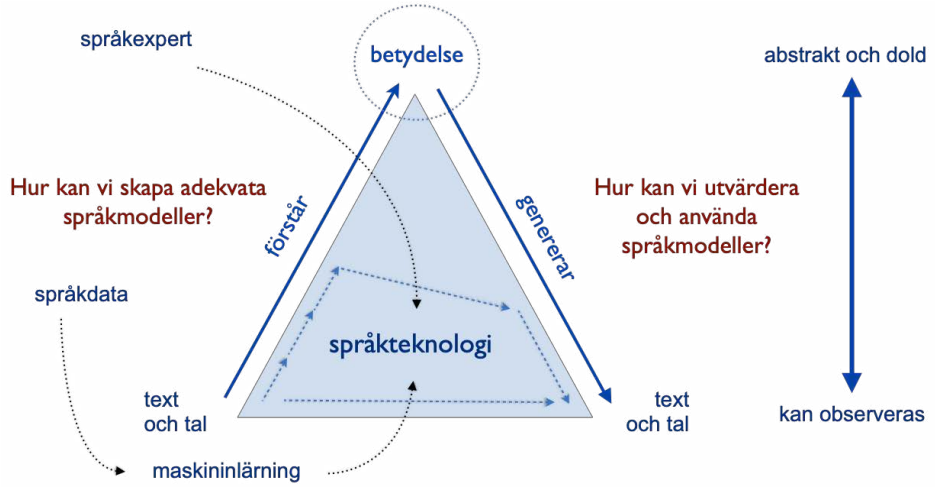
När man betraktar språkens roll är det självklart att vi måste värna om våra språk och hur vi kommunicerar med varandra. Det borde vara självklart att vårt samhälle behöver investera omfattande resurser på språkhantering och effektiv kommunikation. Information måste nå ut till alla för att skapa jämlikhet och för att undvika klyftor, som till stor del bottnar i missförstånd och språkrelaterad diskriminering. Jag vill argumentera för den essentiella roll som öppna språkteknologiska lösningar har i vårt samhälle, där kommunikation sker alltmer globalt och via digitala medier. Och inte bara där, utan lokalt och

inom alla kretsar finns det en utmaning att förmedla information, förhindra misskommunikation och undvika spänningar mellan människor som missförstår varandra. Språkteknologi styr redan nu mycket av vår vardag även om vi inte alltid är så medvetna om det. Det är viktigt att vi kanaliserar denna teknologi till samhällsnytta och rättvisa i stället för exploatering och manipulation.

Jag tar i den här artikeln upp exemplet med maskinöversättning och dess funktion med fokus på informationstillgång och integration och berättar om vårt initiativ att öppna automatiska översättningstjänster så att denna teknologi kan tjäna alla, även små minoriteter som inte finns på radarn för kommersiella produkter. Det finns fortfarande en hel del utmaningar och problem, men de stora fördelarna är tydliga redan nu.

1. Var står vi inom språkteknologin just nu?

Den stora utmaningen inom språkteknologin är att skapa datamodeller som förstår och genererar språk som människor gör. Syftet är att kunna interagera språkligt med en maskin på ett naturligt sätt, men också att förstå hur mänskligt språk fungerar som kommunikationsmedel och hur det relateras till betydelsen och kunskapen vi vill förmedla. Det finns två möjligheter att skapa sådana modeller: (1) En expert lägger till allt vi vet om språk och kommunikation i ett dataprogram så att det kan köras för att hantera text och tal. (2) Vi bygger en maskin som kan lära sig att hantera språk med hjälp av data och exempel och denna maskininlärningsmodell suger in allt den kan hitta i text och tal och kontexten som kan observeras. Båda metoderna används flitigt och kombineras även på många olika sätt.



Figur 1. Målsättningar och metoder inom språkteknologi

En trend som har lett till stora genombrott är maskininlärningsmetoder som bygger på artificiella neurala nätverk (ANN). Allt större nätverk tränas i dag med hjälp av beräkningsmaskiner som blir kraftigare varje år, och den växande datamängden gör det möjligt att inkludera mer och mer av vår kunskap som vi sparar i text och inspelat tal.

Den generella principen är inte särskilt märkvärdig: Ett neuralt nätverk kan betraktas som en universell metod för att approximera vilken matematisk funktion som helst, och det är det vi behöver inom den digitala värld vi jobbar med. En dator behöver en sådan funktion som mappar språkliga signaler till numeriska värden – det är det enda som den kan processa. Till exempel blir all text vi knappar in koderad enligt ett fast schema och bokstäverna vi ser på skärmen avkodar denna funktion för att visa en bild som motsvarar våra förväntningar.

Tyvärr finns det ingen exakt funktion som definierar hur språk mappas till betydelse och där kommer inlärningsmetoden in. Även människor lär sig språk genom att använda det och relaterar dess användning till kontexten. Precis så gör också neurala nätverk och de börjar att lära sig användningen av språkliga tecken genom att jämföra med exempel i olika kontexter. Det krävs mycket data för att förstå hur olika koncept ska användas. Dagens neurala språkmodeller matas med miljontals meningar och dokument för att bygga den komplexa funktion som krävs. Internt har denna funktion miljoner eller miljarder okända parametrar som måste optimeras och det krävs en konkret uppgift som bestämmer i vilken riktning optimeringen ska ske.

Det finns olika uppgifter vi kan definiera. Oftast är de väldigt primitiva, som till exempel lucktextövningar där vår maskin måste gissa vilka ord som saknas inom kontexten. Sådana enkla övningar kan enkelt genereras genom att slumpmässigt ta bort ord ur existerande texter. Sedan körs övningen tills systemet inte längre lär sig någonting. Man förvånas över hur mycket ett system kan lära sig om språk med sådana enkla uppgifter. Det är nämligen det som ligger bakom de flesta stora "AI-lösningar" som omtalas flitigt inom medierna, så som BERT, GPT-3 med flera.

Efter all baskunskap sådana modeller har lärt sig kan vi bygga vidare och träna dem i andra uppgifter med relativt lite träningsmaterial – och plötsligt kan våra artificiella språkmodeller resonera om språkliga frågor, hämta information och generera text som passar in i en kontext. Det är imponerande färdigheter som lätt kan blända oss och få oss att se en allmän intelligens inom systemet som egentligen inte finns där.

Neurala nätverk vet inte vad de gör och varför. Det finns ingen motivation, ingen taktik och inga väl övervägda tankar bakom det som nätet spottar ur

sig. Det reflekterar kunskapen vi lägger in i träningsmaterialet och kan generalisera den på ett effektivt sätt så att det dyker upp mönster och indirekta relationer ur materialet som vi inte förväntade oss att sådana primitiva system skulle kunna hitta.

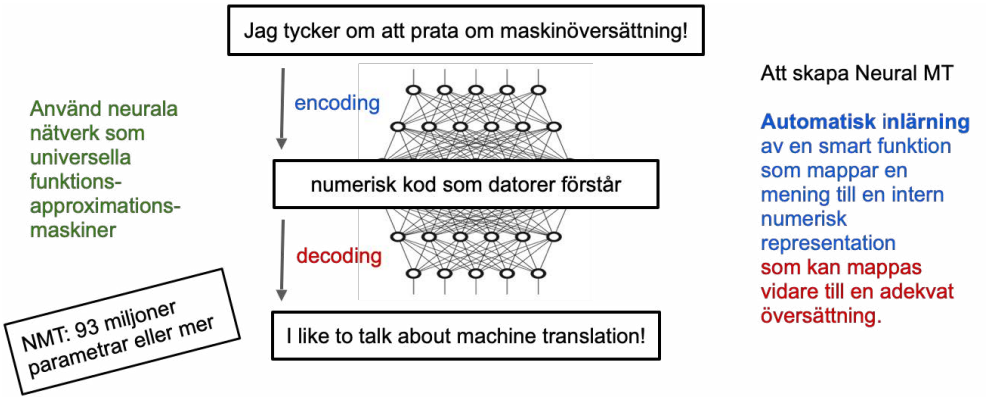
Datadrivna metoder och maskininlärning med neurala nätverk har lett till en revolution som möjliggör tillämpningar vi bara kunde drömma om för några år sedan. Tal och text kan nu hanteras på ett sätt som är flytande och naturligt. Men vi måste inse att vi fortfarande talar om en avancerad signalprocessering och *inte* om intelligens, i alla fall inte på samma sätt som vi betraktar mänsklig intelligens. Om man ser på utvecklingen av språkmodellering kan vi jämföra den med tre olika nivåer av språkförståelse:

1. Jag hör dig
2. Jag förstår ditt språk
3. Jag begriper vad du säger

Nivå två syftar på kunskapen om språkliga uttryck, hur de används och relateras till varandra. Nivå tre däremot kräver en bild av världen och en modell av intention och syfte med kommunikationen. Språkteknologin har nått steg två för resursrika språk, men har fortfarande en hel del kvar för att uppnå steg tre och att täcka många språk som inte ens får stöd på grundnivån. Det är det vi jobbar med.

2. Hur fungerar neural maskinöversättning (NMT)?

Maskinöversättning kombinerar språkförståelse med språkgenering. Den är därför en spännande tillämpning som återspeglar de flesta problem vi måste åtgärda inom språkteknologi. Dessutom översätts det kontinuerligt och vi kan lätt samla in exempeldata som hela tiden produceras av mer eller mindre professionella översättare. På samma sätt som jag beskrev de neurala nätverken behöver vi lära oss en komplex matematisk funktion som mappar språk till en numerisk representation. Den kan sedan användas för att generera text (eller tal), men denna gång på ett annat språk. Modeller som kombinerar en kodningsfunktion och en avkodningsfunktion för språkliga data kallas sekvens-till-sekvens-modeller (förkortas seq2seq) och tränas i princip på samma sätt som andra neurala nätverk – men nu med hjälp av en annan träningsuppgift, nämligen översättning.



Figur 2. Schematisk illustration av neural maskinöversättning

Så i stället för lucktexter övar vår seq2seq-modell att generera översättningar som systemet sedan jämför med det en människa har gjort för att anpassa sina interna parametrar, ända tills det inte längre sker någon förbättring. Återigen kan ett sådant system generalisera och hitta mönster som gör det möjligt att översätta mer än bara allt det som finns i träningsmaterialet. Liksom en människa kan hitta nya sätt att kombinera språkliga element så gör vårt översättningsmodell detsamma. Det hela drivs av statistik, frekvens och kontextuella analogier eller så kallad "distributionell semantik". Den aktuella standardmodellen kallas "transformer" och bygger huvudsakligen på en funktion som kallas för "attention", där systemet lär sig att betrakta olika delar av den språkliga kontexten för att bestämma vad det ska göra som nästa steg (Vaswani et. al, 2017). Miljoner eller miljarder av parametrar styr översättningen och miljontals exempelmeningar med översättning krävs för att få vettiga översättningsmodeller.

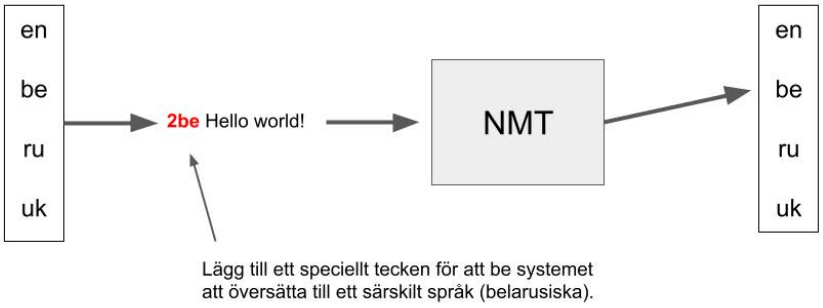
Själva modellen tränas i många iterationer och blir stegvis allt bättre med hänsyn till träningsdatan den använder. Som med alla andra språkmodeller kan vi alltid fortsätta träna när vi får nya data eller behöver förbättra systemet. På det sättet kan ett sådant system anpassas till nya tillämpningsområden. Det är också möjligt att träna på många språk samtidigt. Vi kan skapa en multikulturell polyglott som förstår sig på många språk och kan översätta till många andra språk. Det spännande med denna idé är att flerspråkiga system kan utnyttja likheter mellan olika språk för att överföra kunskap och information från ett språk till ett annat. En efterföljande finjustering kan sedan hjälpa att anpassa modellen till speciella uppgifter och språk. En förutsättning för att det hela fungerar är att vi har tillräckligt med träningsmaterial: Data är därför flaskhalsen.

3. Neural maskinöversättning med lite data

Fungerande översättningsmodeller finns det fortfarande bara för ett fåtal av alla världens språk. Inlärningen är krävande och det behövs mycket träningsmaterial. Vi har ett projekt på gång, OPUS[1] (Tiedemann, 2016), som ska förbättra situationen för många språk genom att systematiskt samla in och förbereda material som vi kan dela med oss till alla som jobbar med maskinöversättning, språkteknologi och allmän språkvetenskap. OPUS växer för varje år, men vi måste ändå hitta andra lösningar för språk där vi aldrig kommer att uppnå de resurser som krävs för de standardlösningar som nämndes tidigare. Mycket forskning pågår och det finns åtminstone två populära strategier:

1. Flerspråkiga modeller och kunskapsöverföring
2. Generering av artificiella data

Att skapa flerspråkiga modeller är ganska enkelt. På källspråkssidan kan man enkelt blanda alla språk och det neurala nätet lär sig självt att känna igen olika språk. För att generera rätt målspråk i översättningen kan man använda ett enkelt trick genom att lägga till ett särskilt tecken som berättar för systemet vilket språk man vill att det ska generera (Johnson et. al, 2017). Det här fungerar som när man först talar om för en mänsklig översättare vilket språk man vill hen ska översätta till, och det fungerar förvånansvärt bra även för neurala översättningssystem. Följande bild illustrerar upplägget:



Figur 3. Flerspråkiga översättningsmodeller

Grundtanken med flerspråkiga modeller är att vi slipper träna många olika system för varje översättningsriktning och att likheterna mellan språk kan hjälpa till att förbättra kvaliteten. Illustrationen visar ett exempel med östslaviska

1 https://opus.nlpl.eu/

språk och engelska. Språk med mindre resurser, som belarusiska (be), kan dra stor nytta av likheterna med ryskan och ukrainskan när vi tränar översättning till och från engelska.

Denna effekt kan åskådliggöras i ett test vi gjorde med olika tvåspråkiga och flerspråkiga system. Följande tabell ger automatiska utvärderingsvärden för översättningen mellan belarusiska och engelska för en testkorpus med 2 500 meningar (Tiedemann, 2020). Vi använder BLEU (Papineni, 2002) som mått för att jämföra automatiska översättningar med mänskliga referensöversättningar och högre siffror betyder bättre kvalitet. Tabellen visar resultat från modeller med ett stigande antal källspråk som inkluderas i träningsprocessen.

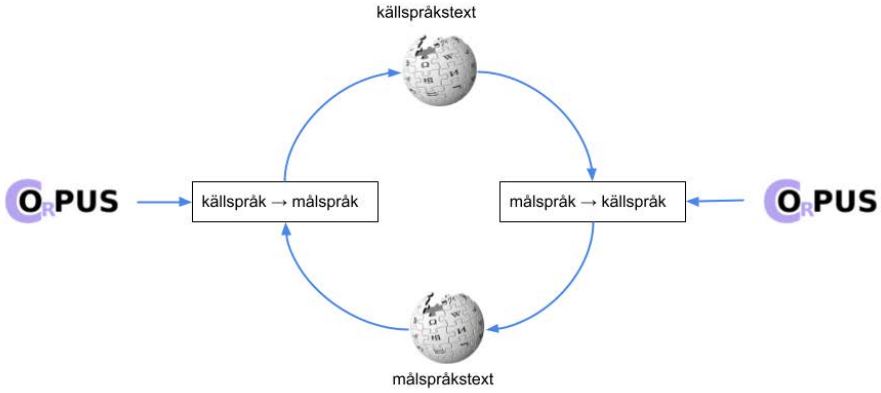
Tabell 1. Automatisk utvärdering av flerspråkiga översättningsmodeller (BLEU värden i %)

modell	belarusiska → engelska	engelska → belarusiska
belarusiska – engelska	10.0	8.2
östslaviska språk – engelska	38.7	20.8
slaviska språk – engelska	**42.7**	**22.9**
indoeuropeiska språk – engelska	41.7	18.1

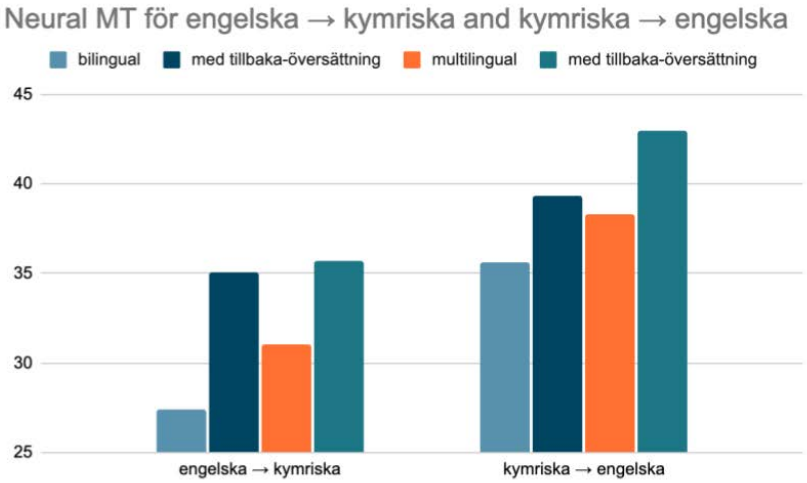
Vi kan se att flerspråkiga modeller leder till en markant förbättring jämfört med den tvåspråkiga grundmodellen (första raden). Dessutom ser vi att det finns gränser för vad sådana modeller klarar av. Att inkludera ett stort antal indoeuropeiska språk blir för mycket för dem, på så sätt att kapaciteten inom modellen inte längre räcker till för att översätta våra testdata lika bra (sista raden).

Ett annat sätt att tackla resursbristen är att skapa artificiella träningsdata. Det låter kanske suspekt, men grundidén är att utnyttja inkompletta resurser för att förbättra vissa aspekter av ett översättningssystem. Det är till exempel möjligt att göra ett system mer flytande i målspråket genom att inkludera enspråkiga målspråkstexter. Dessa kan översättas automatiskt med hjälp av ett översättningssystem i den motsatta riktningen. Följande bild illustrerar denna "tillbaka-översättning" och förklarar hur idén kan leda till en iterativ metod som stegvis kan förbättra våra system.

Frågan är om det verkligen fungerar: Vi har testat denna metod med olika lågresursspråk för att se vilka resultat man kan uppnå. Följande diagram illustrerar översättningskvaliteten (i mån av BLEU värden på ett oberoende testset) när vi lägger till så kallade "back-translations" eller automatiska översättningar av enspråkiga målspråkstexter. Denna gång tar vi engelska-kymriska som exempel.



Figur 4. Att skapa träningsmaterial med iterativ "tillbaka-översättning"

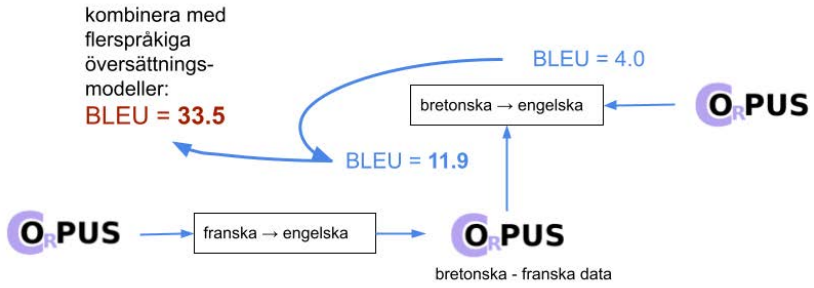


Figur 5. Effekt av automatisk datagenerering och kunskapsöverföring med relaterade språk

Den andra stapeln visar den substantiella ökning vi kan uppnå med sådana automatiskt genererade träningsdata. I synnerhet för översättningen till kymriska kan vi se en rejäl förbättring som visar vikten av språkkunskaper i målspråket särskilt för morfologiskt rika språk. Dessutom kan man lätt kombinera denna datagenereringsmetod med flerspråkiga modeller. I stapeldiagrammet kan vi se effekten av en kombination av båda metoderna med resultat för kymriska till och från engelska när vi använder modeller som är tränade på keltiska språk i stället för enbart kymriska.

Till sist vill jag också presentera ännu en metod för att utöka träningsdata. Vi tar exemplet bretonska till engelska, där en basmodell ger väldigt dåliga

resultat. Om vi mäter i BLEU igen får vi ett värde på bara 4,0 på vårt testset. Det betyder att översättningen i princip är obrukbar. Även bretonska – franska träningsdata ingår emellertid i vår databas och vi har en hyfsat bra maskinöversättning mellan franska och engelska. Det betyder att vi kan översätta franska till engelska och skapa nya automatiskt genererade exempel för vårt bretonska – engelska maskinöversättningssystem. När vi agerar så höjs kvaliteten med över 7 BLEU poäng:



Figur 6. Datagenerering med hjälp av resursrika mellanspråk

Dessutom kan vi kombinera allt detta med flerspråkiga modeller och tillämpa tillbaka-översättningsmetoden, och når till slut till ett värde på över 33 BLEU-poäng, vilket betyder att vår modell helt klart har blivit användbar.

Dessa exempel ger ett liten överblick av olika enkla metoder att hantera en situation med begränsade datamängder. Vår forskning går ut på att vidareutveckla dessa idéer och optimera inlärningsprocessen för många språk, så att vi kan nå en bättre täckning även för minoritetsspråk.

4. Varför ska vi satsa på öppen maskinöversättning?

Automatisk översättning väcker en hel del diskussion. Här är några viktiga punkter att ta hänsyn till:

"Maskinöversättning tar människors jobb"

Ett vanligt argument mot maskinöversättning är att det förstör jobbet för mänskliga översättare och att den samtidigt leder till sämre service för små språk. Det är sant att maskinöversättning eftersträvar automatisering och övertar uppgifter som annars måste utföras av människor. Men hotet mot översättare kommer trots allt från ett annat håll. Vårt samhälle är fullt av ogjorda översättningsjobb. Det finns helt enkelt ingen möjlighet att hantera den gigantiska mängden av information som borde bli tillgänglig för alla på lika villkor. Teknologiska lös-

ningar är tyvärr det enda sättet att förbättra situationen. Det kommer ändå inte finnas någon brist på uppgifter för professionella översättare där tekniken inte räcker till. Där det finns nedskärningar handlar de om politiska beslut och inte om brist på jobb.

Som en jämförelse vill jag ta upp förändringarna inom transportindustrin när bilarna kom med i bilden. Säkert var många inom den branschen rädda när de första bilarna dök upp och det kändes som om transportpersonal inte längre behövdes. Tvärtom har industrin inom logistiken bara exploderat, och jag ser likheter med översättningsbranschen. Tillsammans med teknologin kan vi nu tackla översättningsjobb som vi inte kunde drömma om tidigare för att det inte fanns resurser att ens börja med dessa. Översättningsverktyg på flera språk kommer bara att öka antalet uppdrag. Jobbet kommer såklart se annorlunda ut, på samma sätt som transportindustrin ställde om sina behov. Det här händer redan och det kommer att fortsätta i språkservicesektorn.

Den andra delen av argumentet handlar om kvaliteten som försämras när automatiska verktyg introduceras istället för enbart mänskliga översättare. Även det tycker jag beror på politiska beslut och inte på sådant teknologin förorsakar. Teknologin kan ge stöd till översättningar av hög kvalitet, men man måste naturligtvis satsa på kontroll och bearbetning. Man borde också begrunda om det inte är ännu sämre service att så mycket inte alls blir översatt när resurserna för mänskliga översättare inte räcker till. Automatöversättning kan i så fall fylla en viktig funktion genom att snabbt ge tillgång till information som annars är bortom räckhåll för en viss publik.

"Maskinöversättning kostar ingenting"

Ett argument mot finansiering av offentlig utveckling av maskinöversättning är att detta görs bättre, billigare och snabbare av kommersiella aktörer och att staten och offentliga forskningsinstitut aldrig kommer att kunna konkurrera med stora globala företag och inte heller borde göra det.

De stora it-giganterna har enorma marketingstrategier där de lanserar tjänster som erbjuds smidigt på nätet utan betalning. Inte bara privatpersoner utan också myndigheter bländas av detta erbjudande och inser oftast inte hur stor investeringen är för dessa företag. Massiva resurser satsas på språkteknologi, och maskinöversättning är en av de stora tillämpningar som utvecklas. Det är därför naturligt att myndigheterna inte förstår att man behöver stora resurser för att utveckla liknande tjänster när allt ju finns "gratis" på nätet.

Det finns en bra anledning till varför internetföretag erbjuder tjänster som

verkar vara gratis, men som utelämnar användarna av dem till exploatering av företagen. Anledningen är att de vill hålla kvar sina kunder på plattformen och knyta allt till tjänster där de kan kontrollera innehållet och bombardera användarna med personlig reklam som maximerar vinsten. Dessutom kan de behändigt samla in personliga data och finslipa profilen och sina egna system. Att utelämna samhället och dess informationsbehov till vinstdrivna annonseringsföretag är väldigt riskabelt, och jag tror inte att våra beslutsfattande organ har insett vilken fara detta innebär.

Det enda sättet att undvika riskerna det här medför är enligt mig att vi så snabbt som möjligt utvecklar transparenta och säkra system som är oberoende av vinstintressen för att betjäna alla medborgare med digitala tjänster som erbjuder de funktioner som är nödvändiga för rättvis, jämlik och likvärdig behandling. Det i sin tur kräver ordentliga investeringar och inte bara skrattretande småpengar som inte bygger upp en kraftfull motståndare till de globala företag som odlar sina monopol.

För ett tag sedan utlyste statsrådets kansli i Helsingfors en anskaffning av en pålitlig automatisk översättningstjänst för fyra år framöver som skulle vara anpassad för de interna behoven och kunna hantera alla data på ett säkert sätt. Kostnadstaket fastställdes till 60 000 euro. För mig låter det som ett hån mot dem som utvecklar översättningstjänster, och det visar hur pass lite en myndighet är villig att investera i en otroligt viktig och känslig tjänst.

Vi måste äntligen erkänna att språkteknologiska lösningar är oerhört viktiga i vårt digitala informationssamhälle och att vi inte kan lägga ut känsliga områden på entreprenad till externa företag. Ordentliga resurser borde satsas på en transparent utveckling av säkra tjänster, inte minst för automatisk översättning. Det borde vara självklart och ett krav för en rättvis behandling utan språklig diskriminering.

Ett argument jag också ganska ofta får höra är att ingen ändå kan konkurrera med de stora it-jättarna och att det är bortkastade pengar att investera i egna produkter. Det är synd att höra denna uppgivenhet. Inom vår forskningsgrupp har vi satsat en hel del på att visa vad vi kan uppnå även utan ordentligt statligt stöd. Öppna lösningar kan fungera. Som avslutning vill jag därför presentera vårt projekt OPUS-MT, maskinöversättning som är tillgänglig för alla och som enkelt kan integreras i professionella arbetsflöden.

5. Öppen maskinöversättning med OPUS-MT

Under 15 år har vi samlat in översatta texter för att skapa världens största databas av öppna parallella data med länkade översättningstexter som i nuläget omfattar hundratals språk med totalt över tio miljarder meningar. Denna gigan-

tiska textkorpus finns tillgänglig via OPUS[2] och används flitigt inom språkteknologi och översättningsvetenskap.

För ungefär två år sedan satte vi igång med ett projekt som bygger på våra data och fokuserar på utvecklingen av öppna översättningsmodeller och verktyg som kan användas som en allmän resurs inom forskning och utveckling. Det började med ett projekt som fokuserade på finska och svenska inom ramen för ett pilotprojekt som generöst finansierades av Svenska kulturfonden i Finland under namnet Fiskmö[3] (Tiedemann et. al, 2020). Den väldigt framgångsrika utvecklingen motiverade oss att utvidga omfånget och satsa på många språk och en bred täckning av olika översättningstjänster. Det blev början till OPUS-MT[4] (Tiedemann & Thottingal, 2020) som redan nu har skapat över 1 000 öppet tillgängliga översättningsmodeller samt verktyg som gör det enkelt att integrera dem i praktiska tillämpningar och professionella arbetsflöden. Efter Fiskmö fick vi finansiering från den europeiska språkteknologiinfrastrukturen ELG som ett av deras pilotprojekt.[5] Det gjorde det möjligt att finslipa våra verktyg och förbättra integrationen av OPUS-MT inom populära översättningsplattformer. OPUS-CAT[6] (Nieminen, 2021) har blivit ett stabilt paket som ger professionella översättare tillgång till alla våra system enkelt via mjukvarupaket som kan installeras på vanliga personaldatorer.

Det som är speciellt med OPUS-CAT är att översättningen sker lokalt på användarens dator och att man därmed undviker alla säkerhetsproblem som uppstår med andra tjänster. Inga data lämnar den lokala omgivningen för automatisk översättning. Användaren behöver inte heller vara uppkopplad till en onlinetjänst för att använda våra verktyg. Det är en transparent lösning som utgör en motsats till de vinstdrivna program som försöker exploatera användarens information eller binda kunderna till onlinetjänster där villkor och kostnader snabbt kan förändras. Vi är övertygade om att OPUS-MT har kommit en bra bit på vägen att frigöra användarna från typiska marknadsstrategier där datasäkerhet och integritet är tvivelaktiga och oroväckande. Vi ser OPUS-MT som ett bra exempel på att det går att utveckla översättningsteknologi för allmännyttan och att nationella initiativ kan lyckas.

2 https://opus.nlpl.eu/

3 https://blogs.helsinki.fi/fiskmo-project/

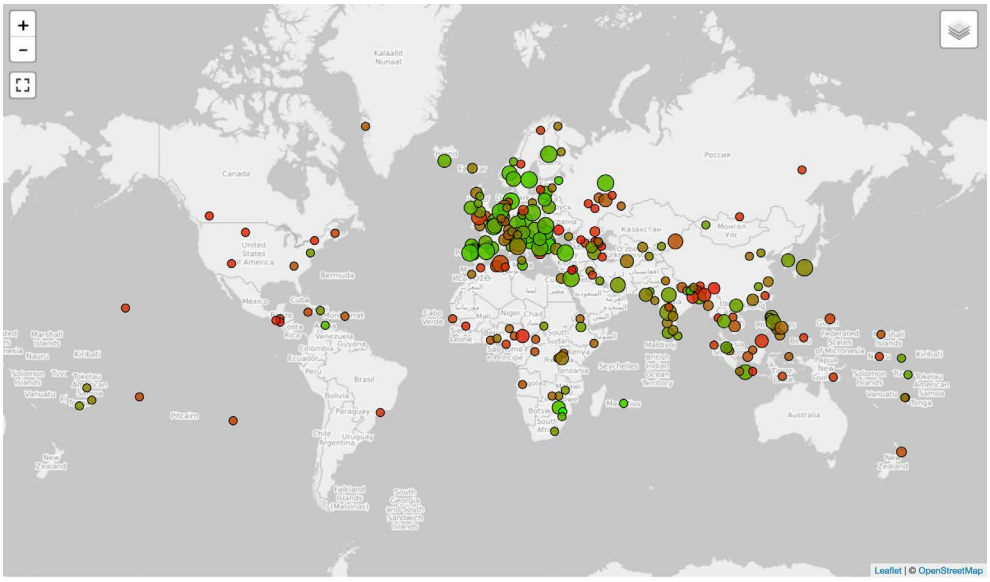
4 https://github.com/Helsinki-NLP/OPUS-MT

5 https://live.european-language-grid.eu/catalogue/project/2866

6 https://helsinki-nlp.github.io/OPUS-CAT/

OPUS-MT är nu tillgängligt via öppna programbibliotek[7] och gratis nedladdningstjänster och all källkod är öppen och transparent. Det har blivit en del av den gemensamma ansträngningen att demokratisera språkteknologiska lösningar så att vi minskar vårt beroende av stora it-företag som monopoliserar digitala tjänster. Vi hoppas på en större satsning på språkteknologisk forskning och akademisk utveckling, så att vi kan fortsätta på den här vägen som är nödvändig för en rättvis hantering av den internationella och mångspråkiga digitala världen.

Motiveringen liknar den bakom världens största encyklopedi Wikipedia: Gemensamt kan vi skapa kollektiv nytta och tackla den stora uppgiften att göra kunskap och information tillgänglig för alla. Det är fortfarande mycket som fattas innan vi är där och vi vet att bara en bråkdel av världens språk kan dra nytta av de stora framsteg som görs inom samhället. Med OPUS-MT satsar vi på mångspråkigt stöd, även om vi fortfarande är mycket begränsade.



Figur 7. Täckning av befintliga översättningsmodeller från OPUS-MT för översättning till engelska

Kartan ovan är en geografisk illustration av täckningen med översättningsmodeller från OPUS-MT för att översätta till engelska. Större cirklar betyder bättre

7 Se t.ex. integrationen i transformer programbiblioteket av huggingface på https://huggingface.co/Helsinki-NLP

resurser. Den gröna färgen symboliserar bättre översättningskvalitet medan rött betyder en otillräcklig kvalitet.

Vi samarbetar förövrigt gärna kring förbättrade översättningstjänster för resursfattiga språk (se t.ex. Vázquez et. al, 2021; Aulamo et. al, 2021). Vi tar också emot donationer av text eller hjälp med inlärning av våra system. Ta gärna kontakt med oss på Helsingfors universitet.

6. Slutsatser

Språkteknologi har kommit en lång väg och praktiska tillämpningar som bygger på språkteknologi finns med när vi surfar på nätet, kollar våra informationskanaler eller använder digitala tjänster. Automatisk översättning är ett måste för att kunna hantera informationstillgång över språkliga gränser och utvecklingen visar vilken nytta vi kan få genom en integration av maskinöversättning i många olika sammanhang. Likvärdig tillgång till information är en förutsättning i en rättvis värld och det krävs stora satsningar för att ge alla världens språk tillräckligt med stöd.

Att överlåta språkhantering till stora vinstdrivna it-företag är mycket riskabelt och vi borde inse att det behövs en strategi för att bygga ett digitalt informationssamhälle där vi inte överlämnar våra privata data villkorslöst till en hungrig annonssmarknad. Därför behöver vi öppna och transparenta lösningar.

Summary

Our society is full of unfinished translation jobs. Without technological solutions, there is simply no way to handle the gigantic amount of information that need to be made available to everyone. The article presents initiatives based on open and secure translation tools that take into account the privacy of users. Our society must support such activities, in order to enable equality without exploitation and we need to develop transparent services for barrier-free information access especially for less commercially interesting minorities and users with certain disabilities.

Författaren

Jörg Tiedemann är professor i språkteknologi vid institutionen för digital humaniora vid Helsingfors universitet. Han doktorerade inom datorlingvistik på Uppsala universitet med forskning om automatisk länkning av översatta texter och maskinöversättning innan han flyttade till univeristetet i Groningen i Nederländerna för ett postdokproject om informationsextraktion och frågebesvarande system. Hans forskningsintressen berör huvudsakligen datadrivna metoder inom språkteknologin och han leder för närvarande ett EU-finanserat projekt om semantisk representationsinlärning och naturlig språkförståelse.

Website: https://blogs.helsinki.fi/tiedeman/

Referenser

Aulamo, M., Virpioja, S., Scherrer, Y. & Tiedemann, J., May 2021: Boosting Neural Machine Translation from Finnish to Northern Sámi with Rule-Based Backtranslation, Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa). Dobnik, S. & Øvrelid, L. (eds.). Linköping: Linköping University Electronic Press, p. 351-356 (Linköping Electronic Conference Proceedings ; no. 78)(NEALT Proceedings Series ; no. 45).

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J., 2017: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. In Transactions of the Association for Computational Linguistics (TACL), Volume 5, p. 339–351

Nieminen, T.: OPUS-CAT: Desktop NMT with CAT integration and local fine-tuning. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021: System Demonstrations, p. 288–294

Papineni, K., Roukos, S., Ward, T., Wei-Jing Zhu, 2002: Bleu: a Method for Automatic Evaluation of Machine Translation, In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, p. 311–318

Tiedemann, J. & Thottingal, S., 1 Nov 2020: OPUS-MT -- Building open translation services for the World,, Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. Martins [et al.], A. (ed.). Geneva: European Association for Machine Translation, p. 479-480

Tiedemann, J., Nieminen, T., Aulamo, M., Kanerva, J., Leino, A., Ginter, F. & Papula, N., 1 May 2020: The FISKMÖ Project: Resources and Tools for Finnish-Swedish Machine Translation and Cross-Linguistic Research, Proceedings of the 12th Language Resources and Evaluation Conference. Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. & Piperidis, S. (eds.). Paris: European Language Resources Association (ELRA), p. 3808-3815

Tiedemann, J., 1 Nov 2020: The Tatoeba Translation Challenge - Realistic Data Sets for Low Resource and Multilingual MT, Proceedings of the Fifth Conference on Machine Translation. Barrault [et al.], L. (ed.). Stroudsburg: The Association for Computational Linguistics, p. 1174-1182

Tiedemann, J., 2016: OPUS -- Parallel Corpora for Everyone, In: Baltic Journal of Modern Computing. p. 384

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., 2017: Attention is all you need. CoRR abs/1706.03762. http://arxiv.org/abs/1706.03762.

Vázquez, R., Scherrer, Y., Virpioja, S. & Tiedemann, J., 1 Jun 2021: The Helsinki submission to the AmericasNLP shared task, Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas. Mager [et al.], M. (ed.). Stroudsburg: The Association for Computational Linguistics, p. 255-264

Nyckelord

maskinöversättning, språkteknologi, informationstillgänglighet