Maskinoversættelse – en sammenligning af to forskellige metoder

Af Eckhard Bick 7. december 2009 • I kategorien Teknologi •

Georgetown-eskperimentet 1954. Et af de første forsøg med maskinoversættelse mellem russisk og engelsk. Maskinen "kunne" 250 ord og 6 grammatiske regler.

En gammel drøm, automatisk oversættelse, det at kunne oversætte sine egne og andres tekster med et enkelt museklik, er efterhånden ved at gå i opfyldelse. Hvem kunne ikke tænke sig at få lavet sine lektier af en automat, at læse en engelsk eller svensk brugsanvisning på dansk og at slippe for besværlige ordbogsopslag ét-ord-ad-gangen.

Som projekt har maskinoversættelse (MT) faktisk en del år på bagen, oprindeligt motiveret af efterretningsmæssige interesser under den kolde krig, ikke mindst med det formål at sætte en engelsktalende person i stand til hurtigt at kunne screene store tekstmængder på sprog som russisk, japansk eller koreansk, gerne i forbindelse med andre sprogteknologiske redskaber som *information extraction (IE)* og *text summarisation*.

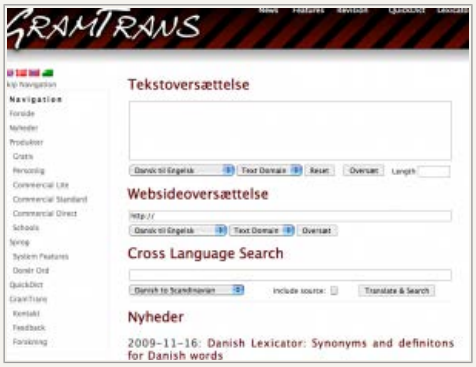
MT-forskningen har været igennem flere faser af entusiastisk offentlig støtte skiftende med en klippefast overbevisning om at projektet aldrig vil kunne realiseres, og udviklingen er igennem tiden blevet præget af en række teknologiske paradigmer, herunder ord-for-ord tilgangen, generativ syntaks, neuronale netværk, kunstig intelligens, hukommelsesbaseret oversættelse, probabilistisk korpusbaseret oversættelse o.a.

Ud fra et brugersynspunkt kan man i dag skelne mellem *machine translation (MT)* og *computer assisted translation (CAT)*. Mens førstnævnte har som mål at levere en egentlig oversættelse, der så skal efterredigeres af en menneskelig oversætter, er sigtet med sidstnævnte (CAT) at understøtte en menneskelig oversættelsesproces igennem automatiske ordbogsopslag, lagring og gentagelse af allerede oversatte tekstbidder m.m. Nærværende artikel fokuserer på MT, og drøfter to rivaliserende metoder, henholdsvis statistisk og regelbaseret maskinoversættelse.

GramTrans og Google Translate

Hvad er status for MT i dag? Er projektet lykkedes? På den ene side findes der efterhånden kommercielle oversættelsessystemer til de store sprogpar som engelsk-tysk og engelsk-japansk, på den anden side har de fleste set smagsprøver på mere eller mindre sjove fiaskoer fra Babelfish og lignende gratissystemer. På den anden side har man som dansker altid skullet leve med det kedelige faktum at så lille et sprogområde er der ingen af de store aktører der vil udvikle sprogteknologi for – hverken godt eller dårligt, hverken dyrt eller gratis – og offentligt støttede projekter har kun i begrænset omfang formået at kompensere for denne mangel på kommerciel interesse.

Denne situation blev grundlæggende ændret da et lille dansk firma, GrammarSoft, i samarbejde med et norsk firma, Kaldera, i 2007 lancerede et internetbaseret system, GramTrans, der håndterer sprogparret dansk-engelsk for fri tekst, uden snævre domænebegrænsninger, og på et niveau der kan konkurrere med de store sprogpar. Servicen dækker også de andre skandinaviske sprog, norsk og svensk, og er gratis ved moderat brug. Hvis man vil oversætte meget, kan man få halvårlige privat- eller firmaabonnementer. GramTrans er ikke god til digte og rim, men kan oversætte stort set alt andet – emails, lektier, reklame, Wikipedia etc. Man kan endda oversætte færdige Word-dokumenter eller bruge sin mobiltelefon til at oversætte med.



Regelbaseret maskinoversættelse. GramTrans oversætter mellem dansk og engelsk, foruden norsk og svensk

Det andet store gennembrud kom i 2008, da Google tilføjede dansk som oversættelsessprog. Google-oversættelser er ekstremt hurtige, velintegreret i andre services, og dækker en stor sprogvifte. Sidstnævnte opnås dog ved at oversætte til

Seneste sprognyheder

19/3	Evolution i sprog: Dovenskab har Ã¦ndret ord | videnskab.dk
30/10	Hvem taler: Kan du hÃ¸re forskel pÃ¥ menneske og maskine? | www.dr.dk
6/3	DansklÃ¦rer: Tosprogede bÃ¸rn bliver sprogligt forsÃ¸mt | www.bt.dk
28/2	Flere og flere ordblinde Ã¦lærter pÃ¥ universitetet - Magisterbladet | magisterbladet.dk
22/2	Keeper eller mÃ¥lmand â€" hvilket ord er bedst? Dansk SprognÃ¦vn | dsn.dk

28/3	Indigenous languages come from just one common ancestor, researchers say - ABC News (Australian Broadcasting Corporation) | www.abc.net.au
27/3	UVic news - University of Victoria | www.uvic.ca
26/3	New York Nonprofit Media: Native American nonprofit aims to boost community health by teaching Lakota language | nynmedia.com
22/3	Teenage business owner aims to revitalize Blackfoot language | globalnews.ca
19/3	Language revival preserving history | www.smh.com.au

FLERE NYHEDER >>>

Verdens sprog på Sprogmuseet.dk på et større kort

Nye kommentarer

Artikler om

Arkiv

Statistisk maskinoversættelse. Google Translate oversætter mellem 50 sprog med engelsk som brosprog

og fra dansk igennem engelsk som brosprog, en teknik der især for ellers nært beslægtede sprog som dansk-svensk giver unødvendige tab. Med undtagelse af GramTrans henter stort set alle andre oversættelsestjenester på nettet deres danske oversættelser fra Google, bare med en ny indpakning.

Statistisk og regelbaseret maskinoversættelse

Indholdsmæssigt er forskellen mellem Google-MT og GramTrans at de som henholdsvis statistiske og regelbaserede systemer står som eksponenter for de to store konkurrerende paradigmer i maskinoversættelsen i dag. Begge metoder forsøger, med vidt forskellige midler, at løse det problem at de fleste ord skal oversættes forskelligt alt efter konteksten, og at kildesprogsord og målsprogsord sjældent svarer til hinanden én-til-én:

Dansk	Engelsk
søjle	column
spalte	to split
	crack
narko	
	narcotics

I statistisk maskinoversættelse (STMT) benytter man såkaldte parallelle korpora – en stor samling tekster der i forvejen er (menneske-)oversat, og maskinen skal så lære at et dansk og et engelsk ord sandsynligvis er oversættelser af hinanden, hvis de oftere forekommer i "parallelle" sætninger (dvs. sætninger der er oversættelser af hinanden) end ellers. Desuden tages der hensyn til naboordene, maskinen søger at vælge, ud af flere mulige, de ord-til-ord oversættelser, der samlet set giver de mest normale (dvs. statistisk verificerbare) kombinationer på målsproget.

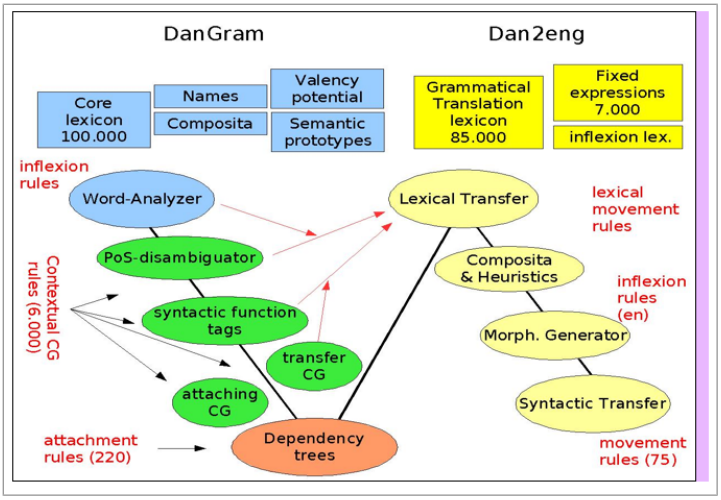
Metoden er tilsvarende god til at vælge mundrette synonymer i snævre kontekster, men kan på den anden side finde på at bytte indholdsmæssigt helt forskellige ord ud med hinanden bare fordi de ofter forekommer i samme kontekst. For eksempel kan den bytte rundt på nationalitetsbetegnelser eller antonymer. Et andet problem ved metoden er at den har svært ved bøjninger og sammensætninger, og svigter for sprog med en stor formrigdom, simpelthen fordi det er umuligt at samle et træningskorpus der er stort nok til at dække alle variationsmulighederne.

I regelbaseret maskinoversættelse (RMT) baseres oversættelsen på en fuld morfologisk analyse af ordene og en syntaktisk analyse af sætningen. Dette gør det så muligt at formulere såkaldte leksikalske transferregler, der vælger en bestemt oversættelse pga ordets klassificering og opbygning, det styrende verbum eller om sætningen har et humant subjekt, for at nævne nogle eksempler.

Et eksempel er det danske verbum *at regne*, der ud over sit meteorologiske oversættelsesækvivalent *(to rain)* også har en række ikke-meteorologiske læsninger *(calculate, consider, expect, convert …)*. Systemet skelner mellem disse oversættelser ved at kræve et situativt subjekt ('det') for *rain*-oversættelsen, mens defaulten for humane subjekter er *calculate*. Tilstedeværelsen af partiklen 'med' fører til oversættelsen *expect*, hvis denne er markeret som præposition, men *include* ved en læsning som inkorporeret adverbialpartikel. Hvis der foreligger både objekt og objektsprædikativ ('regne nogen for noget'), vælges oversættelsen *consider*, osv. Tilsvarende kan den regelbaserede metode generere korrekte bøjningsformer på målsproget udfra de grammatiske kategorier, og ordne ord og ordgrupper syntaktisk korrekt på målsproget (transformationsregler).

Den statistiske og den regelbaserede metode har hver deres ulempe og fordele. Den vigtigste forskel er måske deres *learning curve:* Et STMT-system kan, forudsat man råder over store mængder paralleldata, hurtigt og lønomkostningsfrit trænes for et nyt sprogpar, mens det kræver mange arbejdstimer og megen lingvistisk ekspertise at skabe de nødvendige ordbogs- og grammatik-ressourcer for et nyt RMT-system. På den anden side flader kurven hurtig ud for STMT, de hyppige mønstre dækker over de mere sjældne, og fordi metoden ikke "forstår" sætningen, er den ikke så god til at opløse flertydigheder eller bruge andet end nærkonteksten. RMT, på den anden side, har ikke noget endeligt kvalitetsloft, ethvert problem kan i princippet løses hvis bare man formulerer de nødvendige regler.

Nedenstående ses arkitekturen i GramTrans' RMT-system:



Fremtiden

Fremtiden vil formodentlig byde på en integration af STMT and RMT, fx kan man korrigere bøjningsfejl i STMT vha. en almindelig grammar checker, og et RMT-system kan træffe statistiske valg på områder hvor den mangler regler.

For normalbrugeren vil det under alle omstændigheder kun være slutresultatet der tæller, og som sagen om jægersoldatbogen viser, nærmer vi os hurtigt en situation, hvor det anses for ganske normalt at alt der skrives på ét sprog også kan læses på et andet. Det bliver så interessant at se om eller hvornår maskinoversatte tekster får jurdisk gyldighed på lige fod med translatøroversættelser, eller om der altid vil være brug for menneskelig efterredigering. Talesproget venter også i kulissen, men telefon-MT og talende rejsemoduler bremses foreløbigt af flaskehalsen talegenkendelse – dansk synes at være et af de sprog på kloden der er vanskeligst at forstå, rent fonetisk altså ...

***Eckhard Bick**, dr. phil., Forskningslektor*
ISK/VISL, Syddansk Universitet

Læs også:

1. Kollokationer Hvad er kollokationer? Ordet ser fremmed ud, og et oplagt sted at slå op er da også en fremmedordbog. Slår vi op i Gyldendals Fremmedordbog, kan vi læse at ordet...
2. Hjælpemidler i sprogundervisningen Elle mai s'arranger, quelques phrases sortent amende GoogleTranslate oversætter til fransk, men hvilket fransk? "Elle mai-être vrai, mais il peut aussi se tromper" er oversættelsen dansk-fransk i Google Translate af...
3. Digitale ordbøger Bomærke Kan vi ikke få skruet ned for dommedagsretorikken? Det eneste konstante er forandringen. Og alligevel er der en evindelig ballade hver gang vi tager afsked med noget kendt og...
4. Statistik og journalistik Udsnit af Randi Isagers artikel på dr.dk, 4. jan Den 4. januar 2010 vågnede jeg op om morgenen og var sprogforsker. Det kunne jeg læse i aviserne, at jeg var,...

Tagget med: brugsanvisninger, bøjninger, bøjningsfejl, computer assisted translation, engelsk, flertydighed, generativ syntaks, Google Translate, GramTrans, hukommelse, information extraction, japansk, jurdisk gyldighed, kildesprog, kolde krig, koreansk, korpus, kunstig intelligens, målsprog, maskinoversættelse, morfologi, norsk, objekt, objektsprædikativ, ordbøger, oversættelse, paradigme, paradigmer, præposition, regelbaseret maskinoversættelse, regler, russisk, sammensætninger, skandinavisk, sprogteknologi, statistisk maskinoversættelse, subjekt, svensk, Syddansk Universitet, synonymer, syntaks, talegenkendelse, tekst, telefon, transformationsregler, translatør, verbum

4 kommentarer

Henrik Helbæk
12. december 2009 • 18:59

Utrolig interessant indlæg. Hvor lang tid vil der gå før der maskinoversættelse også tager hensyn til foniske forskelle?

Svar

Per Hagemann
27. februar 2010 • 06:06

I øjeblikket er Google-oversættelse ret enerådende på internet, men oversættelser af god kvalitet må kræve at kildeteksten er skrevet utvetydigt, så maskinen undgår det statistiske gæt på meningen.

Det kan være svært at affatte en utvetydig tekst på så kaotiske sprog som engelsk og dansk, mens det på mere systematiske sprog (tysk, russisk, esperanto) måske er muligt at udtrykke sig stort set utvetydigt. Man kunne også skabe et helt nyt sprog beregnet til at skrive tekster til maskinoversættelse. Sproglege, som netop bygger på dobbeltbetydninger, kan ikke udtrykkes på et utvetydigt sprog, men det er en nødvendig bivirkning, hvis man vil have et praktisk kildesprog til at udføre maskinoversættelser til andre sprog.

Svar

Ole Stig Andersen
27. februar 2010 • 17:46

til Per Hagemann
“oversættelser af god kvalitet må kræve at kildeteksten er skrevet utvetydigt, så maskinen undgår det statistiske gæt på meningen.”

Slet ikke. En god oversættelse af en tekst er også en oversættelse af dens eventuelle mangler. Verden vrimler fx med tvetydige tekster, og de skal vel også kunne oversættes?

Du får svært ved at reformere Google Transæate, hvis du vil reformere dens grundlag, som er ... statistiske gæt.

Så mens du venter på at der bliver skabt et nyt sprog der er mindre tvetydigt end dem vi har i dag, OG venter på at folk i større mængde får lært det og vil bruge det, er vi nok en hel del der godt vil tage til takke med mangelfulde og til tider misvisende, men alligevel meget-bedre-end-ingenting-oversættelser.

Du kommer til at vente adskillige inkarnationer. Og du skal nok heller ikke regne med hjælp fra tysk eller russisk, som naturligvis ikke er et lod mere utvetydige end andre sprog. Det er indholdet af menneskers udsagn der er tvetydige eller utvetydige, ikke den sproglige struktur. At esperanto har en langt mere systematisk (dvs undtagelsesfri) grammatik hindrer naturligvis ikke at der også kan siges det argeste tvetydige, usystematiske sludder på det sprog også.

Forresten, mens Google Translate mærkeligt nok ikke tilbyder esperanto, så gør GramTrans http://gramtrans.com/ som Bicks artikel bl.a. handler om, det, i hvert fald fra dansk til esperanto.

Svar

Per Hagemann
2. marts 2012 • 12:58

Ole Stig skrev:

“Forresten, mens Google Translate mærkeligt nok ikke tilbyder esperanto, så gør GramTrans http://gramtrans.com/ som Bicks artikel bl.a. handler om, det, i hvert fald fra dansk til esperanto.”

For en uge siden kom Esperanto med blandt sprogene i Google Oversæt.

Det mangelfulde ved Googles statistiske model gælder stadig. Det jeg mente med tvetydigheder er når mennesker ud fra en indsigt i tankegangen bag et udsagn kan tolke en sproglig tvetydighed efter den tilsigtede mening. Takket være et stort antal baggrundstekster klarer Google Oversæt ofte disse tvetydigheder.

Tvetydigheder kunne en maskine gengive ved at skrive begge mulige oversættelser. Det gør Google dog ikke. Skriver jeg ordet “mod” vælger Google kun at oversætte den almindeligste betydning.

Svar

Skriv en kommentar

Navn (kræves)

E-mail (kræves)

Hjemmeside

Send mig en e-mail når der kommer flere kommentarer.