Maskinöversättning
Vad och hur?

Anna Sågvall Hein

2016-03-10

Översikt

Vad är maskinöversättning?

Kort tillbakablick

Varför är det så svårt?

Språkvetenskaplig
- regelbaserad maskinöversättning

Statistisk maskinöversättning

Maskinöversättning i praktiken

Forskning i fokus

Kort tillbakablick

De första försöken gjordes redan på 50-talet i USA och Sovjet med en primitiv teknologi.

De gällde översättning mellan ryska och engelska för politiska och militära ändamål och aktörerna var tekniker.

De högt ställda förhoppningarna grusades och efter några års stiltje inleddes en ny period med språkvetare/datorlingvister och datavetare i spetsen.

De språkvetenskapligt baserade metoder som började utvecklas på 60-talet har i modifierad form överlevt till i dag, då de fått stark konkurrens av statistiska metoder.

Varför är det så svårt?

Flertydighet

- ett ord kan representera olika ordklasser, former och betydelser

Variation

- olika ord för samma företeelse – synonymi

Språkskillnader

- ordböjning, ordföljd

Exempel

Boken ligger under *bordet.*
The book is under *the table.*

Studenten somnade under *föreläsningen.*
The student fell asleep during *the lecture.*

Bara ett ***under*** *kan rädda tidningen.*
Only a ***miracle/wonder*** *can save the newspaper.*

Under tiden stannade bilen*.*
Meanwhile, the car stopped*.*

Var snäll och skriv under *här.*
Please, sign *here.*

Regelbaserad översättning

ANALYS

skapar en grammatisk struktur, där översättningsalternativ känns igen och väljs

TRANSFER

översätter den grammatiska strukturen till en grammatisk struktur på målspråket

GENERERING

bygger upp en översättning på den grammatiska strukturen

Alla komponenterna använder sig av omfattande lexikon, grammatiker och transferregler

UPPSALA UNIVERSITET

Convertus

Analysstruktur

Under tiden stannade bilen

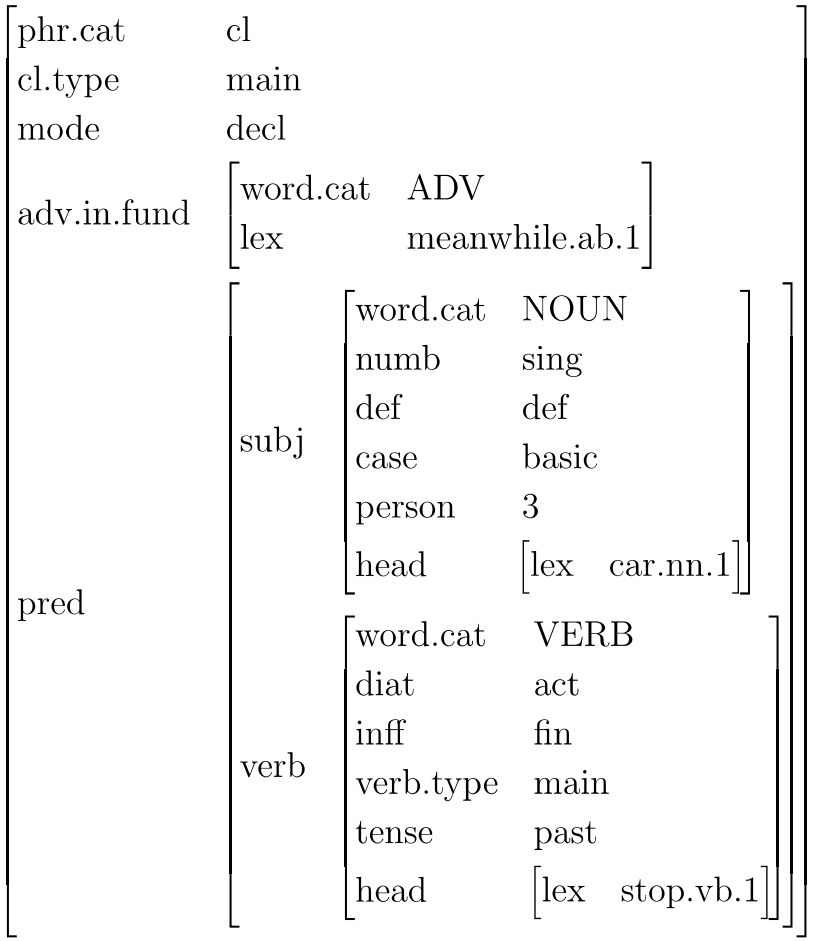
phr.cat	cl	
cl.type	main	
mode	decl	
verb	word.cat	VERB
	diat	act
	inff	fin
	verb.type	main
	tense	past
	head	lex: stanna.vb.1; trglex: stop.vb.1
adv.in.fund	word.cat	ADV
	lex	under_tiden.ab.1
	trglex	meanwhile.ab.1
subj	word.cat	NOUN
	gender	utr
	numb	sing
	def	def
	case	basic
	person	3
	head	lex: bil.nn.1; trglex: car.nn.1

"Under tiden stannade bilen"

UPPSALA UNIVERSITET

Convertus

Transferstruktur



UPPSALA UNIVERSITET

Convertus

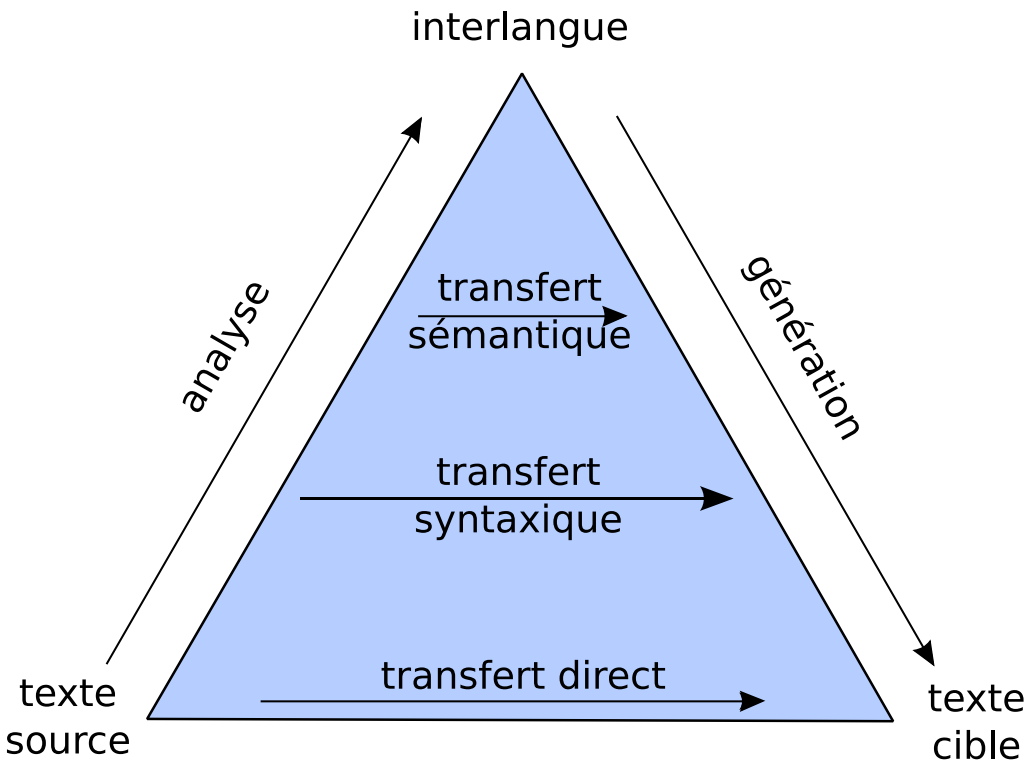
Genereringsstruktur

Meanwhile, the car stopped

$$
\begin{bmatrix}
\textit{"Meanwhile"} \\ \text{ABP}
& \begin{bmatrix} \text{word.cat} & \text{ADV} \\ \text{degree} & \text{pos} \\ \text{lex} & \text{meanwhile.ab.1} \end{bmatrix} \\
\textit{","} \\ \text{SRI}
& \begin{bmatrix} \text{word.cat} & \text{SEP} \\ \text{type} & \text{minor} \\ \text{lex} & \text{comma.sr.1} \end{bmatrix} \\
\textit{"the"} \\ \text{AL}
& \begin{bmatrix} \text{word.cat} & \text{ART} \\ \text{det.type} & \text{det} \\ \text{lex} & \text{the.al.1} \end{bmatrix} \\
\textit{"car"} \\ \text{NNSB}
& \begin{bmatrix} \text{word.cat} & \text{NOUN} \\ \text{numb} & \text{sing} \\ \text{case} & \text{basic} \\ \text{person} & 3 \\ \text{lex} & \text{car.nn.1} \end{bmatrix} \\
\textit{"stopped"} \\ \text{VBRM}
& \begin{bmatrix} \text{word.cat} & \text{VERB} \\ \text{inff} & \text{fin} \\ \text{verb.type} & \text{main} \\ \text{tense} & \text{past} \\ \text{pv} & - \\ \text{numb} & \text{sing} \\ \text{person} & 3 \\ \text{lex} & \text{stop.vb.1} \end{bmatrix}
\end{bmatrix}
$$

"Meanwhile, the car stopped"

Transferöversättning på olika nivåer
Vauquois triangel



Transferbaserad MT

Källspråkstrogna översättningar

Kräver stora lexikon och grammatiker

Tar tid att bygga

Brister som uppdagas kan åtgärdas genom ändringar i lexikon och grammatik

Statistisk maskinöversättning

Ett statistiskt översättningssystem, en SMT, använder sig inte av någon språkkunskap utan bygger helt på tidigare översättningar

En SMT består av två huvudkomponenter

- en sannolikhetsbaserad översättningsmodell
- en modell över målspråket, s.k. språkmodell

Kvaliteten på de översättningar som en SMT producerar beror helt på kvaliteten på de båda huvudkomponenterna

Översättningsmodellen

Modellen består av sannolikhetsbaserade översättningar av ord och längre uttryck.

För varje delöversättning anges ett sannolikhetsvärde.

Översättningar av hela meningar byggs upp genom att delöversättningar kombineras på ett sätt som ger den totalt sett högsta sannolikheten.

Ett stort antal översättningar produceras.

Problem kan uppstå i gränsen mellan olika delöversättningar.

Språkmodellen

Språkmodellen innehåller stora mängder målspråksdata.

De är organiserade som enordingar och flerordingar – s.k. ngram.

De högst rankade översättningarna jämförs mot språkmodellen och den översättning som stämmer bäst väljs.

Träning av översättningsmodellen

Insamling och rensning av träningsdata i form av tidigare översättningar av rätt slag

språkpar, texttyp, ämnesområde

Parallellställning – länkning – av träningsdata på olika nivåer

meningar, ord, fraser

UPPSALA UNIVERSITET

Convertus

Meningslänkade träningsdata Regeringsförklaringen 1988

s1.1	REGERINGSFÖRKLARING .	Statement of Government Policy by the Prime Minister , Mr Ingvar Carlsson , at the Opening of the Swedish Parliament on Tuesday , 4 October , 1988 .	s1.1
s2.1	Eders Majestäter , Eders Kungliga Högheter , herr talman , ledamöter av Sveriges riksdag !	Your Majesties , Your Royal Highnesses , Mr Speaker , Members of the Swedish Parliament .	s2.1
s3.1	Sveriges neutralitetspolitik är av avgörande betydelse för vårt lands fred och oberoende .	Sweden' s policy of neutrality is of decisive importance for our peace and independence .	s3.1
s3.2	Den bidrar också till stabilitet och avspänning i vår del av världen .	It also contributes to stability and détente in our part of the world .	s3.2
s3.3	Kring denna politik finns en bred folklig uppslutning .	There is wide popular support for this policy .	s3.3
s3.4	Den kommer att fullföljas med kraft och konsekvens .	It will be pursued with firmness and consistency .	s3.4
s4.1	Neutralitetspolitiken stöds av ett starkt försvar till värn för vårt oberoende .	Our policy of neutrality is underpinned by a strong defence .	s4.1
		That safeguards our independence .	s4.2
s4.2	Kränkningar av svenskt territorium kommer aldrig att accepteras .	Violations of Swedish territory will never be accepted .	s4.3
s4.3	Armén kommer att reformeras och effektiviseras .	The army will be reorganized with the aim of making it more effective .	s4.4
s4.4	Det är regeringens föresats att söka breda lösningar i frågor som är av betydelse för vår nationella säkerhet .	It is the Government' s intention to seek broad solutions in issues that are of importance for our national security .	s4.5
s5.1	Regeringen har välkomnat överenskommelsen mellan Förenta staterna och Sovjetunionen om att avskaffa de landbaserade medeldistanskärnvapnen .	The Government welcomed the agreement between the United States and the Soviet Union on the elimination of land- based intermediate- range nuclear weapons .	s5.1

Ordlänkning - samförekomst

*Sveriges **neutralitetspolitik** är av avgörande betydelse för vårt lands fred och **oberoende**.*

*Sweden's **policy of neutrality** is of decisive importance for our peace and **independence**.*

***Neutralitetspolitiken** stöds av ett starkt försvar till värn för vårt **oberoende**.*

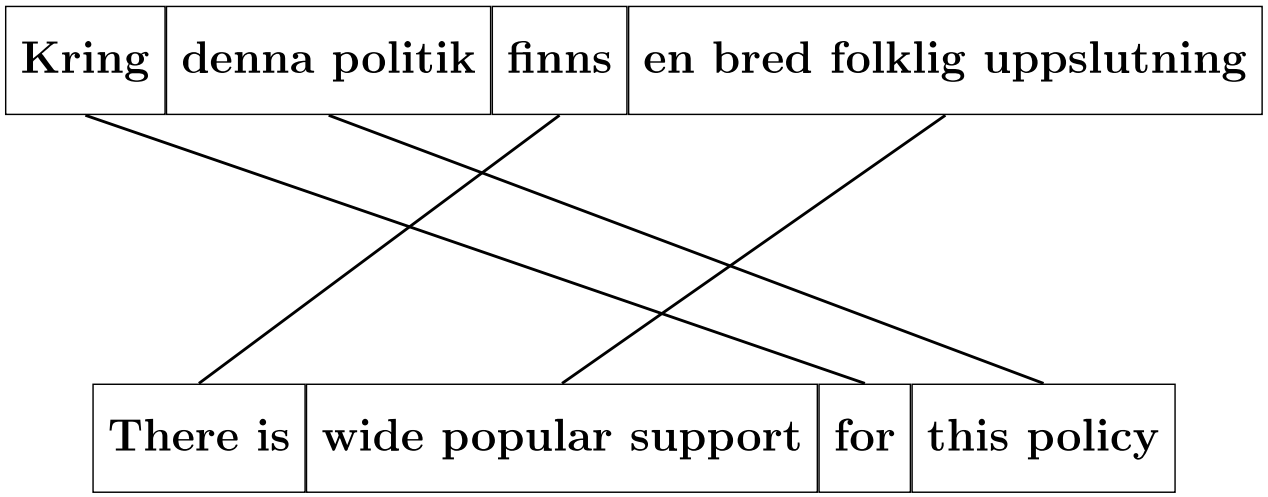
*Our **policy of neutrality** is underpinned by a strong defence. That safeguards our **independence**.*

Fraslänkning

Kring denna politik finns en bred politisk uppslutning.

There is wide popular support for this policy.

Ord- och fraslänkning



Resurser

För ett nytt språkpar behövs en översättningskorpus på minst 1 miljon ord och enspråkiga korpusar på över 1 miljard ord vardera (Franz Josef Och, Google)

Det finns open-source-mjukvara för att träna upp en SMT (t.ex. Moses)

SMT:s beroende av träningsdata

Under tiden somnade studenten.

Meanwhile, the student fell asleep.

Under tiden vaknade studenten.

Meanwhile woke student.

(Google 2016-03-09)

Utvärdering

Manuell

Automatisk

Automatisk utvärdering

Ett facit upprättas i form av en översättningskorpus

Översättningskorpusen hålls utanför träningsdata

En automatisk jämförelse görs mellan översättningen i korpusen och den maskingenererade översättningen

Ett likhetsmått beräknas, t.ex. BLEU som går mellan 0 och 1

Det har sagts att gränsen för att det ska löna sig att redigera en maskinöversatt text ligger vid ett BLEU-värde runt 0.4

Maskinöversättning i praktiken

Maskinöversättning + postredigering för publicering

Maskinöversättning som stöd för mänskliga översättare

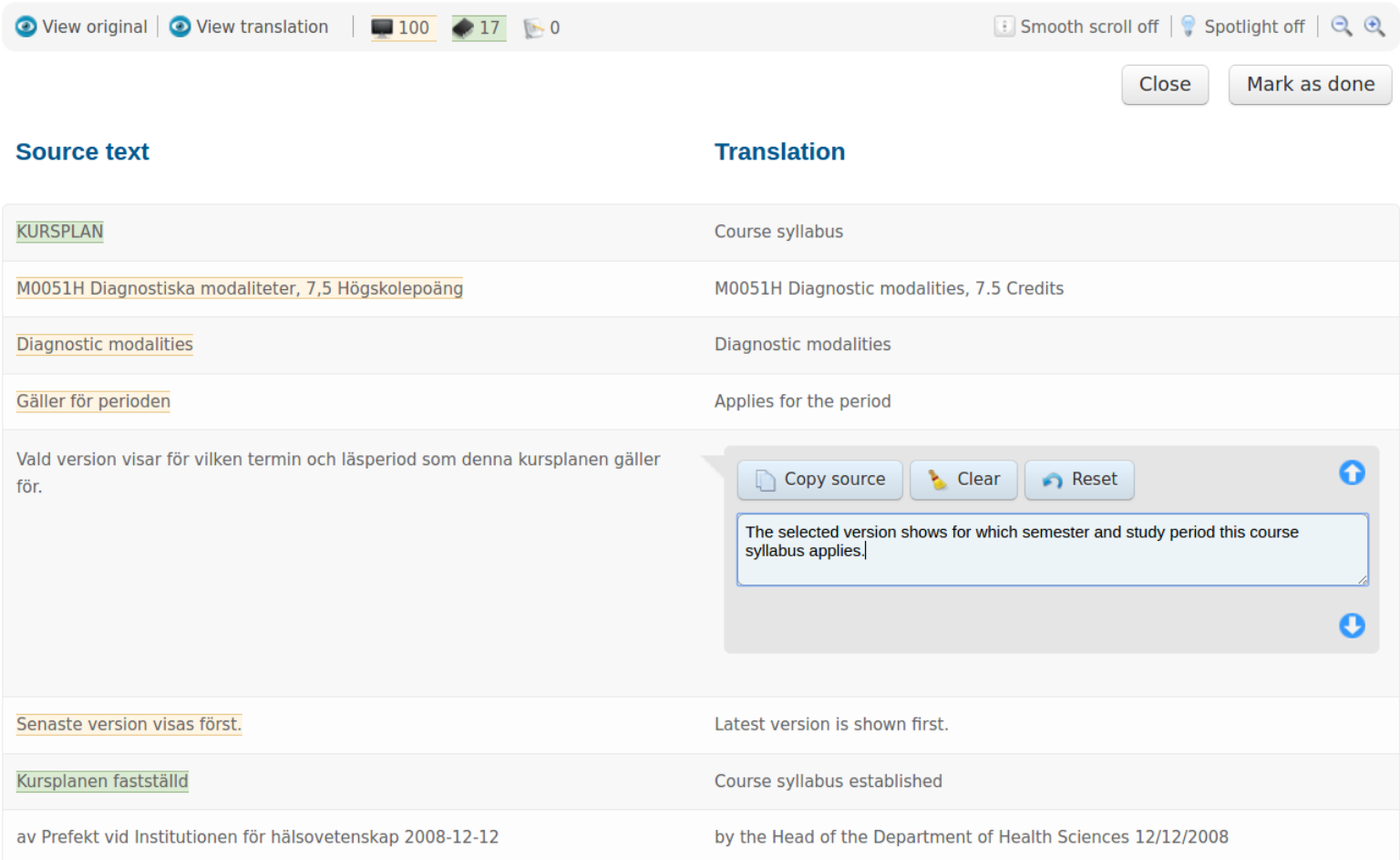
Snabb råöversättning - gisting

UPPSALA UNIVERSITET

Convertus

Post-redigering

Convertus-systemet



Forskning i fokus

Hybridsystem

- kombination av regelbaserade och statistiska system

Beroenden över meningsgränserna

- t.ex. syftningsproblem

Språkanpassning

- Inför träningen förbehandlar man källspråket för att göra det mer likt målspråket