

Introduktion till språkteknologi

OH-serie 8: maskinöversättning

<http://stp.lingfil.uu.se/~matsd/uv/uv08/ist/>



UPPSALA
UNIVERSITET

Mats Dahllöf
Institutionen för lingvistik och filologi
November 2008

1

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Översättning

Maskinöversättning omfattar översättning av texter från ett språk till ett annat, antingen i form av verktyg som hjälper mänskliga översättare, eller program som syftar till att åstadkomma av automatisk översättning. Vanligen ger ett sådant program en första grovöversättning som sedan kan förfinas.

<http://sprakteknologi.se/vad-aer-sprakteknologi/maskinoeversaettning>

2

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Översättning

översättning, överföring av ett budskap från ett språk till ett annat, vanligen i skriftlig form; jfr tolkning. (Nationalencyklopedin).

3

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Översättning

Translation consists in producing in the target language the closest natural equivalent of the text material of the source language, in the first hand concerning meaning, in the second hand concerning style.

Nida, E. (1975) "A Framework for the Analysis and Evaluation of Theories of Translation" in Brislin, R. W. (ed.) *Translation Application and Research*, Gardner Press, New York.

4

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Inblandade språk

Normal är två språk inblandade i en översättningsprocess:

- **Källspråk (KS):** Det språk som källtexten är avfattad på.
- **Målspråk (MS):** Det språk som översättningen skall vara på.

5

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Vad skall en översättning bevara

- "Budskapet" skall bevaras.
- "Betydelsen" (och sekundärt "stilen") skall bevaras.
- Kriterier för detta?
- Hur skall texten användas? Vilket syfte skall översättningen tjäna?

6

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Mer eller mindre fri översättning

Thus conscience does make cowards of us (Hamlet).
(Obs! Skönlitteratur. Metriskt bunden översättning.)

- Formellt strikt översättning:
Så gör oss samvet till pultroner alla
(C.A. Hagberg)
- Friare översättning:
Den inre rösten gör oss alla fega
(Britt G. Hallqvist)

7

Introduktion till språkteknologi — HT 2008 (Mats Dahllöf)

Formell och dynamisk ekvivalens (Nida 1975)

- Formal equivalence focuses attention on the message itself, in both form and content. It aims to allow the reader to understand as much of the source language as possible.
- Dynamic equivalence is based on the equivalent effect, i.e. that the relationship between receiver and message should aim at being the same as that between the original receiver and the source language message.

8

Varför är maskinöversättning intressant?

- Översättning är en dyr tjänst. Tidsödande. Kräver kompetent arbetskraft.
- Mänsklig översättning kan ge en personlig variation där man inte vill ha det.
- Maskinöversättning kan föreligga som en omgående tillgänglig tjänst.
- Maskinöversättning är ett teoretiskt intressant problem, genom vilket många aspekter av språket kan belysas.

9

Maskinöversättning — historia

- Den första icke-numeriska tillämpning av datorer, föreslagen redan under sent 1940-tal.
- Forskning och utveckling av system, främst i USA, under 1950-talet.
- Ca 1958: 200-250 heltider. USA, Sovjetunionen. (UK, Italien).
Bar-Hillel (1960) "The Present Status of Automatic Translation of Languages".

10

Bar-Hillel (1960), viss pessimism

Bar-Hillel (1960) "The Present Status of Automatic Translation of Languages".

Fully automatic, high quality translation is not a reasonable goal, not even for scientific texts. A human translator [...] is often obliged to make intelligent use of extra-linguistic knowledge which sometimes has to be of considerable breadth and depth. Without this knowledge he would often be in no position to resolve semantical ambiguities. At present no way of constructing machines with such a knowledge is known, nor of writing programs which will ensure intelligent use of this knowledge.

11

Vad gäller 50 år senare?

- Språket lika svårt, förstås.
- Enormt mycket mer datorkraft finns tillgänglig.
- Enormt mycket större samlingar av språkligt material finns tillgängliga.
- Forskningen har gått framåt.

12

Vari ligger svårigheterna?

- Flertydigheter
 - Lexikal flertydighet (omfattande svårighet)
 - Grammatisk flertydighet
- Grammatiska skillnader
 - Morfologi
 - Syntaktisk konstruktion av olika innehåll
 - Ordföljd

13

Illustration: translate.google.com

- *They have also been disappointed by strong hints from the Obama team that he is none too keen either on multi-lateral regulatory reforms.* (edition.cnn.com)
- *De har också blivit besvikna över starka antydningar från Obamas team om att inte heller han är särskilt angelägen om multilaterala reformer av regelverken.* (Min övers.)
- *De har också blivit besvikna av starka tips från Obama-teamet att han är inget för angelägna antingen på multilaterala regelverken.* (MT)

14

Några problem i exemplet

- Lexikal flertydighet:
hint — antydning eller tips
either — inte heller eller antingen
- Morfologi, kongruens:
De har också blivit besvikna (borde vara *besvikna*)
- Lexikaliserad fras: *none too keen on*
- Konstruktion:
tips/antydning om att...
regulatory kontra *av regelverken*

15

När vill man använda maskinöversättning?

- Tekniska manualer behöver ofta översättas till många språk. Uppdateras ofta. Stora textsjok behålls.
- Samma sak gäller dagsrapporter om t.ex. väder.
- Samma sak gäller lagar och administrativa regelverk, särskilt inom EU och länder med flera officiella språk.
- När man vill komma åt innehållet i en text utan att ha tillgång till mänsklig översättningshjälp.

16

Olika typer av maskinöversättning

- **Automatiskt översättning:** datorsystem som från en originaltext levererar en översättning (utan mänsklig inblandning).
- **Översättningsminnen/translator's workbenches:** integrerade systemomgivningar för översättare med ett översättningsarkiv som en central komponent.

17

Automatiskt översättning

- Översättningen kan givetvis redigeras av en mänsklig användare efteråt. Den automatiska översättningen blir då mer av ett hjälpmedel.
- Syftet med automatiskt översättning kan vara att ta fram publicerbar text. Efterredigering, eller åtminstone granskning, torde då vara inblandat.

18

Automatiskt översättning (2)

- Ett annat användningsområde är att producera "råöversättningar" som ger läsaren eller beställaren en ungefärlig uppfattning om innehållet i ett originaldokument. De är inte avsedda att spridas, utan används bara för stunden.
- Översättning kan också göras för dokumentökningsändamål (utan att någon människa skall se på resultatet).

19

Översättningsminnen/translator's workbenches

- Hjälpmedel för översättare. Systemet översätter inte text på egen hand.
- Stor databas över befintliga översättningar.
- Slår upp i databasen på ett intelligent sätt. Ordningen hel mening, fraser (som urskiljs enligt någon princip), enstaka ord.
- Kombinerar med terminologistöd, etc.
- Bör vara väl integrerade med ordbehandlaren.

20

Regelbaserade metoder: transfer

- Regelbaserade metoder för MT är den äldre traditionen.
- MT-experten skriver formella språkregler.
- En komponent för analys av KS ger någon typ av grammatiska representationer.
- En komponent (transfer) bygger om KS-representationer till MS-representationer.
- Sedan finns en komponent för att generera MS-meningar från MS-representationer.

21

Transferbaserat system

Per språkpar (KS/MS):

- KS-analys
- KS-MS-transfer (språkparsspecifik anpassning av grammatik och översättning av ord)
- MS-generering

I renskuret fall är KS-analys oberoende av MS och MS-generering oberoende av KS.

22

Transfer, "djup"

- En ytlig transfer kan bygga enklare grammatisk uppmärkning, kanske vissa ordklasser, fraser och satsdelar. Skulle passa för två näraliggande språk, säg svenska och danska.
- En djupare transfer skulle utgå från en mer semantiskt orienterad representation. Skulle passa för mycket olika språk, där saker ofta uttrycks på väldigt skilda sätt, säg svenska och franska.

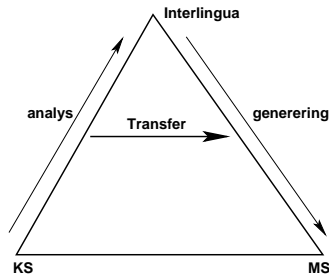
23

Regelbaserade metoder: interlingua

- Interlingua: en representation av betydelse som är språkoberoende (åtminstone relativt den grupp språk som systemet skall stödja).
- KS-till-Interlingua-analys. Interlingua-till-MS-generering.
- Analysen är s.a.s. så djup att den eliminerar behovet av transfer.
- Ekonomiskt, om vi skulle ha många språkpar. Exempel: 10 språk, alla till alla, ger 90 (riktningsberoende) transferkomponenter.

24

"Vaquois triangel"



25

Regelbaserad MT

- Utveckling dyrt, tidsödande insatser av experter.
- Begränsad domän är en fördel. *Météo*, kanadensiskt system för översättning av väderleksrapporter, nämns ofta som en framgång.
- Resultaten ser typiskt ut att vara sisådär.
- *Systran* är det kommersiellt största systemet. (Skall vara motorn i babelfish.yahoo.com.)

26

Datadrivna metoder inom MT

- Datadrivna metoder inom MT bygger på maskininlärning och data i form av parallellkorpusar.
- Parallellkorpusar: ihopparade texter på MS och KS, där ena texten utgör en översättning av den andra.
- Två familjer av ansatser: *Statistisk MT* och *Exempelbaserad MT* (subtila skillnader).

27

Behandling av parallellkorpusar

- Givet: en sammankoppling på dokumentnivå, får man tänka sig. (Beror på slag av dokument.)
- Vad man vill ha (1): En länkning (alignment) mening för mening KS till MS.
Ett språkteknologiskt problem.
Svårighet: relationen kommer inte alltid att vara ett-till-ett.

28

Behandling av parallellkorpusar (2)

- Vad man vill ha (2): En länkning (alignment) ord-för-ord KS till MS.
Ett språkteknologiskt problem (svårare är meningslänkning).
Svårighet: relationen är ofta komplicerad.

29

Ordlänkning (exempel att fundera på)

Jag vill ändå försäkra er om att kommissionen kommer att försöka införliva en sådan analys i den andra rapporten om sammanhållningen, som sannolikt bättre kommer att motsvara denna oro.

Nonetheless, I must assure you that the Commission will make every effort to include an analysis of this type in the second report on cohesion which, no doubt, will respond better to these concerns. (Europaparlamentet 991217)

30

Inlärning i datadriven MT

- Menings- och ordlänkning är förberedande steg. Särskilt ordlänkningen är ett svårt och avgörande steg.
- Inlärningen kan hitta kopplingar mellan uttryck KS till MS, på ordnivå, på frasnivå. Den kan ta hänsyn till kontexter.
- Detta kan givetvis göras på många olika sätt.
- P.g.a. många-till-många-förhållanden mellan ord, många möjligheter att anta fraser, etc. blir antalet möjligheter stort.

31

Inlärning i datadriven MT (2)

- Inlärningen kan även skapa en språkmodell för MS, t.ex. för att anpassa ordföljd och ordval i MS.
- När man betraktar resultat från translate.google.com, så ser det ut som om denna komponent har en stor inverkan.

32

translate.google.com, no–sv

Ta kontakt med den nye banken der du ønsker å bli kunde. Enklest er det å gjøre dette via bankens nettsider. (<http://www1.vg.no/pub/vgart.php?artid=522273>)

Hör med den bank där du vill vara en kund. Det är lättast att göra detta på bankens webbsidor.

33

translate.google.com, no–sv

En smaksrik lammegryte passer selvsagt utmerket som søndagsmiddag. Som drikkefølge velger vi oss en rødvin med masse matchende frukt. (<http://www.aperitif.no/index.db2?id=120989>)

En smaksrik LAMMGRYTA passar naturligtvis utmärkt som söndag middag. Som ett resultat väljer vi dricka en massa matchande rött vin med frukt.

34

translate.google.com, en–sv

Place the butter, sugar, cream and maple syrup into a saucepan and heat slowly until the sugar has dissolved and the mixture is rich and smooth, then add the pecans. (<http://www.guardian.co.uk>)

Placera smör, socker, grädde och lönnsirap i en kastrull och värm sakta tills sockret har löst och blandningen är rik och mjuk, lägg sedan till pekannötter.

35

translate.google.com, en–sv, mer

Following the failure of Lehman Brothers, the turmoil that has affected financial markets over the past year intensified into the most serious financial crisis since the outbreak of the Great War,” said the governor. (<http://www.guardian.co.uk>)

Efter misslyckandet i Lehman Brothers, den turbulensen som har drabbat de finansiella marknaderna under det senaste året intensifierats i den allvarligaste ekonomiska krisen sedan utbrottet av första världskriget, sade guvernören.

36

translate.google.com, da–sv

I USA er hun et stort navn, og har billeder på museer både i Boston og Denver. Med portrættet af Obama er hun blevet endnu mere kendt. (<http://jp.dk/kultur/billedkunst/article1515944.ece>)

I USA, hon är ett stort namn, och har bilder på museer både i Boston och Denver. Med porträtt av Obama, hon har blivit ännu mer känd.

37

translate.google.com, da–sv

Rimfrost gjorde søndag aften vejene omkring Thisted spejlglatte, hvilket kostede en enkelt bilist en tur i grøften i Vilsund. (<http://jp.dk/indland/trafik/article1523004.ece>)

Frost gjorde söndag kväll vägarna runt om i spegel-Örebro smidig, som dödade en bilist på en promenad i diket i Vilsund.

38

Källtextsegenskaper

Källtexterna kan väljas/anpassas så att översättning underlättas med ett visst MT-system.

- Domän: texttyp, typ av innehåll, språktyp.
- Korrekthet (stavning, grammatik).
- ”Kontrollerat språk”: källtexten har utformats i enlighet med strikta föreskrifter rörande t.ex. terminologi och grammatik. (Passar t.ex. i tekniska sammanhang.)

39

Värdering av MT-resultat och MT-system

- Det är svårt att bedöma kvalitet på översättning (både mänsklig och maskinell).
- Två dimensioner: Hur trogen innehållet är översättningen? Hur välformulerad är den?
- En grundläggande aspekt är att värdera enskilda översättningar.
- För att värdera ett helt system måste man värdera översättningen av de texter som kan förekomma med hänsyn till hela användningssituationen, särskilt syftet.

40

Värdering av översättningar

Två sorters metoder:

- Utvärdering där mänskliga bedömningar utgör data. (Kan göras mer eller mindre sofistikerat.) Väl utvalda bedömarpaneler får kanske betygssätta på något genomtänkt sätt. (Dyrt.)
- Automatisk utvärdering mot en ”gold standard” (korpus av facitöversättningar). Hur skall jämförelserna göras? Och hur bra är metoden? (Två bra och jämbördiga översättningar kan ju vara väldigt olika.)

41

Jämförelse 1: källtext (franska)

Källtext:

Un médecin de la cour consigna cependant une autre version dans ses carnets : selon lui, même si Guangxu souffrait de maux divers, il s'était rétabli avant sa mort, qui fut aussi soudaine que suspecte.

(www.lemonde.fr)

42

Jämförelse 1: översättning (engelska)

A doctor of the court however consigned another version in his notebooks: according to him, even if Guangxu suffered from various evils, it s' was restored before its death, which was as sudden as suspect. (babelfish.yahoo.com)

A doctor from the court consigna however, another version in his notebooks: according to him, even if Guangxu suffering from various ailments, it was restored before his death, which was as sudden as suspicious. (translate.google.com)

43

Jämförelse 2: källtext (franska)

Källtext:

Dotée d'un budget de 20 millions d'euros, cette sanglante fresque politique dissèque avec efficacité la naissance, la montée en puissance et la fin pathétique de cette fraction gauchiste, dans l'Allemagne d'avant la chute du Mur.

(madame.lefigaro.fr)

44

Jämförelse 2: översättning (engelska)

Equipped with a budget of 20 million euros, this bloody political fresco dissects with effectiveness the birth, the rise to power and the pathetic end of this gauchist fraction, in Germany of before the fall of the Wall.

(babelfish.yahoo.com)

With a budget of 20 million euros this bloody fresco policy effectively dissects the birth, rise and end this pathetic leftist fraction in Germany before the fall of the Wall.

(translate.google.com)

45

Maskinöversättning

- Svårt problem. Bar-Hillels kommentarer från 1960 gäller i hög grad även idag. Översättning förutsätter normalt förståelse.
- För språkteknologin typisk utveckling från regelbaserade system till datadrivna.
- Ofta imponerande kvalitet, men även ofta knasigt resultat.
- Systemen måste anses användbara. Språkteknologi kan spara pengar.

46