

Maskinomsetjing nordsamisk til norsk - regelbasert, hybrid eller rein LLM?

Først ein konklusjon ...

Først ein konklusjon ...

- Eg er imponert over det Google har fått til med nordsamisk

Først ein konklusjon ...

- Eg er imponert over det Google har fått til med nordsamisk
- Vi skal sjå at dei har best BLEU-skår, mest flytande norsk, og at dei ofte tilbyr eit godt utgangspunkt for å redigere output til å bli ei god omsetjing

Først ein konklusjon ...

- Eg er imponert over det Google har fått til med nordsamisk
- Vi skal sjå at dei har best BLEU-skår, mest flytande norsk, og at dei ofte tilbyr eit godt utgangspunkt for å redigere output til å bli ei god omsetjing
- Dei har med nordsamisk hatt eit svært godt utgangspunkt, til å vere eit språk med så få talarar:
 - dei har digitalisert og OCR-lese dei bøkene som har vorte gjeve ut
 - dei har lasta ned dei samisk-norske og samisk-finske ordbøkene våre (flott!)
 - dei har utnytte dei nevrane MT-modellane dei har laga for andre språk

Først ein konklusjon ...

- Eg er imponert over det Google har fått til med nordsamisk
- Vi skal sjå at dei har best BLEU-skår, mest flytande norsk, og at dei ofte tilbyr eit godt utgangspunkt for å redigere output til å bli ei god omsetjing
- Dei har med nordsamisk hatt eit svært godt utgangspunkt, til å vere eit språk med så få talarar:
 - dei har digitalisert og OCR-lese dei bøkene som har vorte gjeve ut
 - dei har lasta ned dei samisk-norske og samisk-finske ordbøkene våre (flott!)
 - dei har utnytte dei nevrane MT-modellane dei har laga for andre språk
- Dette utgangspunktet har dei laga eit godt program av

Først ein konklusjon ...

- Eg er imponert over det Google har fått til med nordsamisk
- Vi skal sjå at dei har best BLEU-skår, mest flytande norsk, og at dei ofte tilbyr eit godt utgangspunkt for å redigere output til å bli ei god omsetjing
- Dei har med nordsamisk hatt eit svært godt utgangspunkt, til å vere eit språk med så få talarar:
 - dei har digitalisert og OCR-lese dei bøkene som har vorte gjeve ut
 - dei har lasta ned dei samisk-norske og samisk-finske ordbøkene våre (flott!)
 - dei har utnytte dei nevrane MT-modellane dei har laga for andre språk
- Dette utgangspunktet har dei laga eit godt program av
- Det er berre ein handfull minoritets- og urfolksspråk som er i ein slik situasjon

... så ei undring

... så ei undring

- 10-100 milliardar parametrar, 1-10 billionar ord
 - NB 200 milliardar ord (alt som er publisert i Noreg dei siste 500 åra)
 - Nordsamisk 40 millionar ord, sørsamisk 2 millionar
 - -- **vi satsar på å doble!!** (men har framleis eit stykke opp)

... så ei undring

- 10-100 milliardar parametrar, 1-10 billionar ord
 - NB 200 milliardar ord (alt som er publisert i Noreg dei siste 500 åra)
 - Nordsamisk 40 millionar ord, sørsamisk 2 millionar
 - -- **vi satsar på å doble!!** (men har framleis eit stykke opp)
- Det at det likevel fungerer har nok å gjere med at språk er svært like (output for små språk blir betre når vi kastar dei i same gryta)

... så ei undring

- 10-100 milliardar parametrar, 1-10 billionar ord
 - NB 200 milliardar ord (alt som er publisert i Noreg dei siste 500 åra)
 - Nordsamisk 40 millionar ord, sørsamisk 2 millionar
 - -- **vi satsar på å doble!!** (men har framleis eit stykke opp)
- Det at det likevel fungerer har nok å gjere med at språk er svært like (output for små språk blir betre når vi kastar dei i same gryta)
- ... og det fører nok også til at output frå språk til språk blir likare enn det hadde vorte med separate modellar

Og så kan vi starte

Ulike typar maskinomsetjing

1. Omsetjing for å forstå vs. omsetjing for å produsere tekst
2. Omsetjing frå eit majoritetsspråk eller omsetjing frå eit minoritetsspråk
3. Regelbasert omsetjing vs. omsetjing med (i dag) nevrane nett
4. Ein metode (regelbasert/nevral) vs. hybrid omsetjing (regelbasert + nevral)

(1) Omsetjing for å forstå vs. omsetjing for å produsere tekst

(1) Omsetjing for å forstå vs. omsetjing for å produsere tekst

- **Omsetjing for å forstå:** Vi vil vite kva nettsida, artikkelen, replikken på Facebook betyr
 - Viktig: Omsetjinga må ikkje attgje innhaldet feil
 - Vilkår: Vi forstår ikkje kjeldespråket

(1) Omsetjing for å forstå vs. omsetjing for å produsere tekst

- **Omsetjing for å forstå:** Vi vil vite kva nettsida, artikkelen, replikken på Facebook betyr
 - Viktig: Omsetjinga må ikkje attgje innhaldet feil
 - Vilkår: Vi forstår ikkje kjeldespråket
- **Omsetjing for å produsere:** Vi vil gjerne publisere teksten vår på eit anna språk
 - Viktig: Den omsette teksten må vere så nær publisierbar at det er lettare å rette enn å skrive sjølv frå bunnen av
 - Vilkår: Vi forstår målspråket (eller vi lever fårleg)

(2) Omsetjing frå eit majoritetsspråk eller frå eit minoritetsspråk?

(2) Omsetjing frå eit majoritetsspråk eller frå eit minoritetsspråk?

- **Frå eit majoritetsspråk:** Ingen samar les VG med hjelp av Google Translate
 - Minoritetar er tospråklege, og kan i beste fall bruke omsetjing til minoritetsspråket for å:
 - få hjelp til å lage tospråklege publikasjonar (omsetjing for produksjon)
 - få hjelp til rettskriving eller tips om terminologi

(2) Omsetjing frå eit majoritetsspråk eller frå eit minoritetsspråk?

- **Frå eit majoritetsspråk:** Ingen samar les VG med hjelp av Google Translate
 - Minoritetar er tospråklege, og kan i beste fall bruke omsetjing til minoritetsspråket for å:
 - få hjelp til å lage tospråklege publikasjonar (omsetjing for produksjon)
 - få hjelp til rettskriving eller tips om terminologi
- **Frå eit minoritetsspråk:** Skal nordmenn lese Ávvir, treng dei eit maskinomsetjingsprogram.
 - Majoritetar er einspråklege, og treng omsetjing for å forstå (i dei svært få tilfella der ein interessant tekst finst berre på minoritetsspråket)

(2) Omsetjing frå eit majoritetsspråk eller frå eit minoritetsspråk?

- **Frå eit majoritetsspråk:** Ingen samar les VG med hjelp av Google Translate
 - Minoritetar er tospråklege, og kan i beste fall bruke omsetjing til minoritetsspråket for å:
 - få hjelp til å lage tospråklege publikasjonar (omsetjing for produksjon)
 - få hjelp til rettskriving eller tips om terminologi
- **Frå eit minoritetsspråk:** Skal nordmenn lese Ávvir, treng dei eit maskinomsetjingsprogram.
 - Majoritetar er einspråklege, og treng omsetjing for å forstå (i dei svært få tilfella der ein interessant tekst finst berre på minoritetsspråket)
- *Norsk* -> *samisk* for produksjon, *samisk* -> *norsk* for å forstå

(3) Regelbasert omsetjing vs. omsetjing med nevrane nett

- Den gamle vitsen om maskinomsetjing
 - Regelbasert omsetjing er som Brocas afasi, statistiske modellar er som Wernickes

Regelbasert omsetjing vs. omsetjing med nevrane nett

- Tekst: NRK Sápmi om Ordkalotten
- Evaluert etter BLEU-skår ("bilingual evaluation understudy"), eit mål på kor langt unna MT er ei referanseomsetjing

MT	BLEU-skår
Google	0.32
Hybrid Apertium - Giellatekno	0.28
Apertium - Giellatekno	0.26
Helsinki	0.25
Tartu	0.24

La oss sjå på nokre setningar (1)

Orig	Su girji «Suoládeapmi» vuodul ráhkadii Elle Márjá Eira Netflix-filmma «Suoládeapmi».
REF	Etter hennes bok «Stjålet» laget Elle Márjá Eira Netflix-filmen «Stjålet».
Apertium	Hennes boks «Tyveriet» i henhold til laget Elle Márjá Eiras Netflix filmen «Tyveri».
Apertium hybrid	Hennes bok «Tyveriet» har inspirert Elle Márjá Eiras Netflix-film «Tyveri».
Google	Basert på sin bok «The Salt» laget Elle Márjá Eira Netflix-filmen «The Salt».
Helsinki	På grunnlag av Hennes boka «Fornyning» utarbeidet Elle Márjá Eira Netflix-film «Fiskeri».
Tartu	På grunnlag av boka "Slapping" laget Elle Márjá Eira Netflix-filmen "Slapping".

La oss sjå på nokre setningar (2)

Orig	Karen Anne Buljo lea maid mielde Ordkalotten mánáid ja nuoraidprográmmas.
REF	Karen Anne Buljo er også med i Ordkalottens barne og ungdomsprogram.
Apertium	Karen Anne Buljo er også med Ordkalotten på barnas og ungdomsprogrammet.
Apertium hybrid	Karen Anne Buljo deltar også i Ordkalottens barne- og ungdomsprogram.
Google	Karen Anne Buljo er også engasjert i Ordkalottens barne- og ungdomsprogram.
Helsinki	Karen Anne Buljo deltar også i Ordkalottens barn og ungdomsprogram.
Tartu	Karen Anne Buljo deltar også i ordkalotten for barn og ungdom.

La oss sjå på nokre setningar (3)

Orig	Lávvardaga lea son girječálligalledeamis girjerádjosis.
REF	På lørdag er hun på forfatterbesøk på biblioteket.
Apertium	Lørdag er hun på forfatterbesøk i biblioteket.
Apertium hybrid	Lørdag besøker hun biblioteket for et forfattermøte.
Google	Søndag er han på skriveturné på biblioteket.
Helsinki	Det er lørdag under forfatterbesøk i biblioteket.
Tartu	Lørdag er hun på forfatterstund på biblioteket.

Regelbaserte system har grammatiske feil, nevrale system har leksikalske

sme-nob - Utfordringa er å få auge på feilen!

Regelbaserte system har grammatiske feil, nevrale system har leksikalske

sme-nob - Utfordringa er å få auge på feilen!

- *de er analysert godt f.eks. med begrepene etnosentrisme, essensialisme og kravene om likhet.*

Regelbaserte system har grammatiske feil, nevrale system har leksikalske

sme-nob - Utfordringa er å få auge på feilen!

- *de er analysert godt f.eks. med begrepene etnosentrisme, essensialisme og kravene om likhet.*
- dat analyserejuvvojit bures ee. etnosentrismma, essentialismma ja eaktivuoda (= **ekthet**) gáibidusaid doahpágiiguin.

Regelbaserte system har grammatiske feil, nevrale system har leksikalske

sme-nob - utfordringa er å få auge på feilen!

- *de er analysert godt f.eks. med begrepene etnosentrisme, essensialisme og kravene om likhet.*
- *dat analyserejuvvojit bures ee. etnosentrismma, essentialismma ja eaktivuoda (= **ekthet**) gáibidusaid doahpágiiguin.*
- *... undersøker personenes stilling i det samiske samfunnet. når det gjelder etnisitet.*

Regelbaserte system har grammatiske feil, nevrale system har leksikalske

sme-nob - Utfordringa er å få auge på feilen!

- *de er analysert godt f.eks. med begrepene etnosentrisme, essensialisme og kravene om likhet.*
- dat analyserejuvvojit bures ee. etnosentrismma, essentialismma ja eaktivuoda (= **ekthet**) gáibidusaid doahpágiiguin.
- *... undersøker personenes stilling i det samiske samfunnet. når det gjelder etnisitet.*
- ... guorahallá olbmuid sajadaga sámeservodagas sohkaduogáža (= **slektsbakgrunn**) dáfus.

Regelbaserte system har grammatiske feil, nevrale system har leksikalske

sme-nob - utfordringa er å få auge på feilen!

- *de er analysert godt f.eks. med begrepene etnosentrisme, essensialisme og kravene om likhet.*
- *dat analyserejuvvojit bures ee. etnosentrismma, essentialismma ja eaktivuoda (= **ekthet**) gáibidusaid doahpágiiguin.*
- *... undersøker personenes stilling i det samiske samfunnet. når det gjelder etnisitet.*
- *... guorahallá olbmuid sajadaga sámeservodagas sohkaduogáža (= **slektsbakgrunn**) dáfus.*
- *det ser ut til at de er sterkt engasjert i språksamfunnet*

Regelbaserte system har grammatiske feil, nevrale system har leksikalske

sme-nob - Utfordringa er å få auge på feilen!

- *de er analysert godt f.eks. med begrepene etnosentrisme, essensialisme og kravene om likhet.*
- dat analyserejuvvojit bures ee. etnosentrismma, essentialismma ja eaktivuoda (= **ekthet**) gáibidusaid doahpágiiguin.
- *... undersøker personenes stilling i det samiske samfunnet. når det gjelder etnisitet.*
- ... guorahallá olbmuid sajadaga sámeservodagas sohkaduogáža (= **slektsbakgrunn**) dáfus.
- *det ser ut til at de er sterkt engasjert i språksamfunnet*
- čájehuvvo (= **det blir vist**), ahte sii leat čátnasan (= **knytta til**) giellaservodahkii

Pragmatiske faktorar

Orig	Sámi mánáidgárdeoahpaheaddjeoahpus leat vejolašvuodát oažžut liigestipeanddaid. Oahpahus lea sámegillii. Fágasisdoallu lea čohkkejuvvon fáttáid mielde.
REF	På den samiske førskolelærerutdanning er det muligheter for å få ekstrastipend. Undervisningen er på samisk. Faginnholdet er samlet etter tema.
Apertium	På samisk førskolelærerutdanning er det muligheter for å få ekstrastipendene. Undervisning er på samisk. Faginnholdet er blitt samlet etter temaene.
Apertium hybrid	På samisk førskolelærerutdanning er det muligheter for å få ekstrastipend. Undervisningen er på samisk. Faginnholdet er samlet etter temaer.
Google	Det er muligheter for å få tilleggsstipend i samisk barnehagelærerutdanning. Undervisningen foregår på samisk. Faginnholdet er organisert etter emne.
Helsinki	Den samiske barnehagelærerutdanningen har et potensial for tilleggsstipend. Opplæringen er på samisk. Fagholdet er samlet inn i temaer.
Tartu	Den samiske barnehagelærerutdanningen har mulighet for ekstra stipend. Undervisningen foregår på samisk. Faginnholdet er tematisert.

MT som hjelp i terminologisk arbeid (1)

Mål: Omsetje norske termar til samisk

1. Utgangspunkt: Liste over norske termar, manuelt omsett til finsk
 2. MT på 5 ulike måtar til samisk
 3. Deretter manuell omsetjing til samisk, MT som støttefunksjon
- Evaluering:
 - Kor stort er samsvaret mellom MT og den manuelle omsetjinga til samisk?

MT	samsvar m/ manuell omsetjing	samsvar med manuell omsetjing, i %
Apertium fin -> sme	171	25.60 %
Tartu fin -> sme	83	12.43 %
Tartu nob -> sme	69	10.33 %
Google fin -> sme	67	10.03 %
Google nob -> sme	59	8.83 %
alle termpar	668	

MT som hjelp i terminologisk arbeid (2)

ChatGPT-4: «Gi 30 dyrenavn på nordsamisk.»

Fasit: 2/30 rett, 23/30 = samiske ordformer, alle 30 svara *ser samiske ut* (det er vanskeleg å rette feil gjort av nevrane nett)

Merk at ChatGPT kan langt fleire dyre- og fiskenamn(!) på norsk, nesten 30

ChatGPT bruker også finsk til hjelp (riekko = rype, lohi = laks (ikkje fugl!))

- | | | |
|----------------------------|-----------------------|--|
| + 1. guovža - bjørn | - 11. guossa - gjedde | - 21. čavčča - fjellrype |
| - 2. rieban - reinsdyr | - 12. njálla - snømus | - 22. sárgu - sjørret |
| - 3. *gáffa - ravn | - 13. *máhkál - laks | - 23. mánnu - måne (brukt poetisk om ugler) |
| - 4. stuorra - ørn | - 14. *gárgi - gås | - 24. dálvi - vinter (brukt poetisk om snørelaterte dyr som snømus) |
| + 5. boazu - rein | - 15. muorji - skjære | - 25. *bieggolohi - vind (brukt poetisk om fugler som flyr med vinden) |
| - 6. čuoika - frosk | - 16. *orbma - elg | - 26. ruossa - rød (ofte brukt om rødreven) |
| - 7. fálli - hare | - 17. suolu - grouse | - 27. skáidi - grense (brukt om dyr som lever på grensen mellom ulike habitater) |
| - 8. gálgo - ulv | - 18. *juovla - ugle | - 28. čearru - tåke (brukt poetisk om dyr som er sjeldne eller mystiske) |
| - 9. ráhkis - rev | - 19. *riekko - rype | - 29. áibmu - ånd (brukt om dyr som er spesielt æret eller fryktet) |
| - 10. doalli - bever | - 20. *vuolgga - måke | - 30. leavga - laks (ung laks) |

(4) Hybrid omsetjing vs. homogene løysingar

(4) Hybrid omsetjing vs. homogene løysingar

Etter mitt syn sameinar hybride løysingar det beste av to verder for maskinomsetjing der målet er å forstå teksten, i alle fall for språk med så lite tekstmateriale som dei samiske språka

(4) Hybrid omsetjing vs. homogene løysingar

Etter mitt syn sameinar hybride løysingar det beste av to verder for maskinomsetjing der målet er å forstå teksten, i alle fall for språk med så lite tekstmateriale som dei samiske språka

Døme:

Apertium (regelbasert: rett innhald, dårleg pragmatikk

Karen Anne Buljo er også med Ordkalotten på barnas og ungdomsprogrammet.

(4) Hybrid omsetjing vs. homogene løysingar

Etter mitt syn sameinar hybride løysingar det beste av to verder for maskinomsetjing der målet er å forstå teksten, i alle fall for språk med så lite tekstmateriale som dei samiske språka

Døme:

Apertium (regelbasert: rett innhald, dårleg pragmatikk

Karen Anne Buljo er også med Ordkalotten på barnas og ungdomsprogrammet.

Apertium hybrid (ChatGPT-4 har glatta ut norsk output):

Karen Anne Buljo deltar også i Ordkalottens barne- og ungdomsprogram.

(4) Hybrid omsetjing vs. homogene løysingar

Etter mitt syn sameinar hybride løysingar det beste av to verder for maskinomsetjing der målet er å forstå teksten, i alle fall for språk med så lite tekstmateriale som dei samiske språka

Døme:

Apertium (regelbasert: rett innhald, dårleg pragmatikk

Karen Anne Buljo er også med Ordkalotten på barnas og ungdomsprogrammet.

Apertium hybrid (ChatGPT-4 har glatta ut norsk output):

Karen Anne Buljo deltar også i Ordkalottens barne- og ungdomsprogram.

Der målet er å produsere tekst kan nevrane nett vere betre (gje betre idiomatisk språk) enn hybride løysingar, så lenge berre nokon går gjennom og ser at innhaldet er det det skal vere

Konklusjon

Konklusjon

- Nevrale nett lagar omsetjingar med godt språk

Konklusjon

- Nevrale nett lagar omsetjingar med godt språk
- Med språk med så lite tekst som nordsamisk (og andre samiske språk!) er det likevel (framleis?) risikabelt å stole på at dei attgjev innhaldet i originalen på ein god måte

Konklusjon

- Nevrale nett lagar omsetjingar med godt språk
- Med språk med så lite tekst som nordsamisk (og andre samiske språk!) er det likevel (framleis?) risikabelt å stole på at dei attgjev innhaldet i originalen på ein god måte
- For minoritetsspråk (og for nærskyldde språkpar, som bokmål-nynorsk) ser det ut til at hybride modellar er dei mest robuste

Konklusjon

- Nevrale nett lagar omsetjingar med godt språk
- Med språk med så lite tekst som nordsamisk (og andre samiske språk!) er det likevel (framleis?) risikabelt å stole på at dei attgjev innhaldet i originalen på ein god måte
- For minoritetsspråk (og for nærskyldde språkpar, som bokmål-nynorsk) ser det ut til at hybride modellar er dei mest robuste
- Utfordringa med nevrane nett for minoritetsspråk er å finne ut kva dei er flinke til og korleis vi skal handsame svakheitene dei har på best mogleg måte