Íslenska aðferðin:

Hvernig stuðla má að fjölbreytni tungumála og menningar á sviði gervigreindar

Hugmyndir og forsendur fyrir **verkefni**
og **handbók** um stuðning við minni tungumál
og menningarsvæði á sviði gervigreindar,
með hliðsjón af samstarfi Íslands við OpenAI

September 2024

ALMANNARÓMUR

Stjórnarráð Íslands
Menningar- og viðskiptaráðuneytið

MIÐEIND

Menningar- og viðskiptaráðuneytið
Hafnarstræti 5 — 101 Reykjavík
+354 545 9800 | mvf@mvf.is
Í samstarfi við:
Almannaróm
Miðeind

Reykjavík, september 2024

Fyrir frekari upplýsingar, vinsamlegast hafið samband við:
mvf@mvf.is,
almannaromur@almannaromur.is
mideind@mideind.is

https://mvf.is
https://almannaromur.is
https://mideind.is

1. Markmiðið: að stuðla að fjölbreyttri gervigreind

Tækifæri og möguleikar sem felast í nýrri gervigreindartækni eru gríðarlegir, bæði til að auðga daglegt líf fólks og til að auka framleiðni og verðmætasköpun í samfélagsinu. Stór hluti tækninnar, sérstaklega stór mállíkön, hefur þó hingað til **miðast að miklu leyti við ensku** og nokkur stærstu tungumál heims. Þetta skýrist af því gríðarlega magni þjálfunargagna sem stendur til boða á stærstu tungumálunum. Slíkur ójöfnuður í færni nýjustu tæknilausna í mismunandi tungumálum getur leitt af sér menningarslagsíðu og skapað hættu á að gervigreind dragi úr fjölbreytni menningar og tungumála og auki þannig enn á **menningarlega einsleitni**.

Ef gervigreind á að gagnast öllu mannkyni ætti hún að **styðja við og styrkja menningararfleifð okkar**, kunna skil á **fjölmörgum tungumálum** og búa yfir áreiðanlegri og fjölbreyttri þekkingu á misjöfnum **samfélagslegum bakgrunni og sögu**.

Til að taka á vandanum er ljóst að samræmdra og hnitmiðaðra aðgerða er þörf. Við leggjum til **opið, alþjóðlegt verkefni** til að skilgreina vandamálið, safna og koma á gagnreyndum aðferðum, staðla aðferðir, þróa mælipróf, auðvelda gagnasöfnun og geymslu gagna og styðja við rannsóknir á sviði fjölmenningarlegrar og margmála gervigreindar. Slíkt verkefni ætti að fela í sér aðkomu hagsmunaaðila frá gervigreindarfyrirtækjum, rannsókna- og fræðasamfélaginu, ríkisstjórnum og fulltrúum samfélagsins, svo og alþjóðlegum stofnunum á borð við UNESCO.

Með afurðum verkefnisins (sem lýst er hér á eftir) og fræðslu- og miðlunarstarfi er ætlunin að til verði **miðlægur samstarfsaðili fyrirtækja jafnt sem tungumálahópa** sem hefur það markmið að auka veg afskiptra tungumála og menningar í gervigreindarlíkönum og annarri tækni. Verkefnið ætti jafnframt að verða traustur vettvangur og uppspretta **úrræða fyrir tungumál og menningarsamfélög** sem standa illa í heimi tækninnar en vilja leggja af mörkum staðbundin gögn og þekkingu til notkunar á opinn og sanngjarnan hátt svo að greiða megi fyrir auknum stuðningi við þau.

Kveikjan að þessari grein er meðal annars **stafræn vegferð Íslands**, sem hefur gripið til markvissra aðgerða til að vernda og efla tungumál sitt, sem um 350.000 manns tala, á tímum breytinga og áskorana af völdum gervigreindar. Framtak Íslendinga greiddi fyrir árangursríku samstarfi við OpenAI um að styðja við íslensku með GPT-líkönunum og sýndi þannig hvað hægt er að gera þegar þjóð grípur til markvissra aðgerða til að standa vörð um tungumál sitt.

Nýta má reynslu Íslands sem grunn að **handbók** eða sem fyrirmynd sem önnur samfélög geta tileinkað sér og aðlagað til að vernda og styrkja eigið tungumál og menningarlega arfleifð.

2. Verkin látin tala: afurðir verkefnisins

Verkefnið ætti að miða að því að skila einhverju eða öllu af eftirfarandi:

- Viðmiðunarreglum og **gagnreyndum aðferðum við öflun og leyfisveitingu** gagna til notkunar í gervigreindarlíkönum, að teknu tilliti til höfundarréttar, sjónarmiða hagsmunaaðila og annarra takmarkana sem kunna að eiga við.
- Samnýttu, opnu safni einmála gagnasafna og **safn fyrirmæla- og úrlausnargagna** (e. instruction tuning datasets) á ýmsum afskiptum tungumálum, á stöðluðu formi, auk skjalfestra bestu starfsvenja um söfnun og síun slíkra gagna.
- Stöðluðu **safni viðmiða og mæliprófa**, sem verður uppfært reglulega , ásamt meðfylgjandi viðmiðum til sigagjafar til að mæla frammistöðu gervigreindarlíkana á breiðu sviði verkefna og kanna færni í afskiptum tungumálum og þekkingu á margvíslegum menningarheimum og -sögu.
- Safni viðmiða til að mæla **slagsíðu og eitraða orðræðu í úttaki líkana** fyrir margvísleg tungumál, menningu og samfélög. Slík slagsíða getur verið allt frá málfræðilegum kynjahalla til rangnefna og niðrandi orðalags um tiltekna menningarhópa.
- Átaksverkefnum til að **styrkja og styðja við rannsóknir** á sviði fjölbreyttrar tungumálakunnáttu og menningarþekkingar í gervigreind, í gegnum styrki og í samvinnu við háskóla og rannsóknastofnanir um allan heim.

3. Kallað eftir aðgerðum

Hér á eftir verður farið yfir tillögur Íslands að aðgerðum sem miða að því að takmarka hættuna á „gervigreindargjá“ milli tungumála og menningarheima. Við leggjum til að helstu hagsmunaaðilar og þátttakendur á sviði gervigreindar taki höndum saman í opnu alþjóðlegu átaki til að tryggja að gervigreindartækni þróist þannig að tillit sé tekið til og stutt sé við fjölbreytni menningar og tungumála.

Með því að styðja við og taka virkan þátt í umræddu átaki geta hagsmunaaðilar lagt sitt af mörkum við að brúa bil ójöfnuðar í gervigreind, staðið vörð um menningararfleið sína og stuðlað að því að sá ávinningur sem hlýst af gervigreindarbyltingunni muni nýtast öllum samfélögum.

4. Stafræn vegferð Íslands: Dæmi til eftirbreytni

ÁHERSLA Á VARÐVEISLU TUNGUMÁLSINS

Á Íslandi, eyríki í Norður-Atlantshafi með um 400.000 íbúa, er rík bókmenntahefð sem byggist á íslenskri tungu. Hnattvæðing og sterk staða tungumála á borð við ensku er ógn framtíð íslenskunnar.[1] Börn hrærast að miklu leyti í ensku málumhverfi gegnum stafræna miðla og leiki. Sífellt fleiri íbúar hafa íslensku ekki að móðurmáli, fjöldi ferðamanna hefur stóraukist og ný gagnvirk tækni, sem er að mestu leyti á ensku, er orðin hluti af daglegu lífi.

Íslendingar gera sér grein fyrir því að glatist tungumálið glatast jafnframt mikilvægur hluti sjálfsmyndar þeirra. Því hafa þeir sameinast um að varðveita arfleifð tungumáls og menningar stafrænum heimi. Á síðustu tveimur áratugum hafa íslensk stjórnvöld fjárfest verulega í þróun stafrænna málfanga og máltækni. Háskólasamfélagið, einkageirinn og almenningur, með þátttöku í verkefnum þar sem allir leggjast á eitt, hafa stutt við verkefnið.

HELSTU ÁFANGAR Í ÁTT AÐ STAFRÆNU SJÁLFSTÆÐI

Árið 2014 samþykkti Alþingi einróma þingsályktun sem kvað á um aðgerðaáætlun um stöðu íslenskunnar á stafrænni öld. Þetta renndi enn styrkari stoðum undir viðleitni þjóðarinnar til að standa vörð um tungumálaarfleifð sína og markaði upphaf markviss landsátaks til að styrkja stöðu tunumálsins.

Árið 2018 höfðu stjórnvöld útbúið ítarlega aðgerðaáætlun, *Máltækniáætlun fyrir íslensku 2019-2023*[2], og fjárfest talsverðri upphæð í þróun málfanga og mikilvægra hugbúnaðartóla fyrir tungumálið. Komið var á fót samstarfi íslenskra háskóla, opinberra stofnana og einkafyrirtækja til að hrinda áætluninni í framkvæmd með fjárframlögum frá hinu opinbera ásamt mótframlagi frá einkageiranum.

Á meðal málfanga sem þróuð voru samkvæmt máltækniáætluninni var stór einmála málheild með fjölbreyttum tegundum texta á íslensku (*Risamálheildin*), samhliða málheildir fyrir vélþýðingar, raddupptökur og uppskriftir fyrir talgervla, málrýnir til að gá að rit- og málfræðivillum og önnur lykiltækni sem gagnast tungumálinu. Allt þetta var gefið út með opnum og víðtækum leyfum þannig að hver sem er gæti nýtt sér gögnin til að gera íslensku aðgengilega í vörum sínum.

Gagnasöfnin sem þróuð voru undir merkjum máltækniáætlunar hafa staðist tímans tönn og gegndt lykilhlutverki við að tryggja stöðu íslenskunnar gagnvart stórum mállíkönum.

---

[1] John Henley. 2018. *Icelandic language battles threat of digital extinction. The Guardian*, Bretlandi.

[2] Anna Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. *Language Technology Programme for Icelandic 2019-2023*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, Frakklandi. European Language Resources Association.

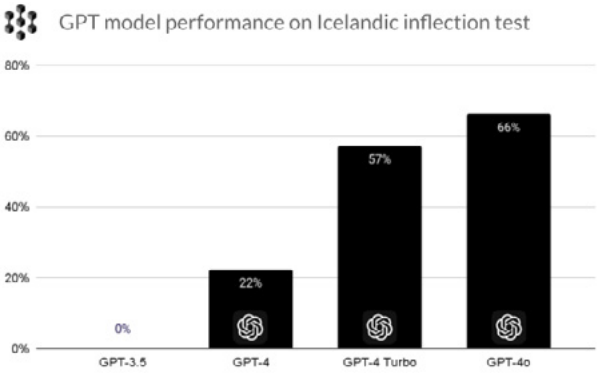
ÍSLAND OG OPENAI: SAGA VERKEFNISINS

Að loknum nauðsynlegum undirbúningi stóðu íslensk stjórnvöld frammi fyrir næsta áfanga: að gera almenningi kleift að njóta ávinnings af máltækni með því að stuðla að innleiðingu hennar í tækni og vörur sem nýtast fólki og fyrirtækjum í daglegu lífi og störfum. Leitast var við að kynna úrræðin og deila þeim með tæknirisum um allan heim, sérstaklega þeim sem standa að helstu stýrikerfum, forritum, öppum og öðrum hugbúnaði.

Forseti Íslands og menningar- og viðskiptaráðherra fóru í maí 2022 fyrir sendinefnd ti Bandaríkjanna þar sem ætlunin var að funda með nokkrum af helstu tæknifyrirtækjum heims. Markmiðið var að kynna tækni og gagnasöfn sem þróuð hafa verið fyrir íslensku, sýna fram á hversu auðvelt væri að samþætta íslensku við lausnir þessara fyrirtækja og bjóða upp á stuðning við slíka samþættingu.

Árangursríkasti fundurinn markaði upphafið að samstarfi íslenska hugbúnaðarfyrirtækisins Miðeindar[3] og OpenAI[4] við þjálfun GPT-líkana, frá og með GPT-4, á íslensku. Þar kom skýrt í ljós hvað þarf að gera til að samþætta tungumál fárra við stærstu mállíkönheimsins.

Samstarfið fól í fyrstu í sér að þjálfa GPT-4 með sýnidæmum um fyrirmæli og úrlausnir á íslensku í ferli sem nefndist „viðgjafarnám með mannlegri endurgjöf“ (e. *reinforcement learning from human feedback*, eða RLHF). Markmiðið var að meta hvað þyrfti til að stórt mállíkan gæti skilið og búið til texta á tungumáli sem fáir tala, svo sem íslensku. Þetta myndi hjálpa til við skipulagningu og fordæmi fyrir önnur smærri tungumál og sýna fram á að með markvissu starfi geta jafnvel tungumál með tiltölulega fáa málhafa blómstrað á stafrænni öld.



Til þessa hefur OpenAI fengið aðgang að rúmlega fjögurra milljarða orða textagögnum á íslensku af háum gæðum, auk mæliprófa fyrir tungumálahæfni. Íslensku gögnin voru notuð í GPT-4 Turbo til að auðvelda markvissar umbætur og sömu hágæðagögnin hafa verið notuð í öllum síðari gerðum.

Þetta hefur hjálpað til við að gera íslensku að gagnlegu sýnidæmi fyrir OpenAI, þar sem fyrirtækið hefur þróað og gert tilraunir með aðferðir til að auka hæfni líkana sinna í afskiptum tungumálum.

[3] See https://mideind.is

[4] See OpenAI's case study: https://openai.com/index/government-of-iceland/

5. Hnattræn hugsun: verkefni og handbók

Það sem hér fer á eftir er fyrsta tilraun til að leggja út af íslensku aðferðinni og nýta hana sem innblástur og grunn fyrir verkefni og handbók sem takast á við sambærileg sjónarmið fleiri mál- og menningarsamfélaga um heim allan.

Þörf er á samræmdu átaki til að varðveita og endurlífga menningarlega fjölbreytni heimsins þegar kemur að gervigreind og annarri stafrænni tækni. Við leggjum til að þetta átak feli í sér staðlaðar aðferðir við söfnun, síun og mat á gögnum fyrir þjálfun gervigreindarmállíkana og mælingar á frammistöðu þeirra. Þetta ætti að gera í opnu samstarfi og samráði við tækni-fyrirtæki til þess að hámarka nytsemi þessara gagna. Verkefnið ætti einnig að hvetja til og styðja við rannsóknir á leiðum til að auka inngildingu á sviði tungumáls og menningar.

Það á enn eftir að koma í ljós hvort stærstu alþjóðlegu gervigreindarlíkönin geti stutt við öll tungumál á fullnægjandi hátt, óháð stærð þeirra eða hvaða tungumálafjölskyldu þau tilheyra. Öll viðleitni til að fylgja tilmælum okkar ætti að hefjast á raunhæfu mati á því hvar viðkomandi tungumál stendur þegar kemur að magni gagna og annarra málfanga, samfélagslegri afstöðu gagnvart verkefninu og möguleika til fjárhagslegs stuðnings við það. Að safna gögnum í málheildir, jafnvel þótt þær séu smáar í sniðum, er þó alltaf mikilvægur þáttur í varðveislu tungumáls og getur auk þess verið undirbúningur fyrir framtíð þar sem yfirfærslunám, gervigagnamyndun og aðrar aðferðir til að styðja við gagnanaum tungumál á sviði gervi-greindar verða lengra á veg komnar.

STAÐFÆRÐ GERVIGREIND GAGNAST SAMFÉLAGINU

Gervigreindarlíkön sem hafa öðlast góða færni í tungumáli geta nýst samfélagi þess á ýmsa vegu: í daglegu lífi, í einkageiranum og hjá hinu opinbera. Þetta hefur verið raunin á Íslandi. Með tækni eins og fjöltyngdum spjallmennum, vélþýðingum og textasamantekt er hægt að koma þjónustu og hugbúnaði, sem annars væri aðeins í boði á ensku (eða öðrum algengum tungu-málum), til skila á íslensku. Eins er hægt að gera þjónustu sem hefur aðeins verið í boði á íslensku aðgengilegri íbúum sem hafa annað móðurmál en íslensku, sem eru 18% íbúa landsins, sem og erlendum ferðamönnum. Ákvörðun um að innleiða innri og ytri viðskiptaferla á íslensku ætti ekki að leiða til þess að fyrirtæki standi höllum fæti samanborið við keppinauta sem „hlaupast undan merkjum“[5] með því að notast við ensku.

Mikilvægt notkunartilvik fjölþættra gervigreindarlíkana sem styðja lítil tungumál er að auðvelda þýðingu í rauntíma á myndefni og leikjum, sérstaklega fyrir börn. Slík virkni, sem þróast frá sjálfvirkri textun yfir í hágæðaraddþýðingar í rauntíma þar sem líkt er eftir upphaflegum raddblæ, styður við máltöku barna á máltökuskeiði.

Gervigreind getur einnig hjálpað fólki með fatlanir að nálgast þjónustu á móðurmáli sínu með eiginleikum á borð við myndgreiningu og lýsingu, einföldun texta, tal í texta, texta í tal og staf-setningarleiðréttingu.

[5] Eða „svíkja“ í skilningi leikjafræðidæmisins Vandamál fangans.

Tækifæri sem þessi, sem geta eflt tungumál og menningu með nýstárlegum tæknilausnum, eru drifkrafturinn í viðleitni okkar til að skapa fyrirmynd sem nýta má til að auka veg afskiptra tungumála í gervigreind.

ÁRANGURSÞÆTTIR: LÆRDÓMUR AF REYNSLU ÍSLENDINGA

Við teljum að mögulegt og æskilegt sé að endurtaka þann hlutfallslega góða árangur sem náðst hefur við að innleiða íslensku í öflugustu gervigreindarlíkön heims fyrir önnur tungumál sem hafa til þessa dregist aftur úr í þróun og nýtingu máltæknilausna. Ýmsir þættir hafa stuðlað að árangri Íslendinga á þessu sviði.

FYRIRKOMULAG FJÁRMÖGNUNAR

Markmið: Að styðja við þróun nauðsynlegra málfanga.

Fyrir mörg afskipt tungumál og samfélög er innlendi markaðurinn of lítill til að bera uppi nauðsynlega þróun málfanga og máltækni sem hluta af arðbærum viðföngum, vörum og þjónustu. Því er nauðsynlegt að tryggja aðrar fjármögnunarleiðir og taka upp miðlægari nálgun. Íslensk stjórnvöld gerðu sér grein fyrir þessu og voru tilbúin að fjármagna máltækniáætluntil langs tíma á grundvelli þarfagreininga og vinnuáætlana máltæknisérfræðinga.

Kjarnaverkefni máltækniáætlunarinnar fengu 100% fjármögnun og komið var á fót mótframlagssjóði fyrir þróun á vörum sem fela í sér íslenska máltækni. Fyrirtæki gátu svo sótt um að fá 50% þróunarkostnaðar endurgreidd úr sjóðnum.

OPINN HUGBÚNAÐUR OG GÖGN

Markmið: Að stuðla að sem víðtækustum stuðningi við íslensku í tæknivörum.

Íslensk stjórnvöld tóku þá lykilákvörðun að birta allar afurðir máltækniáætlunarinnar með opnum og víðtækum leyfum.[6] Með þessum leyfum er hægt að innleiða afurðir verkefnisins án endurgjalds í og þjónustu þriðju aðila, hvort sem er í viðskiptalegum tilgangi eða ekki.

LÝÐVIRKJUN

Markmið: Að safna gæðagögnum á sem hagkvæmastan hátt og gefa almenningi kost á að hafa sitt að segja um tilhögun líkana.

Þátttaka og stuðningur almennings hefur verið mikilvægur hluti af máltækniáætlun Íslands. Forseti Íslands var verndari lýðvirkjunarverkefnis þar sem almenningur var hvattur til að láta í té raddsýni á vefsvæði á vegum máltækniáætlunarinnar. Verkefnið bar ríkulegan árangur, athygli á málefninu var vakin meðal almennings og þúsundum klukkutíma af raddsýnum var safnað til þess að greiða fyrir þróun á íslenskri talgreiningu og talgervlum.

[6] Á meðal leyfanna eru CC-BY, Apache og MIT.

Auk raddsýnanna er nú unnið að undirbúningi lýðvirkjunarverkefnis til að greina og draga úr bjögum og eitraðri orðræðu í textum á íslensku.[7] Þetta er mikilvægt verkefni þar sem það gefur almenningi kost á að láta í ljós hvað honum þyki teljast til eitraðrar orðræðu, slagsíðu af ýmsu tagi og annars slíks sem þjálfa ætti tungumálalíkön til að forðast. Æskilegt er að það sé ekki eingöngu undir tæknifyrirtækjunum komið að taka slíkar ákvarðanir.

SAMSTARF VIÐ RITHÖFUNDASAMBANDIÐ

Markmið: Að tryggja að fjölbreyttar málheildir séu aðgengilegar með opnum leyfum með hliðsjón af sjónarmiðum hagsmunaaðila um notkun á höfundarréttarvörðum textum.

Þar sem takmarkað magn af íslenskum texta er til á stafrænu formi, og bókmenntatextar eru aðeins lítið brot af því, var mikilvægt að finna leiðir til að fella sem mest af slíkum texta inn í opnar málheildir. Hins vegar hafa höfundar (og útgefendur) skiljanlega efasemdir um að opna fyrir aðgang að textum sínum á stafrænu formi. Áhyggjur þeirra snúa bæði að sölutapi vegna afritunar og brota á höfundarrétti, sem og hættunni á að stór tungumálalíkön geti búið til texta sem yrðu stæling á stíl tiltekinna höfunda.

Efnt var til samtals við Rithöfundasamband Íslands innan máltækniáætlunarinnar og samið um málamiðlun: Bókmenntatextum (þ.e. handritum bóka) sem höfundar leggja til hefur verið skipt niður í u.þ.b. 500 orða brot, sem eru síðan tekin inn í málheild í handahófskenndri röð. Jafnframt er unnt að afmá nöfn höfunda úr þjálfunargögnum svo að tungumálalíkön geti ekki tengt texta og stíl við tiltekna höfunda.

Með þessari lausn er hægt að þjálfa tungumálalíkön með umtalsvert meira magni af íslenskum bókmenntatextum af miklum gæðum um leið og komið er í veg fyrir að heilu verkin, eða stórir hlutar þeirra, séu birt eða að búinn sé til texti þar sem ritstíll tiltekinna nafngreindra höfunda er stældur.

GÆÐI, SJÁLFBÆRNI OG VIÐHALD

Markmið: Gögn og viðföng sem standast tímans tönn í síbreytilegu tækniumhverfi.

Með vönduðum, ítarlegum og aðgengilegum málföngum er hægt að þróa og taka í notkun stafræn verkfæri til að styðja við íslenska tungu. Ákvörðun um að notast við samrýmanlega, opna og framtíðarmiðaða staðla, til dæmis fyrir gagnasnið og forritunarmál, hefur reynst mikilvæg til þess að tryggja að máltækniáætlunin standist tímans tönn. Jafnframt er þörf á prófunum, mælingum og stöðugum uppfærslum á málföngum til þess að koma til móts við síbreytilegar þarfir, bæði hvað varðar tækni og tungumál.

[7] Sjá hópvistunarvefsvæðið *Ummælagreining* https://www.xn--ummlagreining-5fb.is/

LEIÐARSTEF: SKREF Í ÁTT AÐ INNGILDANDI GERVIGREIND

Hér á eftir fara tillögur um hvernig minni málsamfélög gætu ráðstafað fjármagni sínu og kröftum til að búa sig undir að takast á við gervigreind. Hverri tillögu fylgir hvatning til þeirra sem þróa líkönin um að leggja sitt af mörkum, veita leiðsögn og, í sumum tilvikum, huga betur að fjölbreytni tungumála við hönnun og smíði líkana.

ÍTARLEG GAGNAHREINSUN FYRIR GÖGN SEM SAFNAÐ ER AF VEFNUM

Veraldarvefurinn er oftar en ekki stærsta safn stafrænna textagagna sem stendur til boða fyrir mörg tungumál. Stór hluti slíkra gagna er þó af litlum gæðum og getur þannig dregið úr hæfni líkana ef þau eru notuð athugunarlaust í þjálfun. Hönnuðum stórra mállíkana standa ýmis verkfæri til boða til að hreinsa og sía gögn en til að ná sem bestum árangri þarf þekkingu á hverju tungumáli. Við mælum því með því að málsamfélög komi sér upp sínum eigin málheildum með gögnum af vefnum sem hafa verið vandlega hreinsuð og síuð með nýjustu og fullkomnustu tækni.[8]

Það sem þróunaraðilar líkana ættu að hafa í huga:

Verkfæri og gagnreyndar aðferðir fyrir **sjálfvirka gæðamælingu og síun** texta, sem hægt er að laga að ýmsum afskiptum tungumálum, ættu að vera skjalfest og gerð aðgengileg. Það hefur sýnt sig að fyrir smærri tungumál eru málheildir sem safnað er saman af vefnum, svo sem CommonCrawl, gjarnan af hlutfallslega minni gæðum en raunin er hjá stærri tungumálum. Oft er þörf á ítarlegri síun til að ná fram sem gagnlegustum þjálfunarmálheildum.

AÐRAR GAGNAUPPSPRETTUR

Stór mállíkön eru aðallega þjálfuð með einmála málheildum sem oft er aflað á vefnum í miklu magni. Það er afar dýrmætt, og góð nýting á fjármunum, að setja saman hágæðamálheild með einmála textum á ýmsum sviðum,[9] eins og Risamálheildin íslenska er dæmi um. Með fjárfestingu í góðri ljóslestrartækni (Optical Character Recognition, OCR) fyrir tungumál er unnt að færa prentuð gögn á stafrænt form og nota fyrir þjálfunarlíkön.

Gervigreindarlíkön verða sífellt margþættari sem þýðir að þau geta lært af annars konar gögnum en texta, svo sem hljóðgönum og hreyfimyndum.

Það sem þróunaraðilar líkana ættu að hafa í huga:

Listi yfir (lágmarks-)kröfur og æskilega eiginleika fyrir tegund, snið, samsetningu og magn gagna sem þarf til að stuðla markvisst að þjálfun margmála, og að lokum fjölþættra, gervigreindarlíkana.

LEYFISVEITINGAR FYRIR GÖGN

Mikilvægur þáttur í allri (opinni) gagnaöflun er að tryggja að afurðir séu gefnar út með opnum leyfum á borð við CC BY, sem gerir að verkum að hægt er að nota þær til að þjálfa gervigreindar-líkön, hvort sem er í viðskiptalegum tilgangi eða öðrum.

---

[8] Sjá A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models til að fá frekari upplýsingar um íslensku nálgunina.

[9] Sjá https://igc.arnastofnun.is/

Það sem þróunaraðilar líkana ættu að hafa í huga:
Samstarf um að setja á fót viðmiðunarreglur og **bestu starfsvenjur um öflun og leyfisskilyrði** gagna til notkunar í gervigreindarlíkönum, að teknu tilliti til höfundarréttar og annarra viðeigandi takmarkana.

MÆLIPRÓF FYRIR TUNGUMÁLAFÆRNI

Mikilvægt er að setja viðmið fyrir getu líkana til að meta og auka hæfni þeirra. Með þessum viðmiðum fást staðlaðir mælikvarðar til að meta hversu vel stóru mállíkönin skilja og búa til texta á tilteknu tungumáli. Íslensku viðmiðin sem Miðeind hefur þróað til þessa hafa vísað þróunarteymum OpenAI veginn, enda hafa orðið stöðugar umbætur með hverri uppfærslu GPT-líkananna. Þær hafa einnig veitt dýrmæta innsýn í það magn og þær tegundir gagna sem þarf til að auka hæfni stórra tungumálalíkana fyrir lítið tungumál eins og íslensku.

Ef fullnægjandi vélþýðingartól er þegar til fyrir tiltekið tungumál er hægt að búa til mælipróf með notkun vélþýðingar, yfirleitt út frá fyrirliggjandi prófum fyrir ensku (t.d. ARC[10], Belebele[11] og MMLU[12]). Það veltur á gæðum tólsins og eðli viðmiðanna hvort þörf sé á mannlegu gæða-eftirliti og leiðréttingum. Hægt er að staðfæra önnur viðmið (t.d. WinoGrande[13]) með mennskri aðkomu svo þau eigi við tiltekin tungumál og menningu. Þýðing og staðfæring fyrirliggjandi mæliprófa getur verið einkar gagnleg við samanburð á svipuðum verkefnum þvert á tungumál.

- Sum mælipróf miðast við tiltekin tungumál, með áherslu á einstaka þætti tungumáls og menningar og það sem greinir þau frá öðrum.
- Þrátt fyrir að tiltekið mælipróf eigi aðeins við um eitt tungumál getur engu að síður verið gagnlegt að leita fyrirmynda í samsetningu og efni þess.
- Mælipróf þurfa að vera krefjandi til að koma að gagni við þróun stórra tungumálalíkana. Tilgangurinn með prófunum er að greina hvar megi gera betur og setja markmið frekar en að staðfesta að stórt tungumálalíkan hafi þegar náð tökum á viðfangsefninu.
- Töflur sem taka saman frammistöðu gervigreindarlíkana eru gagnlegar til að safna niðurstöðum mæliprófa saman á einum stað og stuðla að samanburði og samkeppni.[14]

[10] Abstraction and Reasoning Corpus eftir François Chollet, sjá https://paperswithcode.com/sota/common-sense-reasoning-on-arc-challenge

[11] Belebele er fjölmála gagnasafn fyrir lesskilning, sjá https://arxiv.org/abs/2308.16884

[12] Massive Multitask Language Understanding, sjá https://arxiv.org/abs/2009.03300

[13] "An Adversarial Winograd Schema Challenge at Scale", sjá https://winogrande.allenai.org/

[14] Sjá t.d. stigatöflu Miðeindar fyrir stór íslensk tungumálalíkön: https://huggingface.co/spaces/mideind/icelandic-llm-leaderboard

Það sem þróunaraðilar líkana ættu að hafa í huga:

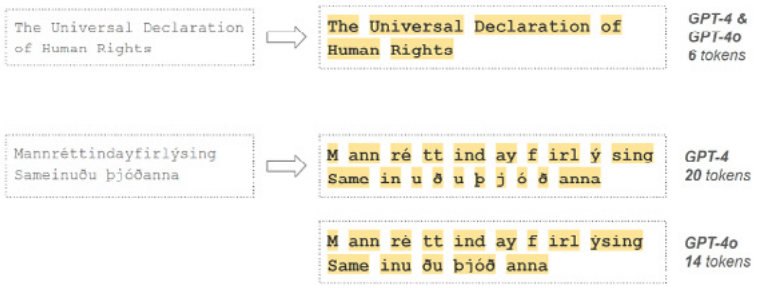
Getan til að meta núverandi færni líkana í skilningi og myndun á gagnanaumum tungumálum er lykilatriði. Nauðsynlegt er að rannsaka og þróa aðferðir og mælipróf á þessu sviði, sem og viðkomandi viðmið og prófanir.

Ákjósanlegt væri að slík próf yrðu fastur þáttur í mati á hæfni stórra mállíkana, sem viðbót við eða hluti af fyrirliggjandi mæliprófum, svo sem MMLU, HellaSwag[15] eða TruthfulQA[16].

Enn fremur kæmi sér afar vel ef útbúin yrði stöðluð aðferð til að reikna út **samantekna hæfnieinkunn** út frá mæliprófunum með tilliti til bæði meðalhæfni og jafnrar frammistöðu þvert á tungumál. Til að fá háa einkunn þyrfti líkan að ná almennum góðum árangri þar sem fá eða engin tungumál væru undanskilin.

TÓKUN

Texti sem er inntak eða úttak í mállíkönum er yfirleitt tókaður í úrvinnslu. Það þýðir að honum er skipt niður í brot eða orðhluta (til dæmis „orð“, „hluta“) og hvert brot fær samsvarandi númer. Úthlutun brotanna er ákvörðuð með tölfræðiaðferðum áður en líkan er þjálfað og er föst eftir það. Dæmigert mengi orðabrota í stóru mállíkani inniheldur 50.000–100.000 einstök brot.



Tokenization efficiency for Icelandic has improved significantly in newer GPT models but still lags behind English.

Mengi orðabrota miðast gjarnan að miklu leyti við enskan texta, sem þýðir að brot sem eru tiltölulega algeng í ensku (og rituð með enska stafrófinu) eru líkleg til að fá eigin númer á meðan jafnvel algeng brot í tungumálum á borð við japönsku eða hindí myndu ekki fá úthlutun, hvað þá brot úr sjaldgæfari tungumálum. Því eru inn- og úttaksrunur af tókum lengri en ella fyrir þessi tungumál. Högun helstu gervigreindarlíkana er þannig að reikniþörf þeirra er í hlutfalli við annað veldi af runulengd (fjölda orðabrota) og því er kostnaður við notkun þeirra á þessum tungumálum umtalsvert hærri. Þetta hefur með öðrum orðum áhrif á bæði færni líkana og notkunarkostnað fyrir smærri tungumál.

[15] HellaSwag prófar eðlilegar ályktanir um fullyrðingar á grundvelli almennrar skynsemi. Sjá https://arxiv.org/abs/1905.07830

[16] TruthfulQA mælir hvernig líkön herma eftir mennskum ósannindum. Sjá https://arxiv.org/abs/1905.07830

Það sem þróunaraðilar líkana ættu að hafa í huga:
Í tilraunum Miðeindar með líkön OpenAI fyrir íslensku komu fram nokkur lykilatriði þar sem gera má betur. Þar má nefna aukinn stuðning við sjaldgæfa bókstafi og hvað varðar skilvirkni tókunar. Viðræðurnar við þróunarteymi OpenAI sem fylgdu í kjölfarið stuðluðu að bættum margmála stuðningi í síðari útgáfum GPT-líkansins, einkum í GPT-4o, þar sem nýr, stærri og fjöltyngdari orðabrotaforði var innleiddur. Frekari rannsókna og prófana er þörf svo að unnt sé að greina og úthluta orðabrotum á sem skilvirkastan hátt til að styðja við margmála vinnslu án þess að það dragi úr getu líkananna á ensku.

SLAGSÍÐA

Við þróun á stórum mállíkönum hefur hingað til verið lögð mikil áhersla á tungumálagetu, þ.e. að þjálfa líkön til að skilja og gefa frá sér málfræðilega rétt úttak ýmsum tungumálum. Nú þarf hins vegar að útvíkka hugsunina og tryggja að þekking á menningu og sögu á hverjum stað sé innbyggð í gervigreindartækni, og að dregið sé úr bjögum og eitraðri orðræðu á grundvelli tungumáls og menningar eftir fremsta megni. Þessi krafa á ekki aðeins við um texta- og raddþætti heldur einnig myndir og myndbönd.

Tungumálatengd slagsíða getur verið í formi málfræðiatriða sem ekki eru til staðar í ensku. Í íslensku fylgja lýsingarorð til dæmis málfræðilegu kyni, karlkyni, kvenkyni eða hvorugkyni, og það hefur sýnt sig að stór tungumálalíkön eru líklegri til að nota kvenkynsmynd lýsingarorða sem tengjast neikvæðum persónueiginleikum.[17]

Gervigreindarlíkön ættu að búa yfir þekkingu á fjölbreyttum menningarlegum og sögulegum staðreyndum og geta byggt svör sín á þessari þekkingu til að lágmarka menningarlega bjaga eða slagsíðu í úttaki. Þetta getur oft verið háð tungumáli viðkomandi samskipta,[18] þó ekki í öllum tilvikum þar sem afar ólíkir menningarheimar geta deilt sama tungumáli.

Málsamfélög geta aðstoðað við þetta verkefni með því að tryggja að þau gagnasöfn sem þau þróa endurspegli menningarvenjur og -sögu þeirra eins og hægt er. Enn fremur þarf að koma upp ítarlegum mælikvörðum til að kanna hvort bjagar og slagsíða séu til staðar, auk almennrar þekkingar á sögu[19] tiltekinnar menningar.

---

[17] Sólmundsdóttir, A., Guðmundsdóttir, D., Stefánsdóttir, L. B., & Ingason, A. K. (2022). *Mean Machine Translations: On Gender Bias in Icelandic Machine Translations*. In N. Calzolari, F. Bechet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), 2022 Language Resources and Evaluation Conference, LREC 2022 (bls. 3113-3121).

[18] Sem dæmi má nefna að svarið við fyrirspurninni „Hver fann Ameríku?“ væri líklega „Leifur Eiríksson“ á íslensku en „Christopher Columbus“ á ensku (og „Cristóbal Colón“ á spænsku).

[19] Sjá til dæmis viðmið Miðeindar, Icelandic Wiki QA: https://huggingface.co/datasets/mideind/icelandic_wiki_qa

Það sem þróunaraðilar líkana ættu að hafa í huga:

Menningarvenjur um ásættanlegt orðfæri, eitraða orðræðu og bjaga eru vitaskuld afar ólíkar milli, og jafnvel innan, samfélaga og líkön þurfa að geta lagað sig að því. Auk þess þarf að ganga úr skugga um að almennar öryggisráðstafanir hvað varðar skaðlegt úttak (leiðbeiningar um sprengjusmíði og þess háttar) séu til staðar fyrir gagnanaum tungumál ekki síður en fyrir ensku.

Þróa þarf aðferðir til að búa til skilvirk mælipróf til að mæla bjaga og eitraða orðræðu þvert á tungumál og menningarheima í samstarfi við samfélögin sjálf, deila slíkum aðferðum með opnum hætti og viðhalda þeim í kjölfarið.

6. Niðurstaða: Ákall um fjölbreyttari gervigreind

Það er ekki bara tæknileg áskorun að viðhalda og efla fjölbreytni tungumála og menningar á sviði gervigreindar, heldur er það jafnframt siðferðileg nauðsyn. Þar sem við stöndum frammi fyrir nýrri öld þar sem gervigreind er leiðandi afl verðum við að tryggja að þessi umbreytingandi tækni þjóni öllu mannkyni - ekki bara ráðandi meirihluta. Lærdómur sem draga má af stafrænni vegferð Íslands og hið fyrirhugaða alþjóðlega samstarfsverkefni sem mun styðja við gagnaöflun og frammistöðumat fyrir tungumál sem hafa staðið illa í nýrri tækni, bjóða upp á vegvísi sem samfélög um allan heim geta notað til að vernda arfleifð sína á stafrænni öld. Með því að efla samvinnu milli þróunaraðila gervigreindar, málsamfélaga og stefnumótenda getum við skapað framtíð þar sem gervigreind eykur menningarlega fjölbreytni á heimsvísu í stað þess að draga úr henni. **Tíminn til aðgerða er núna**. Saman getum við mótað stafrænan heim framtíðarinnar, knúinn með gervigreind, þar sem fjölbreytileikanum er fagnað, tungumál vernduð og allar raddir fá að heyrast.