

Maskinoversettelse

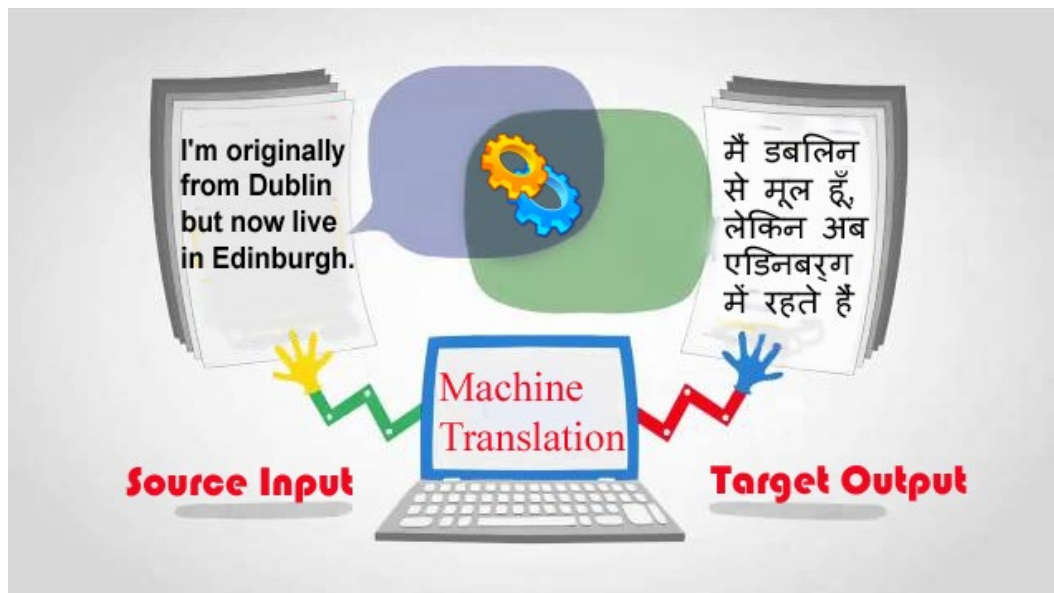
Pierre Lison
plison@nr.no

IN2110: Språkteknologiske metoder
16. april 2024



Maskinoversettelse

- ▶ **Mål:** automatisk oversette tekst eller tale fra et språk (*kildespråket*) til et annet språk (*målspråket*)
- ▶ Hundrevis av millioner brukere på verdensbasis
 - *Google Translate:* > 100 milliarder ord hver dag



Maskinoversettelse

Viktigste bruksområder:

1. «**Gisting**»: forstå den (omtrentlige) betydningen av tekster skrevet på et fremmedspråk
2. **Kommunisere** på tvers av språkbarrierer
3. **Teknologisk støtte** til menneskelig oversettelser



Plan for i dag

- ▶ Maskinoversettelsens historie
- ▶ Hvorfor er det vanskelig å oversette?
- ▶ Tilnærminger
 - Regelbaserte systemer
 - Statistiske metoder
 - Nevrale modeller
- ▶ Evaluering

Plan for i dag

- ▶ **Maskinoversettelsens historie**
- ▶ Hvorfor er det vanskelig å oversette?
- ▶ Tilnærminger
 - Regelbaserte systemer
 - Statistiske metoder
 - Nevrale modeller
- ▶ Evaluering

Maskinoversettelsens historie

- ▶ Forskning på maskinoversettelse startet allerede på 1950-tallet
- ▶ Hovedfokus: russisk → engelsk (kalde krigen!)



- ▶ Disse tidlige MT-systemene var **regelbaserte** (tospråklige ordbøker og enkle grammatiske regler)

Maskinoversettelsens historie

- ▶ 1950-1990: utvikling av (stadig mer avanserte) regelbaserte systemer
- ▶ ALPAC rapport i USA i 1966 kritisk til forskning om maskinoversettelse
 - konkluderte med at MT var "dyrere, mindre nøyaktig og tregere enn menneskelig oversettelse"
 - Drastisk reduksjon i finansieringsmidler i over et tiår
- ▶ 1970: første kommersielle systemer

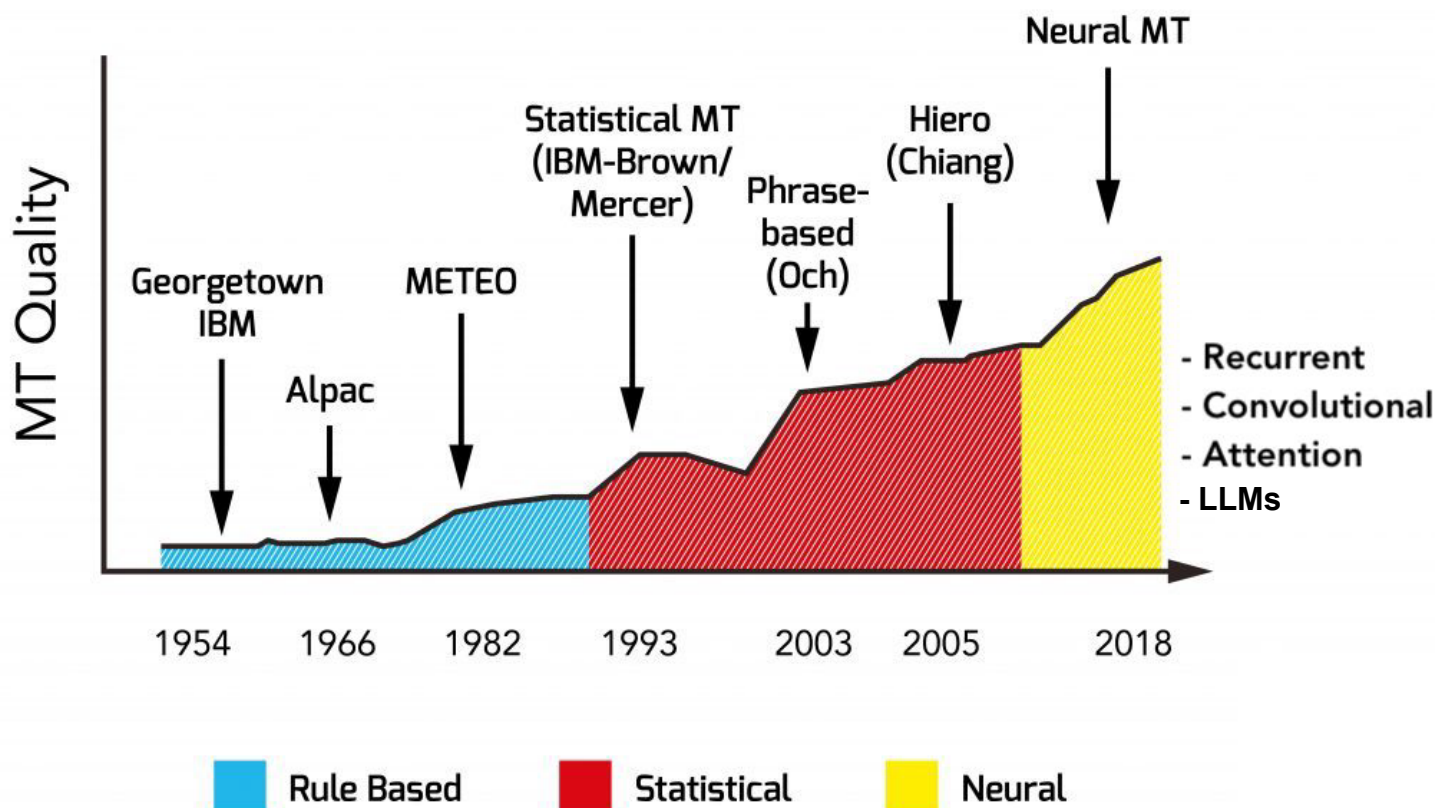


Maskinoversettelsens historie

- ▶ Fra 1990: utvikling av de første statistiske modellene for maskinoversettelse (SMT: «*Statistical Machine Translation*»)
- ▶ 2000-2010: SMT-modeller blir mainstream i forskningsmiljøer og kommersielle systemer
 - 2006: Lansering av Google Translate
- ▶ Fra 2013: Nevrale modeller (NMT: «*Neural Machine Translation*») stadig mer populære, og dominerer feltet
 - Google Translate nå basert på NMT!



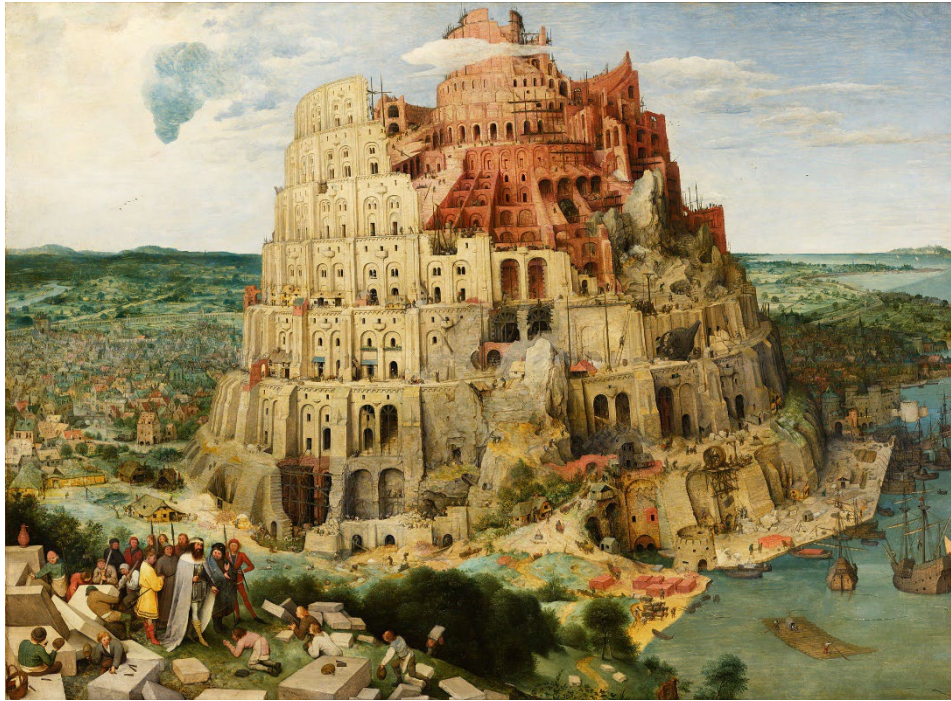
Fremskritt i oversettelseskvalitet



Plan for i dag

- ▶ Maskinoversettelsens historie
- ▶ **Hvorfor er det vanskelig å oversette?**
- ▶ Tilnærminger
 - Regelbaserte systemer
 - Statistiske metoder
 - Nevrale modeller
- ▶ Evaluering

Babels tårn



**Strukturelle
forskjeller mellom
språk:**

- Ordbygging (morfologi)
- Ordstilling (syntaks)
- Idiomatiske uttrykk
- osv.

"AI-complete problem" : for å oversette «riktig» må man forstå hensikten med teksten, utnytte bakgrunnskunnskap og kontekstuelle faktorer, osv.

Utfordringer: morfologi

- ▶ Noen språk har mange sammensatte ord mens andre «fordeler» betydningen på flere ord:
 - **Turkisk:** Avrupalılaştıramadıklarımızdanmışsınızcasına
→ «As if you are reportedly of those of ours that we were unable to Europeanize»
- ▶ **Bøyninger** (verbformer, kasus, osv.) må også handteres:
 - The small table → "der **kleine** Tisch"
 - A small table → Ein **kleiner** Tisch
 - It is on the small table → Es ist auf **dem** **kleinen** Tisch
 - He puts it on the small table → Er liegt es auf **den** **kleinen** Tisch
 - The surface of the small table → die Oberfläche **des** **kleinen** Tisches



Noen språk har veldig rike bøyninger: på *finsk* kan et substantiv har over 2000 ulike former!

Utfordringer: syntaks

Ordstilling kan variere mye fra språk til språk:

Das rote Buch, das er auf den Tisch gelegt hat
| / / / /
Le livre rouge qu'il a mis sur la table

À l'évidence,	son mari	était	un gros fumeur
---------------	----------	-------	----------------

Tydeligvis	var	mannen hennes	en storøyker
------------	-----	---------------	--------------

Diagram illustrating word order differences between French and Norwegian. Lines connect corresponding words: 'À l'évidence,' to 'Tydeligvis', 'son mari' to 'mannen hennes', 'était' to 'var', and 'un gros fumeur' to 'en storøyker'.

Utfordringer

- ▶ En del språk (kalt «pro-drop» eller nullanaforspråk) ikke trenger å uttrykke visse typer pronomener når disse er opplagt fra konteksten, som f.eks. italiensk:
 - "Lei non vuole mangiare" (hun vil ikke spise)
 - = "Non vuole mangiare"
- ▶ Slike setninger er veldig utfordrende å oversette til språk hvor subjektet er obligatorisk (som på norsk)
 - Pronomen "hun" må da genereres i målsetningen selv om den ikke står i kildesetningen

Utfordringer

- ▶ Leksikalske forskjeller mellom språk:
 - "know" → "vite" eller "kjenne" ?
 - "wall" → "vegg" eller "mur"?
 - "bein" → "bone" eller "leg"?
- ▶ Ulike språk setter sammen ord på ulike måter (oversettelsesmønstre ikke alltid *komposisjonelle*):
 - "heavy" → "tung"
 - "smoker" → "røyker"
 - "heavy smoker" → "storrøyker" og ikke "tungrøyker"

Utfordringer

Behov for bakgrunnskunnskap:

The *pen* was in the box

Pennen var i boksen

vs.

The box was in the *pen*

Boksen var i *bingen*



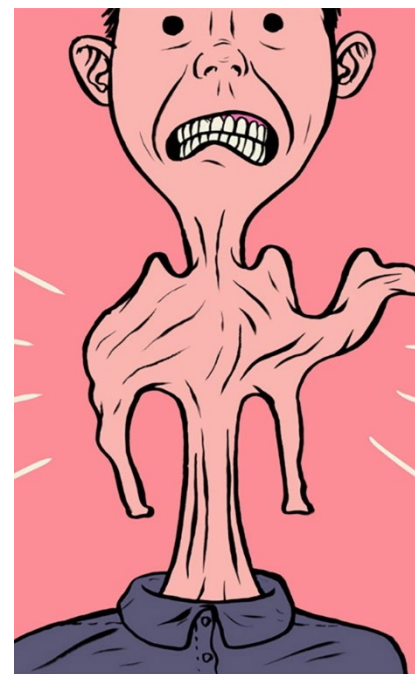
Det engelske ordet «pen» oversettes vanligvis med «pennen» på norsk, men kan også bety «bingen»

→ **Bakgrunnskunnskap:** en boks er selvfølgelig for stor til å være innen i en penn

→ «bingen» er å foretrekke i setning 2

Utfordringer

- ▶ Noen ganger bør man endre setningens struktur for å få et bedre "flyt" i målspråket:
 - "She likes to sing" → "Sie singt gerne"
- ▶ Idiomatiske uttrykk kan definitivt ikke oversettes ord for ord:
 - "Midt i blinken" → "Bull's eye", "right on spot", osv.
 - "Il pleut des cordes" (det regner snorer) → "Det bøtter ned"
 - "Å svelge en kamel" → "To give in"

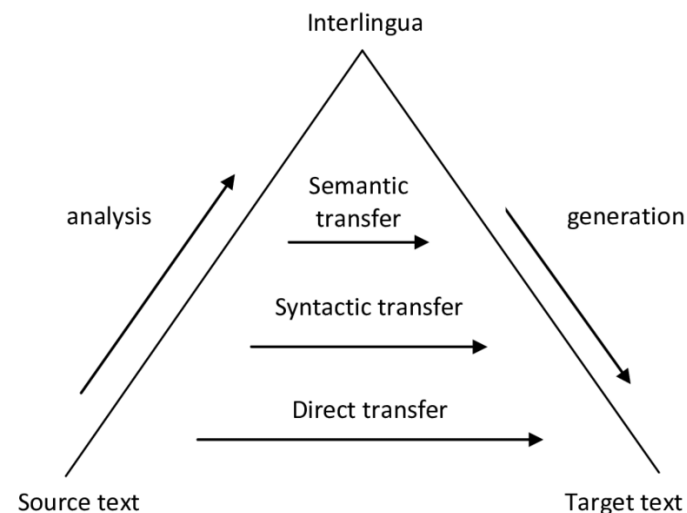


Plan for i dag

- ▶ Maskinoversettelsens historie
- ▶ Hvorfor er det vanskelig å oversette?
- ▶ **Tilnærminger**
 - Regelbaserte systemer
 - Statistiske metoder
 - Nevrale modeller
- ▶ Evaluering

Regelbaserte systemer

- ▶ Basert på tospråklige ordbøker og grammatiske regler
 - Ulike «dybder»
 - Mest radikalt: systemer basert på en «*interlingua*» bygger først en abstrakt, språkuavhengig representasjon av kildeteksten, og konverterer den deretter til en tekst i målspråket
- Veldig ressurskrevende!
- Hvordan kan man handtere **flertydighet**?



[Vauquois-trekant]

Veldig vanskelig å definere...

Regelbaserte systemer

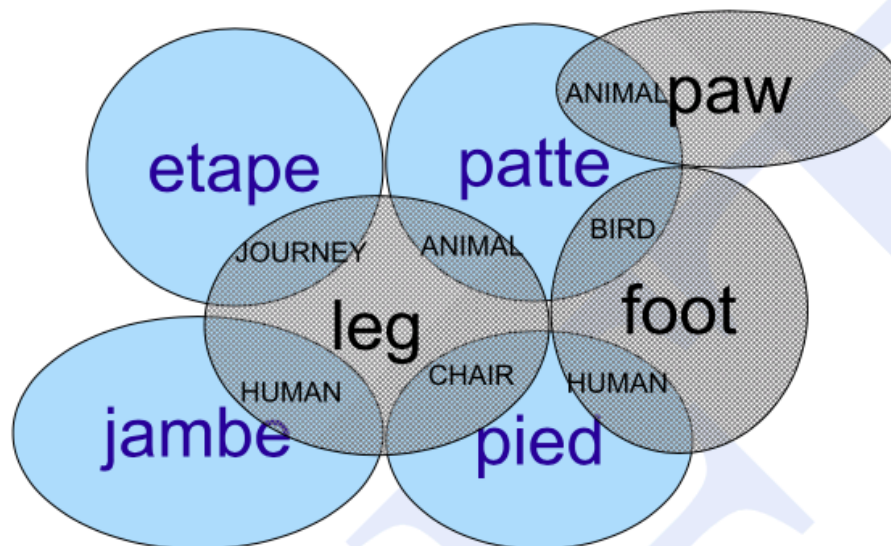
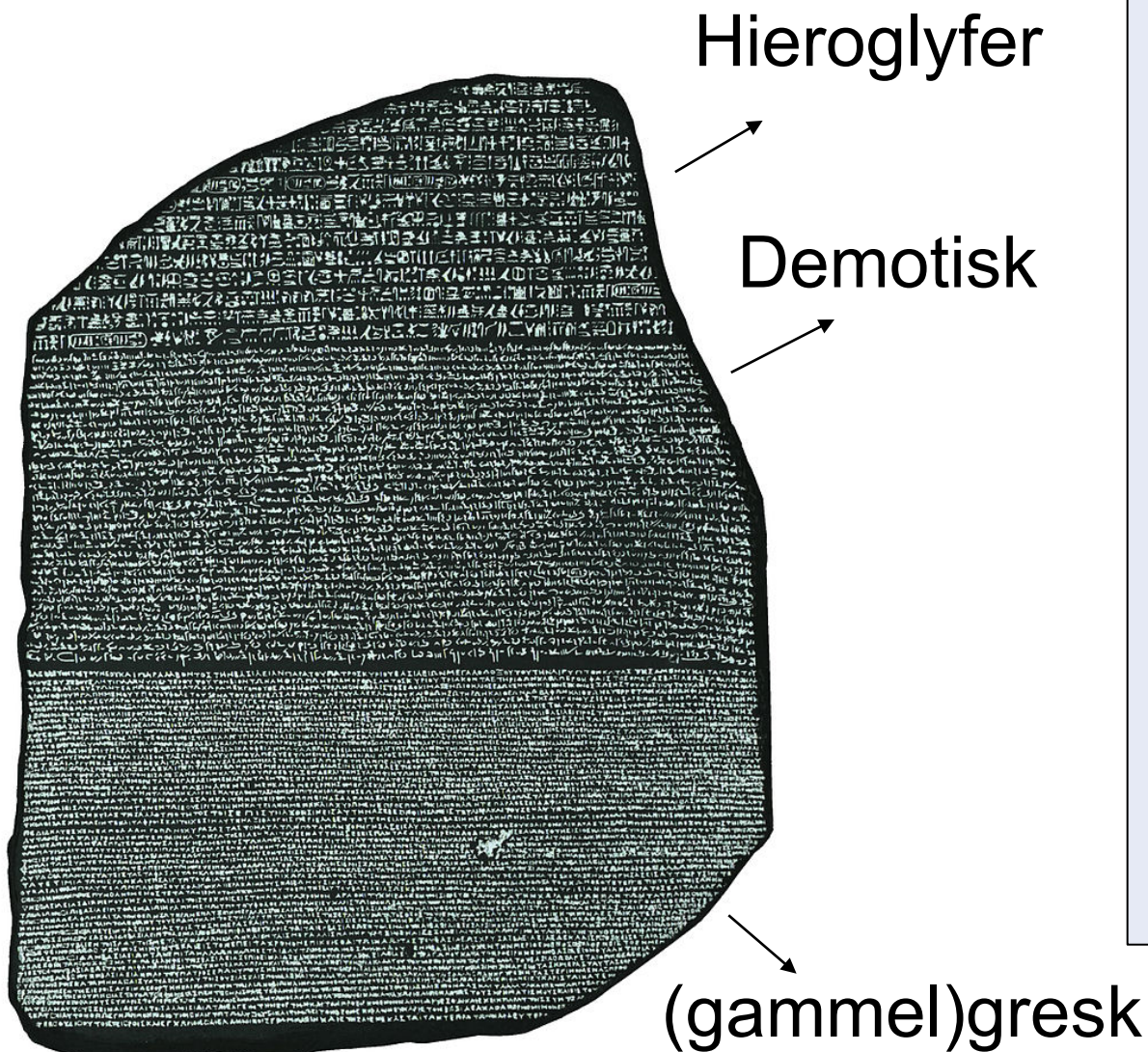


Figure 24.2 The complex overlap between English *leg*, *foot*, etc, and various French translations like *patte* discussed by Hutchins and Somers (1992) .

Statistiske metoder

- ▶ Regelbaserte systemer er veldig krevende å utvikle og er ofte begrenset til bestemte typer dokumenter
- ▶ Er det mulig å automatisk lære oversettelsemønstre ut fra data? → **Ja!**
- ▶ *Parallell korpora* (eller *bitexts*) er samlinger av tekster som er tilgjengelige på (minst) to språk
- ▶ Noen eksempler:
 - Flerspråklige juridiske tekster og parlamentsforhandlinger (EU, FN, osv.)
 - Deler av wikipedia
 - Lokaliseringsfiler
 - Filmtekstinger
 - Bibelen!

Rosettasteinen



- Steinen ble funnet i 1799 under Napoléons felttog i Egypt
- Dekret fra 196 f.Kr. utstedt av Ptolemaios.
- Nøkkelen til dechiffreringen av hieroglyfene
- Tidlig eksempel av et **parallelt korpus**, hvor samme innhold er gjengitt i flere språk

Statistiske metoder

Utvikling av en statistisk maskinoversettelsesmodell:

1. Vi først samler et parallell korpus for språkparet
2. Deretter beregner vi "word alignments" for dette korpuset

	À	I'	évidence	,	son	mari	était	un	gros	fumeur
Tydeligvis										
var										
mannen										
hennes										
en										
storrøyker										

Statistiske metoder

Utvikling av en statistisk maskinoversettelsemodell:

1. Vi først samler et parallell korpus for språkparet
2. Deretter beregner vi "word alignments" for dette korpuset
3. Og utfra disse "word alignments" ekstraherer man såkalte (probabilistiske) frasetabeller:

heavy	tung	0.95
heavy metal	heavy metal	0.61
heavy metal	tungmetal	0.34
smoker	røyker	0.99
heavy smoker	storrøyker	0.99
...	...	...

Statistisk modell

Vi leter etter den «mest sannsynnlige» oversettelse T i målspråket gitt en tekst S i kildespråket:

$$\begin{aligned}\hat{T} &= \arg \max_T P(T|S) \\ &= \arg \max_T \frac{P(S|T)P(T)}{P(S)} \\ &= \arg \max_T P(S|T)P(T)\end{aligned}$$

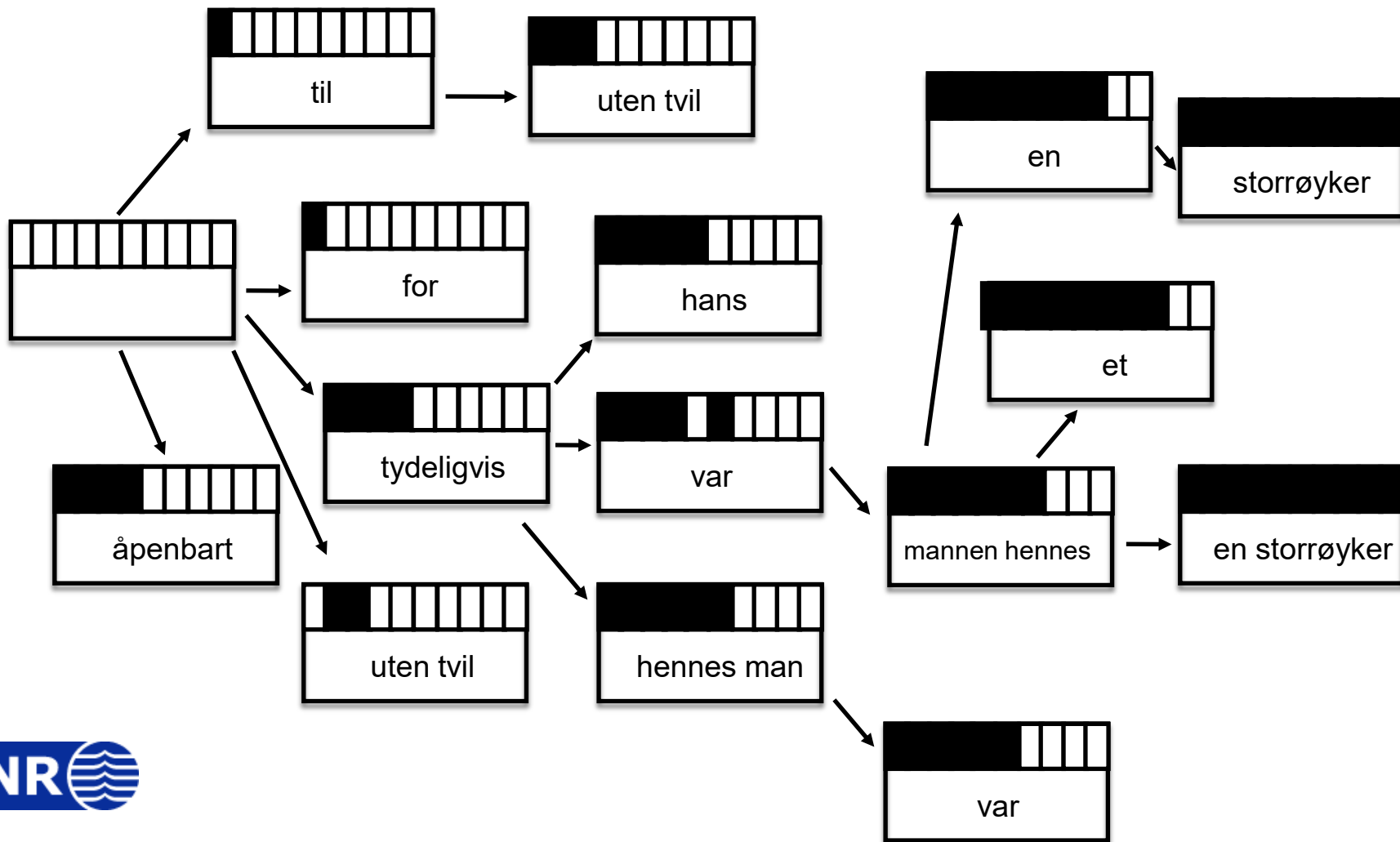
Oversettelsesmodell
(basert på frasetabellen)
Utrykker «*adequacy*» av T som
en oversettelse av S

Språkmodell
Utrykker «*fluency*» av
 T i målspråket

(Dette er den «klassiske» modellen for statistisk maskinoversettelse. Den har en del svakheter og brukes ikke lenger i praksis, men var et godt utgangspunkt på vei til mer avanserte modeller)

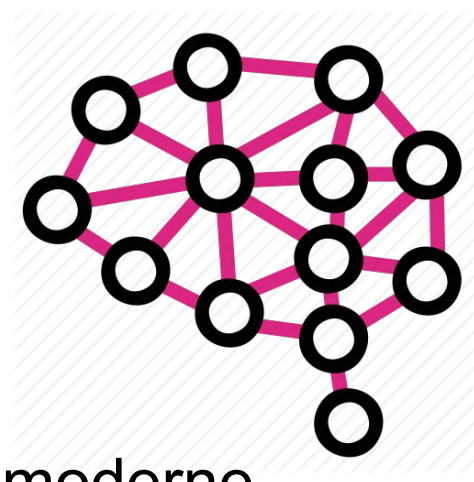
Dekoding (beam search)

À l'évidence, son mari était un gros fumeur



Nevrale modeller

- ▶ 2012:- "Deep learning revolution"
 - Stor påvirkning på maskinoversettelse!
- ▶ Store, nevrale språkmodeller i bruk i alle moderne systemer for maskinoversettelser
- ▶ Oversettelser av høyere kvalitet
 - Bedre "flyt"
 - Bedre bruk av konteksten
- ▶ "end-to-end" optimering av et single, gigantisk nevral nett
- ▶ Kan trene ett system for mange språkpar



Nevrale modeller

- ▶ Vi har allerede sett noen "sekvensmodeller" (bl.a. HMMs) tidligere i kurset
- ▶ I MT bruker man "sequence-to-sequence" modeller, hvor input- og outputsekvenser kan ha ulike lengder
 - "She likes to sing" → "Sie singt gerne"
- ▶ Sekvensene kan være lister med ord, tegn, eller orddel
 - Avhengig av *tokeniseringsmodellen*
- ▶ Språkmodeller genererer oversettelsen token etter token inntil de produserer en spesiell [END] token

Byte-pair encoding

= Enkel måte å bygge vokabularet for en språkmodell

Prosedyre:

1. Vi starter med minimale enheter (bokstaver eller *bytes*)
2. Vi finner enhetene som ofte forekommer etter hverandre i korpuset, og definerer paret som en ny token
3. Vi fortsetter inntil vi kommer til den ønskede størrelsen for vokabularet

Nevrale modeller

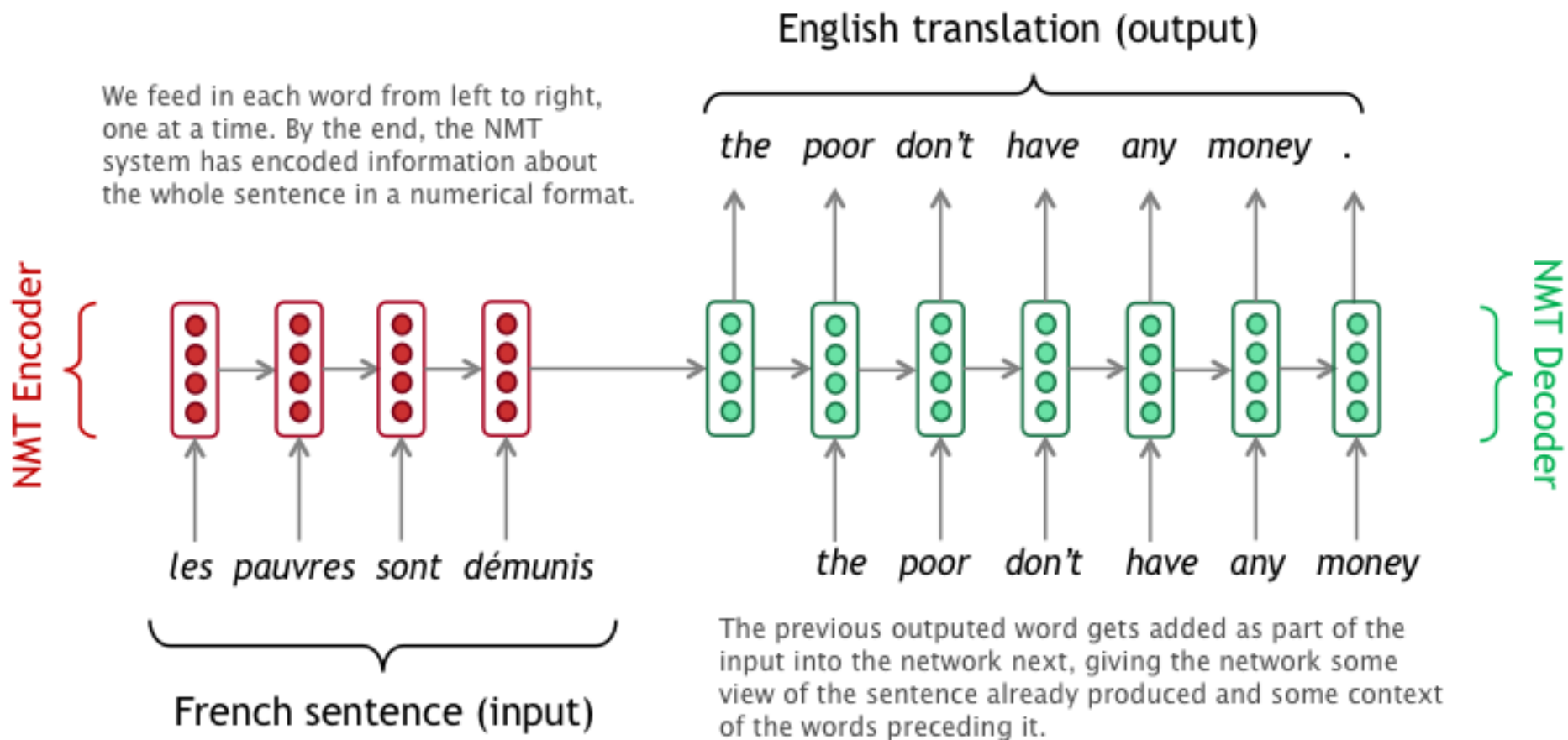
- ▶ **Fase 1** (*encoding*):
nettverket bygger opp
vektorrepresentasjoner av
inputsetningen
- ▶ **Fase 2** (*decoding*):
nettverk genererer
outputsetningen basert på
inputvektorene og det
som er generert så langt



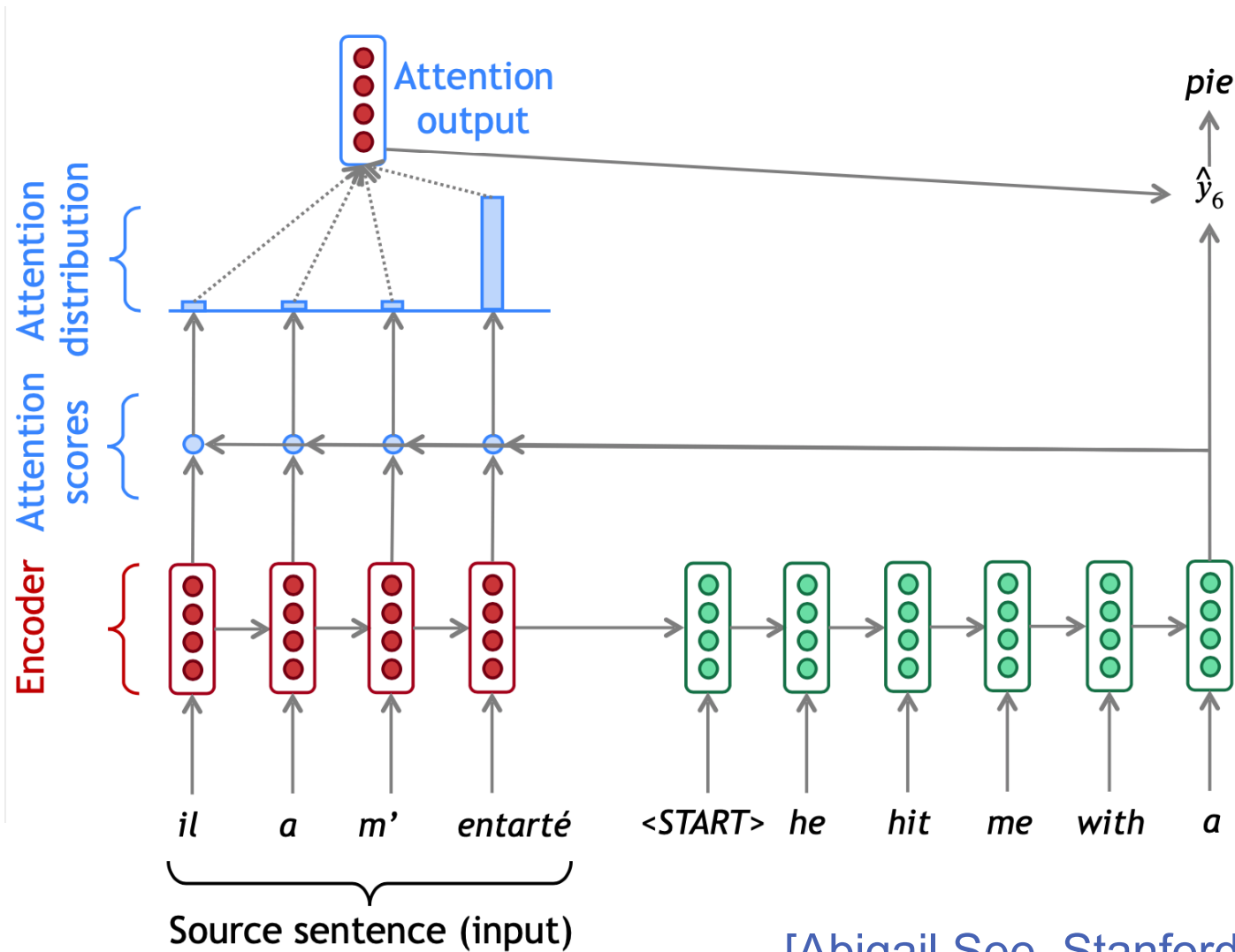
2 strategier:

- Ulike modeller for
fase 1 og 2
(*encoder-decoder*
eller *seq2seq*)
- Én språkmodell for
de to fasene
(*decoding-only*)

Seq2seq modell (eldre variant)



Seq2seq modell med "attention"



[Abigail See, Stanford CS224]

Trening

- ▶ Språkmodellen trenes (eller finjusteres) ved å gi eksempler av (*input, output*) oversettelser, og få prediksjonene fra modellen til å komme stadig nærmere “fasiten”
- ▶ *Cross-entropy loss* brukes ofte til å måle avstanden mellom modellprediksjonen og fasiten i treningsdata
- ▶ Optimering av modelvektene gjøres gjennom varianter av *gradient descent*
- ▶ Tungt regneprosess! Krever GPU-er.

Transformers

[Transformer: A Novel Neural Network
Architecture for Language Understanding:
<https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html>]

Tilnærminger

Typer Modell	Fordeler	Ulemper
Regler	• Ingen behov for treningsdata • Lett å kontrollere • God forklarbarhet	• Tid- og ressurskrevende • Kan ikke dekke all språklig variasjon • Mekanisk oversettelser, dårlig flyt
Statistiske	• Ekstrahere automatisk oversettelsemønstre fra treningsdata	• Pølsefabrikk: avhengig av en rekke pre- og post-prosesseringsmoduler
Nevrale	• State-of-the-art • Enkel, elegant modell	• Krever store regneressurser (GPU-er) • Hallusineringer

Er maskinoversettelse «løst»?

- Langt ifra! Flere uløste problemer gjenstår:
 - Modeller for språk (eller språkpar) med få språkressurser
 - Domenetilpasning når modellen benyttes på andre typer tekster enn de som de ble brukt til trening
 - **Sammenheng** i oversettelser av lengre tekster:

Source	Human Translation	Machine Translation
La película narra la historia de [un joven parisiense] _{c1} que marcha a Rumanía en busca de [una cantante zíngara] _{c2} , ya que [su] _{c1} fallecido padre escuchaba siempre [sus] _{c2} canciones.	The film tells the story of [a young Parisian] _{c1} who goes to Romania in search of [a gypsy singer] _{c2} , as [his] _{c1} deceased father use to listen to [her] _{c2} songs.	The film tells the story of [a young Parisian] _{c1} who goes to Romania in search of [a gypsy singer] _{c2} , as [his] _{c2} deceased father always listened to [his] _{c2} songs.
Pudiera considerarse un viaje fallido, porque [Ø] _{c1} no encuentra [su] _{c1} objetivo, pero el azar [le] _{c1} conduce a una pequeña comunidad...	It could be considered a failed journey, because [he] _{c1} does not find [his] _{c1} objective, but the fate leads [him] _{c1} to a small community...	It could be considered [a failed trip] _{c3} , because [it] _{c3} does not find [its] _{c3} objective, but the chance leads Ø to a small community...

[Lesly Miculicich Werlen (2017), "Towards Document-Level Neural Machine Translation"]

Plan for i dag

- ▶ Maskinoversettelsens historie
- ▶ Hvorfor er det vanskelig å oversette?
- ▶ Tilnærminger
 - Regelbaserte systemer
 - Statistiske metoder
 - Nevrale modeller
- ▶ **Evaluering**

Evaluering

- ▶ Evaluering av MT-systemer er et vanskelig problem!
 - Hva er en "god" oversettelse, egentlig?
 - Flere alternative oversettelser er ofte gyldige
- ▶ En god evalueringsmetode er å be menneskelig eksperter (oversetter) vurdere oversettelseskvaliteten
 - Nøkkelfaktorer: **adequacy** og **fluency**
- ▶ Men slike manuelle evalueringer er dyre og tidskrevende (og må gjentas hver gang systemet endres)

Evaluering

- ▶ Alternativ: bestemme oversettelseskvaliteten ut fra dens avstand til en eller flere «fasiter»
 - skrevet av menneskelige eksperter
- ▶ Det er ikke nok å bare se hvor mange ord er korrekt: vi må også ta hensyn til *ordstillingen*!
- ▶ N-gram overlapp med menneskelige oversettelser

Source: À l'évidence, son mari était un gros fumeur

Reference: Tydeligvis var mannen hennes en storryker

Output 1: Mannen hennes var åpenbart en storryker

Output 2: Hans ektemann var en tung røyker

Unigrams
Bigrams

5 2

2 0

BLEU

- ▶ Det mest kjente evalueringsmetode er **BLEU**, som beregnes ved å se på overlapp mellom N-grams fra fasiten(e) og oversettelsene fra systemet.
- ▶ Mer presist ekstraherer vi for hver setning alle N-grams (med N fra 1 til 4) fra både systemet og fasiten, og beregner hva som er "precision" for $i \in 1,2,3,4$:

$$precision_i = \frac{\text{Antall } i\text{-grams som forekommer i både system og fasit}}{\text{Antall } i\text{-grams i setningene fra systemet}}$$

(beregnes setning for setning!)

BLEU

Deretter beregnes BLEU slikt:

$$BLEU = brevity_penalty * \left(\prod_{i=1}^4 precision_i \right)^{\frac{1}{4}}$$

hvor “*brevity penalty*” brukes til å “straffe” modeller som produserer for korte oversettelser:

$$brevity_penalty = \min\left(1, \frac{\text{Antall ord i systemets setninger}}{\text{Antall ord i fasitens setninger}}\right)$$

BLEU

SYSTEM A: Israeli officials responsibility of airport safety
2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

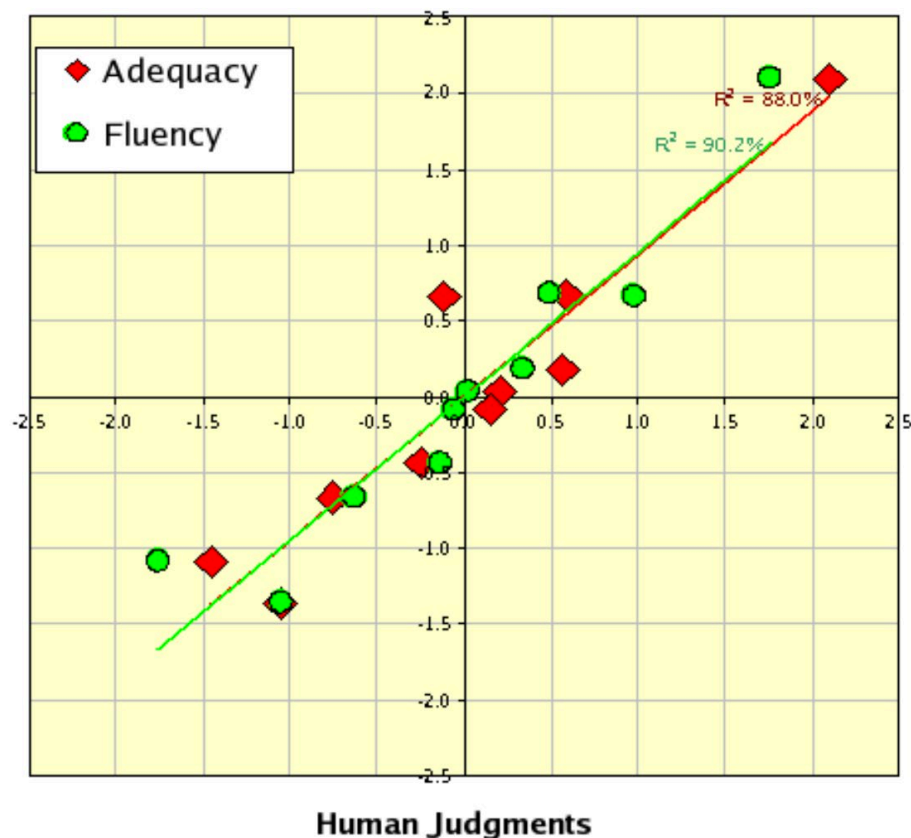
SYSTEM B: airport security Israeli officials are responsible
2-GRAM MATCH 4-GRAM MATCH

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)		
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

BLEU

- ▶ Hovedfordelen med BLEU: kan beregnes automatisk!
- ▶ Ikke perfekt, men korrelert med oversettelses kvalitet

NIST Score (variant of BLEU)



- ▶ Men BLEU har også viktige svakheter:
 - Ignorerer semantikken i setningen ("ikke" er bare ett ord, men et viktig ord i en setning!)
 - Ignorerer den globale koherensen / strukturen i setningen

Oppsummering

- ▶ Maskinoversettelse er en av de viktigste (og eldste) anvendelsene av språkteknologi
- ▶ Å oversette er vanskelig (også for oss!) på grunn av:
 - Strukturelle forskjeller mellom språk
 - Flertydigheter (som må løses ut fra konteksten)
- ▶ Ulike tilnærminger:
 - Regelbaserte systemer
 - Datadrevne systemer (statistiske og nevrale)
- ▶ Evaluering av MT-systemer er langt fra opplagt!