Vad kan en fackö̈versättare vänta sig av maskinöversättning?

Aarne Ranta

Institutionen för data- och informationsteknik
Chalmers tekniska högskola och Göteborgs universitet
SFÖ Konferens, 20 april 2013, Göteborg

Hur jag hamnade här

Född i Tammerfors, Finland

Kunde inte välja mellan språk och matematik

Syntes: logik, grammatik, datalingvistik

Doktorand i Stockholm 1987-1990

1997 inbjuden till Xerox Research Centre Europe, Grenoble, för att starta projekt om ”Multilingual Document Authoring”: skriv ett dokument i ett språk och se det utvecklas i många andra

1999 lektor i datavetenskap i Göteborg, 2005 professor

2010 koordinator i EU-projektet MOLTO = Multilingual Online Translation, med deltagare från 6 länder

En prognos

Five, perhaps three years hence, interlingual meaning conversion by electronic process in important functional areas of several languages may well be an accomplished fact.

En prognos

Five, perhaps three years hence, interlingual meaning conversion by electronic process in important functional areas of several languages may well be an accomplished fact.

IBM press release 1954

http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html

Vad jag vill tala om

Hur maskinöversättning fungerar

Vad som är lätt och vad som är svårt i maskinöversättning

Demo av MOLTO-verktyg

En möjlig ny roll för översättare, högre upp i värdekedjan

Maskinöversättning

Översättning med hjälp av datorer

eller

Datorer som översättare

Tänk på maskiner i allmänhet

Byggarbete som underlättas av maskiner

eller

Maskiner som utför hela byggarbetet

En ingenjörssyn på maskiner

Maskiner kan göra arbeten som är tråkiga ...

... och som är tillräckligt enkla för att mekaniseras

Gräva, lyfta, måla stora ytor ...

... men inte designa, lösa problem, skapa konstverk

Forskningen flyttar ständigt gränsen, men den får inte överskridas för hastigt!

En artificiell intelligens-syn på maskiner

Människan själv är *kanske* en komplex maskin

Maskiner ska simulera människor

Det är inte ännu perfekt, men om fem år ska det vara det!

Synen på maskinöversättning

Ingenjörssynen

- maskiner kan hjälpa till med rutinarbeten: ordboksuppslag, stavningskontroll, översättningsminnen, mm.
- människan gör arbetet, maskinen hjälper till

AI-synen

- maskiner blir bättre och bättre på att göra allting själv: människan gör post-editing
- maskinen gör arbetet, människan hjälper

Är inte det lite konstigt ...

... att just översättning domineras av AI-synen?

Ingen skulle tänka sig något sådant i byggarbete, vård eller journalism

Man kanske tänker att översättning i grund och botten är tillräckligt lätt för att bli helt automatiskt

Nya system spås om och om igen göra översättare överflödiga

Detta har hänt

Optimismen i maskinöversättning resulterade från andra världskrigets framgång med kryptanalys

Självaste Alan Turing föreslog (som *experiment*) att datorer kan göra översättning

Tanken hos amerikanerna: ryska = krypterad engelska

Man behöver bara knäcka koden!

Den första kritiken

Bar-Hillel 1964: FAHQT (Fully Automatic High Quality Translation) är omöjligt - inte bara i den närmsta framtiden, utan i princip.

Exempel: eng. *pen* = sv. *penna* ; *lekhage*

Välj rätt i

the pen is in the box

the box is in the pen

Det behövs intelligens och världskunskap

Översättning är ”AI-komplett” dvs. man kan behöva lösa alla problem i AI för att kunna översätta automatiskt.

Kvarstår Bar-Hillels problem?

Google translate 19 April 2013

The pen is in the box. **Pennan är i lådan.**

The box is in the pen. **Rutan är i pennan.**

The children are playing in the pen. **Barnen leker i pennan.**

ALPAC-rapporten

1966, ALPAC = Automatic Language Processing Advisory Committee

Satsningen på maskinöversättning har varit bortkastade pengar

Maskinöversättning är för dyrt - *dubbelt så dyrt* som att låta människor översätta (!)

- det går snabbare att översätta manuellt från början än att post-editera
- det finns ett tillräckligt stort utbud av översättare för att täcka alla behov

The role of men and machines

Martin Kay, ”The proper place of men and machines in language translation”, 1980 (published 1998)

Flera exempel i Bar-Hillels anda: det är omöjligt att mekaniskt avgöra vilken mening texten har

Flyttade fokus från översättande maskiner till maskiner som hjälper människan att översätta (t.ex. införde idén med översättningsminnet)

Den nya vågen av maskinöversättning

Startade 1989 på IBM

Återupplivande av kryptanalysmetoden - men nu med starkare datorer och mer data

Franska till engelska, baserat på handlingar från parlamentet i Kanada

Inga språkkunskaper behövdes för att utvecka systemet

Google translate, med sina 60 språk, är direkt uppföljning av IBM-systemet

Konsument vs. producent

Fokus i Google translate är översättning för *konsumenter* av information

Konsumenter hittar en webbsida och vill översätta den för eget bruk

Ansvaret ligger hos konsumenten

Ansvaret

Fransk e-handelbutik: *prix 99 euros*

Svensk översättning: *pris 99 kronor*

Ansvaret

Fransk e-handelbutik: *prix 99 euros*

Svensk översättning: *pris 99 kronor*

Kan den svenske konsumenten kräva att få varan för 99 kronor?

Ansvaret

Fransk e-handelbutik: *prix 99 euros*

Svensk översättning: *pris 99 kronor*

Kan den svenske konsumenten kräva att få varan med 99 kronor?

Beror på vem som är ansvarig för översättningen

- den franska butiken, producenten? Ja!
- den svenske användaren, konsumenten? Nej!

Producentverktyg?

Producenter brukar inte lita på maskinöversättning!

Det är människor, fackköversättare som ska göra jobbet!

Hur kan maskiner underlätta jobbet?

Hur mycket går att mekanisera?

Två skolor i maskinöversättning

Statistiska metoder (SMT = Statistical Machine Translation)

- exempel: Google translate
- metoder från kryptanalys
- söker efter ”den sannolikaste översättningen” i ljuset av tidigare data

Regelbaserade metoder (RBMT = Rule-Based Machine Translation)

- exempel: MOLTO
- metoder från lingvistik och dataspråksöversättning
- söker efter ”den rätta översättningen” som ska återskapa meningen

Regelbaserad översättning

Datorn är bra på att **följa regler mekaniskt** - bättre än människan

Vi väntar oss att datorn alltid gör rätt:

- $2 + 2 = 4$
- 365 * 24 * 60 * 60 = 31536000

Datorn är en bra kalkylator

Kompilatorer

Man kan kalkylera med siffror - men också med språkregler

Exempel: **kompilator**, som översätter programkod till maskinkod

```
x = 2 * x + 9    --->  01010110110100010101010100101001
```

Här har faktiskt datorn ersatt människor som översättare: förr i tiden var programmerarna tvungna att skriva maskinkod själva.

Regler i naturligt språk

Morfologi: böjning

- *fena* —> *fena, fenan, fenor, fenorna*
- *känna* —> *känner, kände, känt, känns, kändes, känts*

Syntax: kongruens, ordföljd

- *den + stor + hus* —> *det stora huset*
- *jag + sova + inte* —> *jag sover inte*
- *om + (jag + sova + inte)* —> *om jag inte sover*

Hur datorn kan hjälpa med reglerna

Morfologi och syntax har mekaniska regler

En infödd talare - och en fackö̈versättare - kan dem utantill, *i princip*

Men man kan lätt göra fel, t.ex. följa kongruensen överallt:

- *Den allra största frågan är fortfarande öppen, och det är inte säkert att ledningen någonsin kommer att nå överenskommelse om den.*
- **Det** *allra största* **problemet** *är fortfarande* **öppet**, *och det är inte säkert att ledningen någonsin kommer att nå överenskommelse om* **det**.

Kontrollprogram

Stavningskontroll: ordlista + böjningsregler + sammansättningsanalys

- ganska lätt för datorer
- användaren måste bidra med okända ord

Grammatikkontroll: syntaktisk analys + kongruensregler + ordföljdsregler

- lätt när det sker på nära avstånd
- svårare med längre avstånd
- användaren borde kunna bidra med rätt analys

Grammatikregler i automatisk översättning

Grammatiken varierar från språk till språk

- olika kongruensregler
- olika ordföljdsregler

Men språken kan ha samma **struktur**, s.k. **abstrakt syntax**

Exempel

Abstrakt syntax	`Sats(subjekt,verb,objekt)`	`Sats(I,know,you)`
Engelska	`subjekt + verb + objekt`	*I know you*
Tyska, huvudsats	`subjekt + verb + objekt`	*ich kenne dich*
- bisats	`subjekt + objekt + verb`	*ich dich kenne*
- efter adverbial	`verb + subjekt + objekt`	*kenne ich dich*
Franska, vanligt objekt	`subjekt + verb + objekt`	*je connais cet homme*
- pronomen som objekt	`subjekt + objekt + verb`	*je te connais*

Demo

MOLTO verktyg: GF cloud

- prediktiv parsning
- ordlinjering (word alignment)
- syntax editing
- 26 språk

http://cloud.grammaticalframework.org/minibar/minibar.html

if I know you, you know me

Problem: rätt analys

I ate a pizza with shrimps

Problem: rätt analys

I ate a pizza with shrimps

I ate a pizza with friends

Problem: rätt analys

I ate (*a pizza with shrimps*)

(*I ate a pizza*) *with friends*

Problem: rätt analys

I ate (*a pizza with shrimps*)

(*I ate a pizza*) *with friends*

(*I ate a pizza*) *with my fingers*

Statistisk översättning

Ordlinjering: ord som förekommer parallellt -> översättnigslexikon

n-gram: sekvenser av n ord (n=1,2,3,4,...) -> syntaktiskt möjliga kombinationer

Fördel: rätt analys

Frekventa n-gram fångas väl i modellen

vice president

-> *vice ordförande* (Google translate)

-> *skruvstädspresident* (MOLTO baseline translator)

Glesa data

Det behövs 1 miljon ord av parallell text för bra ordlinjering

Det behövs 10 miljon ord text i målspråket för en bra n-grammodell

Det behövs mer om språket har många böjningar eller varierande ordföljd

Glesa data och böjningar

Alla finska ordformer har inte setts (Google translate 19 April 2013)

yö, yön, yötä, yöksi, yönä, yössä, yöstä, yöhön, yöllä, yöltä, yölle, yöttä, yöt, öiden, öitä, öiksi, öinä, öissä, öistä, öihin, öillä, öiltä, öille, öittä, öine, öin

night, night, night, night, night, night, night, night, night, night, nights, Yotta, night, nights, nights nights, nights, nies, nights, nights, nights, with, company against loss, nities, öittä, öine, night

Problem: ”long distance dependencies”

Svensk kongruens försvinner (Google translate 19 April 2013)

The house is old.	*Huset är gammalt.*
The house is so old.	*Huset är så gammalt.*
The house is not so old.	*Huset är inte så gammal.*

Problem: ”long distance dependencies”, del 2

Lite allvarligare fel: tyska *um* och *bringen* hamnar för långt från varandra (Google translate 19 April 2013)

Er bringt dich um.	*He will kill you.*
Er bringt deinen Freund um.	*He brings to your friend.*

Mellanspråk

Problem: för lite parallell data svenska-bulgariska

Lösning: använd engelska som mellanspråk

- svenska-bulgariska = svenska-engelska + engelska-bulgariska

Problem: fel ordlinjering

Google translate, Summer 2010

Jag är svensk. -> Ich bin ein Amerikaner.

Förklaring

1. Linjering av parallella texter som är lokaliserade, inte bara översatta.

2. Användning av engelska som mellanspråk.

Jag är svensk -> I am American -> Ich bin ein Amerikaner

Hybridmetoder

Kombinera det bästa av RBMT och SMT

RBMT: grammatik

SMT: idiomatiska fraser, automatiskt skapade system

MOLTO

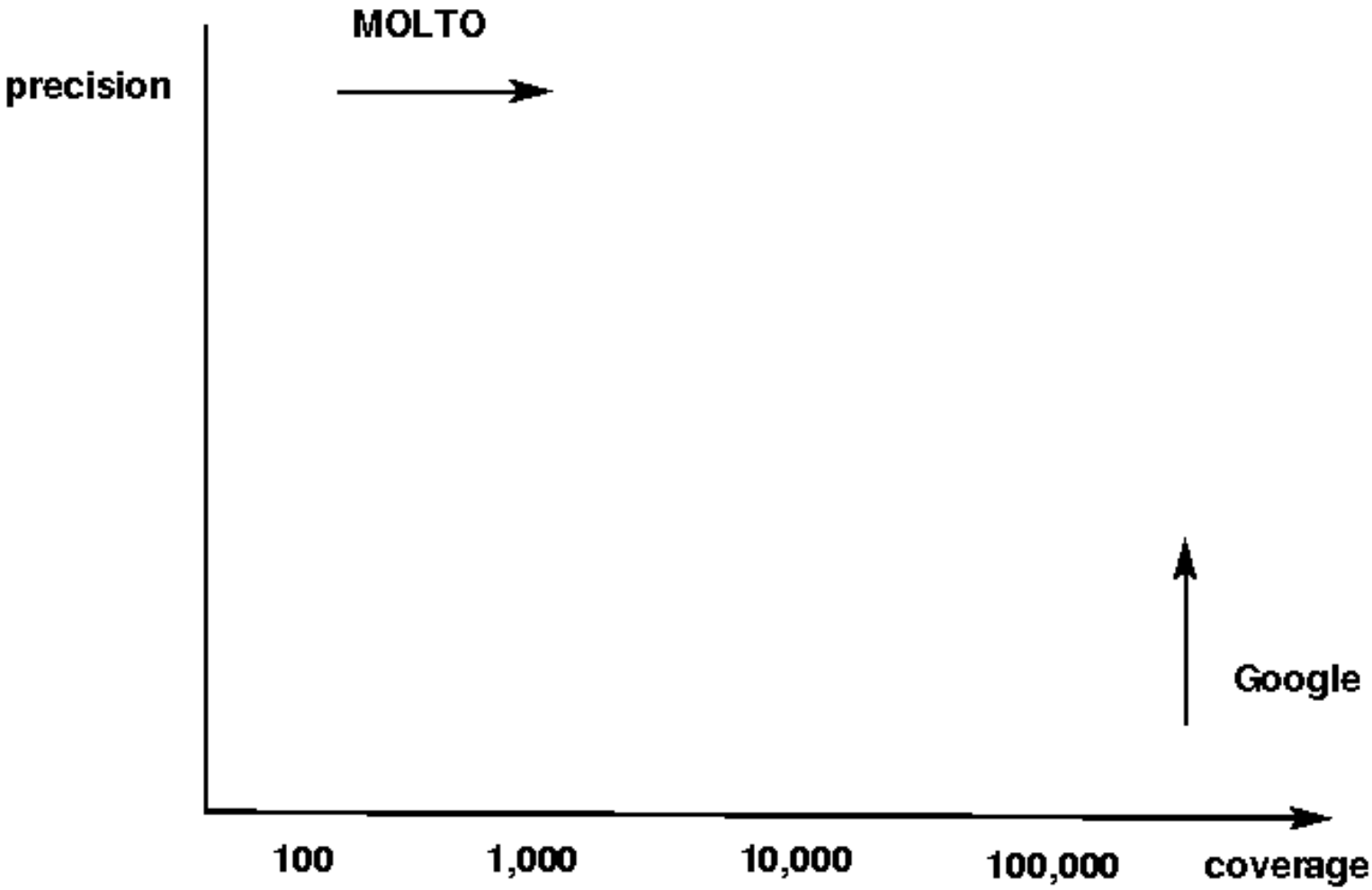
Hög kvalitet, 15 språk

Automatiskt, på webben

Rutinöversättningar med hög volym

Men: med begränsad täckning (s.k. **kontrollerat språk**)

Non multa sed multum.


MOLTO
precision
Google
100
1,000
10,000
100,000
coverage

Exempel från MOLTO

Texter från databas av konstverk (t.ex. Göteborgs stadsmuseum, Wikipedia)

Källa: formell beskrivning (abstrakt syntax)

Mål: 15 språk

Källa: tusentals formella beskrivningar t.ex.

```
MkGenText GSM980019Obj AnnaLindskog OilPainting (MkColour Black) (MkSize (SIntInt 435 365))
  (MkMaterial Canvas) (MkYear (YInt 1885)) (MkMuseum GoteborgsCityMuseum)
```

- PaintingEng: The girl was painted on canvas by Anna Lindskog in 1885. It is of size 435 by 365 and it is painted in black. This oil painting is displayed at the City Museum of Gothenburg.

- PaintingFin: Maalauksen Flickan on maalannut Anna Lindskog kankaalle vuonna 1885. Se on kokoa 435 kertaa 365 ja se on maalattu mustalla. Tämä öljymaalaus on esillä Göteborgin kaupunginmuseossa.

- PaintingFre: Le tableau Flickan a été peint sur toile par Anna Lindskog en 1885. Il est de taille 435 sur 365 et il est peint en noir. Cette peinture à l' huile est exposée dans le musée municipal de Göteborg.

- PaintingIta: Il quadro Flickan è stato dipinto su tela da Anna Lindskog nel 1885. Misura 435 per 365 ed è dipinto in nero. Questo dipinto ad olio è esposto nel museo municipale di Goteburgo.

- PaintingSwe: Flickan målades på duk av Anna Lindskog år 1885. Den är av storlek 435 gånger 365 och den är målad i svart. Den här oljemålningen är utställd på Göteborgs stadsmuseum.

En möjlig ny uppgift för översättare

Översättning av **regler** i stället för dokument

Systemet lägger till de mekaniska konsekvenserna

En generalisering av översättningsminnen

Högre upp i värdekedjan

Demo: grammatikeditor

http://cloud.grammaticalframework.org/gfse/

Uppgift: interaktivt anpassa en given grammatik till ett nytt språk

Slutsatser

Maskinöversättning kan inte ersätta människor inom överskådlig framtid.

Maskiner kan hjälpa till med grammatikkontroll.

Maskiner kan göra rutinöversättningar snabbt.

En möjlig ny roll för översättare: utvecklare av automatiska system.