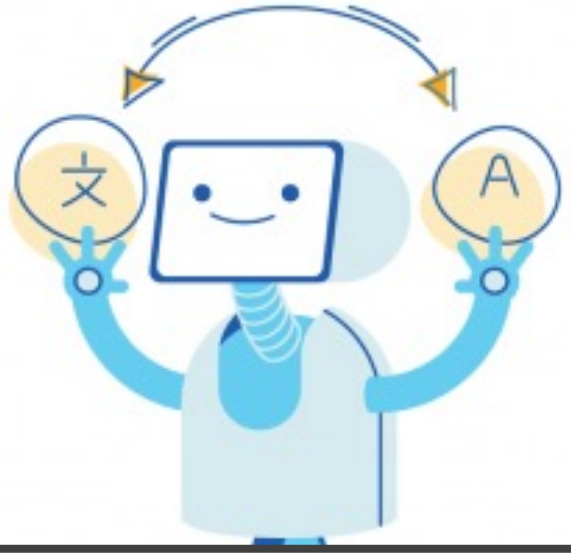


MASKINOVERSETTELSE

Mål: Automatisk oversette tekst eller tale fra et kildespråk til et målspråk

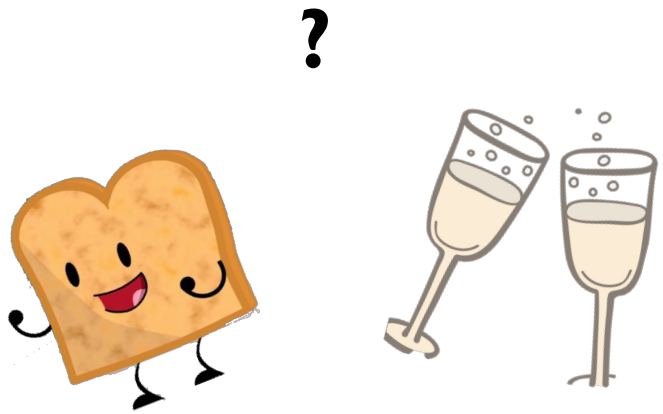
UTFORDRINGER

Strukturelle forskjeller

- morfologi (f.eks. ulike sammensetninger av ord)
- syntaks (ordstilling)

Leksikalske forskjeller

- «toast» → «å skåle» eller «å riste brød»?



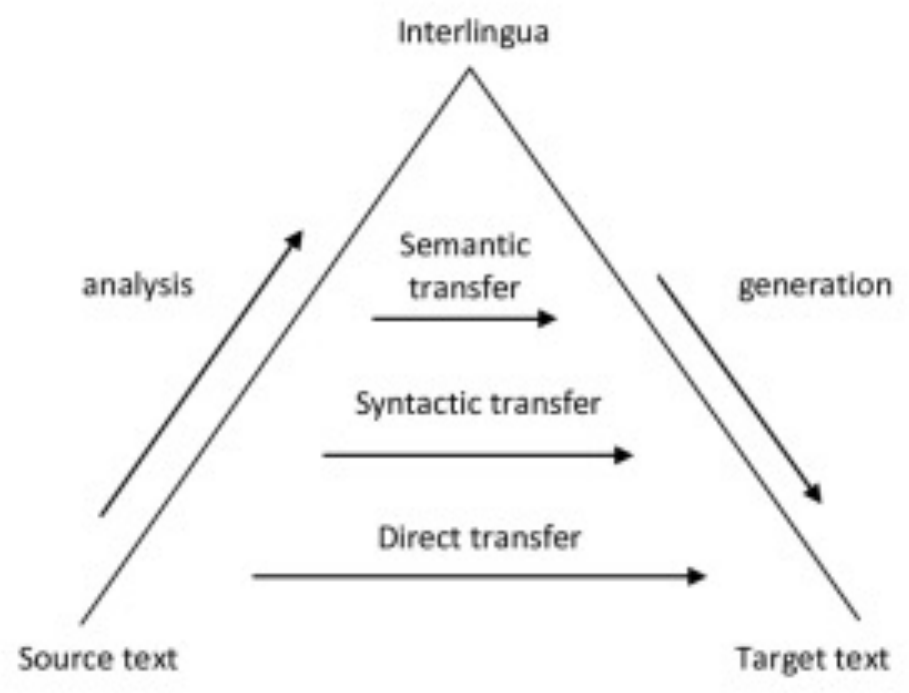
Idiomatiske uttrykk

- «å gå rundt grøten», «ugler i mosen»

Flertydighet

- Krever bakgrunnskunnskap (mennesker)
- «They made a toast with their glasses»
  - Man kan ikke putte glass i brødristeren (og heller ikke skåle med briller?)

VAUQUOIS-TREKANTEN



- Semantisk overføring: overføring av semantiske strukturer
- Syntaktisk overføring: Tar noe hensyn til syntaks
- Direkte overføring: enkleste form, oversetter ord for ord direkte
- Interlingua: radikal form for oversettelse, *språkuavhengig* representasjon av kildeteksten som konverteres til målspråket

ULIKE TILNÆRMINGER

REGELBASERTE SYSTEMER

Basert på tospråklige ordbøker og grammatiske regler

- Enkel å forklare og trenger ikke treningsdata
- Veldig tid- og ressurskrevende

STATISTISKE METODER

- Beregner den *mest sannsynlige* oversettelsen, f.eks. med et *parallellkorpus* (tekster som er tilgjengelige på flere språk)
- Bedre enn regelbaserte, men er avhengig av mye pre- og postprossesering

NEVRALE MODELLER

- «state of the art», sequence-to-sequence-modeller
  - Encoding: Vektorrepresentasjon av inputsetningene
  - Decoding: nettverket genererer outputsetningen basert på inputvektoren og det som er generert så langt
- Krever mye treningsdata og er vanskelig å forklare

EVALUERING: BLEU-SCORE

Hvordan beregner vi kvaliteten på en oversettelse?
Vi kan ikke bare telle antall korrekte oversatte ord. Vi må ta hensyn til ordstillingen!

BLEU

- Sammenligner overlapp av *n*-gram mellom fasiten og systemets oversettelse
- For hver setning ekstraherer vi *n*-gram (*n* er fra 1 og 4) fra systemet og fasiten, og beregner *precision* for hvert *n*-gram
- *Brevity penalty* brukes til å «straffe» modeller som produserer for korte oversettelser
- Kan beregnes automatisk, men ignorerer semantikken

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

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

$$brevity\_penalty = \min(1, \frac{Antall\ ord\ i\ systemets\ setninger}{Antall\ ord\ i\ fasistens\ setninger})$$