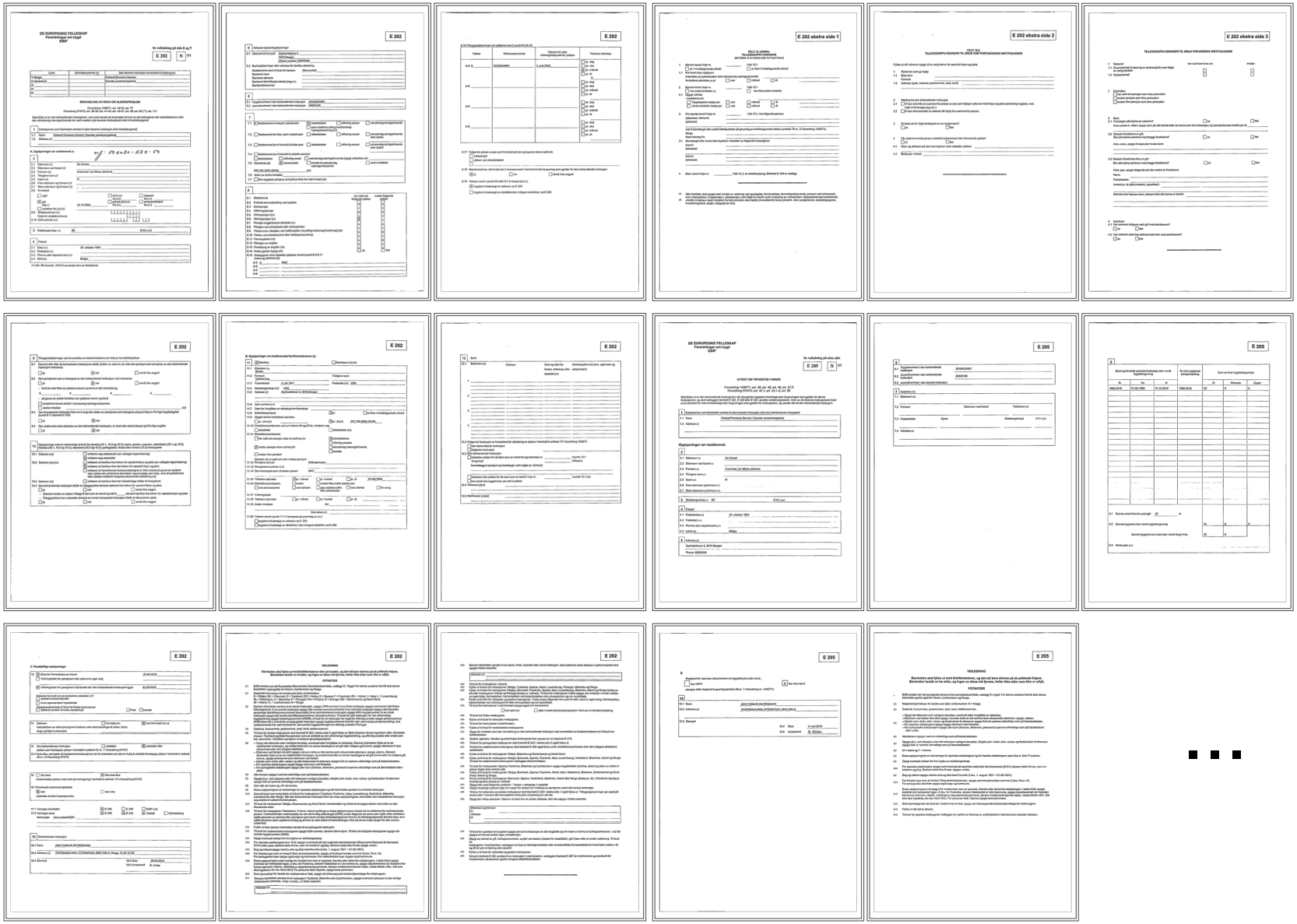
Hvordan virker eTranslation?

Kort om automatisk oversettelse som én av byggeklossene i EUs store digitaliseringssatsning CEF Digital.

Koenraad De Smedt
Universitetet i Bergen / CLARINO

ELRC-workshop, Oslo, 20. november 2018

Pensjon fra utlandet: ca. 100 sider med skjema



...

Leveattest: fortsatt på papir og ekstremt dyr

2 ONDERTEKENING BEVOEGDE INSTANTIE / SIGNATURE AUTHORISED OFFICE

Verklaart hierbij dat de hierboven ingevulde gegevens juist zijn en dat genoemde persoon

The undersigned (name and capacity) hereby declare that the information given above is correct and that the person in question / le/la soussigné(e) (nom et fonction) déclare par la présente que les renseignements ci- avant sont exacts et que la personne concerné(e) / el que suscribe (nombre y función) certifica que los datos de mas arriba son correctos y que el interesando / Der Unterzeichnete (Name und Dienstbezeichnung) bescheinigt, dass die eingetragenen Angaben korrekt sind und dass die genannte Person

☒ heden in leven is
is alive today / est en vie à cette date / vive el dia de la fecha /heute am Leben ist

☐ is overleden op
has deceased on / décéde le / fecha de defuncion / verstorben ist am

Ondergetekende: (naam): SINA FJELDSTAD
Undersigned / Soussigné(e) / El que suscribe / Unterzeichnete

(functie): NOTARY PUBLIC
Function / Fonction / Función / Funktion

Datum: 13/11/2018
Date / Date / Fecha / Datum

Stempel:
Stamp / Timbre / Sello / Stempel

Plaats: BERGEN
Town / Lieu / Lugar / Ort

Handtekening: Sina Fjeldstad

signature and stamp of the authorised authority / signature et cachet de l'autorité compétente / firma y sello de la autoridad competente / Unterschrift und Stempel des zuständigen Beamten

Salgskvittering

Bergen tingrett
Tårnplassen 2, 5012 Bergen
ORGNR: 974737418
TLF: 55 69 97 00

13/11/2018 13:20
Ansatt: Kasse Bruker
Kassepunkt: Bergen Tingrett
Kassepunkt-ID: 1
For kunde: Uregistrert
Kvitteringsnr: 2411

Notarialbekreftelse	282,00
Sum	282,00
Betalingsmiddel: Bank	
Mottatt Kr	282,00
Mva 0,00%	0,00

Grunnlag	Mva	Totalt
282,00	0,00	282,00

Velkommen tilbake!

Timepris = 16 920 kr !
Engelsk som lingua franca

Mål: Sammenkobling av tjenester på tvers av land og med respekt for språklig mangfold i EØS

Bl.a. ODR

CEF Digital

Digital Service Infrastructures (DSI): digitale verktøy og infrastrukturer som hjelper offentlig og privat sektor med levering av digitale tjenester til borgere på tvers av landegrenser.

eDelivery
eID
eInvoicing
eSignature
▶ eTranslation
eArchiving
...

https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

CEF Digital
Connecting Europe

CEF | Building Blocks | Digital Infrastructures | News | Contact us

CEF Digital Home

eTranslation

Making European digital public services multilingual.

Learn about eTranslation

Everything you need to know about eTranslation. +

Use eTranslation

For those requiring translation, either on an individual basis or as an integrated capability. +

Contribute language resources

For language resource holders and collectors helping to improve the translation engine. +

Join the community

Join one or more communities or help promote the uptake of eTranslation. +

Featured

Official Website of 2018 EU Council Presidency Integrates CEF eTranslation

Quick Links

- All eTranslation services
- Monitoring dashboard
- eTranslation grants
- Media Library

Start here

- eTranslation one-pager (PDF)
- eTranslation infographic (PDF)

Latest

Over €18 million requested for CEF eDelivery, eInvoicing & eTranslation

The total available budget for the call is €10.5 million.

Connecting Europe
eTranslation

Translate documents | Translate text | My translation requests

Click here to upload

Supported formats:

From *: EN

To *: BG CS DA DE EL ES ET FI FR GA HR HU IS IT LT LV MT NB NL PL PT RO SK SL SV

Domain: Cutting edge

Output format: Same as source | TMX | XLIFF

☑ E-mail me my translation

☐ Delete after download.

Translate document

Engelsk → norsk

The document defines the scope of the programme as well as the envisaged level of funding which will take the form of procurement, calls for proposals, vouchers and other supporting actions.

The overall aim is to facilitate cross-border interaction between public administrations, businesses and citizens, to stimulate the deployment and modernisation of high-speed internet networks and to provide wireless internet connectivity in local communities thereby contributing to the development of the Digital Single Market and economic growth.

Engelsk → norsk

I dokumentet defineres omfanget av programmet og det planlagte finansieringsnivået som vil ta form av innkjøp, forslagsinnbydelser, kuponger og andre støttetiltak.

Det overordnede målet er å legge forholdene til rette for samspill mellom offentlige myndigheter, foretak og borgere for å stimulere til innføring og modernisering av høyhastighetsnett på tvers av landegrensene, og for å gi trådløst Internett-forbindelser i lokalsamfunn, og dermed bidra til utviklingen av det digitale indre marked og økonomisk vekst.

CEF eTranslation

- Tilgengelig på alle 24 offisielle EU-språk, norsk bokmål og islandsk
- Kan brukes som en frittstående tjeneste (resultatet kan lastes ned eller fås pr epost)
- Kan også integreres i webtjenester (f.eks. Klageportalen ODR)
- Basert på maskinlæring
- Språkmodeller for ulike domener (men ikke for alle språk)
- Kan oversette til flere målspråk samtidig
- Flere dokumentformater
- Trygg nettforbindelse

Norsk → engelsk

Også syvende parti endte remis, og stillingen er nå 3,5-3,5. Carlsen selv tok selvkritikk etter partiet for å spille defensivt.

– Det er ikke ideelt, men det er behagelig ikke å være i tapsfare. Trenden i kampen er ikke fantastisk for meg, men det er ingen grunn til å bli urolig, sier Magnus Carlsen til NRK.

Det var nå han skulle sette inn støtet, med to hvite partier. Det klarte han ikke, og nå er det Caruana som har fordelen av å ha hvit-overtaket. I de fem siste har han tre partier med hvit.

Norsk → engelsk

The seventh lot was, and the position is now 3,5-3,5. The Carlsen itself took a self-criticism after the lot to play it needs in it.

“It is not ideally, but it is comfortable to be at its loss risk. The trend in the fight is not fantastic for me, but there is no reason to be affected, says Magnus Carlsen to NRK.

He was now called into support with two white lots. He was not able to do so, and it is the Caruana, who has the advantage of having white on the roof. During the last five parties, he has three batches with white.

Hvordan funker det?

1. Statistikk

Probabilistisk modell for oversettelse: $P\,(s_m | s_k)$

Hvordan funker det?

1. Statistikk

Probabilistisk modell for oversettelse: $P\,(s_m | s_k)$

Hei! → *Heath! Moorland!*

Hvordan funker det?

1. Statistikk

Probabilistisk modell for oversettelse: $P(s_m | s_k)$

Hei! → Heath! Moorland! Hi!

Hvordan funker det?

1. Statistikk

Probabilistisk modell for oversettelse: $P\ (s_m | s_k)$

Hei! → Heath! Moorland! Hi!

Probabilistisk modell for målspråket: $P\ (w_n \mid w_{n-m} \ldots w_{n-2}\ w_{n-1})$

Hvordan funker det?

1. Statistikk

Probabilistisk modell for oversettelse: $P\ (s_m | s_k)$

Hei! → Heath! Moorland! Hi!

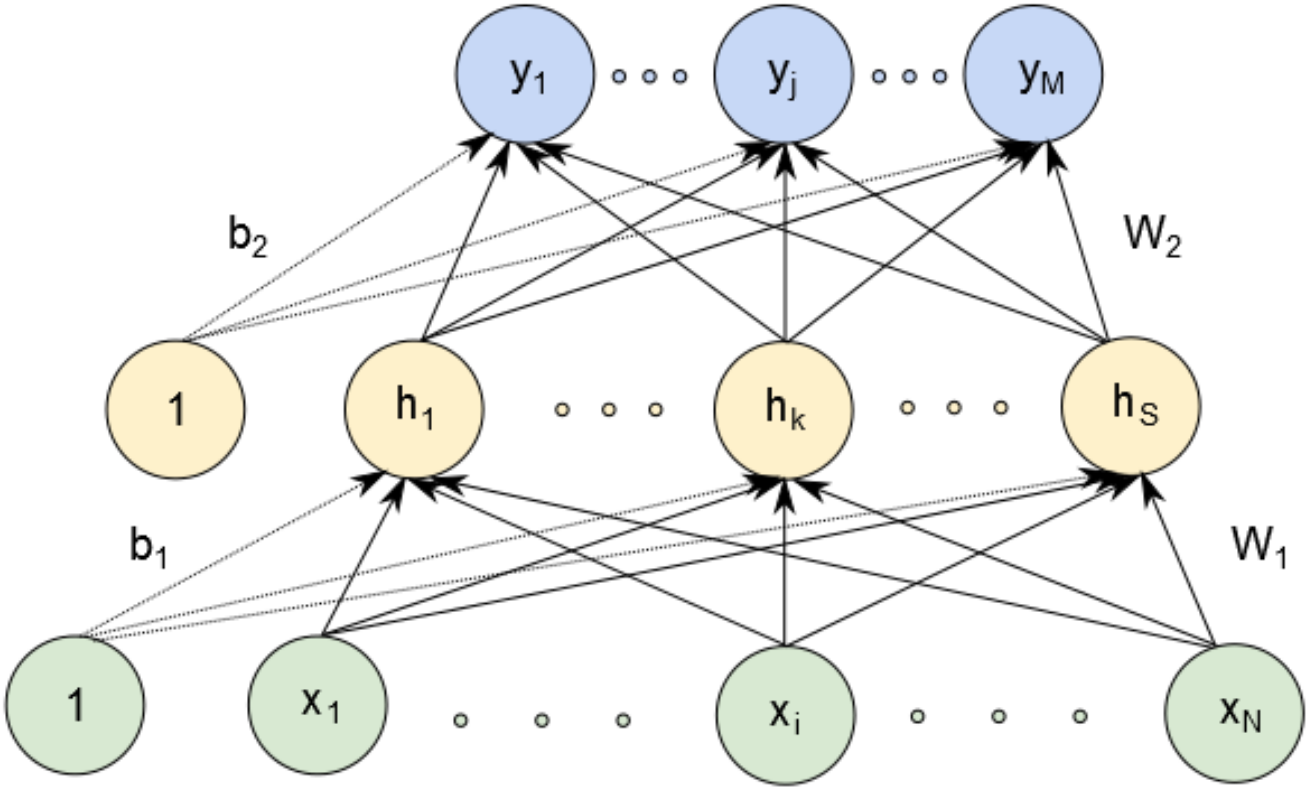
Probabilistisk modell for målspråket: $P\ (w_n \mid w_{n-i} \ldots w_{n-1}, w_{n+1} \ldots w_{n+j})$

He wanted to say hi to you

A heath is a shrubland habitat

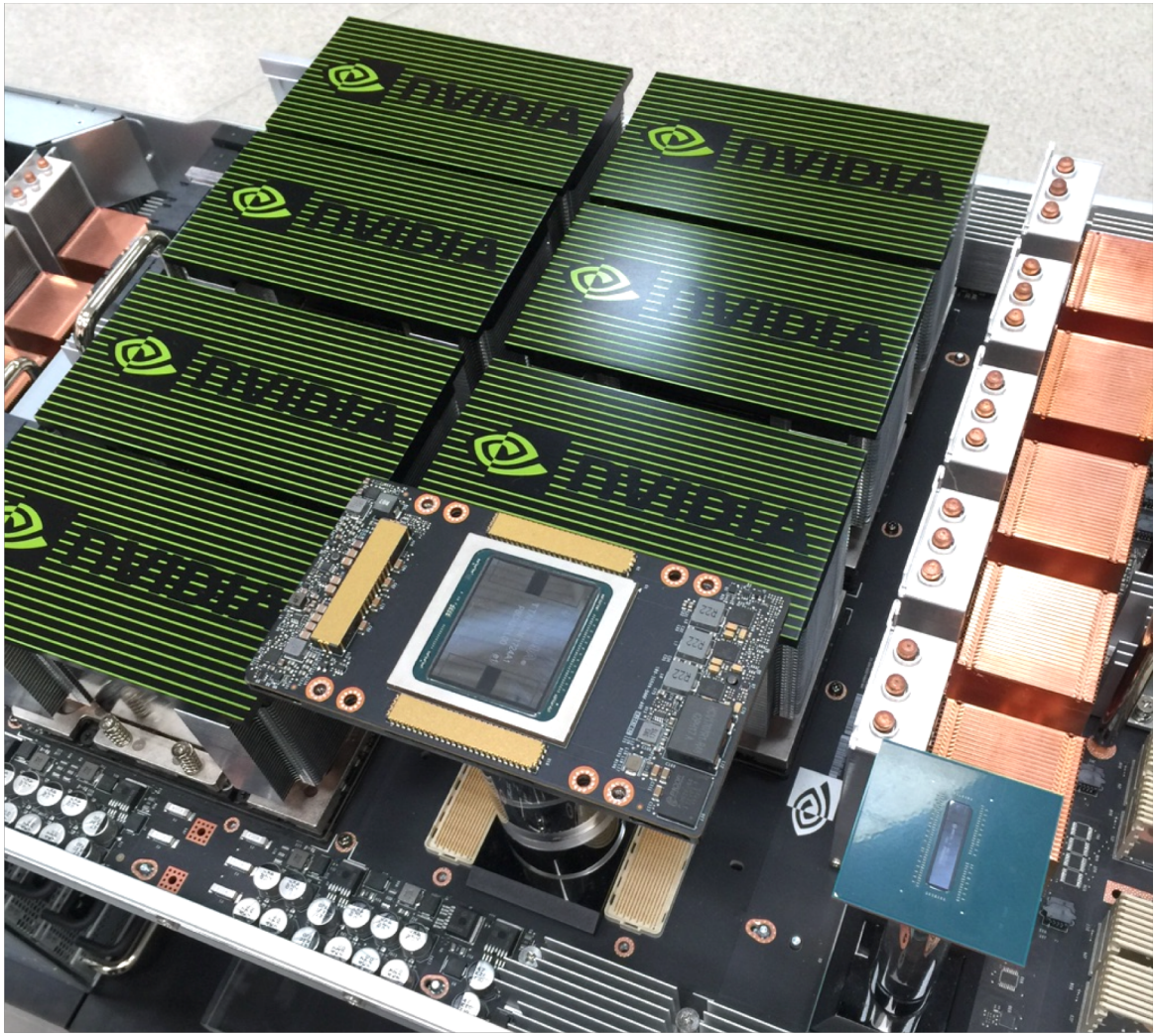
Hvordan funker det?

2. Læring av mønster i nevrale nettverk



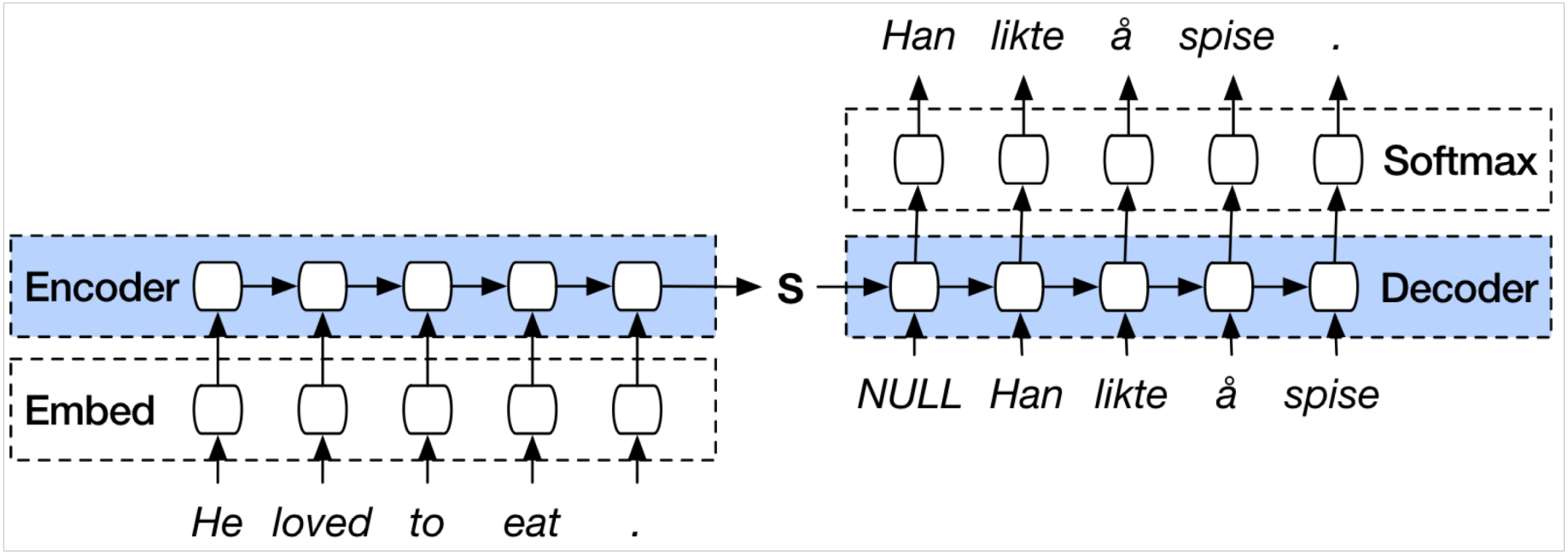
NMT=neural machine translation

Trenger spesialisert programvare og stor regnekapasitet



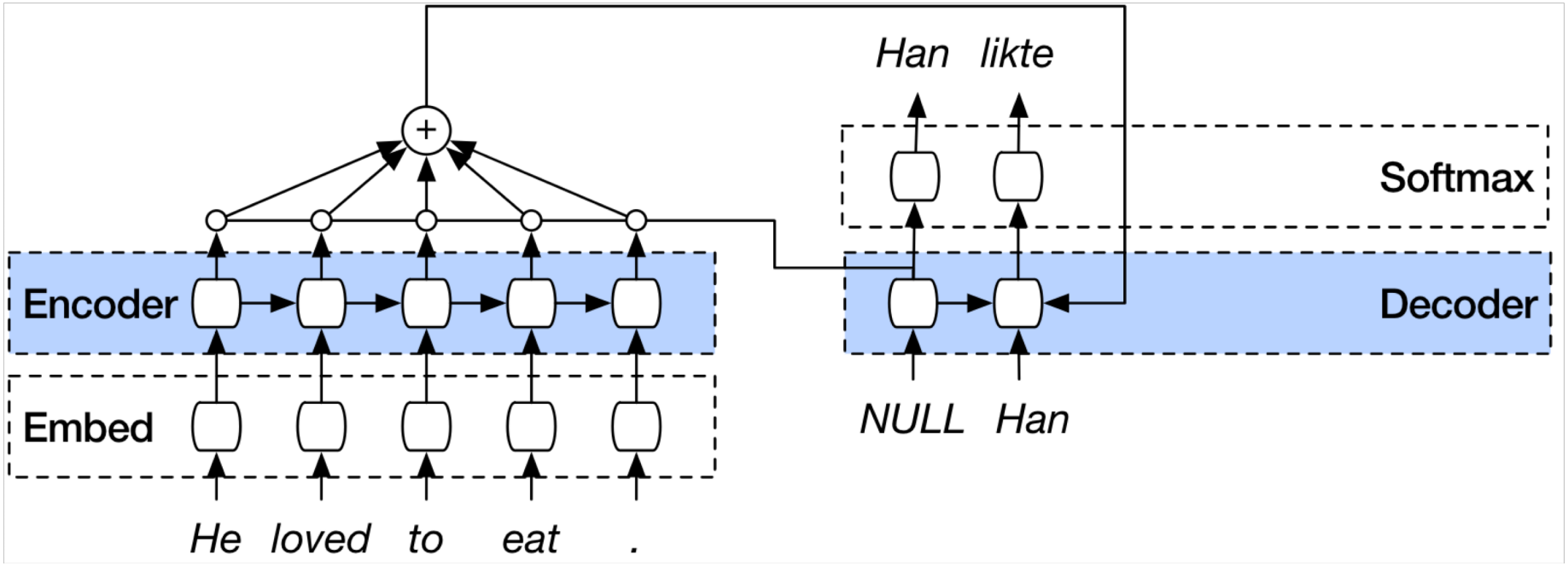
DGX-2

Rekurrente nevrale nettverk: koding og avkoding «husker» tidligere ord i konteksten

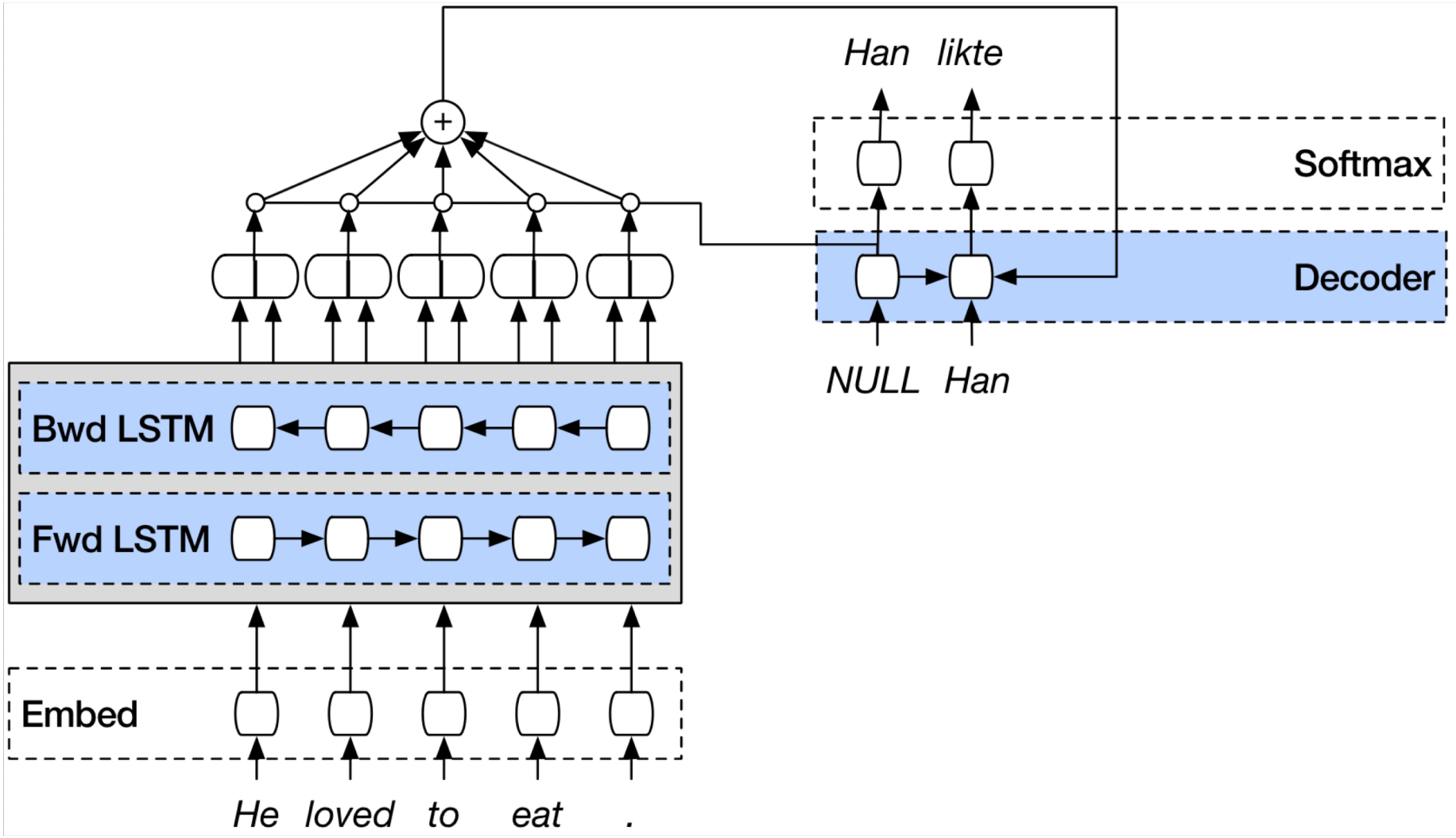


LSTM: lang korttidshukommelse

Rekurrente nevrale nettverk: avkoding «søker» tidligere ord i konteksten



Rekurrente nevrale nettverk: se til venstre og til høyre



Læring fra eksempler: språkressurser er nøkkelen

- Tospråklige treningsdata: parallelltekst, oversettelsesminner som er spesialisert i et kunnskapsdomene
- Énspråklig tekst
- I tillegg: ordlister, termbaser, navnelister, osv.

- Preprosessering (f.eks. sammensetningsanalyse) kan bidra til bedre kvalitet
- Manuell postprosessering er fremdeles billigere enn manuell oversettelse
- Evaluering kan utføres manuelt eller automatisk

Kilder for språkressurser

FoU-ressurser

- Forskningsinfrastruktur: CLARINO Bergen Repository, Tekstlaboratoriet, TROLLing, Giellatekno, Termportalen ...
- Språkbanken, Språkrådet

Rådata

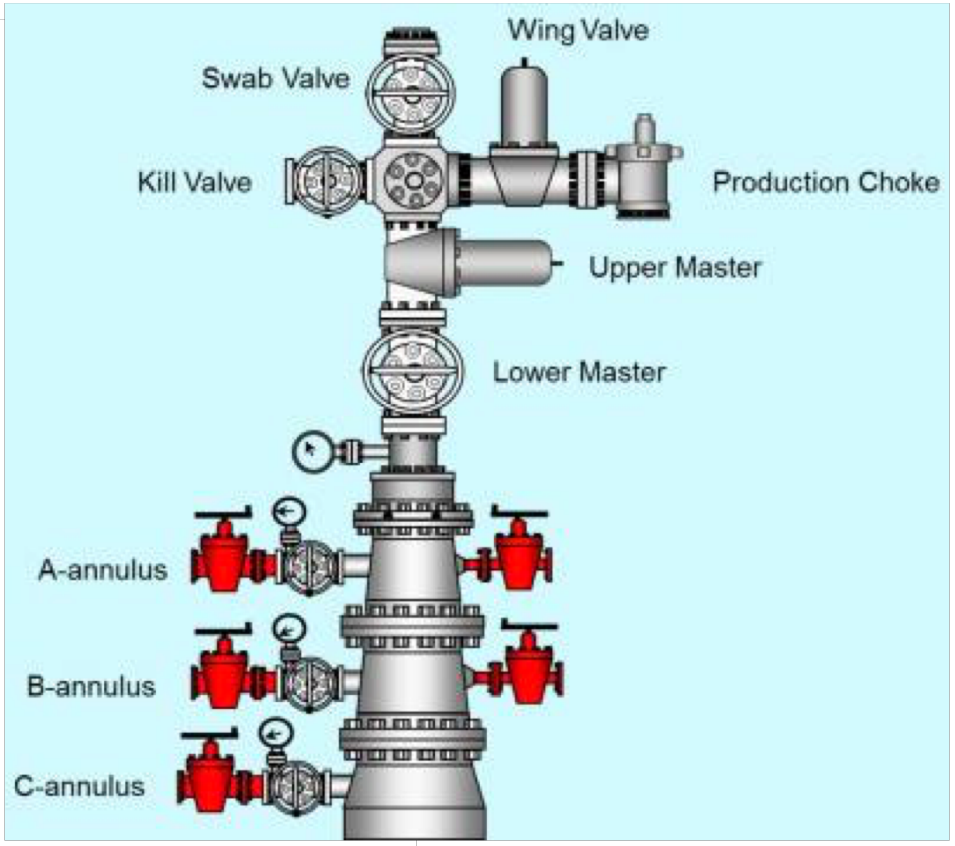
- Offentlig sektor: UD, BLD, IMDi, NAV, UDI, Veivesenet, Difi osv.
- Privat sektor

Termportalen: «Wellhead Christmas tree»

Termbaser utviklet av Norsk termbank

Domene: **BORING**

Språk:	**nob**
Hvd:	**ventiltre**
Syn:	juletre
Språk:	**nno**
Hvd:	**ventiltre**
Språk:	**eng**
Hvd:	**christmas tree**
Syn:	wellhead Christmas tree
Syn:	x-mas tree
Eksempel:	Each well is fitted with a "christmas tree" wellhead valving arrangement which is a piping arrangement with manual and remote-operated valves.
Eksempel:	The wellhead is placed on the maindeck. The "x-mas tree" on the mezzanine deck.



Konklusjon

- Maskinoversettelse er nødvendig for effektiv kommunikasjon over språkgrenser
- CEF eTranslation er EUs løsning basert på ressurser fra deltagende land
- Moderne maskinoversettelse lærer gjennom sofistikerte nevrale nettverk
- Bedre oversettelse er avhengig av språkressurser, spesielt tospråklige tekster
- Norge bør bidra med et større volum og et større mangfold av språkressurser
- Bokmål og nynorsk behandles i denne konteksten som to ulike språk og hver målform trenger nok ressurser.