Samstarf um íslenska máltækni
SÍM

Máltækniáætlun fyrir íslensku

Varða 7 – Lokaskýrsla

1. febrúar 2022

Efnisyfirlit

Formáli

Þessi skýrsla dregur saman við vörðu 7 (M7) afurðir fyrir hvern verkþátt sem skilgreindur var í samningnum milli Almannaróms og SÍM (Samstarf um íslenska máltækni) frá september 2021, í verkefnaáætlun frá september 2021 og í endurskoðuðum vörðum frá janúar 2022.

Á þriðja ári er aukin áhersla lögð á tækniyfirfærslu, þ.e. undirbúningur verkþátta fyrir notkun innan t.d. fyrirtækja og/eða stofnana. Þessar tilraunir munu sýna hvernig nota má ákveðna verkþætti út á við og hversu auðvelt það getur verið. Notendaprófanir halda áfram og þróun með þriðju aðilum. Þessar tilraunir varpa ljósi á tæknilega möguleika afurða og undirbúa þær fyrir þriðju aðila. Þar sem íslenska CLARIN-hirslan hefur stækkað umtalsvert eftir því sem máltækniáætlun hefur þróast verður yfirlit á https://clarin.is gefið út bráðlega, sem mun sýna færslurnar flokkaðar og útgáfur þeirra. Auk þess hafa margar greinar um vinnuna innan máltækniáætlunar verið sendar til LREC 2022.

Allir nema þrír verkþættir náði afurðum M7 að öllu eða mestu leyti. Áætlanir hafa verið gerðar til að ná almennum markmiðum M7 fyrir 1. apríl hið seinasta; sjá töflu að neðan og samsvarandi skýrslur. Eftirstandandi vinna þessara verkefna hefur verið skipulögð þannig að hún hafi ekki áhrif á afurðir M8.

Verkþáttur	**Seinkuð skiladagsetning**
L9 – Málrýni í ritvinnslukerfum	2022-03-01
H2 – Umritun efnis úr sjónvarpi og útvarpi	2022-02-15
T6 – Veflesari	2022-04-01

Auk þess að fjalla um afurðir vörðu 7 eru eftirstandandi seinkaðar afurðir vörðu 6 einnig ræddar í þessari skýrslu, og hún er þannig viðbót við skýrsluna um seinkaðar vörður frá desember 2021. Seinkuðum M6-afurðum þeirra verkþátta sem koma fyrir í töflunni að neðan náðist ekki að ljúka í desember, en hefur nú verið lokið. Þessar afurðir eru ræddar í skýrslum viðeigandi verkþátta. Taflan sýnir einnig hvort viðeigandi verkþáttur heldur áfram á þriðja ári.

Verkþáttur	**Afurðir þriðja árs**
L10 – Aðlögun að máltæknihugbúnaði	Lokið fyrir þriðja ár
H1 – Upptökur með Samrómi	Heldur áfram á þriðja ári
H3 – Umritun samræðna og fyrirspurna	Lokið fyrir þriðja ár
H4 – Umritun fyrirlestra	Lokið fyrir þriðja ár
T13 – Stikaðir talgervlar	Heldur áfram á þriðja ári

Þessi skýrsla inniheldur lýsingar, umræður og niðurstöður vinnu við 30 verkþætti. Vegna umfangs efnisins miðum við að því að gefa skýra yfirsýn yfir stöðu og afurðir hvers verkþáttar á fyrstu síðu viðkomandi kafla.

Við leggjum til að skil M8-vörðu verði á formi kynningarfundar fyrir meðlimi stjórnar Almannaróms og fulltrúa ráðuneyta, sem svipar til afhendingar M5 í júní 2021.

G1 – Risamálheildin

Skýrsla: Stofnun Árna Magnússonar í íslenskum fræðum
Aðilar: Stofnun Árna Magnússonar í íslenskum fræðum

Markmið

Áframhaldandi stækkun RMH. Við lok annars árs greinist málheildin í sjö undirmálheildir. Allar þessar málheildir ættu að vera uppfærðar fyrir lok þriðja árs og meðhöndlaðar með nýjustu tólum. Ný útgáfa RMH ætti að innihalda meira en 1,8 milljarða tóka.

Varða 7 – lýsing

Söfnun fleiri texta frá samstarfsfúsum útgefendum. Vinna við að fá leyfi til að nota texta úr bókum sem hafa þegar verið unnar. Ljúka við lista fræðigreina og safna öllum textum fræðigreina fyrir málheildina.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Við höfum unnið við að safna fleiri textum frá samstarfsfúsum útgefendum bóka og einnig að öðlast leyfi til að nota texta sem þegar hafa verið unnir. Þessi vinna mun halda áfram á næstu vikum og mánuðum. Við búumst þegar við fleiri textum frá bókaútgefendum og vonumst til að ná samkomulagi við menntamálaráðuneytið um notkun allra texta þess. Lokið hefur verið við lista yfir 42 fræðirit og öllum aðgengilegum textum hefur verið safnað.

Skýrsla

Eftirfarandi skýrslu er skipt í tvo hluta fyrir undirmálheildirnar tvær sem eru áhersla þessarar vörðu, málheild bókatexta og málheild fræðirita.

Málheild bókatexta
Söfnun texta
Síðustu mánuðir ársins eru annasamastir fyrir íslenska bókaútgefendur. Eftir reynslu þess að safna bókatextum á síðustu mánuðum ársins 2020 gerum við okkur grein fyrir því að ætlast ekki of mikils á þessum tíma árs. Við héldum sambandi við samstarfsfúsa bókaútgefendur og nú, í janúar 2022, hafa þeir tekið að svara betur og jákvæðar varðandi að senda fleiri texta á komandi mánuðum. Við reynum að leggja áherslu á söfnun texta sem hafa verið skrifaðir af meðlimum Rithöfundasambands Íslands og Hagþenkis, félags höfunda fræðirita og kennslugagna á Íslandi, þar sem við höfum þegar leyfi frá þessum höfundum til að bæta textum þeirra í málheildina. Einn útgefandi, Háskólaútgáfan, hefur t.a.m. gefið út um 250 bækur eftir meðlimi þessara samtaka og hafa þau verið jákvæð varðandi samstarf. Við

eigum von á fleiri bókum frá þeim á komandi vikum. Við höfum einnig fengið svör frá öðrum útgefendum á seinustu vikum sem vilja byrja eða halda áfram að senda texta fyrir málheildina. Söfnun texta úr bókum mun því halda áfram samhliða öðrum verkum fyrir málheildina á næstu mánuðum. Við erum að biðja útgefendur að safna efni og senda það, án endurgjalds, svo við þurfum að vera þolinmóð og reyna að vinna eftir þörfum þeirra eins og hægt er.

Leyfi til að nota texta
Þegar við hófum að skoða þá texta sem við höfum ekki enn fengið leyfi til að nota í málheildina var ljóst að meirihluti þessara texta var gefinn út af menntamálaráðuneytinu. Því var áhersla lögð á viðræður við menntamálaráðuneytið til að reyna að fá leyfi fyrir mörgum textum á sama tíma í stað þess að hafa samband við hvern höfund fyrir sig. Við vonuðumst til þess að útgáfusamningar þeirra gætu með einhverju móti leyft okkur að nota textana án þess að þurfa að fá leyfi frá einstaka höfundum, sérstaklega í ljósi þess að flestar bókanna eru þegar aðgengilegar á vefsíðu ráðuneytisins. Tengiliður okkar í ráðuneytinu ákvað að ráðfæra sig við lögfræðing þeirra áður en hann gæfi endanlegt svar. Við höfum nú bókað fund með þeim í byrjun febrúar þar sem við ræðum þetta nánar. Við höldum áfram að vinna við þetta og finna leiðir til að nota þessa texta.

Málheild fræðirita

Listi yfir 42 fræðirit er fullunninn og sést hann að neðan. Öllum aðgengilegum fræðiritatextum hefur verið safnað en við bíðum svara þriggja útgefenda varðandi aðgengilegar skrár. Í þessum tilfellum hafa ritstjórar átt í erfiðleikum með að safna eða finna skrár sem við þurfum, t.d. vegna nýlegra breytinga í ritstjórn. Þetta kann að taka tíma þar sem við erum að biðja útgefendur að safna skrám og senda án endurgjalds, eins og nefnt var að ofan varðandi söfnun bókatexta. Vegna þessa þurfum við að senda reglulega áminningar og biðja oft um hverja sendingu. Í þessum tilfellum erum við að reyna að fá skrár á aðgengilegu sniði, t.d. Word- eða InDesign-skrár. Meirihluti fræðiritanna á listanum er aðgengilegur á vefnum í PDF-skrám eða -sniði en við reynum að fá aðgang að öðrum sniðum til að gera vinnslu skráanna fljótlegri og auðveldari.

Fræðirit	**Útgefandi**	**Efni**
Andvari	Hið íslenzka þjóðvinafélag	íslensk menning og bókmenntir
Árbók Hins íslenzka fornleifafélags	Hið íslenzka fornleifafélag	fornleifafræði, listasaga, þjóðfræði, örnefni
Bliki, tímarit um fugla	Náttúrufræðistofnun Íslands og Flækingsfuglanefnd, Fuglavernd, Líffræðistofnun háskólans	fuglafræði
Gripla	Stofnun Árna Magnússonar í íslenskum fræðum Háskólaútgáfan	íslensk og forn norræn fræði
Hugur: Tímarit um heimspeki	Félag áhugafólks um heimspeki	heimspeki
Hugrás	Hugvísindasvið Háskóla Íslands	hugvísindi

Heimspekivefurinn	Heimspekistofnun Háskóla Íslands	heimspeki
Iðjuþjálfinn: Fagblað iðjuþjálfa	Iðjuþjálfafélag Íslands	iðjuþjálfun
Íslenska þjóðfélagið	Félagsfræðingafélag Íslands	þjóðfræði
Íslenskt mál og almenn málfræði	Íslenska málfræðifélagið	málfræði
Jón á Bægisá	Þýðingasetur Háskóla Íslands, Ormstunga	þýðingafræði
Jökull: Ársrit Jöklarannsóknafélags Íslands og Jarðfræðafélags Íslands	Jöklarannsóknafélag Íslands og Jarðfræðafélag Íslands	jarðfræði
Kvarði: Fréttabréf Landmælinga Íslands	Landmælingar Íslands	landmælingar
Ljósmæðrablaðið	Ljósmæðrafélag Íslands (LMFÍ)	ljósmóðurfræði
Læknablaðið	Læknafélag Íslands	læknisfræði
Lögfræðingur: Tímarit laganema við Háskólann á Akureyri	Lögfræðingur	lögfræði
Lögmannablaðið	Lögmannafélag Íslands	lögfræði
Málfregnir og ályktanir Íslenskrar málnefndar	Íslensk málnefnd	málskipulag, íslenskufræði, málvísindi
Milli mála: Tímarit um erlend tungumál og menningu	Stofnun Vigdísar Finnbogadóttur í erlendum tungumálum (SVF) við Háskóla Íslands	tungumál, bókmenntir, málvísindi, þýðingafræði og tungumálakennsla
Náttúrufræðingurinn	Hið íslenska náttúrufræðifélag	náttúrufræði
Netla: Veftímarit um uppeldi og menntun frá menntavísindasviði Háskóla Íslands	Menntavísindasvið Háskóla Íslands	menntavísindi
Orð og tunga: Ársrit Stofnunar Árna Magnússonar í íslenskum fræðum	Stofnun Árna Magnússonar í íslenskum fræðum	málfræði, orðabókarfræði
Ritið: tímarit Hugvísindastofnunar	Hugvísindastofnun Háskóla Íslands	hugvísindi
Ritröð Guðfræðistofnunar	Guðfræðistofnun - Skálholtsútgáfan	guðfræði
Saga: tímarit Sögufélags	Sögufélag	sagnfræði
Sjúkraþjálfarinn	Félag sjúkraþjálfara	sjúkraþjálfun
Skíma	Samtök móðurmálskennara	íslenska sem móðurmál
Skírnir	Hið íslenska bókmenntafélag	bókmenntafræði, málvísindi, sagnfræði, heimspeki, stjórnmál, myndlist, þýðingar
Són: Tímarit um óðfræði	Óðfræðifélagið Boðn	skáldskaparfræði

Stjórnmál og stjórnsýsla	Stofnun stjórnsýslufræða og stjórnmála við Háskóla Íslands	stjórnmála- og stjórnsýslufræði
Tannlæknablaðið	Tannlæknafélag Íslands	tannlækningar
Tímarit Alzheimersamtakanna	Alzheimersamtökin	Alzheimer-sjúkdómur
Tímarit félagsráðgjafa	Félagsráðgjafafélag Íslands	félagsráðgjöf
Tímarit hjúkrunarfræðinga	Félag íslenskra hjúkrunarfræðinga	hjúkrunarfræði
Tímarit Kennslumiðstöðvar Háskóla Íslands	Kennslumiðstöð Háskóla Íslands	menntunarfræði, uppeldisfræði
Tímarit lífeindafræðinga	Félag lífeindafræðinga	lífeindafræði
Tímarit lögfræðinga	Lögfræðingafélag Íslands	lögfræði
Tímarit Lögréttu: Félags laganema við HR	Lagadeild Háskólans í Reykjavík	lögfræði
Tímarit um uppeldi og menntun	Menntavísindasvið HÍ, Hug- og félagsvísindasvið HA og Félag um menntarannsóknir	uppeldisfræði, menntunarfræði
Tímarit um viðskipti og efnahagsmál	Viðskipta- og hagfræðideild HÍ, viðskiptadeild HR og Seðlabanki Íslands	viðskiptafræði, hagfræði
Verktækni: Tímarit Verkfræðingafélags Íslands	Verkfræðingafélag Íslands	verkfræði
Öldrun: tímarit um öldrunarmál	Öldrunarfræðifélag Íslands	öldrunarfræði

I5 – Þáttari

Skýrsla: Háskóli Íslands
Aðilar: Háskóli Íslands, Miðeind, SÁM

Markmið

Þessi (litli) verkþáttur felur í sér áframhaldandi aukningu og uppfærslu reglubyggðra þáttara og tauganetsfullþáttara fyrir íslensku.

Auk þess verða útbúin nauðsynleg gögn til að þjálfa UD-þáttara (e. UD parsers), með UD-varpara (e. UD converter tool) sem getur varpað af liðgerðarformi (e. constituency structure) yfir á venslamálfræðiform (e. UD structure). Undirsetti Greynismálheildarinnar frá vörðu M5 í þessu undirverkefni verður varpað úr liðgerðartrjám (e. full constituency trees) yfir í venslamálfræðitré (e. UD trees), þar sem það mun gegna hlutverki frumþjálfunargagna.

Fullþáttun (e. full constituency parsing) er lykilforsenda málrýnieiningar. Þar er reglubyggður þáttari nauðsynlegur, vegna þess að mögulegt er að bæta sérstaklega merktum „villureglum“ við samhengisfrjálsa málfræði hans, þar sem algengum málfræðivillum er lýst. Reglubyggði þáttarinn er líka notaður til að framleiða þjálfunargögn fyrir tauganetsbyggða fullþáttarann. Þess má geta að utan máltækniáætlunar er þegar farið að nota reglubyggða þáttarann í hugbúnaði, sem ekki eru á vegum Máltækniáætlunarinnar, svo sem í talþjóni.

Reglubyggði þáttarinn hefur þegar verið gefinn út í heild sinni, en hægt er að setja hann upp sem Python-pakka ásamt meðfylgjandi skjalabúnaði, og sem opna GitHub-hirslu með MIT-leyfi. Þessi verkþáttur mun skila uppfærslum í Python-pakkann og hirsluna.

Tauganetsþáttarinn, ólíkt reglubyggða þáttaranum, umber málfræðivillur og útbýr áætlað þáttunartré jafnvel fyrir rangar setningar. Hann gæti því nýst betur en reglubyggði þáttarinn þegar kemur að því að þátta raddfyrirspurnir og annað inntak sem kemur beint frá notanda, eins og það kemur fyrir. Hins vegar er ekki hægt að aðlaga málfræði hans að sérstökum notkunardæmum; slík aðlögun krefst markverðs (e. nontrivial) þjálfunarsett fyrir hvert notkunardæmi.

Varða 7 – lýsing

Algengustu villurnar í UD-varpara hafa verið lagaðar.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð.

- Algengustu villurnar í UD-vörpunartóli hafa verið lagaðar, og nákvæmni tólsins endar í gildinu 81,39 fyrir merkta nákvæmni (e. LAS, labeled accuracy score).
- Tólið hefur verið gefið út á CLARIN á http://hdl.handle.net/20.500.12537/169 með Apache 2.0-leyfi. Það er einnig aðgengilegt á GitHub á https://github.com/thorunna/UDConverter.
- Reglubyggði þáttarinn var uppfærður, bæði almennt og sérstaklega fyrir málrýni í verkþætti L7. Uppfærða útgáfan hefur verið gefin út á CLARIN á http://hdl.handle.net/20.500.12537/175 og er einnig aðgengileg á GitHub á https://github.com/mideind/GreynirPackage.
- Að auki var undirliggjandi tilreiðarinn uppfærður á svipaðan hátt. Uppfærða útgáfan er aðgengileg á CLARIN á http://hdl.handle.net/20.500.12537/178 og er einnig aðgengileg á GitHub á https://github.com/mideind/Tokenizer.

Skýrsla

UD-vörpunartól

UD-vörpunartólið sem um ræðir er UDConverter, sem var upprunalega þróað utan máltækniáætlunar, sem hluti af Universal-trjábankagerðarverkefninu fyrir IcePaHC, sem var fjármagnað af Markáætlun í tungu og tækni. Tólið varpar trjábönkum frá liðgerðabyggðu PPCHE-merkingarsniði yfir á venslabyggða UD-sniðið. Úttakssnið er CoNLL-U, sýnt á Mynd 1. Annar dálkurinn sýnir tókann, sjöundi dálkur sýnir haus tókans og áttundi dálkur sýnir tegund vensla milli hauss og viðkomandi tóka.

```
# sent_id = 2008.OFSI.NAR-SAG,.25
# X_ID = 2008.OFSI.NAR-SAG,.41
# text = En þetta var auðvitað grátlegt og fáránlegt.
1       En      en      CCONJ   CONJ    _       5       cc      _       IFD_tag=c
2       þetta   þessi   DET     D-N     Case=Nom|Gender=Neut|Number=Sing|PronType=Dem   5       nsubj   _       IFD_tag=fahen
3       var     vera    AUX     BEDI    Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act 5       cop     _       IFD_tag=sfg3eþ
4       auðvitað        auðvitað        ADV     ADV     _       5       advmod  _       IFD_tag=aa
5       grátlegt        grátlegur       ADJ     ADJ-N   Case=Nom|Definite=Ind|Degree=Pos|Gender=Neut|Number=Sing        0       root    _       IFD_tag=lhensf
6       og      og      CCONJ   CONJ    _       5       cc      _       IFD_tag=c
7       fáránlegt       fáránlegur      ADJ     ADJ-N   Case=Nom|Definite=Ind|Degree=Pos|Gender=Neut|Number=Sing        5       amod    _       IFD_tag=lhensf|SpaceAfter=No
8       .       .       PUNCT   .       _       7       punct   _       IFD_tag=.
```

Mynd 1. Dæmi um CoNLL-U úttak varparans.

Nákvæmni tólsins var 72,82 fyrir merkta nákvæmni (e. LAS, labeled accuracy score) og 80,79 fyrir ómerkta nákvæmi (e. UAS, unlabeled accuracy score). Merkt nákvæmni samsvarar merktu F1-gildi setningafræðilegra tengsla og tekur tillit til þess hversu margir tókar hafa fengið bæði réttan haus og rétta venslamerkingu þar á milli. UAS telur með fjölda réttra hausa en skoðar ekki merkingarnar.

Greint er frá nákvæmni tólsins í óútgefinni grein, þar sem hugsanlegar endurbætur bæði við val hauss og venslamerkinga eru ræddar. Alls eru fimm endurbætur ræddar, sem áætlað er að auðveldast verði að laga og muni jafnframt auka merkta nákvæmni mest. Villur við val hausa eru þrjár, en þær tengjast ‘punct’, ‘cop’ og ‘cc’ venslum, og villur við venslamerkingar eru tvær, venslin ‘acl’ og ‘obl:arg’. Vinna við endurbætur þáttarans sem hluti af verkþætti I5 er byggð á þessari umfjöllun.

Endurbætur á UDConverter

Endurbætur á vali hausa
Eins og áður var nefnt eru endurbætur við val á hausum þrjár: ‘punct’, ‘cop’ og ‘cc’. Villur við hausaval þýða að tóki stjórnast af röngum tóka, en í þessum tilfellum er sambandið milli hauss og þess sem það stjórnar rétt merkt. ‘Punct’ er merkingin fyrir greinarmerki, ‘cop’ er samband kerfisorðs sem notað er til að tengja frumlag við sagnfyllingu, og ‘cc’ er samband liðar við aðaltengingu á undan.

Fyrir endurbætur var ‘punct’-merkingin sögð stjórnast af röngu aðalorði í 75,63% allra tilvika. Þessi háa villutíðni var að hluta til vegna þess að greinarmerki við enda setningar (e. end-of-sentence, EOS) stjórnuðust í flestum tilfellum af röngu aðalorði. Leiðbeiningar um UD-þáttun gera ráð fyrir því að greinarmerki við enda setningar stjórnist af rót setningar, en svo var ekki fyrir flest greinarmerki við enda setningar, þar sem flest stjórnuðust af aðalorði seinustu aukasetningar. Eftir að hafa lagað flestar þessara villna er villutíðni ‘punct’ nú 28,36%, eins og sést í Töflu 1.

Næst var ‘cop’-merkingin lagfærð. Tengisagnir (e. copulas) eru meðhöndlaðar á mismunandi hátt í liðgerðabyggða inntakinu og UD-byggða úttakinu; í inntakinu er tengisögnin, sem er gjarnan sögn, merkt á sama hátt og aðrar sagnir, en í UD-úttaki getur tengisögn ekki verið rót setningar, ólíkt öðrum sögnum. Þessi grundvallarmunur flækir vörpunina, og meðhöndlun tengisagna var ekki fullgerð í varparanum. Þetta orsakaði 21,86% villutíðni við val aðalorða fyrir sambönd merkt ‘cop’. Meðhöndlun fyrir ‘cop’ var bætt með því að skoða hverja og eina villu, og bæta þeim við reglur varparans, sem orsakaði á endanum 7,40% villutíðni, eins og sést í Töflu 1.

Þriðja venslamerkingin sem var löguð er ‘cc’. Merkingin ‘cc’ er gefin samtengingum og sambandið er í flestum tilfellum frá hægri til vinstri, sem þýðir að haus samtengingar kemur á eftir henni. Í einföldu dæmi samanstendur samtengingarliður af þremur orðum, tveimur nafnorðum og samtengingu sem tengir þau saman. Samkvæmt leiðbeiningum um UD-þáttun er fyrra nafnorðið haus liðarins, seinna nafnorðið er háð því, og samtengingin er háð seinna nafnorðinu. Úttak varparans fylgdi þessum reglum ekki alltaf, svo hann var bættur með því að skoða einstök tilfelli og reglur varparans bættar þar sem hægt var. Þessi endurbót lækkaði villutíðni sambandsins frá 18,52% í 3,69%, eins og sést í Töflu 1.

Merking	**Fyrri villutíðni (%)**	**Núverandi villutíðni (%)**
punct	75,63	28,36
cop	21,86	7,40

cc	18,52	3,69

Tafla 1. Venslamerkingar sem velta á röngum haus, ásamt fyrri og núverandi villutíðni.

Endurbætur á venslamerkingum
Endurbætur á venslamerkingum sem rætt er um í greininni eru tvenns konar: ‘acl’ og ‘obl:arg’. Í þessum tilfellum er hausinn réttur, en venslin milli háðs orðs og hauss þess er rangt merkt. ‘Acl’ stendur fyrir persónuháttar- eða nafnháttarsetninu sem stýrir nafnlið. Haus eða aðalorð ‘acl’-sambandsins er nafnorðið sem stýrist, og orðið sem er háð nafnorðinu er haus setningarinnar sem stýrir nafnorðinu, sem er í flestum tilfellum sögn. ‘Obl:arg’, aukafallsrökliður, greinir aukafallsrökliði frá viðhengjum, og er notað til að greina forsetningar sem virka eins og rökliðir.

Fyrst var ‘acl’-merkingin bætt. Villutíðni merkingarinnar var 72,01%, og flestar þeirra villna voru í nafnháttarsetningum, þar sem sambandið var merkt sem ‘acl’ í stað ‘xcomp’ (aukafallsliður sagnar eða lýsingarorðs sem er fyllilið). Einfaldar reglur í varparanum tengja saman aðalorð liða við ákveðnar venslamerkingar, og sumar þeirra reglna voru lagaðar til að lækka villutíðni. Aðalorð ákveðinna undirflokka nafnháttarsetninga, t.d. beinnar ræðu, stigsnafnháttar og nafnháttar sem er eins og frumlag voru áður merkt sem ‘acl’ en eru nú merkt sem ‘xcomp’. Þessi endurbót orsakar villutíðni upp á 31,25%, eins og sést í Töflu 2.

Síðasta endurbótin var gerð á ‘obl:arg’. Þessi venslamerking er undirflokkur ‘obl’, og kom fyrst fyrir í útgáfu 2 af UD. Hún var ekki notuð í upprunalega varparanum, en var notuð þegar prófunarsettið var merkt handvirkt til að samræmast íslensku PUD-málheildinni[1]. Munurinn milli ‘obl’ og ‘obl:arg’ er merkingarlegur, og því reyndist erfitt að bæta reglum þessara merkinga við varparann. Aðalorð ýmissa nafnliða hafa merkinguna ‘obl’, og eftir tilraunir með að breyta þeim í ‘obl:arg’ varð ljóst að þessari merkingu er ekki hægt að bæta í varparann með þessum aðferðum. Varparinn reiðir sig algerlega á upplýsingar frá þáttuðum inntaksskrám fyrir þennan hluta vörpunar, og þar sem þessi merkingarlega skilgreining er ekki til staðar í inntakinu yrði að afla utanaðkomandi upplýsinga til að geta notað ‘obl:arg’. Villutíðni merkingarinnar var 27,73%, eins og sést í Töflu 2, og helst því óbreytt.

Merking	**Fyrri villutíðni (%)**	**Núverandi villutíðni (%)**
acl	72,01	31,25
obl	27,73	27,73

Tafla 2. Rangar venslamerkingar, ásamt fyrri og núverandi villutíðnum.

[1] https://universaldependencies.org/treebanks/is_pud/index.html

Þessar endurbætur sem greint er frá að ofan orsökuðu hærri LAS og UAS, eins og sést í Töflu 3. LAS fór frá 72,82 í 81,39, sem er aukning um 8,57 stig, og UAS fór frá 80,79 í 88,32, sem er aukning um um 7,53 stig.

	Fyrra gildi	Gildi niðurstöðu
LAS	72,82	81,39
UAS	80,79	88,32

Tafla 3. LAS og UAS fyrir og eftir endurbætur.

I8 – Merkingargreining – BERT-lík mállíkön

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að gera frekari tilraunir með margmála Transformer-líkön, ásamt gagnaukningu fyrir margmála málheildir. Stækkun einmála málheilda fyrir tungumál með litlu eða meðalmiklu magni gagna (e. low and medium-resource languages) er ekki alltaf fýsileg nálgun. Annar möguleiki er að nota margmála líkön í staðinn, en nóg er til af forþjálfunargögnum fyrir þau. Tilraunir á öðru ári hafa sýnt að stór, margmála líkön, svo sem mBert (sem er þjálfað á 104 tungumálum), standa sig ekki betur en álíka stór einmála líkön fyrir íslensku. Nýlegar rannsóknir benda til þess að fyrir tungumál með litlu eða meðalmiklu magni gagna, þá kunni að vera ákjósanlegra að framkvæma forþjálfun á jafnari margmála málheildum sem samanstanda af skyldum tungumálum.
Annar valkostur er að auka einmála málheildina með gagnaaukningaraðferðum. Slíkum aðferðum hefur verið beitt við verk á borð við vélþýðingar, með góðum árangri.
Gengið verður frá viðmiðum NLP-verka og þau gefin út opinberlega.

Varða 7 – lýsing

Margmála Transformer-byggt mállíkan þjálfað á málheild Norðurlandamála.

Afurðir

Öllum markmiðum hefur verið náð. Við gefum út tvö forþjálfuð mállíkön Hugging Face-hirsluna: https://huggingface.co/jonfd:

- **electra-small-is-no**: Tvímála íslenskt/norskt ELECTRA-Small-líkan
- **electra-small-nordic**: Margmála íslenskt/norskt/danskt/sænskt ELECTRA-Small-líkan

Skýrsla

Síun texta

Það hefur verið sýnt fram á það með sannfærandi hætti að stækkun forþjálfunarmálheildar hefur jákvæð áhrif á árangur forþjálfaðra mállíkana „neðanstreymis“ (e. downstream performance). Því er orðin venja að auðga slíkar málheildir með texta sóttum af netinu. Þessir textar eru svo síaðir á ýmsan máta til að minnka magn lággæða texta.

Við höfum farið yfir fjölda textasía sem voru notaðar voru við gerð nokkurra þekktra líkana, þ.á.m. GPT-3, T5, mT5 og Gopher. Eftirfarandi síur voru metnar:

Lágmarkslengd skjals
Skjalið verður að innihalda minnst 50 tóka.

Lágmarkshlutfall af þekktum orðum
Minnst 60% orðanna (tókar sem innihalda bókstafi) verða að vera í orðasafni þekktra orða (t.d. í Beygingarlýsingu íslensks nútímamáls).

Lágmarkshlutfall bókstafa
Hlutfall bókstafa verður að vera 80% eða hærra.

Grunsamleg tákn og tókar
Skjalið má ekki innihalda ákveðna stafi eða tóka sem gefa til kynna að það innihaldi gögn sem eru ekki málfræðileg eða gölluð (t.d., Ã,  og).

Hámarkshlutfall tákna miðað við orð
Hlutfall hvaða tákns sem er (tákn önnur en bók-/tölustafir eða punktur) miðað við fjölda orða (tóka sem innihalda bókstafi) má ekki vera hærra en 30%.

Lágmarksfjöldi stopporða
Skjalið verður að innihalda minnst fjögur ólík stopporð.

Lágmarks-/hámarksmeðallengd tóka
Meðallengd tóka í skjalinu verður að vera minnst 3 og mest 10.

Síurnar eru metnar með því að nota þær á íslenska málheild sem var samsett úr mörgum uppsprettum:

- Risamálheildin (IGC) (1,69 milljarðar tókar)
- The Icelandic Common Crawl Corpus (IC3) (824M tókar)
- The Icelandic Crawled Corpus (ICC) (922M tókar)
- Íslenski hlutinn af Multilingual Colossal Clean Crawled Corpus (mC4) (793M tókar, aðeins íslensk svið)

Sameinuð málheild inniheldur samtals 4,16 milljarða tóka þegar tvítekin skjöl eru fjarlægð. Taflan að neðan sýnir fjölda fjarlægðra tóka með hverri síu. U.þ.b. 3,71 milljarður tóka stendur eftir eftir síun.

Sía	**Tókar**
Lengd	327.234.616
Hlutfall þekktra orða	60.963.714
Hlutfall bókstafa	40.908.094

Grunsamleg tákn og tókar	13.670.135
Hlutfall tákna miðað við orð	5.550.795
Stopporð	2.418.173
Meðallengd tóka	135.373
Samtals síaðir tókar	**450.880.900**
Samtals tókar eftir	**3.705.368.923**

Við forþjálfum ELECTRA-Small-líkön á ósíuðum og síuðum málheildum og metum þau á þremur neðanstreymisverkum: málfræðilegri mörkun (e. POS, part-of-speech), nafnakennslum (NER) og venslaþáttun (e. DP, dependency parsing). Við notum WordPiece-tilreiðara með orðaforðastærð 32.000. Árangur þessara tveggja mállíkana er borinn saman við þriðja líkanið sem var aðeins forþjálfað á RMH.

- Fyrir málfræðilega mörkun (POS) er líkanið fínstillt á MIM-GOLD-málheildinni (útgáfu 21.05) fyrir 20 mynstrasöfn með námshraðann 5e-5 og skammtastærð 16, og metið með tífaldri krossprófun.
- Fyrir nafnakennsl (NER) er líkanið fínstillt á MIM-GOLD-NER-málheildinni fyrir 10 mynstrasöfn með námshraðann 5e-5 og skammtastærð 16, og metið með tífaldri krossprófun.
- Fyrir venslaþáttun (DP), er líkanið fínstillt á Universal Dependencies-útgáfu IcePaHC-málheildarinnar fyrir 200 mynstrasöfn með námshraðann 2e-3 og skammtastærð 5000.

Sía	**Tókar**	**POS**	**NER**	**DP**	**Meðaltal**
RMH	1,69 millja.	96,84%	91,23%	84,90%	90,99%
RMH+ (ósíuð)	4,16 millja.	96,92%	91,08%	84,85%	90,95%
RMH+ (síuð)	3,71 millja.	**96,99%**	**91,29%**	**85,01%**	**91,10%**

Líkanið sem var forþjálfað á síuðu málheildinni nær bestum meðalárangri, betri en hin tvö líkönin á öllum þremur verkum. Þó að kostirnir við að sía forþjálfunarmálheildina virðist mjög skýrir er munurinn á meðalárangri líkananna þriggja frekar lítill. Það er möguleiki á að frekari síun málheildarinnar skili betri árangri. Í framtíðinni ætlum við að gera tilraunir með flokkun á gæðum texta (þ.e. að flokka skjöl sem há- eða lággæða eftir flækjustigi, auk flokkara undir eftirliti sem þjálfaðir eru á gagnasetti há- og lággæða skjala).

Margmála líkön

Mörg margmála líkön, eins og mBERT, XLM-R og mT5, eru forþjálfuð á gríðarstórum málheildum sem innihalda texta á yfir 100 tungumálum. Sýnt hefur verið fram á að líkön þjálfuð á málheild sem samanstendur af fáum skyldum tungumálum ná stundum betri árangri

neðanstreymis fyrir tungumál með með lítil og meðalmikil gögn. Af þessari ástæðu forþjálfum við og metum tvö ELECTRA-Small-líkön á íslensku auk annara Norðurlandamála. Eitt líkan er þjálfað á málheild íslenskra og norskra texta á meðan hitt inniheldur einnig danska og sænska texta.

Íslenska hluta málheildarinnar er lýst í fyrri hluta (RMH+ síuð). Fyrir hin tungumálin notum við texta fenginn frá málheildinni mC4. Hver af hinum fjórum málheildum Norðurlandamála hafa verið hreinsaðar af tvítekningum á skjalastigi, auk þess að vera síaðar samkvæmt atriðunum að ofan (fyrir utan síun á lágmarkshlutfalli þekktra orða fyrir tungumál önnur en íslensku). Hvert tungumál hefur jafnt vægi í forþjálfunarmálheildinni. Við notum orðaforðastærð 64.000 fyrir íslensk/norsku málheildina og 96.000 fyrir málheild Norðurlandamála (is/no/da/sv). Til að gera líkaninu kleift að klára tvær umferðir á forþjálfunarmálheildinni forþjálfum við íslensk/norska líkanið á 1,1 milljón skrefa (í stað venjulegra 1,0 milljón skrefa). Fyrir líkan Norðurlandamála aukum við þess í stað skammtastærð úr 128 í 256. Niðurstöður matsins koma fram í eftirfarandi töflu.

Líkan	**Tókar**	**POS**	**NER**	**DP**	**Meðaltal**
RMH+ (síuð)	3,71 millja.	**96,99%**	91,29%	**85,01%**	**91,10%**
íslensk/norskt	7,41 millja.	96,68%	91,36%	84,56%	90,86%
Norðurlandamála (is/no/da/sv)	14,82 millja.	96,64%	**91,71%**	84,84%	91,06%

Margmála líkönin ná einhverjum árangri, þar sem líkan Norðurlandamála nær bestum árangri á NER-verkum með nokkrum mun. Hins vegar nær einmála líkanið besta árangri fyrir hin verkin tvö, auk besta meðalárangri. Að því sögðu er munurinn á meðalárangri einmála líkansins og líkani Norðurlandamála frekar lítill. Aðrar líkanahaganir sem hannaðar eru til að þjálfa margmála líkön, eins og XLM-R, gætu hugsanlega náð betri árangri ef þau eru forþjálfuð á málheild Norðurlandamála. Í framtíðinni ætlum við að gera tilraunir með mismunandi líkanahaganir, stærðir orðaforða og hlutföll tungumála í forþjálfunarmálheildum. Við munum einnig gera tilraunir með notkun málheilda af hærri gæðum, eins og Norwegian Colossal Corpus, í stað vefskrapaðra skjala frá mC4.

V2b – Bakþýðingar

Skýrsla: Miðeind ehf
Aðilar: Miðeind ehf

Markmið

Lykilmarkmið í vélþýðingum á þriðja ári (sjá undirverkefni V4) er að styðja þýðingarlíkön fyrir vítt samhengi (á stigi margra setninga og/eða á skjalastigi). Búa þarf til nýjar bakþýðingagervimálheildir sem eru síðan notaðar sem hluti af þjálfun lokaútgáfu þýðingarlíkans á skjalastigi (e. document-level translation model).

Þetta undirverkefni felur einnig í sér hugbúnaðarvinnu til að þjappa/einfalda líkanið til að búa til þýðingar við ályktun (e. at inference time) á hraða sem endanlegir notendur geta sætt sig við, sérstaklega fyrir atvinnuþýðendur (sjá kafla um þýðingarvélar fyrir sérsvið) og fyrir notendur sem þurfa þýðingar í rauntíma á t.d. fréttum með viðbót í vafra eða sambærilegum aðferðum. Líkön á skjalastigi eru í eðli sínu stærri og hægari en líkön sem þýða aðeins setningu-fyrir-setningu, og þess vegna kalla þau á meiri hugbúnaðarvinnu fyrir hraða og samþjöppun.

Varða 7 – lýsing

Einmála gögn með víðu samhengi valin og undirbúin fyrir bakþýðingar. Vinnsluhraði og grunnlínur fyrir nákvæmni og markmið skilgreindar.

Afurðir

Markmiðum þessa undirverkefnis hefur verið náð, gagnasettum hefur verið safnað og grunnlínur fyrir framleiðslustig settar eins og skýrt er að neðan.

Skýrsla

Gögn með löngu samhengi hafa verið skilgreind fyrir bakþýðingu. Við höfum valið og undirbúið gögnin sem talin eru upp í töflunni að neðan, sem er blanda af ritstjórnartextum og völdum textum af vefnum. Bakþýðingar með löngu samhengi munu byrja þegar fyrsta kynslóð íslenskra ↔ enskra þýðingarlíkana á skjalastigi er aðgengileg.

Einmála gagnasett	**Upprunatungumál**	**Tókar (milljónir)**
Risamálheildin (án íþrótta) (IGC)	íslenska	1.388

The Icelandic Common Crawl Corpus (IC3)	íslenska	824
Ritgerðir nemenda	íslenska	367
Greynir News	íslenska	76
Wikipedia	íslenska	50
Íslenskar rafbækur	íslenska	1,7
Books3	enska	108 851
NewsCrawl	enska	59 204
SCOTUS opinions	enska	2.522
Wikipedia	enska	542
EuroPARL	enska	337
Reykjavik Grapevine	enska	70
Iceland Review	enska	43

IC3 og síun krossóreiðu (e. cross entropy)
The Icelandic Common Crawl Corpus er safn íslenskra texta dregið úr Common Crawl.[2] Gæði gagnanna skiptir sköpum fyrir bakþýðingar; við höfum því þjálfað tvö lítil generatíf mállíkön (GPT-2) til að meta gæði texta. Hið fyrsta er þjálfað á öllum tiltækum íslenskum einmála gögnum óháð gæðum, og hið seinna er þjálfað eins nema fínstillt á völdu hágæða undirsetti RMH. Þessi líkön eru svo notuð til að gefa skjölum í málheildinni einkunn (eftir mælingum flækjustigs) og þau sem ná betri einkunn með líkaninu sem þjálfað er á hágæða gögnum eru geymd fyrir bakþýðingar.

Ensk gögn með íslensku samhengi
Þar sem þýðingar milli íslensku og ensku innihalda gjarnan íslenskt samhengi (eins og örnefni, stofnanir og íslensk nöfn) höfum við stefnt að því að taka með enska texta um Ísland. Fréttasíður á ensku, *Grapevine* og *Iceland Review*, voru því sérstaklega teknar með. Við teljum að þessi gagnasett reynist nytsamleg þar sem þau kynna vel uppbyggt enskt mál sem inniheldur sérstaklega íslenskar nafneiningar sem við höfum sýnt að ollu vandamálum við þýðingar á fyrri árum máltækniáætlunar.

Úrtak
Gagnasettin sem innihalda mjög mikið magn gagna verða ekki notuð að fullu; við tökum úrtak skjala frá NewsCrawl og Books3 (safn bóka) fyrir endanlega stærð bakþýðingamálheildar með enskan uppruna til að vera jöfn hinni íslensku.

[2] https://commoncrawl.org/

Útgáfa bakþýðinga
Þar sem mikið af gögnunum sem við þýðum eru höfundarréttarvarin getum við ekki gefið út allar bakþýðingamálheildirnar á vefnum. Þó að þetta gæti hafa talist ásættanlegt fyrir einstaka stokkaðar setningar finnst okkur ekki boðlegt að gefa út texta með löngu samhengi og þýðingar úr ritgerðum nemenda, bóka, Iceland Review og Grapevine.

Framleiðslustigshraði og nákvæmnisgrunnlínur

Tauganet eins og vélþýðingarkerfi eru reiknislega þung í vinnslu og best er að vinna þau á skjákortum. Þau eru einnig nokkuð stór hvað varðar geymslu- og minnisnotkun.

Stærð þessara neta hefur mikla fylgni við gæði úttaks. Það er freistandi að halda áfram að stækka stærð líkansins, og jafnframt auka vélbúnaðarkröfur, en þetta er ekki alltaf hægt í reynd. Stærðin hefur bein áhrif á tímann sem það tekur að mynda þýðingar, sem þarf þá að vega á móti með enn öflugri (og dýrari) vélbúnaði.

Með þetta í huga höfum við sett markmið til að geta boðið líkön sem geta þýtt lengri texta á hraðanum 100–200 orð á sekúndu á tileinkuðum netþjóni með að lágmarki eitt 1080 eða sambærilegt skjákort (með 8GB skjákortsminni). Kortið ætti að geta keyrt tvö líkön á saman tíma svo það geti þýtt í báðar áttir. Þessi frammistaða tryggir að upplifun notenda sé ásættanleg og að búast megi við því að tæknin virki vel í uppsetningu sem flestir aðilar hafa tök á. Þessari frammistöðu er þegar náð hjá https://velthyding.is, þar sem þýðingarlausn byggð á tauganeti var þróuð sem hluti af máltækniáætlun og sett í notkun. Verk okkar fyrir komandi vörðu er því að smíða flóknari líkön sem geta þýtt lengri þýðingar án verulegrar rýrnunar á frammistöðu.

Þegar líkönin eru aðeins keyrð á örgjörva, sem krefst ekki vélbúnaðar sem erfitt er að útvega, er þýðingarhraði töluvert lægri. Markmið um hraða og nákvæmni er að líkanið sé nothæft með örgjörva fyrir styttri þýðingar. Þó að þetta sé óljóst markmið sem getur verið ólíkt fyrir mismunandi notkun er markmiðið að ná 50 orðum á sekúndu og jafnframt missa ekki meira en 5 BLEU-stig á prófunarsetti okkar innan sviðs.

V4a – Opin þýðingarvél fyrir íslensku

Skýrsla: Miðeind ehf
Aðilar: Miðeind ehf

Markmið

Markmið þriðja árs fela meðal annars í sér að pakka og gefa út á GitHub og CLARIN þýðingarlíkan sem nota má áreiðanlega og án sérþekkingar umfram almenna forritunarfærni. Það er, þekking á tauganetum og gagnaúrvinnslupípum ætti ekki að vera skilyrði til að taka í notkun framleiðslukerfi með notkun þýðingarlíkans.

Til að kerfið verði gagnlegt, afkastamikið og samkeppnishæft að minnsta kosti næstu árin, höfum við stigið skref í átt að þýðingum sem spanna fleiri en eina setningu í einu og viðhalda þar með samhengi og bæta samheldni.

Að lokum verður tekið á þoli með því að skilgreina mælikvarða, grunnlínur og markmið fyrir afköst líkans á ófullkomnum heimildartexta (þ.e. textum með stafsetningar- eða málfræðivillum; textum skrifuðum af börnum, lesblindu fólki og íslenskumælandi fólki sem hefur annað móðurmál) og stilla greypingaraðferðir (e. embedding strategies) til að bæta mælanlegan árangur í slíkum dæmum.

Varða 7 – lýsing

Gagnasetti fyrir þýðingar á skjalastigi safnað. Gagnasett með náttúrulegu suði safnað. Mælikvarðar og grunnlínur fyrir mælikvarða þols.

Afurðir

Afurðum undirverkefnisins var náð með því að a) safna gagnasettum á skjalastigi, b) safna gagnasettum með “suði” og skoða handvirkt úrtak þýðinga þeirra og c) skilgreina grunnlínur árangurs og þols eins og skýrt er að neðan.

Skýrsla

Gagnasett þýðinga á skjalastigi

Við höfum safnað samhliða gagnasettum sem nýtast við þjálfun og mat á kerfum með lengra samhengi. Greint er frá þeim í eftirfarandi töflu:

Nafn	Grófleg stærð (orð)	Samraðað
EES-reglugerðir	17.000.000	Nei
Biblían	1.400.000	Já
IPAC[3]	1.800.000	Já
WMT21	40.000	Já
Skýrsla rannsóknarnefndar Alþingis	140.000	Nei
Skýrslur SÍM	160.000	Nei
Vottar Jehóva[4]	7.600.000	Já
Stúdentablaðið	200.000	Nei

Gagnasettin sem eru ekki samröðuð þegar (á setningastigi) verður samraðað með Bleualign[5] með sömu aðferðum og fyrir IPAC.

Gagnasett náttúrulegs suðs

Til að meta þol og árangur vélþýðingarlíkana á ófullkomnum upprunatextum höfum við safnað gagnasetti texta með náttúrulegu suði frá ýmsum áttum. Þetta er texti sem ekki hefur verið lesinn yfir og inniheldur ýmis frávik frá hefðbundnum ritstýrðum texta, eins og málfræði- og prentvillur, sjaldgæf tákn og óformlegt mál sem gjarnan er notað í tali á netinu.

Við skilgreindum eftirfarandi gerðir suðs sem við viljum sjá í náttúrulegum textum, bæði fyrir íslensku og ensku:

- stafsetningarvillur
- málfræðivillur
- óhefðbundin notkun falla/greinarmerkja/há- eða lágstafa
- orð sem vantar í texta
- bil vantar í texta (algengt þegar texti er afritaður/límdur)
- kommur vantar á stafi í orðum
- styttingar (algengar, en getur verið erfitt að þýða utan samhengis)
- mjög langar setningar (sum tauganetslíkön hafa hámark á lengd)
- óhefðbundin stafsetning orða sem líkja eftir tali (“looooong”, “hurru”, “kjellinn”)
- tákn
- útlenskir bókstafir
- emojis
- snið

[3] https://arxiv.org/abs/2108.05289
[4] Gögn hafa verið fjarlægð úr uppruna; það á eftir að staðfesta hvort þau náist með öðrum hætti.
[5] https://github.com/rsennrich/Bleualign

- útlensk orð eða bútar í texta
- málskipti
- hlekkir á vef eða html-tög í texta
- mállýskur, hreimar (enska)

Setningarnar voru valdar af handahófi og flestar línur, en ekki allar, innihalda suð. Flestar villurnar að ofan koma fyrir í söfnuðum gögnum.

Fyrir íslensku söfnum við texta með suði frá þessum upprunum:

Nafn	**Gerð texta**	**Uppruni**
Icelandic Common Crawl Corpus (IC3)	Gögn úr Common Crawl á íslensku	https://huggingface.co/datasets/mideind/icelandic-common-crawl-corpus-IC3
Íslenska lesblinduvillumálheildin	Textar frá lesblindum notendum	http://hdl.handle.net/20.500.12537/132
Bland.is	Spjallborð á netinu	https://github.com/pallih/spjallbord/tree/master/bland
Hugi.is	Spjallborð á netinu	https://github.com/pallih/spjallbord/tree/master/hugi
Jeppaspjall.is	Spjallborð á netinu	https://github.com/pallih/spjallbord/tree/master/jeppaspjall
Twitter.com	Samfélagsmiðill (IGC-Social)	http://hdl.handle.net/20.500.12537/138

Fyrir ensku notum við viðmiðunargagnasett MTNT[6]. Það samanstendur af athugasemdum með suði frá Reddit, og hefur verið faglega þýtt yfir á frönsku og japönsku (við notum enska upprunann). Eftir skoðun og samkvæmt meðfylgjandi skjali inniheldur þetta gagnasett svipað suð og í íslenska gagnasettinu.

Við höfum safnað 2.000 línum af texta frá þessum upprunum, fyrir hvora þýðingarátt. Þar sem mannlega viðleitni þarf til að meta vélþýðingar þessara gagna höfum við takmarkað fjölda lína við það sem við teljum gerlegt að meta handvirkt.

Eftirfarandi eru dæmi úr íslensku gögnunum, sem sýnir suð sem kemur fyrir:

- hvar er þessi á landinu
- hann er eikvað riðgaður og eikver göt komin á hann ran í gengum skoðum samnt fyrir um ári síðan og hefur lítið verið notaður eftir það filgja með honum húd spísar glóðakerti lokur annað grill og eikvað fleira smá dótt
- Ég var að kaupa dælu sem heitir Marco UP 2 dælir 10l/min og kostaði undir 20 Þús
- Góðan daginn…

[6] https://aclanthology.org/D18-1050.pdf

- -þegar rakspíralykt elskhugans er enn þá í sængurfötunum
- 11 ára pjakkur að splæsa í shot shaping... bara.
- vá 😅🤩
- 11:05 Þórólfur: Bólusetning allt næsta ár, Pfizer tefst og annað seint
- ókei kannski aðeins meiri pening í heterodox ráðstefnuna?
- ég er með *nokkrar* hugmyndir…
- Elska ⚽fimmtudagskvöld hvort sem er 🤢
- KÞBAVD að svara rannsóknum.
- Afi: það er lúxusfiskur hah!

Grunnlína og mælingar þols

Markmiðið er að kerfið séð þolið gagnvart suði í inntaki, þ.e. stafsetning og málfræði ætti ekki að hafa stórvægileg áhrif á hvort þýðingin skilar réttri merkingu (við ætlumst ekki til þess að þýðingarlíkanið sé sérfræðingur í leiðréttingu málfræðivillna). Greinarmerki sem eru óhefðbundin eða vantar, villur hástafa, tjámyndir o.s.frv. ættu ekki að hafa stórvægileg áhrif á gæði þýðinga.

Til að prófa þetta munum við þýða fyrrnefnd gagnasett með þýðingarkerfum okkar, og meta handvirkt gæði þýðinga með og án mælikvarða þols (þróað fyrir M8 og M9).

Að auki munum við nota sjálfvirka mæliaðferð, handahófskennt táknsuð (e. random character noise), til að meta almennt þol yfirfarinna prófunarsetta eins og WMT. Við munum þýða matssett okkar eftir suðun texta (1–5% af inntaki) og meta BLEU fyrir a) líkön sem fengu ekki suðað inntak við þjálfun og b) líkön sem fengu suðað inntak við þjálfun.

Til að skilgreina grunnlínu til bráðabirgða fyrir náttúrulega suðað gagnasett höfum við metið vélþýðingar á sýni af gögnunum (200 setningar fyrir hvora þýðingarátt). Setningarnar voru þýddar með nýjasta vélþýðingarlíkaninu sem keyrir á velthyding.is.

Hver setning var metin á skalanum 1-4, þar sem 1 þýðir að öll eða mest af upprunalegri merkingu tapaðist við þýðingu, og 4 þýðir að merking setningarinnar hélst óbreytt. Niðurstöður þessa mats sjást í eftirfarandi töflu.

	is-en	**en-is**
4 Öll merking helst	42,5%	37,0%
3 Einhver merking tapast	34,0%	29,0%
2 Mikil merking tapast	18,0%	25,0%
1 Öll/mest merking tapast	5,5%	9,0%

Niðurstöður eru mjög svipaðar fyrir þessi tvö gagnasett, þar sem 42,5% (is-en) og 37% (en-is) þýðinganna ná að halda merkingu upprunatexta. Eftir standa merkingarfræðilegar villur, allt frá vægum til alvarlegra villna sem brengla mjög merkinguna.

V4b – Opin þýðingarvél fyrir íslensku – þýðingarvél fyrir sérsvið

Skýrsla: Miðeind ehf
Aðilar: Miðeind ehf

Markmið

Á þriðja ári einbeitum við okkur að umbótum á kerfinu með þýðingar á EES-reglugerðum fyrir utanríkisráðuneytið í huga. Þetta samstarfsverkefni er eins konar leiðarljós til að læra að aðlaga og innleiða vélþýðingarkerfið fyrir raunverulega, faglega notkun. Það verður sniðmát á hvernig skal pakka fínstilltu pípunni til að nota á öðrum sviðum, og fínstilla „neðanstreymis“ (e. downstream) fyrir ákveðin aðgengileg notkunarsvið með lágmarks fyrirhöfn.

Umbætur á vélþýðingarkerfinu fela í sér að safna yfirförnum (e. post-edited) þýðingum frá þýðendum utanríkisráðuneytisins til frekari fínstillingar líkansins, og svo það styðji sérhæfðar þýðingar fyrir undirsvið innan EES-reglugerðasviða (fjármála-, umhverfis- o.s.frv.), þar sem hugtök kunna að hafa staðlaða, undir-sviðsbundna (e. sub-domain specific) þýðingu sem líkanið verður að læra.

Að auki verður komið á samvinnu við Stafrænt Ísland með það að markmiði að þróa þýðingarkerfi fyrir eina eða tvær gerðir þjónustunnar. Þessi hluti tengist Evrópuverkefninu National Language Technology Platform (NLTP), sem er að hluta til fjármagnað af CEF-Telecom.

Varða 7 – lýsing

Undirsvið valin og líkön fínstillt.

Afurðir

Markmið afurða undirverkefnisins náðust, undirsvið voru valin og líkön þjálfuð. Niðurstöður eru kynntar í skýrslunni að neðan.

Skýrsla

Við snúum okkur nú að því að aðlaga þýðingar að ákveðnum sérsviðum innan EES, og sú vinna byggir á reynslu okkar og þekkingu sem við fengum með samstarfi við Þýðingamiðstöð utanríkisráðuneytisins (ÞMST). Í könnun sem lögð var fyrir notendur við M6 bentu þýðendur ÞMST á að helsti ókostur þess að nota vélþýðingarkerfi í þýðingarumhverfi þeirra væri að

hugtök sem tengjast undirsviði þeirra vantaði stundum, þar sem hugtökin eru mismunandi milli sviða.

Við höfum aðlagað vélþýðingarlíkön okkar fyrir 20 undirsvið, sem hluta af vinnu til að gera þýðingar hæfari fyrir ákveðin svið, eins og lýst er að neðan.

Aðferð

Alls höfum við næstum milljón búta frá þýðingarminnum ÞMST í 55 flokkum. Sumir flokkar eru algengari en aðrir.

Viðbúið er að árangur fari eftir fjölda búta, en merkið fyrir smærri flokka getur þó verið nytsamlegt þegar merking er notuð til að sýna sviðið.

Nálgun okkar við að veita sérstaka þýðingu eftir undirsviði byggir á því að nota stakt þýðingarlíkan sem hefur verið útfært til að taka við sérstökum tóka sem gefur til kynna undirsvið inntakssetningar. Þetta er ólíkt því að fínstilla einstaka líkön fyrir hvert undirsvið, sem er reiknilega dýrt í þjálfun og sérstaklega hýsingu. Ein vél með 1080 nVidia-skjákorti getur eins og er keyrt tvö líkön í einu. Notkun sviðstilgreinis gefur möguleika á einu líkani, fyrir hvora þýðingarátt, sem má nota fyrir öll tilfelli.

Niðurstöður

Til að meta þessa nálgun skoðum við hvort þetta valdi minnkandi ávinningi eða hafi slæm áhrif á undirsvið með fáa búta. Við gerum þetta með því að raða 40 stærstu undirsviðunum í flokka eftir stærð. Við gefum upp fjölda búta og BLEU-mælikvarða fyrir undirsviðin á bilunum 1–5, 11–15, 21–25 og 31–35. Við útilokum bútapör þar sem báðar tungumálahliðar hafa ekki 3 eða fleiri orð (tóka). Alls berum við saman fjórar gerðir líkana:

1) Almennt ensk-íslenskt líkan byggt á MBART án nokkurra EES-fínstillinga (*Engin fínstilling*)
2) Fínstilling líkansins úr (1) á öllum aðgengilegum EES-gögnum, ekki aðeins þau 40 völdu undirsvið (*EES*)
3) Aðskilin fínstilling líkansins úr (2) fyrir hvert tiltekið undirsvið (Aðeins undirsvið)
4) Fínstilling líkansins úr (2) einu sinni á öllum 40 stærstu undirsviðunum sameinuðum með notkun sviðstilgreinis (e. subdomain specifier) (*Merkt*).

Undirsvið	Bútar	Merkt	Aðeins undirsvið	EES	Engin fínstilling
Landbúnaður	110749	80,3	**80,6**	77,5	71,5
Flutningur	84638	77,9	**79,6**	76,5	72,3
Fjármál	77546	**66,8**	66,4	62,8	56,3
Umhverfi	62892	75,1	**76,3**	73,4	68,7
Vélar	52002	73,3	**73,8**	70,9	66,4

Orka og iðnaður	19343	76,3	**77,2**	75,1	71
Hættuleg efnasambönd	17052	83,8	**84**	81,1	75,2
Upplýsinga- og fjarskiptatækni	11708	65,9	**67**	65,1	60
Lyfjavörur	11637	75,9	**77,6**	75,9	71.9
Tækni og iðnaður	9650	81,8	**83,4**	80,9	78
Innkaup	7419	70,3	**72,2**	66,1	60,1
Innflytjendamál	6424	66,1	**66,8**	60,7	56,4
Varnarmál	6206	64,8	**65,6**	55	42,8
Félagsleg réttindi	6012	77,4	**80**	74,8	69,2
Hagkerfi	5504	54,1	**55,7**	48,4	41,9
Skattar	3478	55,2	**58,4**	48,2	41,1
Ýmislegt	2788	42,8	**44,9**	39,8	33
Hugverkaréttur	2499	62,6	**65,5**	59,7	51,9
Réttindi launafólks	2424	71,2	**74,8**	72,9	66,2
Vín	2048	65,6	**68,9**	63,9	58,5
Meðaltal		69,4	**70,9**	66,4	60,6

Munurinn á BLEU-mælikvarða á hráa grunnlíkaninu án nokkurra fínstillinga er birtur að neðan. Í stuttu máli er marktækur ávinningur í BLEU-gildum þegar líkanið er fínstillt fyrir hvert undirsvið (dálkurinn “Aðeins undirsvið”). Þetta er þó dýr nálgun, bæði hvað varðar forþjálfun og á ályktunartíma. Þess í stað má ná fram stærstum hluta þess ávinnings (m.v. almennt grunnlínulíkan) með fínstillingu einstaks líkans á EES-textum og veita auðkenningartóka undirsviðs sem hluta af inntaki þegar skjöl eru þýdd úr því undirsviði (dálkurinn “Merkt”).

Undirsvið	**Merkt**	**Aðeins undirsvið**	**EES**
Landbúnaður	8,8	9,1	6
Flutningur	5,6	7,3	4,2
Fjármál	10,5	10,1	6,5
Umhverfi	6,4	7,6	4,7
Vélar	6,9	7,4	4,5
Orka og iðnaður	5,3	6,2	4,1
Hættuleg efnasambönd	8,6	8,8	5,9
Upplýsinga- og fjarskiptatækni	5,9	7	5,1
Lyfjavörur	4	5,7	4
Tækni og iðnaður	3,8	5,4	2,9

Innkaup	10,2	12,1	6
Innflytjendamál	9,7	10,4	4,3
Varnarmál	22	22,8	12,2
Félagsleg réttindi	8,2	10,8	5,6
Hagkerfi	12,2	13,8	6,5
Skattar	14,1	17,3	7,1
Ýmislegt	9,8	11,9	6,8
Hugverkaréttur	10,7	13,6	7,8
Réttindi launafólks	5	8,6	6,7
Vín	7,1	10,4	5,4
Meðaltal	8,7	10,3	5,8

V4f – Opin þýðingarvél fyrir íslensku – Stofnanaheiti, nefndir, embætti o.s.frv.

Skýrsla: Stofnun Árna Magnússonar í íslenskum fræðum
Aðilar: Stofnun Árna Magnússonar í íslenskum fræðum

Markmið

Markmiðið er að safna þýðingum á nafneiningum úr stjórnsýslu til þess að bæta gæði þýðinga opinberra skjala og vefsíðna, en mun þó einnig nýtast á öðrum sviðum, eins og fréttum.

Varða 7 – lýsing

Safna listum þýðinga á nöfnum sem notuð eru til að vísa til samtaka og stofnana í stjórnsýslu eða hópum innan þeirra.

Afurðir

Markmiðum vörðunnar hefur verið náð og lista fleiri en 3.800 þýðinga á íslenskum og alþjóðlegum samtökum, fyrirtækjum og opinberum titlum hefur verið safnað og hann gefinn út á CLARIN á http://hdl.handle.net/20.500.12537/184

Skýrsla

Lista þýðinga var safnað úr ýmsum áttum, m.a. frá skrá yfir þýðingar stofnana og fyrirtækja sem gefin var út af Hagstofu Íslands árið 1989, vefsíðu ríkisstjórnar Íslands, utanríkisráðuneytinu og opinberum lénum íslenskra sveitarfélaga. Þýðingum var raðað á eftirfarandi hátt (sjá að neðan):

1	Íslenskt heiti	Önnur íslensk heiti	Íslensk skst.	Eldri ísl. skst.	Erlent tungumál	Erlent heiti	Önnur erlend heiti	Erlend skst.	Eldri erlendar skst.		Aths.
856	Safn Ásgríms Jónssonar					Ásgrímur Jónsson collection					
857	Safnahús Borgarfjarðar				EN	Borgarnes Museum					
858	Safnahús Vestmannaeyja				EN	Vestmannaeyjar Culture House					
859	Safnahúsið				EN	The Culture House					
860	Safnasafnið				EN	The Icelandic Folk and Outsider Art Museum					
861	Sálfræðingafélag Íslands				EN	Icelandic Psychological Association					
862	Samband félaga flutningaverkamanna í Evrópu				EN	European Transport Workers' Federation		ETF			
863	Samband garðyrkjubænda		SG		EN	Icelandic Association of Horticulture Producers					
864	Samband íslenskra kvikmyndaframleiðenda		SÍK		EN	Association of Icelandic Film Producers					
865	Samband íslenskra myndlistarmanna		SÍM		EN	Association of Icelandic Visual Artists					
866	Samband íslenskra sveitarfélaga		SÍS		EN	Icelandic Association of Local Authorities					
867	Samband samgönguráðherra í Evrópu				EN	European Conference of Ministers of Transport		ECMT			
868	Samband tónskálda og eigenda flutningsrétta		STEF		EN	Composers' Rights Society of Iceland					
869	Sameiginleg rannsóknarmiðstöð	Sameiginleg rannsóknarmiðstöð framkvæmdastjórnarinnar			EN	Joint Research Centre	Joint Research Centre of the European Commission	JRC			
870	Sameiginlegt evrópskt fyrirtæki um alþjóðlegar tilraunir með kjarnaofn og þróun samrunaorku				EN	European Joint Undertaking for ITER and the Development of Fusion Energy	Fusion for Energy				
871	Sameinað sílikon hf.	Sameinað sílikon			EN	United Silicon					
872	Sameinuðu þjóðirnar		SÞ		EN	United Nations		UN			
873	Sameinuðu þjóðirnar og sérstofnanir þeirra	SÞ og sérstofnanir þeirra			EN	United Nations and the Specialized Agencies	UN and the Specialized Agencies				

V5 – Viðmót fyrir þýðingarvél

Skýrsla: Miðeind ehf
Aðilar: Miðeind ehf

Markmið

Markmiðið á þriðja ári er að tryggja langlífi vélþýðingarlausnanna, umfram máltækniáætlunina, með því að skjala og pakka biðlara fyrir þýðingarlíkanið fyrir beina notkun hönnuða og rannsakenda í gegnum skipanalínuviðmót og sem Python-einingu. Á sama tíma halda áfram að styðja veflægt þýðingarkerfi (viðmiðunarútfærsla sem er opinber á velthyding.is) með áframhaldandi umbótum, sem tól fyrir almenning til að nota við styttri þýðingarverkefni og sem sýnidæmi um getu kerfisins.

Varða 7 – lýsing

Viðvarandi uppfærslur á veflæga viðmótinu, með viðbótum úr öðrum undirverkefnum.

Afurðir

Markmiðum undirverkefnisins hefur verið náð. Uppfærslur og breytingar hafa verið gerðar á vélbúnaði í notkun og hugbúnaði eins og lýst er í skýrslunni að neðan.

Skýrsla

Vefþýðingarkerfið fyrir almenning á https://velthyding.is hefur verið fært yfir á skjákortsþjón fyrir aukin afköst. Þetta er sérstaklega mikilvægt fyrir samvinnu við utanríkisráðuneytið.

Í vefviðmóti þýðingarmats hafa verið gerðar endurbætur fyrir aðferðir samanburðar þar sem einföldu vali milli tveggja þýðinga hefur verið skipt út fyrir stiku fyrir nákvæmari samanburð.

Villumeðhöndlun hefur verið bætt, sérstaklega þegar tenging rofnar. Aðrir hlutar viðmótsins hafa verið bættir.

Að lokum er frumútgáfa afurðar fyrir M9: pakkaður biðlari fyrir staðbundna þýðingu hefur verið hannaður og gefinn út á PyPI í pakkanum `greynirseq`. Dæmi um notkun pakkans má sjá að neðan. (Athugið að keyrsla pípunnar sækir nokkur gígabæti af þýðingarlíkönum.)

```
$ pip install greynirseq

# For en->is translation
$ echo "This is an awesome test that shows how to use a pretrained translation
```

```
model." | greynirseq translate --source-lang en --target-lang is

Þetta er æðislegt próf sem sýnir hvernig nota má forprófað þýðingarlíkan.

# For is->en translation
$ echo "Þetta er æðislegt próf sem sýnir hvernig nota má forprófað
þýðingarlíkan." | greynirseq translate --source-lang is --target-lang en

This is an awesome test that shows how a pre-tested translation model can be
used.
```

Þegar skjalaþýðingar verða tilbúnar munum við uppfæra pakkann til að styðja nýju líkönin.

V6 – Grunnlína fyrir opnar margmála þýðingar yfir á og frá íslensku

Skýrsla: Miðeind ehf
Aðilar: Miðeind ehf

Markmið

Markmið þessa undirverkefnis er að koma á fót og sýna almenna skilvirkni þeirra aðferða sem þróaðar hafa verið fyrir þýðingar milli íslensku og ensku, með því að þýða margmála líkön til að þýða milli íslensku og annarra tungumála en ensku. Mörg gagnasöfn eru til í þessum tilgangi, frá sömu gagngjöfum og voru nýttir við þýðingar milli ensku og íslensku. Þó við spáum því að ekki náist sömu gæði og fyrir ensku, þá mun þetta verkefni tryggja að þekkingin og aðferðirnar sem þróaðar eru innan Máltækniáætlunar hafi eins mikil áhrif og mögulegt er á framtíð vélþýðingarverkefna á og frá íslensku. Afurðirnar verða stökkpallur fyrir slík framtíðarverkefni og þróun fleiri tungumálapara til viðbótar. Grunnvinnan sem innt er af hendi í þessu undirverkefni gæti einnig liðkað fyrir framtíðarfjármögnun máltækniverkefna í gegnum ESB/EES-styrki.

Sem dæmi, og líklega forgangsmál, myndi opið vélþýðingarkerfi milli íslensku og pólsku hafa mikil áhrif og mikinn samfélagslegan ávinning, þar sem nærri 10% vinnuafls á Íslandi í dag hefur pólsku að móðurmáli.

Varða 7 – lýsing

Samhliða og einmála gögnum safnað til þjálfunar og gæðamats.

Afurðir

Markmiðum þessa undirverkefnis hefur verið náð. Samhliða og einmála gögnum hefur verið safnað eins og skýrt er að neðan.

Skýrsla

Samhliða gögnum hefur verið safnað til að meta og þjálfa margmála þýðingarkerfi. Einmála gögnunum hefur verið safnað til að búa til bakþýðingar og þjálfa undirliggjandi (margmála) mállíkan.

Þrátt fyrir að við teljum upp mörg tungumál og gagnasett að neðan er þetta metnaðarfullt verkefni og við gætum þurft að takmarka rammann með undirsetti tungumála fyrir komandi vörður.

Margmála líkan

Við höfum farið yfir fræði um margmála líkön og teljum DeltaLM (Ma et al. 2021)[7] vera ákjósanlegan kost fyrir tilraunir okkar. Þessi líkanahögun getur notað **bæði einmála og margmála gögn í forþjálfun.** Þetta gerir DeltaLM kleift að ná betri árangri en mBART (fyrri grunnhögun okkar) með helmingi minni líkanastærð. DeltaLM nær einnig góðum árangri á ýmsum öðrum verkum þvert á tungumál, sem gæti greitt fyrir áhugaverðum möguleikum síðar. Við stefnum á að nota DeltaLM-líkanið sem aðgengilegt er almenningi, sem hefur þegar verið þjálfað á 100 tungumálum, m.a. ensku, íslensku og pólsku, og forþjálfa það síðan á gagnasettunum sem lýst er að neðan. Síðan munum við fínstilla það fyrir þýðingu milli ýmissa tungumála, m.a. íslensk-pólska parið. DeltaLM-líkanið er þjálfað á verkefnum spilltra spanna (e. span corruption tasks) og spilltra þýðinga (e. translation corruption tasks) eins og sýnt er í dæmunum að neðan (frá upprunalegu greininni).

(a) Span corruption task

(b) Translation span corruption task

Val tungumála

Við höfum íhugað hvaða tungumál (önnur en íslensku) forþjálfun og margmála þýðingar okkar skulu innihalda. Tungumálin sem við íhugum eru réttlætt eftir fjórum viðmiðum:

a) Líkindi við íslensku, þ.e. Norðurlandamál
   i) danska (da)
   ii) færeyska (fo)
   iii) norska (no, nn, nb)
   iv) sænska (sv)
b) Algeng tungumál, mikið magn gagna aðgengilegt
   i) enska (en)
   ii) þýska (de)
   iii) spænska (es)
   iv) franska (fr)

[7] https://arxiv.org/abs/2106.13736

c) Tengd tungumál með sérstaka tengingu við íslenskt samfélag[8]
   i) pólska (pl)
   ii) litáíska (lt)
d) Tungumál skyld c) sem gætu reynst nytsamleg ef kemur til skorts á gögnum
   i) rússneska (ru)
   ii) tékkneska (cs)
   iii) slóvakíska (sk)
   iv) lettneska (lv)

Samhliða gögn

Samhliða gögn eru sótt frá JW300, WMT, Wikimatrix, TildeMODEL, Flores, EMEA, Europarl og CCMatrix. Samhliða gögnin eru ekki alltaf í boði milli allra tungumálapara, eins og sýnt er annaðhort með röðum eða dálkum sem vantar eða auðum reitum í töflunni að neðan. Flores-gögnin og WMT-gögnin eru notuð fyrir prófanir. JW300-,[9] Wikimatrix-, TildeMODEL-, Europarl-, EMEA- og CCMatrix-gögnin eru sótt frá OPUS-safninu.[10] WMT-gögnin eru málheildir notaðar í keppnum í vélþýðingum.

Gögnunum hefur verið safnað, en við eigum eftir að sjá að hvaða leyti óáreiðanlegri gögn eins og CCMatri- gögnin verða notuð. Fyrstu tilraunir sýna að þau má nota, a.m.k. þegar þau eru dregin niður eða síuð frekar en hágæða málheildir.

Flores[11]
Flores-gagnasettið inniheldur 3.001 setningu á 101 tungumáli, er sérstaklega gert til að árangursmæla vélþýðingarlíkön og er hágæða. Við munum nota þessi gögn fyrir prófanir og mat. Því miður inniheldur það enga færeysku. Þetta gagnasett hentar vel til að meta árangur þýðinga milli allra tungumála.

WMT[12]
Þessi gagnasett eru notuð fyrir mat í keppnum í vélþýðingum. Við notum þau í sama tilgangi. Þau eru að mestu ætluð ensku, þ.e. þau mæla árangur milli ensku og einhvers annars tungumáls.

EMEA[13]
EMEA er samhliða málheild sem samanstendur af PDF-skjölum frá Lyfjastofnun Evrópu. Hún inniheldur hágæða samraðanir u.þ.b. milljón setningapara fyrir öll tungumál sem við skoðum, utan færeysku, rússnesku og, furðulega, norsku.

---

[8] Mörg önnur tungumál eins og arabíska, tælenska og víetnamska gætu talist áhugaverðir kostir vegna innflytjenda en við takmörkum okkur við þessi tungumál þar sem þau eru skyldari og hafa flesta mælendur.
[9] JW300-málheildin er ekki lengur aðgengileg en við notum afrit af þeim hluta hennar sem við höfum
[10] https://opus.nlpl.eu/index.php
[11] https://github.com/facebookresearch/flores
[12] https://www.statmt.org/
[13] https://opus.nlpl.eu/EMEA.php

Europarl[14]
Europarl er samhliða málheild dregin úr vefsíðu Evrópuþingsins af Philipp Koehn (University of Edinburgh). Gagnasettið inniheldur 0,5–2 milljónir paraðra setninga milli allra tungumála nema norsku, íslensku, rússnesku og færeysku.

Wikimatrix-gögn[15]
Þetta eru gögn sem safnað hefur verið frá mismunandi Wikipedia-síðum. Gögnin innihalda þó töluvert af slæmum þýðingarpörum, eins og skýrt er frá í *Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets* (Kreutzer et al., 2021).[16] Á mörgum tungumálum er meira en helmingur samraðaðra setninga merktur sem rangur. Þetta er eina gagnasettið með texta merktan sem færeyskur, en við skoðun á samröðunum virðist það ekki passa. Íslensku samraðanirnar passa ekki heldur.

CCmatrix-gögn[17]
Þetta eru gögn sem hafa verið sótt og samröðuð úr Common Crawl, stór demba af góðum hluta internetsins. Það er líklegt að þau innihaldi töluvert magn vélþýðinga, eins og önnur gagnasett sem sótt eru af vefnum. Við skoðun íslensku gagnanna sést að þessar samraðanir eru mikið betri en Wikimatrix. Það eru einhverjar vélþýðingar í gögnunum, en heilt yfir virðast samraðanirnar að mestu réttar. Tafla 1 að neðan sýnir fjölda samraðaðra setninga á hvert tungumálapar.

tungumál	cs	da	de	en	es	fr	is	la	lt	lv	no	pl	ru	sk	sv
cs		5.3M	32.9M	56.3M	33.4M	34.2M			7.5M	5.3M	8.0M	31.8M	28.0M	37.4M	22.6M
da	5.3M		28.3M	52.3M	24.1M	24.3M	1.3M		5.8M	4.2M	17.7M	5.9M	8.5M	7.9M	23.1M
de	32.9M	28.3M		247.5M	112.4M	136.5M	3.4M		13.1M	9.6M	23.8M	44.6M	45.9M	25.0M	43.9M
en	56.3M	52.3M	247.5M		409.1M	328.6M	8.7M	1.1M	23.3M	16.7M	47.8M	74.1M	139.9M	38.1M	77.0M
es	33.4M	24.1M	112.4M	409.1M		285.9M		0.6M	23.7M	9.2M	20.6M	43.6M	55.9M	22.5M	38.0M
fr	34.2M	24.3M	136.5M	328.6M	285.9M			0.6M	24.6M	9.3M	20.7M	44.0M	48.3M	22.7M	36.6M
is		1.3M	3.4M	8.7M							1.3M				1.7M
la				1.1M	0.6M	0.6M									
lt	7.5M	5.8M	13.1M	23.3M	23.7M	24.6M				5.3M	4.6M	9.2M	14.8M	7.0M	6.5M
lv	5.3M	4.2M	9.6M	16.7M	9.2M	9.3M			5.3M		3.4M	6.0M	10.1M	5.1M	4.8M
no	8.0M	17.7M	23.8M	47.8M	20.6M	20.7M	1.3M		4.6M	3.4M		5.0M	7.5M	6.7M	23.3M
pl	31.8M	5.9M	44.6M	74.1M	43.6M	44.0M			9.2M	6.0M	5.0M		36.6M	26.0M	27.0M
ru	28.0M	8.5M	45.9M	139.9	55.9M	48.3M			14.8M	10.1M	7.5M	36.6M		21.7M	21.4M

[14] https://opus.nlpl.eu/Europarl.php
[15] https://ai.facebook.com/blog/wikimatrix/
[16] https://arxiv.org/abs/2103.12028
[17] https://ai.facebook.com/blog/ccmatrix-a-billion-scale-bitext-data-set-for-training-translation-models/

				M											
sk	37.4M	7.9M	25.0M	38.1M	22.5M	22.7M			7.0M	5.1M	6.7M	26.0M	21.7M		12.1M
sv	22.6M	23.1M	43.9M	77.0M	38.0M	36.6M	1.7M		6.5M	4.8M	23.3M	27.0M	21.4M	12.1M	

Tafla 1. *Fjöldi samraðaðra CCmatrix setninga í tungumálapörunum sem við skoðum*

TildeMODEL[18]
TildeMODEL-gögnunum er safnað frá vefsíðum á almennum markaði í Evrópu. Þetta eru hágæða samraðanir. Af þeim tungumálum sem við skoðum, fyrir öll tungumál nema tékknesku, dönsku, rússnesku og færeysku, eru 0,3–5,1 milljón samraðana milli allra tungumála og ensku, frönsku og þýsku.

Biblían
Þetta eru textar sem safnað er úr trúarlega textanum, einnig hágæða, u.þ.b. 30.000 setningapör. Því miður er færeyska ekki til staðar í gögnunum.

JW300
JW300 er hágæða safn af samröðuðum texta frá vefsíðu Votta Jehóva. Því miður er þetta gagnasett ekki lengur opinberlega aðgengilegt. Ef það verður aðgengilegt aftur á komandi mánuðum mun því vera bætt við. Eins og er höfum við aðeins aðgengi að hluta gagnanna, sem innihalda íslensku – ensku, norsku – ensku og ensku – pólsku.

Einmála gögn

Einmála gögn eru aðgengileg frá skröpuðum heimildum eins og mC4 og C4, OSCAR. Aðrar heimildir ákveðinna tungumála eru að mestu ritskoðaðar og af hærri gæðum og þær eru ákjósanlegri ef ekki á að framkvæma mikla hreinsun. Þó að þessar skröpuðu heimildir séu gjarnan notaðar fundum við, og aðrir[19], að slíkar heimildir reynast oft óáreiðanlegar fyrir tungumál með litlar heimildir sem hafa mikið hlutfall rangt merktra texta.

Skröpuð gögn

Við munum nota skröpuð gögn eftir þörfum, aðallega fyrir tungumál önnur er Norðurlandamálin, og beita jusText-[20] og fastText-tungumálasíunum[21] sem aukaráðstöfun í síun.

Oscar
Oscar-gagnasettið[22] er aðgengilegt í gegnum Hugging Face-gagnasettagáttina.

C4 og mC4
Enska C4-[23] og margmála C4-gagnasettin[24] eru einnig aðgengileg á Hugging Face.

---

[18] https://opus.nlpl.eu/TildeMODEL.php
[19] https://arxiv.org/abs/2103.12028
[20] https://pypi.org/project/jusText/
[21] https://fasttext.cc/docs/en/language-identification.html
[22] https://huggingface.co/datasets/oscar
[23] https://huggingface.co/datasets/c4
[24] https://huggingface.co/datasets/mc4

Ritskoðuð gögn

Við munum leggja áherslu á notkun gagnasetta sem aðeins eru ritskoðuð fyrir Norðurlandamálin til að takmarka umfangið.

Íslenska

Við reiðum okkur á íslensku gagnasettin, the Icelandic Common Crawl Corpus og Risamálheildina ásamt nokkrum smærri gagnasettum eins og kemur fram í töflunni að neðan.

Gagnasett	Stærð	Tókar
IGC (ritstjórnartexti)	8,2GB	1.388 milljón
IC3 (hreinsuð vefskröpun)	4,9GB	824 milljón
Ritgerðir nemenda	2,2GB	367 milljón
Greinar frá Greynir News	456MB	76 milljón
Læknisfræðibókasafn	33MB	5,2 milljón
Opnar íslenskar rafbækur	14MB	2,6 milljón
Íslendingasögur	9MB	1,7 milljón
Samtals	15,8GB	2.664 milljón

Danska

The Danish Gigaword Corpus[25] inniheldur fjölbreytt ritskoðuð gögn, samtals yfir milljarð orða.

Færeyska

Þó að fáar færeyskar málheildir séu til notum við það sem við finnum á internetinu, sem er m.a. eftirfarandi:

- Faroese Text Collection[26] – 1.100.000 orð
- Corpuseye Faroese corpus (Sosialurin newspaper) [27] – 200.000 orð
- The Faroese Parsed Historical Corpus[28] – 50.000 orð
- Biblían er aðgengileg á færeysku.[29]

Þessi gagnasett eru lítil en öll færeysk gögn eru vel þegin.

Norska

[25] https://gigaword.dk/
[26] https://spraakbanken.gu.se/en/resources/fts
[27] https://live.european-language-grid.eu/catalogue/corpus/13581
[28] http://hdl.handle.net/20.500.12537/92
[29] https://www.bible.com/bible/374/GEN.1.FB

Stórt norskt gagnasett er aðgengilegt, the Norwegian Colossal Corpus.[30] Það er safn smærri gagnasetta sem innihalda mörg mismunandi tungumál, en aðallega norsku. Það inniheldur mikið magn dönsku (957 milljón orð), en flestir textarnir eru úr norskum dagblöðum.

Sænska
The Swedish Gigaword Corpus[31] inniheldur ýmsa sænska texta frá fréttum, Wikipedia og bókum.

Enska
Books3-málheildin[32] inniheldur 160GB bóka og verður notuð að hluta.

Önnur tungumál
Spænsku, frönsku, pólsku og rússnesku má finna í hundruðum milljóna skala í mC4-málheildinni og valda okkur ekki áhyggjum.

Fyrir hin tungumálin höfum við gert smá rannsókn. Einhverjar heimildir eru til (til dæmis í tilfelli pólsku eða litáísku) en þær eru ekki auðveldlega aðgengilegar þannig að við veljum að nota þær heimildir á netinu sem lýst er að ofan.

Til að takmarka umfang verkefnisins látum við einnig nægja að nota mC4- og Oscar-málheildirnar fyrir þessi tungumál, þ.e. fyrir litáísku, tékknesku, slóvakísku og lettnesku.

[30] https://huggingface.co/datasets/NbAiLab/NCC
[31] https://spraakbanken.gu.se/en/resources/gigaword
[32] https://github.com/soskek/bookcorpus

L2 – Sérhæfðar villumálheildir

Skýrsla: Háskóli Íslands
Aðilar: Háskóli Íslands

Markmið

Safna textum frá tilteknum undirhópum innan þess hóps fólks sem notar íslensku sem annað mál, og stækka lesblindumálheildina. Stærðarmarkmiðum annars árs hefur verið náð en vinna við þau hefur sýnt að málheildirnar þarf að stækka enn frekar.

Varða 7 – lýsing

Stórir innflytjendahópar skilgreindir og söfnun texta frá þessum hópum hafin.

Afurðir

Öllum markmiðum hefur verið náð, stórir hópar innflytjenda hafa verið skilgreindir og söfnun texta frá þeim hópum er hafin. Við höfum þó ekki enn fengið texta frá Litáum. Við erum að vinna í því að koma á sambandi við þann hóp, svipað því sambandi sem við höfum við mælendur frá Póllandi, og við búumst við textum frá þeim brátt.

Skýrsla

Þrír stórir hópar innflytjenda voru skilgreindir: fólk frá Póllandi, Litáen og Filippseyjum. Hóparnir voru valdir eftir upplýsingum frá Hagstofu Íslands (https://hagstofa.is/utgafur/frettasafn/mannfjoldi/mannfjoldi-eftir-bakgrunni-1-januar-2021/) um hlutfall innflytjenda á Íslandi eftir þjóðerni. Þrír stærstu innflytjendahóparnir árið 2021 eru frá Póllandi (35,9%), Litáen (5,7%) og Filippseyjum (3,7%).

Þar sem fólk frá Póllandi er langfjölmennasti hópur innflytjenda leggjum við til að byrja með áherslu á söfnun texta frá þessum hópi. Við náðum góðum tengslum við *Retor*, kennslustofnun sem kennir íslensku fyrir innflytjendur og heldur kennslustundir sem eru sérstaklega ætlaðar pólsku fólki á fimm stigum, eftir kunnáttu þeirra á íslensku. Frá því á vorönn þessa árs hefur söfnunarátak okkar verið hluti af námsskrá þeirra, og því mun fjöldi texta aukast eftir því sem líður á önnina. Þetta samstarf mun vonandi veita okkur marga texta. Við höfum auglýst í nágrenni við Háskóla Íslands auk auglýsinga á samfélagsmiðlum (hópar fyrir nemendur íslensku á Facebook o.s.frv.), og höfum fengið mörg jákvæð viðbrögð. Við höfum einnig verið í sambandi við kennara íslensku sem annars máls við Háskóla Íslands og hafa þeir kynnt söfnunina fyrir nemendum sínum. Markmiðið nú er að koma á tengslum við fólk frá Litáen og Filippseyjum og er sú vinna í gangi.

Eins og er höfum við 84 texta alls í villumálheild annars máls, af þeim eru 7 frá fólki frá Póllandi og 12 frá fólki frá Filippseyjum, en við höfum enn ekki fengið texta frá fólki frá Litáen. Hinir textarnir eru skrifaðir af fólki með önnur tungumál, sem kemur til vegna þess að upprunalega söfnuðum við textum frá öllum annarsmálshöfum. Af þessum 84 útgefnum textum í villumálheild annars máls hafa 76 textar verið lesnir yfir tvisvar og 61 texti lesinn yfir þrisvar. Þetta var gert til að tryggja að villur og leiðréttingar væru réttar og samræmdar.

Taflan að neðan sýnir villutíðni í villumálheild annars máls í heild sinni, í textum frá Pólverjum og í textum frá Filippseyingum. Athugið af textar frá þeim tveimur síðarnefndu eru með í villutíðni annars máls (L2).

	L2	Pólskir	Filippseyskir
Fjöldi endurskoðana	15.369	467	872
Fjöldi villna	22.423	607	1.479
Fjöldi orða	146.267	4.086	5.319
Villur á 1.000 orð	153,30	148,56	278,06

L7 – Kerfisbundin þróun málrýnis

Skýrsla: Miðeind
Aðilar: Miðeind, Háskóli Íslands

Markmið

Verkefnið má skilgreina gróflega út frá þremur þáttum: A) villuflokkar, B) notendahópar, og C) villumeðhöndlun/prófunaraðferðir.

Í vinnu þriðja árs verður C) í brenndiepli. Gerðar verða tilraunir fyrir mismunandi notendahópa í undirverkefni L14. Til að fá nytsamlegustu og heildstæðustu afurðina fyrir stærsta notendahópinn fyrir lok áætlunarinnar er markmiðið á þriðja ári að bæta leiðréttingar, tillögur og lýsingar fyrir villuflokka sem áður hafa verið meðhöndlaðir í almennu villumálheildinni. Slíkar upplýsingar eru notendum mjög dýrmætar.

Tveir almennir notendahópar eru skilgreindir; fjölmiðlafyrirtæki, og höfundar fræðilegra ritgerða, rannsóknargreina o.s.frv. Á öðru ári ljúkum við notendaprófunum fyrir fjölmiðlafyrirtæki. Til að mæta kröfum annarra hópa verða gerðar notkunarprófanir meðal háskólanema sem eru að skrifa lokaritgerðir.

Varða 7 – lýsing

F-gildi villuleiðréttinga fyrir samhengisháðar og samhengisfrjálsar villur hefur verið bætt um sem nemur 3 prósentustigum.

Afurðir

Öllum markmiðum um afurðir var náð.

- F-gildi villuleiðréttingar fyrir samhengisfrjálsar villur og samhengisháðar villur er nú 43,63%, sem er bæting um 3,23 prósentustig.
- Ný útgáfa af GreynirCorrect-pakkanum er aðgengileg hjá CLARIN á http://hdl.handle.net/20.500.12537/174, sem og á GitHub á https://github.com/mideind/GreynirCorrect/releases/tag/3.2.1.
- BinPackage var einnig uppfærður og er aðgengilegur hjá CLARIN á http://hdl.handle.net/20.500.12537/177.
- Málrýnirinn er aðgengilegur á vefnum á https://yfirlestur.is/.

Skýrsla

1. Yfirlit

Áherslan fyrir M7 var að bæta villuleiðréttingu. Við lögðum einnig grunn að frekari stuðningi við notendur, til að vera undirbúin þörfum undirverkefna L8 og L9, ásamt almennri þörfum.

2. Mat og niðurstöður

2.1 Heildarniðurstöður

Taflan að neðan sýnir tíðni og F0.5-gildi fyrir greiningu og leiðréttingu villna fyrir samsvarandi yfirflokka, bæði fyrir M6 og M7. Athugið að aðeins tveir yfirflokkar eru skilgreindir sem innan ramma eins og er, *málfræði (e. grammar)* og *réttritun (e. orthography)*, þar sem aðrir yfirflokkar hafa að gera með óljósari vandamál. Eins og sést hefur F-gildi villuleiðréttinga fyrir réttritun, yfirflokk sem inniheldur samhengisfrjálsar og samhengisháðar villur, hækkað um 3,23 prósentustig, sem uppfyllir markmið M7. Að auki hefur F-gildi villureiðrétinga fyrir málfræðivillur hækkað um 3,4 prósentustig, og heildar F-gildi villuleiðréttinga hefur hækkað um 1,95 prósentustig.

Yfirflokkar	**Tíðni**	**M6-greining**	**M7-greining**	**M6-leiðrétting**	**M7-leiðrétting**
coherence	4	8,33	8,33	8,33	8,33
grammar	182	19,89	24,64	8,98	12,38
orthography	1182	61,5	62,32	40,4	43,63
other	256	19,74	83,33	7,35	0
style	4	41,67	0	0	0
vocabulary	34	44,09	45,05	21,11	15,25
Samtals	1662	49,98	50,82	31,3	33,25

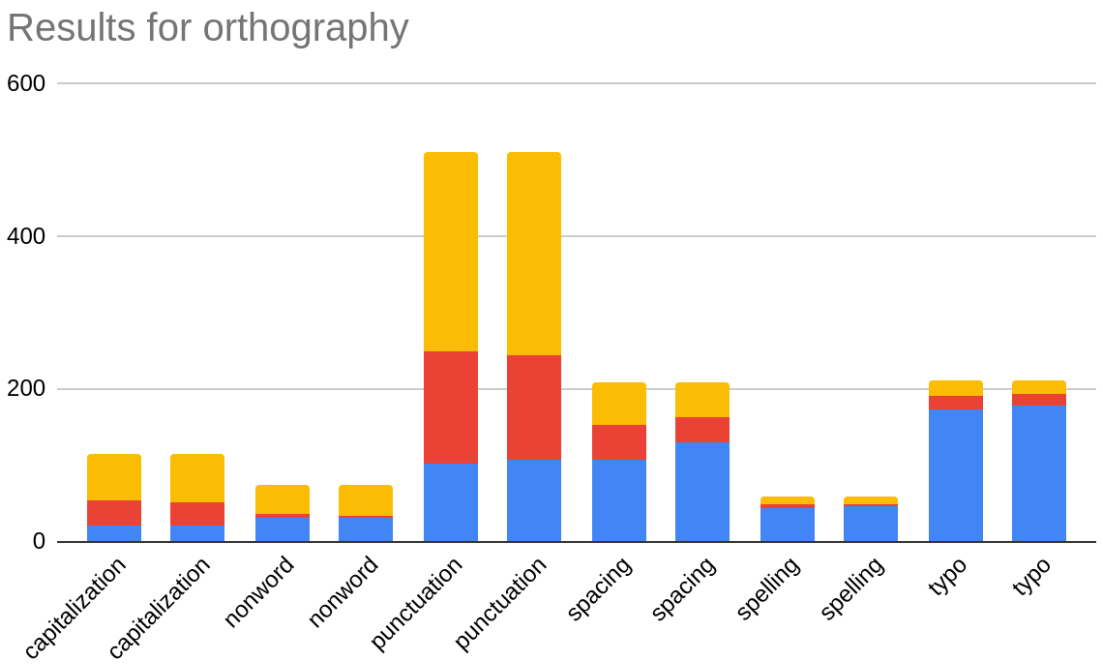
2.2 Niðurstöður fyrir réttritun

Þar sem áherslan fyrir M7 var að bæta villuleiðréttingu samhengisfrjálsra og samhengisháðra villna lítum við nánar á yfirflokkinn *réttritun* og undirflokka þess flokks. *Réttritun* er langstærsti flokkurinn í IceEC. Taflan að neðan sýnir niðurstöður villugreiningar fyrir hvern undirflokk. Athugið að undirflokkurinn *foreign* kemur ekki fyrir í prófunarsettinu og er því ekki hafður með í niðurstöðum.

Undirflokkar	**Tíðni**	**M6-greining**	**M7-greining**	**M6-leiðrétting**	**M7-leiðrétting**
capitalization	114	46,36	44,35	17,83	17,83
nonword	74	46,97	44,9	41,57	40,89
punctuation	511	48,56	47,56	20,04	21,19
spacing	209	72,55	77,62	51,16	62,28
spelling	60	79,51	80,95	71,27	74,8

typo	210	91,37	92,66	83,02	84,33

Súluritið að neðan sýnir niðurstöður fyrir alla undirflokka yfirflokksins *réttritun*. Blái liturinn táknar villuleiðréttingu, þ.e. fjölda villna sem við greinum og leiðréttum rétt. Rauði liturinn táknar villur sem við greinum en leiðréttum ekki. Þetta þýðir að bláu og rauðu súlurnar saman tákna villugreiningu. Að lokum táknar guli liturinn villur sem voru hvorki greindar né leiðréttar. Saman tákna bláu, rauðu og gulu súlurnar tíðnina fyrir hvern flokk. Hver flokkur hefur tvær súlur. Fyrri táknar niðurstöður við M6, seinni núverandi niðurstöður.



Undirflokkurinn *punctuation* er langstærstur, og hefur notið góðs af breytingum á tilreiðaranum eins og rætt er í 5. kafla, þó að áherslan hafi ekki verið á greinarmerkjavillur. Ný villugögn sem rætt er um í kafla 3.1 hafa leitt til aukningar bæði á greiningu og leiðréttingu fyrir undirflokkana *spacing*, *spelling*, og *typo*, en *nonword* og *capitalization* hafa minnkað aðeins.

2.3 Umræða

Nákvæmni kerfisins er háð hreiðruðum villum í villumálheildinni. Þetta eru villur sem eru innan annarra villna, t.d. ‘They are newer not happy’ > ‘They are always happy’. Önnur villan tengist leiðréttingunni ‘never not’ > ‘always’, sem er orðalagsvilla, og eru þær í IceEC. Hin villuleiðréttingin er ‘newer’ > ‘never’. Kerfið gæti merkt rétt villuna ‘newer’ en ekki ‘never not’, og því týnist önnur villumerkingin. Þetta virðist lækka nákvæmni kerfisins sem er þó í raun hærri. Hvernig þessar hreiðruðu villur eru merktar í villumálheildinni verður endurskoðað fyrir M8.

Einnig skal taka fram að kerfið hefur komist á skrið meðal almennings, og við höfum fengið góða endurgjöf frá notendum.

3. Endurbætur í villuleiðréttingu

3.1 Réttritunarvillur (e. orthography errors)

Nýjum villugögnum á tókastigi frá *Ritmyndum* (hluti af undirverkefni G4) var bætt við kerfið og þau meðhöndluð. Nýju gögnin innihalda upprunalegu villuna, leiðrétta gildið, uppflettimyndir, tög, og upplýsingar um eðli villunnar. Þetta gaf góða hækkun í árangri, bæði fyrir greiningu og leiðréttingu.

Ýmsar hástafavillur voru einnig bættar, villur þar sem hástöfun vantaði eða var ofaukið, í bæði samhengisháðum og samhengisfrjálsum villum. Samhengisfrjálsar villur voru bættar með því að yfirfara orðalista og bæta orðum í BinPackage þar sem þurfti. Samhengisháðar villur voru villur þar sem hástöfun var alhæfð vegna samhengisfrjálsra aðferða. Villumerkingu fyrir hástöfun er bætt við á tókastigi, og meðhöndlun var bætt við svo að þessi merking er fjarlægð í ákveðnum tilfellum eftir að setning hefur verið þáttuð.

3.2 Málfræðivillur (e. grammar errors)

Til að bæta nákvæmni og villuleiðréttingu yfirfórum við allar aðgerðir sem tengjast málfræðivillum. Þetta nær yfir reglur setningarliða fyrir villumálfræðina, og öll leitarmynstur fyrir setningafræðitrén. Við gefum leiðréttingargildi þar sem hægt er, og settum takmark á ramma villunnar. Í einhverjum tilfellum merktum við heilu undirtrén sem villur þegar aðeins hefði átt að merkja eitt orð. Í einhverjum tilfellum ætti að eyða öllum setningarliðnum og í örfáum tilfellum var ekkert rétt gildi í boði. Öll slík tilfelli voru skjöluð í kóðanum og meðhöndluð á skilmerkilega mismunandi hátt. Að auki voru tillögur fyrir þessar villumerkingar bættar í undirbúningi fyrir M8.

Þar sem notendaprófanir sýndu nokkuð hátt hlutfall rangra jákvæðra var aðgerðum sem finna tvöfalda ákveðni í nafnliðum, þ.e. nafnliðum sem innihalda bæði nafnorð með greini og ákveðnilið, gefinn takmarkaðri rammi. Meðhöndlun var einnig bætt við fyrir samtenginguna *né* sem getur aðeins birst í tvískiptu margorða samtengingunni *hvorki né*, en ekki stök sem undirskipuð samtenging.

3.3 Aðrar villur

Nýju málsniðsupplýsingarnar í BinPackage, sem rætt er um frekar í hluta 5, gerðu okkur kleift að merkja stílsvillur, þ.e. tilfelli þar sem orðið telst úrelt, gamaldags eða einfaldlega rangt. Að auki innihalda bannorðavillur nú betur skilgreind leiðréttingar-/tillögugildi, sem bætir villuleiðréttingu.

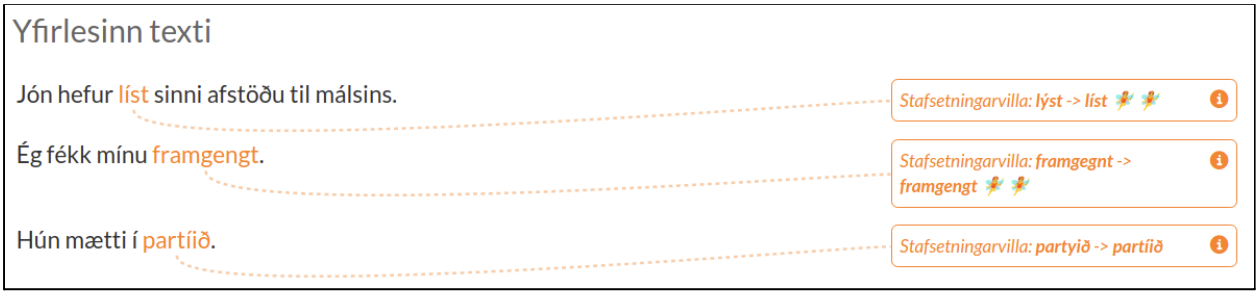
4. Endurbætur í tillögum og lýsingum

Í undirbúningi fyrir vörðu M8 byrjuðum við að bæta tillögur og lýsingar villna.

Við bættum við sérstökum villuskilaboðum fyrir alla flokka úr Ritmyndum, byggt á skjölun hvers og eins. Skilaboðin eru geymd sem f-strengir (e. f-strings), sem leyfir strengi sem tákna upprunalegt gildi, leiðrétt gildi, og uppflettimyndina, sem má setja inn á ákveðnum stöðum.

Margir villuflokkarnir í Ritmyndum eru tengdir við sérstakar reglur í íslenskum málstaðli í skjöluninni. Við bættum þessum strengjum við gögnin fyrir hvern villuflokk í Ritmyndum, og útfærðum aðgerð til að sækja rétta vefslóð frá hverjum streng. Þetta leyfir okkur að bæta hlekkjum á tilvísanir fyrir hverja villu í vefviðmótinu. Að neðan er dæmi um slík villuskilaboð. Hlekkurinn er táknaður sem emoji, eins og er álfur (e. fairy).

Við hlökkum til þess að bæta fleiri slíkum tilvísunum við villuflokka eða ákveðnar villur til að leiðbeina notandanum, eftir því sem fleiri tilvísunum í fleiri heimildir er bætt í Ritmyndir.



Mynd 1. Hvert álfstákn er tilvísunarhlekkur,
sem dæmi https://ritreglur.arnastofnun.is/#6.2.

5. Aðrar endurbætur

BinPackage (undirverkefni G4b) var uppfærður til að innihalda nýjar upplýsingar úr BÍN (undirverkefni G4). Nýr reitur, málsnið (e. register), gerir okkur kleift að forgangsraða færslum í leitaraðferðum; færslur merktar sem úreldar/gamaldags/rangt málsnið eru nú hafðar síðastar í lista hugsanlegra skilgreininga. Annar reitur, millivísun (e. cross-reference), inniheldur auðkenni (e. id) annarrar færslu, í flestum tilfellum ákjósanlegri. Við útfærðum nýja aðgerð sem leit eftir auðkenni. Þetta gerir okkur kleift að sækja millivísaða færslu sem rétt gildi fyrir villumerkingu. Málfræðireglur í þáttaranum voru uppfærðar til að nota þessar nýju upplýsingar.

Þó nokkur sérstök tilfelli voru löguð sem sneru að kerfishönnun og stoðtól voru einnig uppfærð samhliða. Í tilreiðaranum voru strengir eins og „5 mars“ tilreiddir sem dagsetningar, en engin villa merkt þar sem tilreiðarinn gefur ekki kost á villumerkingu. Á sama hátt merktu sum lög í kerfinu villur áður en farið er aftur í þáttarann, sem gefur ekki heldur kost á villumerkingu. Þetta þýddi að ef tókinn var uppfærður í millilagi þáttara, eins og í sameiningu tóka eða skiptingu, þá týnist villumerkingin. Til að taka vel á þessum tilfellum þyrfti að flækja stoðtólin óþarflega. Þess í stað bættum við normaðri mynd, eins og „5. mars“, fyrir tilfellin í tilreiðaranum, og skjöluðum hvert tilfelli í þáttaranum. Við innleiddum svo nýtt villugreiningarlag í kerfið sem leitar sérstaklega að tókum án villumerkinga sem innihalda misræmi milli upprunalegs og núverandi texta, og merkjum þá með almennari villumerkingu.

Í einhverjum tilfellum getur einn tóki innihaldið fleiri en eina villu. Dæmi um þetta er setningin „eg beið þín“. Hér byrjar kerfið á því að leiðrétta „eg“ yfir í „Eg“, þar sem það er byrjun setningarinnar. Þetta er þó ekki gilt orð, en þegar kerfið leitar seinna að stafsetningarvillum skoðar það ekki tóka sem þegar innihalda villur. Til að taka á þessu leyfum við frekari skoðun á tókum sem innihalda hástöfunarvillur, og leiðrétta þannig „Eg“ yfir í „Ég“. Eins og er skrifar

þetta yfir upplýsingar um hástöfunarvillurnar, sem vekur spurningar um hvort tókar ættu að geta haft margar villumerkingar og hvernig það skal vera meðhöndlað. Þetta kallar á frekari hönnunarumræðu, þannig að eins og er leyfum við að skrifa sé yfir.

6. Valkostir

Til að koma til móts við þarfir annarra verkefna, eins og talgervinga og viðbóta, hönnuðum við í einingum kóðann sem skilaði upphaflega aðeins skipanalínuúttaki, og bættum við valkostum eða stikum til að styðja mismunandi notkun. Nýju hjúpeininguna (e. wrapper module) má nota beint í öðrum verkefnum. Þessa valkosti má svo senda í málrýniaðgerðir til að hafa áhrif á kerfisflæði. Að neðan er yfirlit valkosta.

Til að skilgreina inntak er valkosturinn *infile* notaður. Inntakið getur verið ReadFile-hlutur eða ítrandi hlutur eins og gjafi (e. generator), þar sem sys.stdin er sjálfvalið. Ef valkosturinn *one_sent* er sannur veit kerfið að inntakið inniheldur aðeins eina setningu.

Þó nokkur mismunandi snið úttaks eru studd, með valkostinum *format*. Ef gildið er *text* inniheldur úttakið aðeins leiðrétta útgáfu inntaksins. Með *textplustoks* fáum við færslu með leiðréttu inntaki og tókunum. Við fáum JSON-streng með *json* og CSV-snið með *csv*. Gildið *m2* skilar úttakinu á sama sniði og m2scorer (https://github.com/nusnlp/m2scorer), sem notað er í CoNLL14. Aðskilinn valkostur, *annotations*, virkar með *text* og *textplustoks* og skilar merkingunum sem lista við enda úttaksins.

Valkosturinn *all_errors* skilgreinir stig leiðréttingar, og gefur merkingar á setningastigi ef hann er sannur, en aðeins merkingar á tókastigi ef ósannur.

Setningar sem ekki er hægt að þátta eru merktar sem villur í heild sinni, en þetta olli of mörgum röngum jákvæðum í notendaprófunum. Ef valkosturinn *annotate_unparsed_sentences* er ósannur er engin óþáttanleg merking gefin.

Ef valkosturinn *generate_suggestion_list* er sannur getur merkingin í ákveðnum tilfellum innihaldið lista mögulegra leiðréttinga fyrir notandann að velja úr. Þetta er sérstaklega nytsamlegt fyrir ritvinnsluforrit. Þetta er aðeins gert eins og er með úttaki þrístæðulíkansins.

Valkosturinn *ignore_wordlist* leyfir okkur að skilgreina sett strengja þar sem hver strengur táknar orð sem ekki skal merkja sem villu eða leiðrétta. Þetta er aftur til undirbúnings fyrir viðbætur í ritvinnsluforrit.

Valkosturinn *suppress_suggestions* leyfir okkur að hafna langsóttari eða ólíklegum sjálfvirkt sóttum leiðréttingum, þannig að engri villu er bætt við. Notendaprófanir sýndu að þær voru oftast rangar, en gætu reynst nothæfar fyrir notendur sem vilja merkja eins margar villur og hægt er, jafnvel þótt sumar villumerkingarnar gætu verið rangar. Það er svo í höndum notanda að meta merkingarnar; samþykkja eða hafna þeim.

Ef valkosturinn *apply_suggestions* er sannur erum við djarfari í að breyta tókum með tillögum til leiðréttinga, þ.e. ekki bara leggja þær til. Þetta hjálpar þegar þetta er eining í flóknara kerfi, þar sem við þurfum bara leiðréttingar, ekki tillögur sem notandi metur.

Valkosturinn *suggest_not_correct* fer í hina áttina, og breytir öllu í tillögur sem notandinn metur. Ekkert er leiðrétt sjálfkrafa. Þetta gæti hæft mannlegum notendum sem eru mjög hæfir í yfirlestri, eða fyrir viðkvæm gögn sem innihalda jafnvel beinar tilvitnanir.

L8 – Málrýni í snjalltækjum

Skýrsla: Grammatek ehf
Aðilar: Grammatek ehf

Markmið

Að gera málrýni mögulega á íslenskum lyklaborðum, sem og sjálfvirka útfyllingu (e. autocompletion), í Android og iOS tækjum. Við byrja að undibúa þetta undirverkefni með því að skoða hugbúnaðarhögun og umhverfi og hvernig þurfi hugsanlega að aðlaga kjarna málrýnis til að hægt verði að nota hann í snjalltækjum. Þessi vinna mun einnig nýtast öllum öðrum verkefnum sem þurfa að keyra á snjalltækjum og verður hún unnin í nánu samstarfi við H9 (Talgreining í snjallsímum). Vegna minnkandi tilfanga munum við einbeita okkur að því að afhenda lausnir fyrir Android-tæki, hins vegar verða kröfur og vegvísir fyrir iOS innleiðingu afhentir sömuleiðis fyrir innleiðingu í framtíðinni.

Varða 7 – lýsing

Vegvísir fyrir Android- og iOS-þróun tilbúinn. Skilgreiningar fyrir vörður M8 og M9 endurskoðaðar.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Greint er frá vegvísi fyrir Android- og iOS-þróun að neðan. Skilgreiningar varða M8 og M9 hafa verið endurskoðaðar.

Skýrsla

Yfirlit: Android og iOS
Í dag er megináhersla þróunar fyrir snjalltæki á tvö umhverfi: Android og iOS. Önnur umhverfi eru notuð af fáum og það lítur ekki út fyrir að það muni breytast í bráð.

Umhverfin eru ekki samhæfanleg, hvorki hvað varðar sameiginlegt forritunarmál, eða sameiginleg högunarhugtök eða viðmót. Til eru kerfi frá þriðju aðilum sem reyna að ná þessum ólíku umhverfum undir eitt þak með því að varpa almennu forritunarmáli yfir á óhlutbundin forritunarviðmót sem sameina mörg almenn/lík hugtök í báðum umhverfum. Það eru þó mörg högunarhugtök sem eru í grundvallaratriðum ólík milli Android og iOS. Að reyna að brúa það bil er ógerlegt á almennan máta og oft þarf að forrita hugbúnað með því að aðlaga sig að þessum sérhæfðu högunarhugtökum, með notkun aðferða þess umhverfis og jafnvel staðlað forritunarmál þess umhverfis. Undirliggjandi hugtök Android og iOS varðandi þróun málrýnieiningar eru jafnvel ólík að öllu leyti.

Málrýni í Android og iOS
Android er mikið opnara kerfi og leyfir útfærslu sérsmíðaðrar málrýniþjónustu, sem er í boði heilt yfir, þ.e. fyrir öll forrit sem vilja nota málrýni í hugbúnaði sínum. Það eru hrein hugbúnaðarviðmót fyrir málrýni á grundvelli annaðhvort orðs eða setningar. Notandinn getur sótt smáforrit og valið uppáhaldsmálrýniþjónustu sína í stillingum.

Í samanburði veitir iOS ekki forritunarviðmót og býður því engar sérsmíðaðar málrýniþjónustur sem valkost utan lokaðs málrýnikerfisins sjálfs. Því miður styður nýjasta útgáfa málrýnis iOS ekki einu sinni íslensku.

Því þyrftu iOS-forrit sem gætu viljað stuðning við íslenska málrýni að útfæra sérstaka hugbúnaðareiningu/safn í hugbúnaði sínum. Annar kostur fyrir íslenska málrýni í iOS gæti verið á formi málrýnismáforrits sem fær aðgengi að texta með eftirfarandi aðferðum:

- Texti á klippiborði
- Skrár í geymslu á tæki eða í skýi
- Hlekkir vefsíðna (e. URLs)
- Texti með því að nota OCR á myndir

Hins vegar þyrfti notandinn að opna appið sérstaklega og annað hvort afrita og líma texta handvirkt, opna hlekki vefsíðna eða undirbúa og velja skrár. Niðurstöður málrýni þyrfti einnig að afrita aftur í uppruna sinn, sem gerir ferlið allt mjög þunglamalegt, ef það er yfirhöfuð ásættanlegt.

Sjálfvirk útfylling texta í Android og iOS
Auk málrýni er þörf á góðri sjálfvirkri útfyllingu þegar texti er skrifaður með lyklaborði á skjá. Oftast er slík virkni útfærð beint í lyklaborðssmáforriti án aðkomu kerfisþjónustu.

Þó að venjulegt iOS-lyklaborð sé þegar með frumstæðan stuðning við íslensku eru vankantar á þeim stuðningi. Ekki er hægt að breyta því og það er almennt talið vera gagnslaust. Það er þó möguleiki á að setja upp lyklaborð þriðju aðila, t.d. býður Microsoft Swiftkey upp á góðan stuðning fyrir íslensku.

Fyrir Android eru aðstæður almennt betri, vegna þess að flest sjálfvalin lyklaborð veita þegar nothæfa sjálfvirka útfyllingu fyrir íslensku. Einnig eru mörg lyklaborð þriðju aðila í boði, og sum þeirra hafa stuðning við íslensku, t.d. Microsoft Swiftkey. Eitt opið (e. open-source) skal nefna sérstaklega: AnySoft-lyklaborðið (https://github.com/AnySoftKeyboard/AnySoftKeyboard), sem við leggjum áherslu á til að útfæra virkni sjálfvirkrar útfyllingar.

Almenn högunarhugtök
Eins og nefnt var áður myndum við vilja geta notað allar hugbúnaðareiningar fyrir bæði Android og iOS. Huglægt er kjarnavirkni málrýni eða sjálfvirkrar útfyllingar svipuð á báðum kerfum. Samtvinnun þessara eininga í viðeigandi þjónustum eða hugbúnaði verður að gera á ólíkan máta.

Rammi / forritunarmál
Í leitinni að ramma sem gefur möguleika á að búa til einingu fyrir málrýni og sjálfvirka útfyllingu fyrir Android og gefur á sama tíma möguleika á notkun sömu hugbúnaðareininga síðar í iOS, ákváðum við að nota Kotlin Multiplatform.

Kotlin er nútímalegt og viðkunnanlegt forritunarmál og hefur nú tekið við af Java sem aðalforritunarmál í Android. Þar er það notað bæði fyrir notendaviðmót og rekstrarrökfræði (e. business logic). Eins og Java keyrir Kotlin á Java-sýndarvél og getur því haft samskipti við Java og önnur mál sem keyra einnig á JVM. Nýlega öðlaðist Kotlin stuðning við heimakóða (e. native code) og aðrar inningar í formi Kotlin Multiplatform. Þetta gerir rekstrarrökfræði skrifaðri í Kotlin kleift að keyra í öðrum umhverfum en Android, eins og iOS, Linux, MacOS, Windows og í vafranum svo lengi sem öll pökkuð ákvæði eru í römmum Kotlin Multiplatform eða Kotlin-kóða.

Það sem lætur Kotlin Multiplatform skara fram úr öðrum römmum eru færri skorður högunar, t.d. þegar kemur að fjölþráðavinnslu, og það sem er mikilvægara er að það er þegar komið í Android. Á hinn bóginn gefur það ekki almenn viðmót fyrir forritun notendaviðmóts, heldur samþættist vel römmum í hverju viðkomandi umhverfi fyrir sig. Það er einnig mögulegt að flytja út Kotlin Multiplatform-einingu í safn fyrir viðkomandi kerfi svo það megi nota í hvaða appi sem er ótengt Kotlin. Kotlin í Android er stöðugt forritunarmál af framleiðslustigi, á meðan Kotlin Multiplatform er á beta-stigi. Þrátt fyrir það hefur Kotlin Multiplatform þegar verið tekið upp hjá stórum fyrirtækjum eins og Netflix, VmWare, DropBox o.s.frv. Þar sem megináhersla okkar í L8 er að skila útfærslum sem virka fyrir Android skoðum við eins og er að nota Kotlin, með það í huga að nota stöðuga útgáfu Kotlin Multiplatform seinna á árinu þegar við gefum út einingar okkar. Þetta þýðir að vinna fyrir iOS í framtíðinni getur verið byggð á vinnu okkar.

Kjarnatækni málrýni
Til að gera gögn sem voru þróuð áður innan máltækniáætlunar fyrir málrýni nýtanleg í Android eða iOS eru þau innleidd með netþjónustu. Málrýnieiningin býr til málrýnibeiðnir og sendir þær til vefþjónustu (https://yfirlestur.is) og sýnir notanda svörin. Þessi keyrsla yfir vefinn er nauðsynleg því hugbúnaðurinn innan verkþáttar L7 var skrifaður í Python, sem er ekki stutt í Android og iOS og er því ekki hægt að nota beint á snjalltækjum. Þróun eigin málrýnipípu í Kotlin er þó utan umfangs þessa verkefnis.

Kjarnatækni sjálfvirkrar útfyllingar, lyklaborð
Sjálfvirk útfylling er í fyrstu útfærð í Android með Lucene Java-safninu. Lucene hefur mjög góðan stuðning við sjálfvirka útfyllingu. Það verður skoðað hvort Lucene getur síðar unnið með ósnortinni Kotlin-útfærslu, t.d. algrím byggt á n-stæðum, svo þessi eining yrði flytjanleg.

Í Android verður einingin innleidd í kvísl okkar af AnySoft-lyklaborðinu. Unnið verður að þessari kvísl í samstarfi við verkefni H9 – Talgreining í snjallsímum. Þetta lyklaborð hefur þegar eigin útfærslu fyrir sjálfvirka útfyllingu, og grunnstuðningur við íslensku með orðabók

er í boði. Við munum fyrst bæta þennan grunnstuðning með betri orðabók og seinna skipta honum út fyrir okkar einingu fyrir sjálfvirka útfyllingu.

Í iOS höfum við ekki fundið tilbúið opið lyklaborð sem myndi henta fyrir íslensku, en við fundum opna lyklaborðsrammann KeyboardKit (https://github.com/KeyboardKit/KeyboardKit) sem grunn til að búa til okkar eigið sérsmíðaða lyklaborð. Það samræmist þegar íslensku lykilkorti (e. keymap) og mætti auðveldlega bæta með sjálfvirkri útfyllingu.

Atriði verkþáttar

- **Málrýni**
  - safn/eining, byggð á vefforritaskilum Yfirlesturs
  - Viðmót:
    - Merkja orð
    - Merkja setningu
    - Leiðrétta orð => Listi af tillögum
    - Leiðrétt setning
    - Bæta við skilgreindum orðum frá notanda

- **Sjálfvirk útfylling**
  - Safn/eining, Lucene/byggt á þrístæðum og orðabók
  - Viðmót:
    - Útfylling => listi af tillögum
    - Bæta við orðum (orðakort => einkunn)
    - Bæta við skilgreindum orðum frá notanda

Vegvísir/Vörður:

M8

- Frumgerð málrýnieiningar
  - Byggð á vefþjónustu Greynir Correct
  - Beta-útgáfa á Google Play
- Anysoft Keyboard
  - Bætt íslensk orðabók
  - Togbeiðnir til upprunalegrar hirslu opnaðar

M9

- Eining sjálfvirkrar útfyllingar fyrir Android
  - Byggð á Lucene
  - Innbyggð í Anysoft Keyboard-kvísl
  - Anysoft Keyboard-kvísl 1.0 gefin út á Google Play
- Málrýnieining 1.0 gefin út á Google Play

L9 – Málrýni í ritvinnslukerfum

Skýrsla: Háskóli Íslands
Aðilar: Háskóli Íslands

Markmið

Að eiga heildarsamþættingu af íslenska málrýninum í Google Docs, MS Office og MacOS. Verkþátturinn felur einnig í sér notendaprófanir háskólanema á Google Docs-samþættingunni.

Varða 7 – lýsing

Fyrsta útgáfa Google Docs-samþættingarinnar tilbúin og notendaprófanir háskólanema undirbúnar.

Afurðir

Afurðum var ekki náð að fullu.

Við höfðum áætlað að:

- Innleiða GreynirCorrect inn í Google Docs
    - Leiðréttingar gerðar með forritaskilakalli á Yfirlestur.is
    - Virkni notendaviðmóts eins víðtæk og mögulegt er með Google-viðbót þriðja aðila
    - Virkni úr undirverkefni Kjarnans L7 frá M6 notuð aftur þar sem hægt er
- Skipuleggja notendaprófanafasa
    - Í samstarfi við Háskóla Íslands
    - Helst lokaðar prófanir, gerðar af háskólanemendum
    - Tímarammi: febrúar–maí

Innleiðing í Google Docs er vel á veg komin en tafðist stórlega vegna veikinda. Þessi töf seinkaði einnig skipulagi notendaprófana. Því er hvorugt þessara atriða tilbúið.

Núverandi staða:

- Þróun viðbótar
    - Þróunarumhverfi Google komið
    - Grunnvirkni innleidd (utanaðkomandi forritaskilakall, gangavinnsla, breyting texta í skjali, o.s.frv.)
    - Umfang notendaviðmóts skilgreint og vinna hafin. Einhver virkni innleidd.
- Notendaprófanir

- Ekki byrjaðar fyrir utan fyrstu skref

Afurðum M7 verður náð mánuði seinna, við lok febrúar, og að neðan er áætlun til að klára þessar afurðir.

Skýrsla

Google býður upp á staðlað þróunarumhverfi fyrir viðbætur í Google Docs í formi Google Apps Script IDE, sem er samþætt vafra. Þetta umhverfi er líka notað til að hýsa kóðann fyrir útgefnar Google-viðbætur, sem verður reyndin með okkar verkefni. Í stað þess að kóða í vafra, sem er ætluð notkun þróunarumhverfisins, fer kóðun fram á eigin vél, og kóðinn er þýddur og sendur í Google-umhverfið með NodeJS.

Snemma í þróunarferli Google Docs-samþættingarinnar komumst við að því að hið staðlaða þróunarumhverfi Google fyrir viðbætur er takmarkaðra en við héldum, sérstaklega hvað varðar notendaviðmót og víðari notendaupplifun. Í stað þess að nota línuviðmót fyrir villumerkingar, eins og var upprunalega áætlað, samanber innbyggðum málfræði- og stafsetningarmerkingum í Google Docs, notum við innbyggða hliðarstiku fyrir villumerkingar, sem er opinberlega stutt af Google-umhverfinu. Smá tími var notaður í að finna leiðir til að komast hjá þessum takmörkunum, en áhrif þessarar tafar voru á endanum lítil.

Þróun tafðist mikið í desember vegna veikinda starfsmanns. Þetta tafði bæði þróun viðbótarinnar sjálfrar og skipulag á notendaprófanafasa. Vegna ónægs tíma fyrir vörðu M7 verður skilum á afurðum seinkað um einn mánuð.

Til að tefja ekki komandi M8-vörðu verður skipulag notendaprófana unnið af öðrum starfsmanni en var áætlað. Þetta tryggir að þróun Google Docs-samþættingar getur haldið áfram óhindruð og verður tilbúin til notkunar um miðjan febrúar, og notendaprófanafasi byrjar við lok febrúar.

Google Docs-samþættingin sjálf er eins og er óklárað, þó að grunnvirkni sé til staðar með tengingu við utanaðkomandi forritaskil og útdrátt texta í skjali. Helsta verkið sem stendur eftir fyrir útgáfu er að klára notendaviðmót viðbótarinnar, í formi innbyggðar hliðarstiku, eins og lýst er að ofan, og óformlegar prófanir innanhúss. Umfang eftirstandandi vinnu passar vel innan þeirrar eins mánaðar töf sem lýst er að ofan.

L10 – Aðlögun að máltæknihugbúnaði

Skýrsla: Grammatek
Aðilar: Grammatek, Hljóðbókasafnið, Miðeind

Markmið

Að hafa útgáfu málrýnisins sem má nota sem einingu (e. module) innan stærri máltæknihugbúnaðar. Einingin verður að bjóða upp á sveigjanlegar stillingar til að koma til móts við þarfir mismunandi hugbúnaðar. Við munum innleiða útgáfu málrýnis til að samþætta í talgervil, en á sama tíma kanna þarfir annars hugbúnaðar sem gæti notið góðs af sjálfvirkum leiðréttingum stafsetningarvillna, t.d. vélþýðingar og leitar. Helsti munurinn á virkni málrýnis sem aðstoð við skrif og einingar til sjálfvirkrar leiðréttingar í máltæknihugbúnaði, er að máltæknieining þarf að leiðrétta án aðkomu notanda.

Seinkuð varða 6 – lýsing

Niðurstöður MOS-prófana á talgervilskerfi með og án málrýnieiningar tilbúnar. Málrýnieining fyrir talgervil, viðmið fyrir aðlögun einingarinnar fyrir aðra notkun tilbúin og gefin út á CLARIN.

Afurðir

Öllum markmiðum hefur verið náð. Í stað MOS-prófana (e. Mean-Opinion-Score), sem mæla hversu „viðkunnanleg“ eða „náttúruleg“ talgervilsrödd hljómar, framkvæmdum við prófanir á skiljanleika, þar sem notendur skrifuðu niður prufusetningar sem talgervilskerfið las. Sjá niðurstöður að neðan.

- Aðlöguð málrýniskvísl, GreynirCorrect (L7) er aðgengileg á GitHub á https://github.com/grammatek/GreynirCorrect4LT
- Prófunarsett fyrir skiljanleika talgervla aðgengilegt á CLARIN á http://hdl.handle.net/20.500.12537/182

Skýrsla

Helsta markmið málrýnisverkefnisins innan máltækniáætlunarinnar er að þróa lausn fyrir aðstoð við skrif með athugasemdum til notenda. Þessi verkþáttur skoðar notkun málrýni í máltæknihugbúnaði, sem hefur að hluta til mismunandi þarfir. Við leggjum áherslu á tvær þarfir sem eru sérstaklega mikilvægar í þessu samhengi: a) sjálfvirkar leiðréttingar og b) aðlögunarhæfni einingarinnar fyrir hönnuði sem vilja nota málrýni í hugbúnaði sínum. Auk þess að hafa málrýni í framendaþjónustu talgervilsins sem notkunardæmi prófuðum við mikilvægi réttrar stafsetningar fyrir skiljanleika í talgervilskerfinu (utanaðkomandi prófun

málrýnis). Ein prufa var einnig keyrð á röngum fyrirspurnum í BÍN til að meta notagildi málrýnis í fyrirspurnum.

GreynirCorrect4LT

Þar sem þarfir fyrir málrýni sem innbyggða einingu í máltæknihugbúnaði eru að einhverju leyti ólíkar þörfum málrýnis sem aðstoð við skrif kvísluðum við GreynirCorrect (sjá verkþátt L7) til að hefja aðskilda þróun. Notendur þessarar einingar verða aðallega hugbúnaðarhönnuðir og þeir ættu að hafa möguleika á að aðlaga virkni málrýnis fyrir hvaða verk sem er. Slíkar breytingar gætu hins vegar stangast á við almenna þróun GreynirCorrect sem tól fyrir aðstoð við skrif, og því verður nálgun okkar að uppfæra máltækniútgáfuna með endurbótum frá GreynirCorrect og á sama tíma vinna að aðlögunarhæfni tólsins. Fyrir talgervla er t.d. mikilvægt að geta látið málrýni hunsa skilgreind tög, eins og SSML-tög[33], sem gætu hafa verið sett inn áður en málrýnir vinnur textann. Enn fremur myndum við vilja nota málrýninn sem aðra umferð í normun, þ.e. til að auka nákvæmni í vali á réttu falli við víkkun talna og styttinga. Aftur á móti ætti málrýnir fyrir talgervla ekki endilega að eiga við málfræði upprunalegs texta, en það er vissulega hlutverk málrýnis sem aðstoð við skrif.

Við lítum á GreynirCorrect4LT sem hluta af framendapípu talgervils (T9+T10) og verður þróun haldið áfram samhliða því verki það sem eftir er af þriðja ári.

Skiljanleikaprófanir talgervils

Til að meta áhrif málrýnieiningar í framendapípu talgervils framkvæmdum við skiljanleikaprófanir. Uppsetning þeirra var svipuð klassískri uppsetningu *Semantically Unpredictable Sentences* (ísl. merkingarfræðilega ófyrirsjáanlegar setningar) (sjá Benoît o.fl (1996)[34]) en með nokkrum breytingum fyrir þetta verk, og ber þar helst að nefna prófanir á áhrifum stafsetningarvillna á skiljanleika.

Prófunarsett
Prófunarsettið samanstendur af 2x50 setningum: fyrir hverja rétt stafsetta setningu var sama setning búin til með stafsetningarvillu. Svo voru tveir listar af 50 setningum búnir til, hvor með 25 rétt stafsettum setningum og 25 setningum með villu. Ef rétt útgáfa setningar er í öðrum listanum verður ranga útgáfan í hinum listanum, þ.e. hver setning kemur aðeins einu sinni fyrir í hvorum lista.

Í prófunarsettinu eru tvær gerðir setninga: a) *Semantically Unpredictable Sentences (SUS)* og b) setningar úr raunverulegum gögnum, aðallega úr IceErrorCorpus (sjá undirverkefni L1) en einnig nokkrar prófunarsetningar úr normara talgervils (sjá T9). *SUS* eru byggðar upp þannig að þær eru setningafræðilega réttar en merkingarfræðilega vitleysa. Sem dæmi er fræg SUS frá Chomsky: *Colorless green ideas sleep furiously (ísl. litlausar grænar*

[33] https://www.w3.org/TR/speech-synthesis11/
[34] Benoît, Christian, Martine Grice, and Valerie Hazan (1996): The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. Speech Communication. 18. 381-392.

hugmyndir sofa tryllingslega). Í samræmi við Benoît o.fl. (1996) bjuggum við til setningar úr fjórum setningarmynstrum: Frumlag - áhrifslaus sögn; Frumlag - áhrifssögn - andlag; Boðháttur - Nafnliður/Boðháttur; Spurnarorð. Nokkur dæmi: *Kalda liðið hljómar ljóst, Punktum finnst árið lítið, Veiddu fyndið klósett, Viltu þrífa fundinn?*.

Val orðaforðans var gert með hliðsjón af eftirfarandi atriðum: stafsetningarvillur voru teknar úr lista misheppnaðra fyrirspurna í BÍN, aðallega tvö atkvæði, algeng orð (með tilvísun í tíðnilista Risamálheildar), og stafsetningarvillurnar þurftu að vera einfaldar: ein innsetning, eyðing, útskipting eða bókstafaskipti, þ.e. aðallega algengar innsláttarvillur (t.d. vegna aðliggjandi takka á lyklaborði). Að auki þurfti sennilegasta leiðréttingin að vera ótvíræð. Fyrir orðin sem voru rétt skrifuð í báðum útgáfum setninganna voru það eiginleikarnir *stutt lengd* og *tíðni* sem réðu mestu fyrir val. Þetta ferli við val er mikilvægt svo að setningarnar reyni ekki of mikið á vitsmunalegt vinnuálag prófara vegna langra og/eða sjaldgæfra orða.

Prófarar og nákvæmnisútreikningur
Prófararnir voru nokkuð einsleitur hópur 20 Íslendinga án heyrnar- eða skriftarskerðinga. Þeim var skipt upp í tvo hópa og hver hópur hlustaði á annað tveggja prófunarsetta og hafði engan aðgang að hinu settinu. Þeim var falið að hlusta á hverja setningu einu sinni og skrifa hana niður, með það að markmiði að skrifa raunveruleg orð jafnvel þó að þau heyrðu einhver „brengluð“ orð (eins og búist var við vegna stafsetningarvillna). Ef þeim fannst þau þurfa að hlusta aftur var það leyfilegt og þau merktu í reit á prófunarblaðinu. Prófurum var aldrei gerð grein fyrir stafsetningarvillunum eða þeim gefinn upp tilgangur prófunarinnar.

Útreikningur á nákvæmni er gerður á eftirfarandi hátt: a) ein eða fleiri villur með tilliti til viðmiðunarsetningar valda því að setningin er talin sem villa, þ.e. nákvæmni byggir á setningum, b) ef viðmiðunarsetningin inniheldur stafsetningarvillu verður skrifaða setningin að vera leiðrétt útgáfa, þ.e. án stafsetningarvillunnar. Ef setning með villu er hins vegar skrifuð óbreytt, með stafsetningarvillunni, er setningin merkt sem „skrifuð óbreytt“ (en telur samt sem villa í heildarreikningi), c) stafsetningarvillur frá prófurum sem hafa ekki áhrif á framburð eru ekki taldar sem villur. Þar á meðal eru i/y-ruglingur, stakir eða tvöfaldir samhljóðar (ef það hefur ekki áhrif á framburð, t.d. *aðallega* á móti *aðalega*) eða þegar bil vantar, t.d. *sjö þúsund* á móti *sjöþúsund*. Mögulegar stafsetningarvillur frá prófurum sem myndu orsaka annan framburð teljast þó sem villur, þar sem ómögulegt að ákveða eftir á og án þess að ræða við prófarana um hvort tiltekin stafsetning er stafsetningarvilla eða ekki.

Niðurstöður

Prófunarsett	**Nákvæmni**	**Fjöldi setninga í**

		prófunarsetti
Samtals	43,30%	1000
Réttar setningar	64,40%	500
Setningar m. stafsetningarvillum	22,20%	500
Réttar setningar raunveruleg gögn	77,20%	250
Stafsetningarvillur raunv. gögn	30,00%	250
Rétt SUS	51,60%	250
SUS m. stafsetningarvillum	11,10%	250

Niðurstöðurnar sýna skýrt hvernig einfaldar stafsetningarvillur eins og í prófunarsettinu (aðeins ein villa í setningu og aðeins smávægileg breyting frá réttu orði) hafa mikil áhrif á skiljanleika talgervilskerfis. Það verður því stöðug viðleitni framendapípu talgervils að bæta málrýninn með hliðsjón af sjálfvirkum leiðréttingum óorða.

Prófanir málrýnis fyrir leitarfyrirspurnir

Við teljum að málrýnir geti hjálpað gríðarlega í ýmsum máltæknihugbúnaði. Sem annað dæmi en fyrir talgervil keyrðum við prófanir á villulista samsettum úr árangurslausum fyrirspurnum í BÍN (https://bin.arnastofnun.is/). Listinn inniheldur 11.609 orð, og af þeim leiðrétti GreynirCorrect 9.745 eða 84%. Erfitt er að segja að hversu miklu leyti þessar leiðréttu fyrirspurnir samræmast því sem notandinn var í raun að leita að, sérstaklega þar sem fyrirspurnirnar eru aðeins eitt orð án samhengis, en það gæti hjálpað notendum að bæta „áttirðu við“-eiginleika við síðuna.

L14 – Málrýni með djúpum tauganetum

Skýrsla: Miðeind ehf
Aðilar: Miðeind ehf

Markmið

Meginmarkmið þessa verkþáttar er að þjálfa tauganet til að þýða „ófullnægjandi“ íslenska texta yfir á réttari texta. Annmarkarnir geta verið margra flokka, þar á meðal textar með „dæmigerðum“ stafsetningar- og málfræðivillum sem fullorðnir móðurmálhafar gera, textar skrifaðir af lesblindum, textar skrifaðir af börnum, og textar eftir fólk sem hefur annað móðurmál en íslensku.

Málrýnar byggðir á tauganetum munu líklega skila betri afköstum en reglubyggðir rýnar (sbr. L7), og viðbúið er að þeir ráði betur við hærri villutíðni, t.d. í mjög ófullnægjandi textum. Aftur á móti verður geta þeirra til að útskýra leiðréttingar takmörkuð. Þeir henta því vel til að leiðrétta mikið magn af texta og þar sem ekki er þörf á útskýringum.

Þetta undirverkefni byggir á reynslu sem safnast hefur saman í undirverkefnum um vélþýðingar. Við höfum komist að því að vélþýðingaafkóðarar (e. MT decoders) sem búa til íslenskar setningar úr enskum gjafatexta (e. source text), með hjálp vel þjálfaðs mállíkans, eru nokkuð góðir að búa til rétta íslenska stafsetningu og málfræðilega rétta texta. Við erum því bjartsýn á að net sem þýðir frá „slæmri“ íslensku yfir á „góða“ íslensku gæti náð góðum árangri, að því gefnu að kóðarahlutinn af netinu sé þjálfaður með viðeigandi hætti, þar á meðal með vandlega stilltu brottfalli (e. dropout) og með greypingum sem eru að hluta á bókstafsstigi (e. character level) (öfugt við orðflísar/BPE eingöngu).

Netin verða þjálfuð fyrir hvert villuóðal (almennt, lesblindra, barna, útlendinga) með blöndu af raunverulegum gögnum sem hefur verið safnað í öðrum undirverkefnum og gervigögnum sem verða búin til með því að líkja eftir villumynstrum sem koma fyrir í raunverulegu gögnunum.

Aðferðir og reynsla úr þessu undirverkefni gæti orðið til þess að bæta niðurstöður við umritun texta með talgreini (sbr. undirverkefni H7), sem má flokka sem eina gerð af þýðingu frá „ófullnægjandi“ íslensku til „góðrar“ íslensku.

Aukamarkmið þessa undirverkefnis er að þjálfa flokkara ofan á leiðréttingarkóðarann. Þessi flokkari úthlutar líkum á því að inntakssetning sé málfræðilega rétt. Ef setningin er mjög líklega rétt, þá þarf ekki að þátta hana og athuga með (hægari) reglubyggða rýninum, og þannig er ferlinu flýtt – sem er sérstaklega mikilvæg fyrir lengri skjöl, svo sem skýrslur, lokaritgerðir og bækur.

Varða 7 – lýsing

Tauganetsflokkari fyrir málfræðilegan réttleika (e. grammatical correctness) tilbúinn og gefinn út á CLARIN. Almennir villuflokkar valdir til prófana.

Afurðir

Undirverkefnið náði góðum árangri og öllum afurðum var náð. Tauganetsflokkari villna fyrir setningar hefur verið gefinn út á CLARIN á http://hdl.handle.net/20.500.12537/183.

Skýrsla

Auk flokkara fyrir málfræðilegan og stafsetningarlegan réttleika (þ.e. tauganet sem metur líkindi þess hvort setning er í heildina röng) byrjuðum við að undirbúa nákvæmari flokkara sem greinir villutegundir og spannir, sem er afurð seinni varða.

Íslensk villumálheild[35] (IceEC) úr undirverkefni L1 var notuð til að fínstilla íslenska mállíkanið IceBERT[36] fyrir setningaflokkun. Markmiðið var að þjálfa málfræðileg villugreiningarlíkön sem gætu flokkað eftir því hvort setning inniheldur ákveðna tegund villu.

Íslenska villumálheildin inniheldur um 58.000 setningar, og 23.000 þeirra innihalda villumerkingar. IceEC-gögnin voru dregin úr XML-skrám og þeim varpað yfir á liðtækara JSON-snið. Þetta snið samræmist setningaflokkaratólunum sem eru hluti af hinu opna Hugging Face Transformers[37]-safni.

Nokkrar þjálfunarítranir voru keyrðar yfir aðalflokkana, undirflokkana og villumerkingarnar eins og skilgreint er í skema IceEC. Setningaflokkari var einnig þjálfaður til að meta líkindi þess að setning sé villulaus (sem er afurð M7-vörðu fyrir þetta undirverkefni).

Eins og er munum við prófa alla villuflokka og öll þrjú stig (aðalflokka, undirflokka og villukóða) til að öðlast betri skilning á því hvaða stig hentar best taugalíkönunum til að læra af.

Niðurstöður tilrauna okkar eru settar fram og ræddar að neðan. Þær eru byggðar á opinbera IceEC-prófunarsettinu.

Fínstillt líkan fjölmerkinga fyrir flokkun er aðgengilegt á CLARIN[38].

	Atriði	F1	Nákvæmni	Heimt	$F_{0.5}$	Hittni

[35] https://github.com/antonkarl/iceErrorCorpus/
[36] https://huggingface.co/mideind/IceBERT
[37] https://github.com/huggingface/transformers
[38] http://hdl.handle.net/20.500.12537/183

		fjölvameðaltal (e. macro avg)				
í heildina	17621	63,78%	72,23%	61,13%	67,40%	91,85%
coherence (skiljanleiki)	174	59,36%	62,29%	57,58%	60,92%	99,35%
grammar (málfræði)	3165	58,19%	67,90%	56,17%	62,02%	90,52%
orthography (réttritun)	8665	76,37%	82,76%	73,35%	79,65%	85,13%
other (annað)	1287	58,73%	73,86%	55,97%	64,02%	94,36%
style (stíll)	2818	64,25%	74,43%	61,11%	68,80%	88,70%
vocabulary (orðaforði)	1512	65,76%	72,13%	62,62%	68,98%	93,01%

Tafla 1. Flokkari fjölmerkinga á setningastigi fínstilltur á aðalflokkunum 6.
Greint er frá niðurstöðum í heildina og fyrir einstaka flokka.

Samantekt á niðurstöðum fyrir aðalvilluflokkana 6 er sýnd í Töflu 1. Í heildina er F1-gildið 63,75%, og nákvæmni töluvert hærri en heimt, sem þýðir að líkanið nær oft ekki að finna setningar með villum, en þegar það finnur þær nær það oftast að flokka þær rétt að mestu. Flokkurinn *orthography* nær besta árangrinum, sem gert var ráð fyrir, þar sem hann er með auðveldustu og reglubundnustu villurnar, eins og mistök í greinarmerkjasetningu, hástöfun, eða einföldum stafsetningarvillum. Á hinn bóginn er árangur villna í erfiðari flokkum eins og *grammar* og *coherence* í lægri kantinum. Þessar villur eru mjög fjölbreyttar og tauganetið nær líklega ekki að sjá nægilega mörg dæmi á þjálfunartíma til að læra fyllilega að greina allar mismunandi villur sem koma fyrir í prófunarsettinu. Athugið að háu tölur hittni koma að mestu til vegna þess að flestar setningar voru flokkaðar sem án villu, sérstaklega *coherence*.

	F1 fjölvameðaltal	Nákvæmni	Heimt	$F_{0.5}$	Hittni
í heildina	63,39%	74,74%	60,49%	67,79%	91,84%

Tafla 2. Flokkari fjölmerkinga á setningastigi fínstilltur á undirflokkunum 33.

	F1 fjölvameðaltal	Nákvæmni	Heimt	$F_{0.5}$	Hittni
í heildina	69,66%	69,57%	69,76%	69,61%	99,71%

Tafla 3. Flokkari fjölmerkinga á setningastigi fínstilltur á villukóðunum 258.

Töflur 2 og 3 sýna niðurstöður fyrir undirflokkana og raunverulega villukóða, hver um sig. Þó að við greinum þá ekki frekar hér getum við staðfest svipað mynstur eins go fyrir aðalflokkana í Töflu 1, þar sem auðveldast er að greina prentvillur.

	F1 fjölvameðalta	Nákvæmni	Heimt	$F_{0.5}$	Hittni

í heildina	70,56%	77,82%	69,55%	73,85%	75,79%

Tafla 4. Tveggja valkosta flokkarinn á setningastigi.

Tveggja valkosta flokkarinn á setningastigi (Tafla 4) er fínstilltur til að greina einfaldlega hvort setning inniheldur villu eða ekki. Hér er F1-gildið 70,56%, með heimt rétt undir 70%. Fyrir raunheimsbeitingu myndum við þurfa hærri heimt, svo að færri setningar með villum sleppi í gegn. Góður setningaflokkari innbyggður í málrýni GreynirCorrect (sem er aðgengilegur almenningi á vefsíðunni Yfirlestur.is) gæti hraðað vinnslunni með því að greina setningar sem eru líklega réttar, og því sleppa þörfinni á fullþáttun (e. full constituency parsing) til að merkja hugsanlegar villur.

Gögn um málfræðilegar villur eru erfið í notkun (og erfitt að búa til) af ýmsum ástæðum. Villurnar eru mjög fjölbreyttar, þar sem engin mörk eru fyrir því hvaða villur fólk skrifar í textum. Jafnvel innan hvers villuflokks er ekki endilega líklegt að sama villa komi fyrir aftur í nákvæmlega sama orði, sem þýðir að líkanið þarf að sjá mörg dæmi við þjálfun til að geta framreiknað út frá þeim.

Taka skal fram að IceEC hefur nokkur einkenni sem takmarka notagildið fyrir þjálfun villuflokkunarlíkana. Þegar margar villur eru til staðar og innan sömu spannar hefur aðeins lengsta spönnin verið merkt. Villur eru líka oft aðeins skiljanlegar sem slíkar þegar þær eru skoðaðar í röð; svona eru gögnin ekki unnin á þjálfunartíma, þar sem allar villur eru skoðaðar á sama tíma. Betri nálgun gæti verið að skoða hvert skref í merkingunum sem sjálfstætt eða sjálfsaðhverft merki. Að lokum eru villuflokkarnir sjálfir ekki alltaf vel skilgreindir og það er nokkuð um ósamræmi í notkun þeirra.

Þar til nú höfum við aðeins notað almennu villumálheildina fyrir þjálfun, og ekki þær sérhæfðu (þ.e. textar skrifaðir af annarsmálshöfum, lesblindum eða börnum), þar sem sú vinna er áætluð fyrir M8. Þó að þessar sérhæfðu málheildir hafi aðra eiginleika munu þær nýtast til að fá fleiri dæmi um villur raunverulegra notenda.

Heilt yfir, þó að tölurnar sem greint er frá hér séu kannski ekki nógu háar fyrir beina notkun, nýtast þær sem sía fyrir vinnslu neðanstreymis. Líkanið má t.d. nota til að meta hvort setningar skal senda áfram í reglubyggða kerfið í hluta L7, sérstaklega ef þröskuldur fyrir flokkun er lækkaður til að lágmarka falsk-neikvæða svörun. Við munum einnig byggja á þessum niðurstöðum í áframhaldandi vinnu, t.d. þegar gögnin hafa verið auðguð með tilbúnum villum (e. synthetic errors). Auk þessara tilrauna höfum við hafið vinnu við þjálfun flokkara á tókastigi, þar sem bráðabirgðaniðurstöður sýna mikinn mun á árangri milli mismunandi villuflokka, þar sem *orthography* nær besta árangri og *coherence* lægsta.

Þó að réttritunarvillur séu auðveldastar fyrir tauganet að greina eru þær einnig tiltölulega auðveldar fyrir reglubyggð kerfi til að greina og leiðrétta. Við höfum einnig áhuga á því að nýta eðlislæga tungumálaskilningshæfileika stórs mállíkans til að takast á við erfiðari villur sem reglubyggt kerfi gæti átt erfitt með.

H1 – Upptökur með Samrómi

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að ljúka skilgreindri söfnun upptaka stuttra segða fyrir þjálfun talgreinis. Verkvangurinn verður notaður til að safna gögnum frá fullorðnum, unglingum, málhöfum með íslensku sem annað mál, og gagnasöfnuninni verður haldið áfram til að afla gagna frá þessum hópum. Enn fremur hefur söfnunarappið (e. crowd-sourcing app) verið aðlagað til að safna radddæmum frá fólki sem les ekki skrifaðar setningar, en þannig ætti að vera hægt að safna radddæmum frá ungum börnum.

Seinkuð varða 6 – lýsing

• 50.000 segðir annars máls mælenda teknar upp
• 30.000 segðir teknar upp af fullorðnum með því að endurtaka talaðar segðir
• 30.000 segðir teknar upp af börnum með því að endurtaka talaðar segðir

Þessi verkþáttur hefur enga vörðu 7.

Afurðir

- 50.000 segðir annars máls mælenda teknar upp
  -> náð við M6
- 30.000 segðir teknar upp af fullorðnum með því að endurtaka talaðar segðir
  -> náð, 51.515
- 30.000 segðir teknar upp af börnum með því að endurtaka talaðar segðir -> náð, 41.787

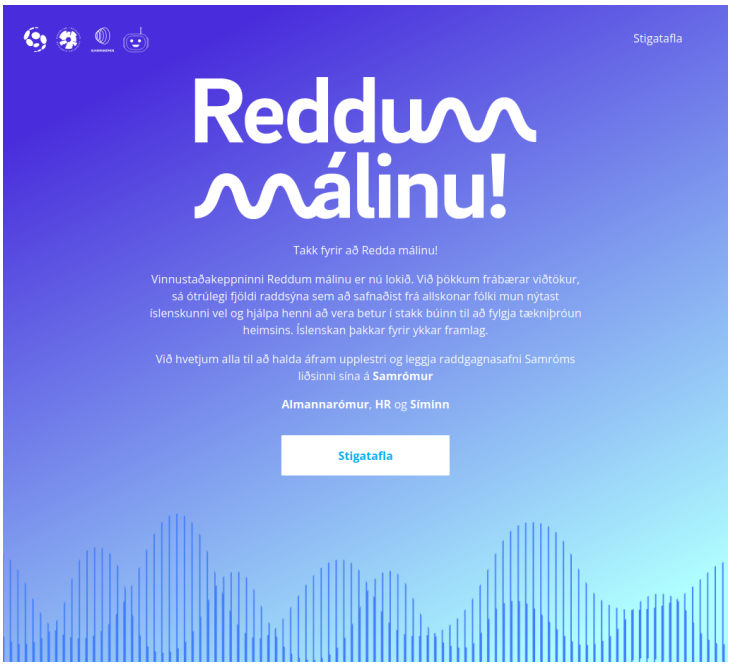
Öllum markmiðum hefur verið náð. Við M6 hafði tveimur markmiðum ekki verið náð: 30.000 segðir teknar upp af fullorðnum með því að endurtaka talaðar segðir, og 30.000 segðir teknar upp af börnum með því að endurtaka talaðar segðir. Á meðan á M7 stóð náðust bæði þessi markmið. Markmiði fyrir fullorðna var náð í nóvember með keppninni „Reddum málinu“ og nokkur framvinda náðist fyrir börn í sömu keppni (22.700 af 30.000 safnað). Þriðja grunnskólakeppnin var haldin dagana 20.–26. janúar 2022 og síðasta markmiðið, 30.000 segðir teknar upp af börnum með því að endurtaka talaðar segðir, náðist á fyrsta degi keppninnar.

Skýrsla

Í þessari vörðu höfum við lagt áherslu á söfnun segða með því að endurtaka talaðar segðir til að klára M6-markmið. Þessi skýrsla greinir aðeins frá seinkuðum M6-afurðum þar sem engar afurðir eru fyrir M7 í þessum verkþætti. Tvær nálganir voru notaðar. Sú fyrri var að slökkva á venjulegum möguleika þátttöku svo einu möguleikar þátttakenda voru að endurtaka talaðar segðir eða sannreyna fyrirliggjandi gögn. Við ákváðum að slökkva ekki á þeim möguleika þar sem við þurfum alltaf að sannreyna. Seinni nálgunin var keppni haldin milli fyrirtækja kölluð „Reddum málinu“.

Reddum málinu

Við héldum keppni í samstarfi við Símann og vefsíðan fyrir hana var búin til úr hirslukvísl Samróms. Síminn hjálpaði mikið með því að borga fyrir auglýsingar í dagblöðum, stoppistöðvum strætó og á samfélagsmiðlum og með því að aðstoða við almannatengsl. Fyrir þessa keppni milli fyrirtækja þurftum við uppfærslur í umhverfið. Nýtt útlit var ákveðið og fyrirtæki þurftu að geta skráð sig sjálf og nýrri stigatöflu og tölfræði í rauntíma fyrir keppnina var bætt við. Hér beindum við einnig notendum að „taka þátt með vinnuflæði endurtekinna segða“ þar til markmiði vörðu 6 var náð. Þetta gerði okkur kleift að ná markmiði fyrir fullorðna fljótt, og þrátt fyrir einhverja þátttöku barna var það ekki nóg. Við gerðum allt aðgengilegt á bæði https://samromur.is og https://reddummalinu.is svo að hver sem skoðaði aðra hvora síðuna tæki þátt og öll gögnin voru send í gagnasafn Samróms. Keppnin heppnaðist mjög vel og við söfnuðum 360.000 segðum. Eftir seinustu grunnskólakeppnina, fyrir keppnina „Reddum málinu“, hallaðist aldursdreifing í safninu nær börnum, en flestar þeirra 360.000 nýju segða eru frá fullorðnum, sem hjálpar aðeins til við aldursdreifingu.



Forsíða fyrir „Reddum málinu“

Þriðja grunnskólakeppnin

Til að halda uppi vitund almennings og ná vörðunni sem vantaði skipulögðum við og undirbjuggum þriðju grunnskólakeppnina. Elko, íslenskt fyrirtæki, lagði til þrjá þrívíddarprentara fyrir sigurvegara í hverjum flokki. Forseti Íslands og forsetafrúin munu halda litla verðlaunaafhendingu á Bessastöðum og veita vinningshöfunum verðlaun, en það veltur þó á samkomutakmörkunum vegna COVID. Fyrir grunnskólakeppnina aðlöguðum við umhverfið frá Reddum málinu, og ber þar helst að nefna notkun stigatöflunnar, tölfræði í rauntíma og breytingar á gagnasafni. Við veitum skólum einnig forritaskil sem gera þeim kleift að senda samþykkisbeiðnir til foreldra/forráðamanna án þess að þurfa að skrá inn handvirkt kennitölur nemendanna og netfang foreldris/forráðamanns í umhverfið.

Keppnin sjálf heppnaðist mjög vel. Næstum 1,3 milljónum segða var safnað í þeirri viku. Síðustu tveir dagarnir einkenndust af keppnisæði, og seinasti dagurinn var langstærstur, þar sem 487.936 setningar voru lesnar af þátttakendum þann dag. Samtals tóku 118 skólar þátt og 5.652 manns tóku þátt fyrir hönd skóla sinna í þessari keppni.

Gögn

Umhverfið hefur safnað samtals 2,8 milljónum segða, og af þeim hafa 333.000 verið metnar gildar og 160.000 ógildar. Í M7 var 1.678.000 segðum safnað. Af þeim gögnum sem hafa verið yfirfarin höfum við hlutfallið 67.5% gild á móti 32.5% ógild. Þetta er lítilsháttar framför síðan í byrjun M7, þegar hlutfall gildra gagna var 64%. Kynjadreifing er enn ójöfn, með fleiri segðum frá konum (66%) en körlum (34%). Tölfræði í rauntíma er aðgengileg á https://www.samromur.is/gagnasafn og https://stats.samromur.is.

Útgáfa og aðgengi

Við stefnum á að gefa gögnin út á þremur útgáfuvettvöngum, Linguistic Data Consortium (LDC), OpenSLR og CLARIN. Þrjár málheildir eru í útgáfuferli, og staða hverrar þeirra kemur fram í töflunni að neðan.

	LDC	OpenSLR	CLARIN
Samrómur 21.05	Í útgáfuáætlun LDC	Gefið út[39]	Gefið út[40]
Samrómur Children	Í útgáfuáætlun LDC	Gefið út[41]	Gefið út[42]
Samrómur Queries	Dreifingarleyfi undirritað	Gefið út[43]	Gefið út[44]

[39] http://openslr.org/112/
[40] http://hdl.handle.net/20.500.12537/189
[41] https://openslr.org/117/
[42] http://hdl.handle.net/20.500.12537/185
[43] http://openslr.org/116/
[44] http://hdl.handle.net/20.500.12537/180

Samantekt og frekari vinna

Keppnir eru enn áhrifarík leið til að safna gögnum og Reddum málinu sýndi einnig mátt markaðssetningar með meiri fjármunum, en nafnið Samrómur er nú nokkuð þekkt. Þó að við höfum ekki aðgang að sambærilegum fjármunum höfum við séð grunnskólakeppnina blómstra og ná gríðarlegu magni gagna.

Lítilsháttar framför gildra á móti ógildum segðum var viðbúin. Við M6 sýndi sjálfvirka yfirferðin margar segðir sem gildar, þó að þær þyrftu enn gilt atkvæði frá notanda umhverfisins til að vera staðfestar gildar. Við vonum að sú tala muni halda áfram að hækka.

Útgáfuferlið er vel á veg komið og nú þegar þessar þrjár málheildir eru útgefnar höfum við gefið út meirihluta aðgengilegra gildra gagna. Áskorunin nú er að fara yfir það mikla magn gagna sem við höfum og eftir þriðju grunnskólakeppnina munum við skoða hvernig við getum hvatt notendur umhverfisins til að yfirfara meira.

H2 – Umritun efnis úr sjónvarpi og útvarpi

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík, Creditinfo, Tiro ehf, RÚV

Markmið

Að eiga umritanir sjónvarps- og útvarpsefnis til þjálfunar talgreina fyrir fjölmiðla. Á fyrsta ári er markmiðið að umrita 100 klukkustundir af umræðuþáttum úr útvarpi og sjónvarpi, umrita eða framkvæma samröðun hljóðs-til-texta með for-umröðuðum gögnum frá textavarpssíðu 888 (fréttir), og umrita 25 klukkustundir af afþreyingarefni.

Varða 7 – lýsing

Gefa út öll tiltæk gögn sem hafa verið umrituð.

Afurðir

Ekki hefur öllum markmiðum vörðunnar verið náð. Á meðan á pökkun gagna stóð rákumst við á tölvuvandamál sem ollu því að við gátum ekki gefið gögnin út fyrir lok M7. Hins vegar teljum við að þessi vandamál verði leyst í næstu viku, sem gerir okkur kleift að klára pökkun gagnasettanna. Því verður útgáfu gagnanna frestað til 15. febrúar 2022.

Gagnasettið RÚV_TV_unidentified_speakers verður fyrst gefið út á OpenSLR og svo á CLARIN, vegna stærðartakmarkana við að gefa út beint á CLARIN.

Útvarpsgögnin frá Creditinfo hafa verið samröðuð og öllum nauðsynlegum upplýsingum til útgáfu aflað.

Skýrsla

Útvarpsgögnin frá Creditinfo hefur verið samröðuð. Við fengum 250 klukkustundir umritaðra gagna. Af þeim 250 klukkustundum voru aðeins 74 þeirra nothæfar fyrir þjálfun talgreina. Hinar 176 klukkustundirnar voru samraðaðar með aðferðum sem þróaðar voru við HR.

Aðferð
Aðferðin byggir á því að hafa ekki fullkomna umritun af hljóðskrá með tímastimpla fyrir byrjun og enda fyrir hvert orð. Þessar umritanir fengum við frá Tiro. Fyrir hvern bút í umritun Creditinfo er líklegasta staðsetning í umritun Tiro fundin. Þegar sú staðsetning er fundin finnum við fyrsta orðið í umritun Creditinfo sem passar fullkomlega við orð í umritun Tiro.

Tímastimpill byrjunar þess orðs er svo geymdur sem byrjunartími segðarinnar. Hið sama er gert fyrir seinasta orðið sem passar fullkomlega milli aðferðanna og endatímastimpils þess orðs, sem er geymdur sem lok segðarinnar.

Til að auka nákvæmni aðferðarinnar vildum við taka með til skoðunar málfræðilegt fall orða. Til dæmis, ef Tiro myndi umrita orð sem „Morgunvaktinni“ en rétt umritun er „Morgunvaktina“, ættu þessi orð að teljast sama orðið þar sem umritun Tiro er í grundvallaratriðum rétt. Til að gera þetta var notaður listi allra íslenskra orðmynda frá BÍN til að greina hvort orð eru orðmyndir af sama orði.

Forvinnsla
Þessi aðferð krafðist nokkurrar forvinnslu á umritun Creditinfo. Í hverjum bút umritunar Creditinfo gátu margir mælendur sagt margar setningar, sem hentar ekki þjálfun talgreina. Ofan á það voru upplýsingar um mælendur oft geymdar í umritunartexta sem bætir miklu suði í gögnin. Ennfremur var ekkert samræmi í því hvernig upplýsingar um mælendur voru geymdar í umritunartexta. Minnst þrjár ólíkar aðferðir til að tilgreina upplýsingar um mælendur fundust við forvinnslu. Upplýsingar um mælendur var ekki hægt að virða að vettugi þar sem mikilvægara er fyrir gagnasettið að hafa upplýsingar um mælendur með þar sem hægt er.

Niðurstöður
Úr þessari samröðun voru 126 klukkustundir af hágæða þjálfunargögnum fyrir talgreina. Erfitt er að meta raunverulegt gagnatap þar sem eitthvað suð er til staðar í gögnunum frá Creditinfo til að byrja með. Ástæða þess er að fyrir bútana sem gefnir eru í umritun Creditinfo eru gjarnan margir mælendur að eiga samtal. Milli þess sem fólk mælir eru þagnir sem ætti að sía út.

Sumarið 2021 var nokkrum sumarnemendum gefið það verkefni að samraða gögnin handvirkt. Þetta skilaði 14,13 klukkustundum af vel samröðuðum gögnum, á meðan umritun Creditinfo fyrir sömu hljóðskrár voru sagðar hafa 15,54 klukkustundir segða. Við sjáum því 9% tap með því að klippa aðeins út þagnir. Þær 126 klukkustundir gagna sem dregnar voru úr upprunalegu 176 klukkustundunum sýna þá alls 28% tap, á meðan tap yfir þær 14,13 klukkustundir af vel samröðuðum gögnum frá sumarnemendunum var aðeins 11%.

Gagnasett Creditinfo er alls 200 klukkustundir og samanstendur af 45.553 segðum frá 1.360 mælendum. Gagnasettið RÚV TV unidentified speakers (ísl. ótilgreindir mælendur) er 287 klukkustundir. Þetta eru samtals 487 klukkustundir. Á meðan á bútun í hljóðskrár fyrir hverja segð stóð komu upp ófyrirséð tölvuvandamál sem töfðu pökkun gagnasettanna. Tveggja vikna frestur ætti að nægja til að senda gagnasettin til útgáfu.

H3 – Umritun samræðna og fyrirspurna

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að umrita samræður og fyrirspurnir til að eiga talmálsgögn með frjálsara flæði. Magn safnaðra gagna verður skilgreint þegar upptökubúnaður hefur verið settur upp og fyrstu upptökum lokið. Samtöl verða tekin upp á fyrsta og öðru ári, en upptökur fyrirspurna hefjast og þeim lýkur á öðru ári.

Seinkaðar vörður 5 og 6 – lýsing

M5: 20 klst. af samræðugögnum teknar upp og umritaðar. Fyrirspurnamálheild búin til.
M6: 20 klst. af lesnum fyrirspurnum teknar upp.

Afurðir

Öllum markmiðum bæði vörðu 5 og 6 hefur verið náð. Fyrirspurnamálheild M6 hefur verið send inn og samþykkt til útgáfu hjá Linguistic Data Consortium og OpenSLR: http://openslr.org/112/. Fyrirspurnagagnasettið hefur verið gefið út á CLARIN á http://hdl.handle.net/20.500.12537/180.

Samræðugögnin hafa verið umrituð og prófarkalesin, og eru samtals 21 klst. Gagnasettið hefur verið gefið út á CLARIN á http://hdl.handle.net/20.500.12537/187.
Spjallsetrið er enn uppi til að safna nýjum gögnum: https://spjall.samromur.is.

Skýrsla

Þar sem gagnasett fyrirspurna og samræðna eru ólík í eðli sínu mun skýrslan fyrst greina frá fyrirspurnum, og svo samræðum.

Yfirlit fyrirspurna

Formáli

Tilgangur þessa hluta verkefnisins var að búa til sérhæfða málheild fyrirspurna. Við upphaf vörðunnar voru 3 klukkustundir gildra gagna og 28 klukkustundir óyfirfarinna gagna. Helsta markmið þessarar vörðu var að fara yfir næg gögn til að hafa 20 klukkustundir gildra gagna tilbúin fyrir útgáfu, og gefa gagnasettið út.

Yfirferð

Ferli yfirferðar átti sér stað á spjallsetri Samróms (www.samromur.is). Meðlimum hópsins voru gefnar leiðbeiningar varðandi hvað telst gild upptaka og hvað telst ógild, og þeim svo gefnir ofuraðgangar (e. super user accounts) að setrinu. Með þessum aðgöngum er eitt atkvæði nóg til að merkja upptöku sem gilda eða ógilda, en annars þarf tvö atkvæði. Ofuraðgangarnir hafa einnig aðgang að yfirferðarpökkum. Í þessu tilfelli voru allar 28 klukkustundir óyfirfarinna gagna settar í slíkan pakka svo að þeir sem vinna að yfirferð gætu auðveldlega farið í gegnum ferli yfirferðar.

Útgáfa

Málheildin hefur verið send inn og samþykkt til útgáfu hjá Linguistic Data Consortium, og OpenSLR, sem eru bæði gagnadreifingarvettvangar sem eru mikið notaðir af talsamfélaginu. Málheildin hefur einnig verið gefin út á CLARIN[45].

Yfirlit samræðna

Vegna tæknilegra tafa við M6 og tafa vegna faraldursins við M7 lauk ferli umritana og yfirferðar samræðugagnasettsins ekki fyrr en í þriðju viku janúars. Þá var seinasta gagnasettið yfirfarið, persónulegum nöfnum eytt, og sett í tvær aðskildar möppur: heil samtöl og hálf samtöl. Heil samtöl hafa hljóð og umritun fyrir báða mælendur. Hálf samtöl hafa aðeins hljóð og umritun fyrir annan tveggja mælenda. Þrátt fyrir að hafa aðeins hljóð annars tveggja mælenda nýtast hálf samtöl fyrir samræðumállíkön í sjálfvirkri íslenskri talgreiningu. Málheildin hefur verið gefin út á CLARIN[46].

[45] http://hdl.handle.net/20.500.12537/180
[46] http://hdl.handle.net/20.500.12537/187

H4 – Umritun fyrirlestra

Skýrsla: Háskólinn í Reykjavík
Aðilar: Tiro, Háskólinn í Reykjavík

Markmið

Að hafa safn umritaðra fyrirlestra til þjálfunar sjálfvirkra talgreinikerfa (ASR) fyrir það svið. Þessari vinnu verður lokið á tveim árum. Markmiðið á fyrsta ári er að fá kerfið í gang og fyrstu umritun sem sönnun á gildi hugmyndar (e. proof of concept).

Seinkuð varða 5 – lýsing

50 klukkustundir fyrirlestra umritaðar frá fimm fræðasviðum.

Afurðir

Þessi verkþáttur var hluti af seinkuðum vörðum M6 og var fjallað um hann í skýrslunni um þau markmið. Öllum markmiðum hefur nú verið náð og gagnasafnið hefur verið gefið út á CLARIN á http://hdl.handle.net/20.500.12537/171.

H5 – Samröðun á stórum málheildum

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að safna miklu magni af hljóðupptökum frá opinberum og einkareknum stofnunum. Opinbert svið (e. public domain) gæti innihaldið upptökur af Alþingisræðum, opnum nefndarfundum og réttarhöldum. Einkarekið svið (e. private domain) inniheldur einnig gögn sem gætu verið aðgengileg, svo sem hljóðbókasöfn og símafyrirspurnir. Þessa gagngjafa þarf að bera kennsl á og afla viðeigandi leyfa.

Varða 7 – lýsing

Tiltækir gagngjafar skilgreindir með greinargerð um tilganginn í tilviki hvers gjafa.

Afurðir

Öllum markmiðum vörðunnar hefur verið náð. Við höfum viljayfirlýsingar frá *Mannviti*, *Hvað er málið* og *Neðanmáls*.

Skýrsla

Við fundum 13 hlaðvörp sem verða helstu gagnagjafar okkar. Auk þeirra voru borin kennsl á aðra gagnagjafa: frá RÚV, hljóðbækur, og samstarfsaðila innan HR. Við lögðum áherslu á hlaðvörp með handriti þar sem líklegra er að til séu bæði textar og hljóðgögn. Fyrsta samskiptabréf hefur verið skrifað og sent til 14 af þeim gagnagjöfum. Þrír þeirra gagnagjafa hafa gefið sýnigögn, svo við getum hafið vinnu við M8 með því að hanna gagnavinnslupípu okkar utan um þau gagnasnið sem þau hafa gefið.

H9 – Talgreining í snjallsímum

Skýrsla: Tiro ehf.
Aðilar: Tiro ehf.

Markmið

Að þróa talgreinisviðmót fyrir Android-lyklaborð. Á fyrri hluta þriðja árs verður miðpunktur vinnunnar þróun á frumgerð Android-lyklaborðs með einföldu talviðmóti. Unnið verður í samstarfi við L8 - Málrýni í snjalltækjum (3 mánuðir úr H9 og 3 mánuðir úr L8). Á þriðja ári munum við kanna aðferðir til að bæta talgreini og málrýni við snjallsímalyklaborð, bæði sem vefþjónustu og sem einingu í tæki. Við munum búa til frumgerð af lyklaborði með rödd-til-texta hæfni (e. voice-to-text facility) fyrir Android, með því að nota vefþjónustu. Vegna minnkandi tilfanga munum við einbeita okkur að því að gefa lausnina út fyrir Android, en einnig skilgreina kröfur fyrir iOS fyrir framtíðarþróun (sjá einnig L8).

Varða 7 – lýsing

Vegvísir fyrir Android- og iOS-þróun tilbúinn. Endurskoðaðar vörður M8 og M9 í samræmi við vegvísinn.

Afurðir

Öllum markmiðum vörðunnar er náð. Skýrt er frá vegvísi fyrir Android- og iOS-þróun að neðan. Skilgreiningar varða fyrir M8 og M9 hafa verið endurskoðaðar.[47]

Skýrsla

Talgreining á Android og iOS
Tvö helstu stýrikerfi síma sem rætt er um í markmiði þessa undirverkefnis hafa bæði innbyggðan[48] stuðning við talgreiningu margra tungumála. Þetta felur í sér ýmsa notkun talgreiningar sem inntak eins og raddritun, raddleit, og talþjónn (e. voice assistant).

Android-umhverfið gefur kost á viðmótum fyrir talgreiningu sem hjálpar við útfærslur þriðju aðila sem má setja upp og nota á hátt sem er í raun á jöfnum grunni og innbyggða útfærslan. Enn fremur leyfir það þróun á útfærslu þriðju aðila á eigin meðhöndlara fyrir hinar ýmsu stillingar raddinntaks (þ.e. raddritun, leit, o.s.frv.). iOS-umhverfið gefur ekki kost á neinu af þessu, og þróendur verða að nota innbyggða talgreiningu[49] eða að nota sérbyggða útfærslu í hverju og einu forriti.

[47] Sjá einnig skjalið Endurskoðaðar vörður 3. árs.
[48] Í reynd fyrir Android. Útvegað af Google Services, sem fylgir langflestum Android-símum.
[49] Enginn stuðningur við íslensku.

Á öðru verkefnisári þróuðum við Heyra, óútgefna frumgerð Android-talgreiningarþjónustu. Frumgerðin er útfærð með Kotlin og tengist vefþjónustu Tiro Speech úr undirverkefni H8 og sendir gögn milli tækis og þjóns til að útfæra Android-þjónustuna. Þjónustan verður aukin, útgefin og notuð til að bæta möguleikum talinntaks við sýndarlyklaborð.

Í iOS er mögulegt að nota annaðhvort gRPC- eða REST-viðmótin í vefþjónstu Tiro Speech til að veita talgreiningarmöguleika fyrir einstaka forrit í stað þess að nota innbyggða viðmótið.

Talinntak í Android og iOS
Í Android eru nokkur mismunandi viðmót, eða verknaðir, sem forrit geta notað til að biðja um taltextun, t.d. fyrir einhvern innsláttarreit. Þessir verknaðir nota eina af tiltækum kerfisþjónustum talgreiningar til að greina tal notandans og mögulega gefa möguleika á því að breyta svarinu á einhvern hátt áður en greindur texti er sendir til baka í tiltekið forrit.

Til að uppfylla markmið verkefnisins er sýndarlyklaborð með möguleika á talinntaki nauðsynlegt. Við munum kvísla og byggja vinnu okkar á AnySoftKeyboard-verkefninu (https://github.com/AnySoftKeyboard/AnySoftKeyboard) í samstarfi við undirverkefni L8 – Málrýni í snjalltækjum. Eins og nefnt var að ofan munum við fyrst bæta frumgerð Android-talgreiningarþjónustunnar sem þróuð var á öðru ári. Við munum svo nota þá þjónustu til að útfæra verknað sem meðhöndlar Android-ásetning (e. intent), `ACTION_RECOGNIZE_SPEECH`, sem kemur frá sýndarlyklaborðinu sem fyrsta skref. Sem annað skref munum við útfæra InputMethodService með notkun þjónustunnar, þar sem það gefur kost á betri samþættingu við lyklaborðið.

Fyrir iOS gæti möguleg útfærsla byggst á opna lyklaborðsrammanum KeyboardKit (https://github.com/KeyboardKit/KeyboardKit).

Vegvísir / Vörður:

M8
- Gefa út alfaútgáfu Heyra, sem er útfærsla á talgreiningarþjónustu Android með því að nota vefþjónustu á Google Play
- Taltextun (e. speech-to-text) í AnySoftKeyboard með því að nota meðhöndlara ásetnings

M9
- Gefa út útgáfu 1.0 af Heyra á Google Play
- Taltextun útfærð í AnySoftKeyboard með innsláttaraðferðarútfærslu (e. input method implementation)

H10 – Raddstýring, fyrirspurnir

Skýrsla: Háskólinn í Reykjavík, Tiro ehf
Aðilar: Háskólinn í Reykjavík, Tiro ehf

Markmið

Að þróa Kaldi-forskriftir fyrir raddstýringar- og spurningasvörunarkerfi á íslensku. Raddstýring vísar til einhliða samskipta milli manns og kerfis, með það að markmiði að biðja um einhverja aðgerð. Frá sjónarhóli talgreiningar krefst þetta þess að mállíkanagerðinni sé veitt sérstök athygli, þar sem ekki er víst að segðin lúti hefðbundnum reglum samfellds máls. Talgreining fyrir spurningasvörun er háð sambærilegum skilyrðum með sérhæfða uppbyggingu segða, en í spurningaformi. Þetta þýðir að markmiðuð tungumálalíkön mætti þróa fyrir verkefnið. Nálgunin hér snýst um að skilgreina sértæk verkefni innan almennra flokka talgreinikerfa og búa til Kaldi-forskriftir fyrir þau verkefni. Væntur fjöldi þróaðra verkefna er 3 í hverjum flokki.

Varða 7 – lýsing

Verkefni skilgreind og málfanga aflað fyrir að minnsta kosti tvö verkefni.

Afurðir

Öllum markmiðum hefur verið náð. Við höfum skilgreint lista verkefna sem njóta góðs af sérhæfðu mállíkani. Málföngum hefur verið aflað fyrir tvö verkefnanna.

Skýrsla

Fyrsta verkið var að finna og skilgreina lista verka sem henta flokkunum tveimur sem lagðir eru til fyrir þetta verkefni. Til þess settum við saman vinnuhóp með meðlimum frá öllum þátttakendahópum. Saman bjuggum við til lista verka til að velja úr. Listanum er skipt í tvö sett eftir því hvaða flokk þau falla í, sjá töfluna að neðan.

Raddstýring	**Spurningasvörun**
Hjálparskipanir - Heimilisstjórn - Neyðarhnappur - Spila tónlist - Ljósastýring - Setja áminningar - Sjónvarpsfjarstýring - Leiðsögn Útfylling dálka - Heimilisföng - Símanúmer - Full nöfn - Kennitölur	Spurningar úr spurningaleikjum Hjálparspurningar - Biðja um upplýsingar um veður - Umbreyting eininga - Umbreyting gjaldmiðla - Biðja um leiðsögn og staðsetningu - Spyrja um tímann - Opnunartímar - Reiknivél - Bóka pöntun

Fyrir þessa vörðu ákváðum við að byrja á einu verki fyrir hvorn flokk. Valin verk eru íslensk heimilisföng og spurningar úr spurningaleikjum, undirstrikuð í töflunni að ofan. Seinni hluti þessarar vörðu fól í sér öflun gagna fyrir þessi tvö verk.

Spurningar úr spurningaleikjum

Fyrir þetta verk vildum við safna eins mörgum spurningum og mögulegt var frá nokkrum mismunandi gagnagjöfum. Við fundum samtals þrjá gagnagjafa:

- **Spurningar á íslensku:** Safn spurninga á íslensku með mörgum erfiðleikastigum og flokkum.
- **Spurningar.is:** Safn spurninga og svara fyrir náttúrulega málvinnslu.
- **Gettu betur:** Íslenskur spurningaleikjaþáttur sem hefur verið á RÚV í nokkra áratugi.

Allir þessir gagnagjafar innihalda bæði spurningar og svör en í öllum tilfellum þurfum við aðeins spurningarnar. Samtals söfnuðum við um 36.000 spurningum.

Spurningar á íslensku
Þetta er opið gagnasett spurninga á íslensku. Það er gefið með opnu leyfi á GitHub á https://github.com/sveinn-steinarsson/is-trivia-questions. Málheildin inniheldur 11.610 spurningar í 7 mismunandi flokkum.

Spurningar.is

Þetta er verkefni sem er enn í gangi við Háskóla Íslands. Gögnin hafa enn ekki verið gefin út opinberlega en þau verða það í framtíðinni. Við fengum góðfúslega leyfi til að fá aðgang að 20.378 spurningum sem hefur þegar verið safnað. Þessar spurningar koma frá almennum gjöfum og eru um hvað sem finna má á netinu.

Gettu betur
Við höfðum samband við marga spurningahöfunda þáttarins. Nokkrir svöruðu og gáfu okkur takmarkaðan aðgang að spurningum þeirra. Við völdum að nota aðeins spurningar úr hraðaspurningum þar sem aðrar spurningar geta verið allt að því nokkrar málsgreinar, á meðan hraðaspurningar eru að mestu stuttar setningar. Þetta sett inniheldur rétt yfir 4.000 spurningar. Þessar spurningar eru ekki almenningseign og verða ekki gefnar út en má nota sem hluta af þessu verkefni til að búa til mállíkan.

Heimilisföng

Fyrir þetta verk þurftum við annaðhvort að safna öllum mögulegum heimilisföngum eða búa þau til út frá opnum gögnum. Við völdum seinni aðferðina þar sem auðveldara er að endurframkvæma hana þegar nýjar götur er byggðar eða þær breyta um nafn. Til þess að gera þetta notuðum við lista löglegra heimilisfanga úr Fasteignaskrá (https://skra.is). Við búum heimilisföng til út frá nokkrum algengum mynstrum þar sem póstnúmer og hverfi eru annaðhvort til staðar eða þeim sleppt. Að auki erum við að vinna að reglubyggðum gjafa sem getur annaðhvort búið til gögn út frá setti málfræðireglna eða búið til endanleg stöðuferjöld (e. finite-state transducers, FST) beint úr reglunum.

H11 – Sérhæfð talgreining

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að þróa Kaldi-forskriftir fyrir börn, unglinga og þá sem hafa íslensku sem annað mál. Þessi vinna er áframhald þróunar talgreinaforskrifta fyrir íslensku sem beinist að notendahópum sem þarf að aðlaga bæði hljóð- og mállíkön fyrir með einhverjum sérstökum hætti.

Varða 7 – lýsing

Kaldi-forskrift fyrir börn.

Afurðir

Röðun varða í H11 hefur breyst[50] þar sem gögn barna eru nú tilbúin en ekki gögn frá þeim sem hafa íslensku sem annað mál. Upprunaleg röðun fer frá toppi til botns; nýja röðunin er táknuð með tölu í sviga.

- Þróa Kaldi-forskrift fyrir þá sem hafa íslensku sem annað mál (3)
- Þróa Kaldi-forskrift fyrir unglinga (2)
- Þróa Kaldi-forskrift fyrir börn (1)

Samkvæmt því hefur öllum markmiðum vörðunnar verið náð. Kaldi-forskrift fyrir tal barna hefur verið búin til með nýju málheildinni „Samrómur Children“. Forskrift fyrir börn (4 til 12 ára gömul) er aðgengileg á GitHub á https://github.com/cadia-lvl/samromur-asr/tree/s5_children_lstm.

Skýrsla

Eins og er er stærð málheildarinnar Samrómur Children 131 klukkustund gildra gagna frá börnum á aldrinum 4 til 17 ára. Þrátt fyrir að virðast mikið er þetta lítið magn gagna, þar sem þeim þarf að dreifa milli mismunandi aldurshópa, sem eru ekki jafnir í málheildinni.

Til að tryggja kjöraðstæður fyrir tilraunir var prófunarhluti málheildarinnar gætilega valinn til að samanstanda af nákvæmlega sama magni fyrir alla aldurshópa.

Þessi forskrift er byggð á Kaldi-forskriftinni sem búin var til í M5, þar sem hún inniheldur LSTM-þjálfun. Þrátt fyrir það voru nokkrar nýjar aðferðir notaðar til að gera ferlið sem almennast, til að spara tíma og vinnu. Eftirfarandi er yfirlit yfir þær nýjungar:

[50] Sjá Endurskoðaðar vörður 3. árs.

- Python3-skriftu var bætt við fyrir söfnun gagnanna úr „metadata.tsv“-skránni. Áður var það bash-skrifta (allar útgefnar málheildir koma með lýsigagnaskrá).
- Skriftuna sem nefnd er að ofan má nota fyrir hvaða gögn sem er á TSV-gagnasniði Samróms.
- Forskriftirnar taka nú við hljóðskrám á flac-sniði, sem er sniðið sem notað er til að gefa Samrómsgögnin út.
- Skilaboð í skelinni gefa notanda til kynna þegar ákveðnu þrepi forskriftarinnar er lokið.
- Í skriftunni run.sh má taka fram hvaða þrep skal framkvæma með breytunum „from_stage“ og „to_stage“.
- Mállíkönin og framburðarorðabækurnar eru ekki búnar til á ferðinni eins og áður. Slóðir þeirra verður að skilgreina í skriftunni og mállíkönin verða að vera á ARPA-sniði. Þessi breyting veldur því að hægt verður að gefa mállíkönin og orðasafnið út á CLARIN, OpenSLR, o.s.frv., í stað þess að gefa út hráa textann[51].

Eftirfarandi tafla sýnir niðurstöður sem náðust með notkun HMM ásamt LSTM:

Forskrift	**Aldursbil**	**HMM villuorðahlutfall**	**LSTM villuorðahlutfall**
s5_children_lstm (dev)	4–12 ára	22,08%	11,24%
s5_children_lstm (test)	4–12 ára	34,95%	18,06%

Taka skal fram að Kaldi-forskriftirnar sem þróaðar voru í þessari vörðu eru almennustu forskriftirnar sem við höfum enn sem komið er, svo það verður ekki vandamál að nota þær fyrir gögn annarsmálshafa, þegar þau eru tilbúin, og fyrir gögn fullorðinna í M8.

[51] Mállíkönin og framburðarorðabækurnar fyrir forskriftirnar hafa verið gefnar út á CLARIN á http://hdl.handle.net/20.500.12537/172

T4 – Vefgáttir fyrir talgervla

Skýrsla: Tiro ehf.
Aðilar: Tiro ehf.

Markmið

Að þróa talgreiningarviðmót fyrir Android lyklaborð. Áherslan á fyrri hluta þriðja árs er að þróa frumgerð Android lyklaborðs með einföldu talviðmóti. Vinnan verður gerð í samstarfi við L8 – Málrýni í snjalltækjum (3 mánuðir frá H9 og 3 mánuðir frá L8). Á þriðja ári munum við skoða aðferðir til að bæta talgreiningu og málrýni við lyklaborð snjallsíma, bæði sem vefþjónustu og sem einingu á tæki. Við munum búa til frumgerð lyklaborðs með eiginleika tals-til-texta fyrir Android, með vefþjónustu. Vegna þverrandi fjárframlaga munum við leggja áherslu á útgáfu lausnar fyrir Android, en við munum einnig skilgreina þarfir fyrir iOS fyrir þróun í framtíð (sjá einnig L8).

Varða 7 – lýsing

Fyrsta áfanga hugbúnaðarprófana er lokið.

Afurðir

Fyrsta áfanga hugbúnaðarprófana er lokið. Fyrsti áfangi samanstóð af innleiðingu einfaldrar sjálfvirkrar einingaprófunar í hugbúnaðarþróunarferlið ásamt því að búa til prófanir fyrir fyrirfram ákveðinn hluta kóðagrunnsins. Þessu markmiði hefur verið náð.

Hugbúnaðurinn er aðgengilegur á GitLab á https://gitlab.com/tiro-is/tiro-tts.

Skýrsla

Innleiðing hugbúnaðarprófunar
Einingaprófunarhugbúnaður hefur verið innleiddur í verkefnið. Prófunarferlið er sjálfvirkt og í hvert skipti sem verkefnið fær uppfærslu keyrast prófanirnar, og uppfærslurnar eru ekki gefnar út nema þær standist allar prófanir. Þetta þýðir að minni líkur eru á því að hvaða breytingar eða uppfærslur sem koma í framtíð geti skemmt eitthvað eða orsakað ófyrirséða hegðun.

Einingaprófanir
Við byrjun vörðunnar var ákveðinn hluti kóðagrunnsins skilgreindur til prófunar. Þessi hluti samanstendur af mörgum minni hlutum þar sem nokkrar einingaprófanir eru útbúnar fyrir hvern hlut. Einingaprófanir eru, í einföldu máli, þegar eitthvað inntak er gefið og búist er við

ákveðnu (réttu) úttaki. Inntakið sem er gefið fyrir þessar prófanir nær yfir helstu notkunardæmi fyrir hvern hluta ásamt óalgengum dæmum sem gætu komið upp.

Hver hluti af skilgreindum stærri hluta kóðagrunnsins hefur verið prófaður ítarlega. Eftir því sem við höldum áfram að þróa þennan hugbúnað stefnum við á að auka dekkun prófana og bæta við einingaprófunum fyrir fleiri hluta kóðagrunnsins. Það mun gera hann enn þolnari og minnka líkur á villum í raunverulegri notkun.

Að lágmarka þarfir og ályktunartíma líkans
Þó að það sé ekki strangt til tekið hluti af M7-vörðu var talið nauðsynlegt að bregðast við endurgjöfum varðandi ályktunartíma frá tveimur helstu notendum talgervilsþjónustunnar, WebRICE (undirverkefni T6) og Símaróms (undirverkefni T5), og gera tilraunir með bestunaraðferðir.

Tauganetsraddirnar tvær sem þjálfaðar eru af Háskóla Reykjavíkur samanstanda af fjórum líkönum, hljóðlíkani[52] til að breyta fónemrunum í mel-rófrit og raddkótari[53] til að breyta mel-rófritum í hljóðdæmi fyrir hverja rödd. Líkönin eru þjálfuð með PyTorch og ekki send út eða bestuð fyrir ályktun. Þetta er minna vandamál þegar ályktun er keyrð fyrir talgervilsþjónustuna á vefþjóni heldur en fyrir kraftminni tæki, eins og snjallsíma. PyTorch-ramminn skilgreinir TorchScript,[54] sem er í grunninn lítið undirsett Python til að standa fyrir PyTorch-líkani og raða því til að hlaða í öðrum frammistöðumiðaðri forritunarmálum. Að senda út líkan til TorchScript er líka fyrsta skrefið fyrir bestun. Vinna okkar við að breyta og besta líkönin er í samstarfi við undirverkefni T5. Viðmiðið er ályktun x86_64-þjóns með einhvern sveigjanleika fyrir þetta undirverkefni. Í T5 er viðmiðið ályktun Android-snjallsíma á Android/ARM. Sjá skýrslu T5 fyrir nánari upplýsingar.

[52] FastSpeech2-útfærsla: https://github.com/cadia-lvl/FastSpeech2
[53] MelGAN-útfærsla: https://github.com/seungwonpark/melgan
[54] TorchScript: https://pytorch.org/docs/1.10.1/jit.html

T5 – Talgervlar í snjallsímum

Skýrsla: Grammatek ehf
Aðilar: Grammatek ehf, Háskólinn í Reykjavík, Tiro ehf

Markmið

Að hanna talgervla sem styðja aðgengisviðmót snjallsíma (e. smartphone accessibility interface). Eldri Android-raddir, Karl og Dóra, er verið að úrelda og leggja af, þannig að framtíðarnotendur skjálesara munu ekki geta notað þær. Enn fremur vantar iOS stuðning fyrir íslensku svo það verður að leysa, annaðhvort með því að hanna hann sjálf, eða með samstarfi við erlend fyrirtæki.

Vefbyggðir talgervlar geta nýtt sér meira vinnsluafl en talgervlar í tæki og þar af leiðandi búið til betra tal. Að bæta hágæða vefröddum við báða talgervlana er því ákjósanlegt í notkunardæmum þar sem náttúrulegir eiginleikar eru mikilvægari (í samvinnu við T4).

Varða 7 – lýsing

Endurbætt útgáfa Android-talgervilsappsins frá öðru ári.

Afurðir

Öllum markmiðum hefur verið náð. Android-smáforritið Símarómur inniheldur tauganetsraddir varpaðar á viðeigandi snið til að keyra þær á símanum.
Uppfærð opin betaútgáfa sem inniheldur nýju eiginleikana verður aðgengileg á Google Play 14. febrúar. Tveimur vikum seinna munum við gefa út útgáfu 1.0 af Símarómi. Kóðinn er aðgengilegur á https://github.com/grammatek/simaromur.

Skýrsla

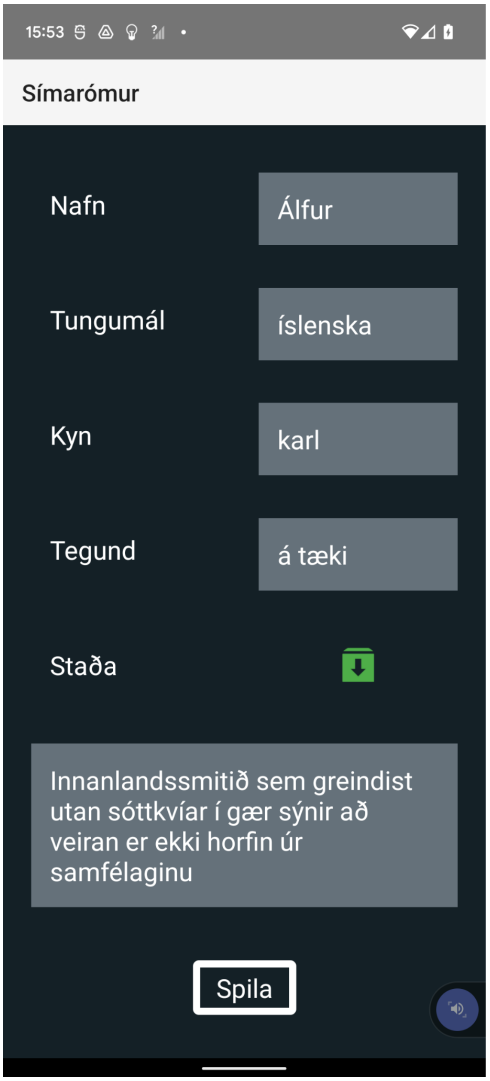
Innleiðing raddar á tæki

PyTorch-rammi snjallsíma er notaður til að keyra talgervilsraddir beint á snjallsímanum. Til að gera þetta verður að breyta raddlíkönunum á annað snið. Annars vegar er það til að gera líkönin samræmanleg við keyrslulag á Android, og hins vegar sparast geymslupláss á tækinu með skömmtun (e. quantization). Einnig eykst keyrslu-/álykturnarhraði með breytingu líkans. Greint er nánar frá þessum breytingarskrefum að neðan.

Við höfum tvö líkön fyrir raddir á tæki: Álf og Diljá. Til að greina betur í sundur raddir á tæki og raddir úr netforritaskilum eru allar raddir af netinu með viðskeytið „ net“, þ.e. „Álfur net“, „Diljá net“, o.s.frv., en raddir á tæki heita einungis „Álfur“, „Diljá“. Einnig sýnir smáforrit

Símaróms gerð raddar á upplýsingaskjá. Fyrir þessar raddir er til staðar framendapípa á tæki sem inniheldur normun og G2P.

Raddaskjár, sýnir raddir á tæki og á neti



Skjár sem sýnir nánari upplýsingar netraddar

Breyting tauganetslíkans

Hvor raddanna tveggja sem eru til staðar í netforritaskilunum samanstendur af tveimur líkönum: FastSpeech2[55]-hljóðlíkani, sem breytir fónemsrunu yfir á mel-rófrit, og MelGAN[56] -raddkótari, sem breytir mel-rófriti yfir á bylgjuform. Líkönin eru útfærð og þjálfuð með PyTorch. Meginforritaskil PyTorch og inning (e. runtime) leggja áherslu á auðvelda smíði, þjálfun og tilraunir líkana frekar en ályktun í framkvæmd. Til að leysa þetta vandamál er til staðar stuðningur til að senda út þjálfuð líkön fyrir ályktun með öðrum inningum. PyTorch hefur TorchScript, sem er í grunninn takmarkað undirsett af Python með kyrrlega gagnagerð sem má nota til að lýsa líkanahögunum og raða til að hlaða inningum í öðrum málum (t.d. C++). Skömmtun og bestun fyrir markumhverfin má svo framkvæma.

[55] FastSpeech2-útfærsla: https://github.com/cadia-lvl/FastSpeech2
[56] MelGAN-útfærsla: https://github.com/seungwonpark/melgan

Sem fyrsta skref þess að lágmarka þörf á geymsluplássi og keyrslutíma tauganetsradda fjarlægðum við allt úr þjálfuðu líkönunum sem þarf ekki fyrir ályktun og sendum út í TorchScript. Í tilfelli raddkótaranna fól þetta í sér að fjarlægja allt nema gjafalíkanið, sem minnkaði stærð líkansins frá u.þ.b. 240MiB í 17MiB. Fyrir hljóðlíkanið var þessi minnkun frá u.þ.b. 330MiB í 110MiB. Frekari minnkun á stærð kom til vegna skömmtunar á stillingum líkansins frá 32 bita kommutölu í 8 bita heiltölu. Þetta krafðist nokkurra tilrauna og meðhöndlunar þar sem ekki var mögulegt að skammta öll lög og aðgerðir. Við náðum að skammta hljóðlíkanið, en ekki MelGAN-raddkótarana. Þegar hljóðlíkanið er skammtað og bestað fyrir x86 er stærðin u.þ.b. 30 MiB og fyrir ARM (snjallsíma), er hún u.þ.b. 60MiB.

PyTorch-rammi snjallsíma hefur takmörkuð forritaskil sem hentar ekki beinlínis til meðferðar á þinum (e. tensors) sem ætlaðir eru sem annaðhvort inntak í eða úttak frá líkönunum. Því bættum við einfölduðum aðferðum við viðmót snjallsíma til að senda líkönin út með algjöru lágmarki inntaks- og úttaksstillinga.

T6 – Veflesari

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að gera talgervil aðgengilegan fyrir lestur á vefsíðum. Fyrsta markmiðið er að þróa tól sem gerir talgervil á íslensku aðgengilegan vefhönnuðum til að nota á vefsíðum sínum. Tvö aukamarkmið eru að búa til viðbót fyrir vafra og að kynna opna íslenska talgervla sem verða í boði fyrir utanaðkomandi veflesarahönnuði.

Varða 7 – lýsing

SSML (Speech Synthesis Markup Language)[57]-stuðningur í WebRICE, vafraviðbót.

Afurðir

Afurðirnar voru ekki kláraðar að fullu. Viðbótin var að fullu útfærð fyrir Firefox-vafrann en hefur ekki verið bætt í viðbótarverslun Firefox. Þetta verkefni hefur ekki áhrif á M8 eða M9 þar sem það er ekki háð öðrum verkþætti, og hefur auk þess ekki fyrirfram ákveðið markmið sem myndi skarast innan M8. Afurðirnar verða kláraðar tveimur mánuðum seinna en áætlað var, 1. apríl.

WebRICE-viðbótirnar eru aðgengilegar á GitHub á https://github.com/cadia-lvl/Webrice_extensions.

Afurðir sem ekki er búið að klára eru:

- Bæta við stuðningi fyrir SSML.

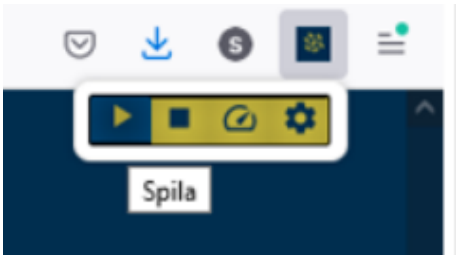
Áætlun til að klára afurðirnar:

- Bæta við stuðningi fyrir SSML.
- Gefa út uppfærða WebRICE-viðbót í Firefox-verslunum.

Skýrsla

Webrice-viðbótin er tól sem notendur geta sótt úr viðbótarverslun og innleitt í vafra til að geta hlustað á texta. Í M7 var minnsta raunhæfa afurð fyrir Firefox-viðbótina útfærð. Vegna tímaskorts gátum við ekki tengt viðbótina við aðra vafra.

[57] https://www.w3.org/TR/speech-synthesis/



Viðmót WebRICE-viðbótarinnar

Viðbótin hefur eftirfarandi eiginleika:

- Spilunarhnappur
    - Spilar valinn texta eða heila síðu ef enginn texti er valinn
    - Tengt við bakenda Tiro (undirverkefni T4)
- Hléhnappur (birtist við spilun)
    - Gerir hlé á/heldur áfram með núverandi hljóð
- Stopphnappur
    - Stöðvar núverandi hljóð og endursetur það til að spila valinn texta
- Stillingahnappur til að sýna notendaleiðbeiningar
- Hraðahnappur
    - Sýnir fellivalmynd með hraðavalmöguleikum fyrir spilun

Áherslan við þessa vörðu hefur verið á þróun viðbótar fyrir Firefox, og ná tilbúinni lágmarksraunhæfri afurð. Nokkur vandamál komu upp, t.d. villur í Firefox-vafranum sjálfum vegna uppfærslna. Við unnum að því að gera viðmótið þægilegt og auðvelt í notkun með með móttækilegum hnöppum.

Til að stjórna ákveðnum eiginleikum tals eins og framburði, hljóðstyrk, tónhæð, hraða, o.s.frv. fyrir ólík umhverfi með möguleikum talgervils má nota SSML sem almenna skilgreiningu af ætluðu úttaki. Fyrstu áætlanir innihéldu valkosti til að stjórna þessum eiginleikum ítarlega í umhverfinu, en vegna tímaskorts var ákveðið að hafa aðeins hraðastillingar spilunar í viðbótinni. Þó að hægt sé að stjórna hraða beint með SSML-málskipan kusum við að breyta beint spilunarhraða HTML-hljóðstaksins. Þetta hefur þann mikla kost að við þurfum ekki að biðja um hljóðið frá bakendanum í hvert skipti sem notandi breytir hraðastillingunni.

Næstu skref eru að taka með greiningu á völdum texta og bæta SSML-málskipan við hann áður en hann er sendur til bakenda (verkefni T4 sem styður SSML-inntak). Markmiðið er að meðhöndla spurningar og stutt hik milli búta. Við munum einnig vinna að því að bæta viðbótinni við í viðbótarverslun Firefox.

T8 – Mat á gæðum talgervla

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík

Markmið

Að veita nauðsynlegan stuðning við þróun talgervla með því að meta frammistöðu fyrir ýmsar útfærslur talgervla. Eftir bakgrunnsrannsókn verða þrjár til fimm aðferðir valdar og innleiddar í hugbúnaði. Prófanir eru svo skipulagðar með því að undirbúa raddirnar, tímasetja prófanir og fá hlustendur. Niðurstöðurnar eru einnig metnar og greint frá þeim. Hugbúnaður „neðanstreymis“ (e. downstream) eins og t.d. afurðir T5, eru notendaprófaðar og skýrslur gerðar fyrir hverja og eina.

Varða 7 – lýsing

Huglægt mat á (T7. M6), matsverkvangur gerður aðgengilegur þeim sem þróa talgervla, ásamt viðeigandi skjölun.

Afurðir

Öllum markmiðum hefur verið náð. Við komum á fót matsverkvangi talgervla sem við hýsum á vefþjóni okkar https://eyra.ru.is/mosi/login?next=%2Fmosi%2F. Kóðinn fyrir verkvanginn er gefinn út með Apache 2.0 leyfi á CLARIN á http://hdl.handle.net/20.500.12537/186.
Til notkunar getur fólk sett upp eigið tilvik verkvangsins með því að sækja kóðann eða hafa samband við okkur til að fá aðgang að okkar vefþjóni í einhverjum tilfellum.
Við höfum framkvæmt mat fyrir talgervla með nýja verkvangnum. Greining á niðurstöðum verður gerð við næstu vörðu.

Skýrsla

MOSI

Þróun
Þróun verkvangsins hefur reynst krefjandi. Markmiðið er að þróendur geti sent inn talgervilshljóðskrár og á auðveldan máta keyrt matsprófanir.

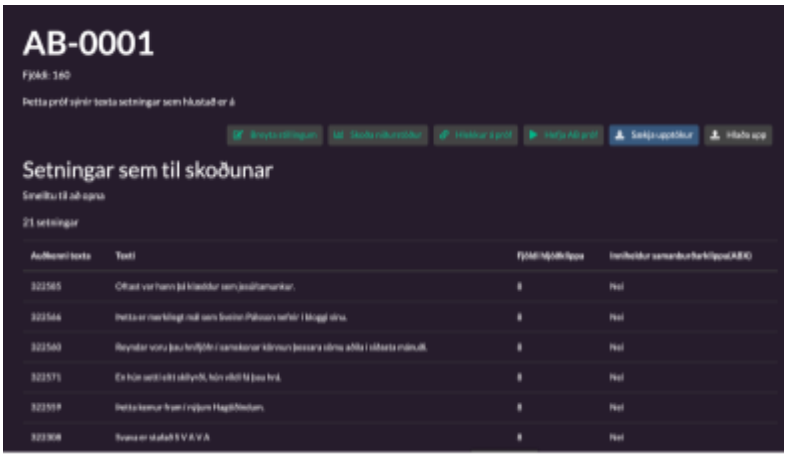
Við byrjuðum með grunnútgáfu Flask-forrits sem kallast LOBE, sem er upptökubiðlari fyrir talgervla sem við útfærðum möguleika MOS-prófana í. Við ákváðum að skilja upptökubiðlarann frá matsverkvangnum.

Eftir að hafa sett upp grunnsnið og síað út gamlan LOBE-kóða stóð eftir verkvangur sem getur framkvæmt MOS-prófanir. Við ákváðum að gefa verkvangnum nafnið MOSI, sem vísar bæði til MOS-prófana og íslenska orðsins „mosa“, sem hljómar líka vel. Við bættum við möguleika fyrir AB- og ABX-prófun hljóðbúta. Þetta krafðist jafnvægis milli þess að vilja gefa möguleika á mörgum afbrigðum prófunaruppsetninga og þess að gera verkvanginn auðveldan í notkun. Nýlega bættum við grunnstuðningi fyrir SUS-prófanir við. SUS-prófun má aðallega nota til að gefa þátttakendum aðgang að hljóðskránum, hugsanlega hýsingu hljóðskráa. Hugsanlega getum við bætt fullum SUS-stuðningi við ef þurfa þykir í framtíðinni.

Mynd: Skjáskot af því hvernig AB-prófun birtist

Prófanirnar gera notandanum kleift að senda inn hljóðbúta, velja samanburð eða hljóðbúta sem eru viðeigandi og senda svo hlekki til prófara. Þegar búið er að kynnast verkvangnum er hægt að gera þetta á örfáum mínútum. Fyrir flóknari innsendingar gæti stilling tekið lengri tíma. Höfundur getur þá haldið utan um niðurstöður og sótt gögnin eða skoðað þau í verkvangnum ásamt viðeigandi gröfum.

Fyrsta hugmynd okkar var að setja verkvanginn upp á vefþjóni þriðja aðila eins og Amazon. Þetta myndi hjálpa okkur að fá tilfinningu fyrir því hvernig aðrir notendur gætu viljað sett upp verkvanginn. Ef auðvelt er að setja hann upp á einhverju eins og Amazon getur hver sem er gert það án vandræða. Þrátt fyrir að þetta sé enn gott markmið virtist það of flókið og dýrt eins og er. Við notuðum þess í stað eigin þjón sem nefnist *Eyra*. Eftir örlítil vandræði með geymsluplass útfærðum við verkvanginn án vandræða.



Mynd: Skjáskot af því hvernig rannsakendur sjá uppsetnigarskjá fyrir AB-prófanir eftir innsendingu

Að gera Mosa aðgengilegan talgervilssamfélaginu
Eins og áður kom fram er stefnan að gera þennan verkvang aðgengilegan talgervilssamfélaginu. Því gerum við kóðann aðgengilegan öðrum til að setja upp verkvanginn. Eins og stendur er uppsetningarferlið tiltölulega einfalt fyrir hvern sem hefur smá reynslu í vefforritun og Python, sérstaklega Flask. Markmiðið er að einfalda það enn frekar svo að hver sem er ætti að geta sett hann upp fyrir eigin rannsókn. Annar valkostur væri að gefa rannsakendum kost á að biðja um aðgang að vefþjóni okkar til að keyra prófanir tímabundið.

Mikilvægur hluti af því að deila kóða og vinna á þennan máta er skjölun og leiðbeiningar. Það er auðvitað huglægt, hvenær er til nóg skjölun? Við reynum að skjala það sem við teljum viðeigandi, en mögulega þarf að vinna meira að þessu. Við munum halda áfram að skjala eftir því sem þróun miðar áfram.

Margmálastuðningur er annar mikilvægur eiginleiki sem ætti að íhuga. Eins og stendur er íslenska eina tungumálið sem er stutt í verkvangnum, en þetta má bæta nokkuð auðveldlega, fyrst með ensku. Hugsanlega gæti opna samfélagið (e. the open-source community) hjálpað við þetta fyrir önnur tungumál ef verkvangurinn nær einhverjum vinsældum.

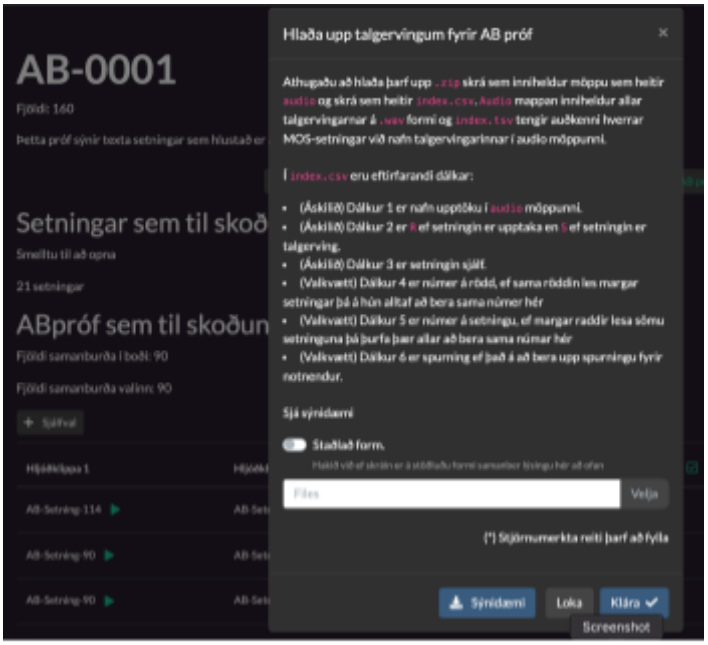


Image: Screenshot of a window where researchers can upload clips

Verkvangurinn er enn á því sem við getum kallað betafasa. Við erum að pússa gróf horn og finna villur. Þetta er áframhaldandi verkefni ætlað máltækni fyrir framtíðina, og mun því halda áfram að þróast.

Leiðbeiningarnar eru geymdar ásamt kóðanum í formi .md-skráa, og einnig eru leiðbeiningar innan verkvangsins fyrir notendur að lesa.

Mat á talgervlum (T7. M6)

Uppsetning tilraunar
Í fyrstu AB-matsprófunum okkar þjálfuðum við þrjú mismunandi Parallel WaveGAN[58]-líkön, sem eru:

- Karlkyns líkanið: Þetta líkan var þjálfað með öllum hljóðdæmum karlkyns radda úr gagnasettinu Talrómur 1.
- Kvenkyns líkanið: Þetta líkan var þjálfað með öllum hljóðdæmum kvenkyns radda úr gagnasettinu Talrómur 1.

[58] Parallel WaveGan https://github.com/kan-bayashi/ParallelWaveGAN

- Almenna líkanið: Þetta líkan var þjálfað með bæði hljóðdæmum karlkyns og kvenkyns radda úr gagnasettinu Talrómur 1.

Við þjálfuðum 800.000 skref fyrir hvert líkan, þar sem hvert skref jafngildir 256.000 hljóðtökutíðni sinnum sex, sem samsvarar u.þ.b. sjö sekúndum. Að auki höfðum við með grunnsannleik (e. ground truth) í AB-matsprófunum okkar.

Fyrir hverja prófun verður uppsetning innsendra mappa að innihalda hljóðmöppuna, sem inniheldur öll líkönin sem undirmöppur, og svo wav-skrár í hverri undirmöppu. Hún verður einnig að hafa stílblað lýsigagna sem inniheldur eftirfarandi upplýsingar fyrir hverja hljóðskrá, sýnt í töflu 1.

Slóð	S/R	Texti	GT/Synth	model_id	utterance_id	voice_id	question
model1/2021-05-27T093747.464Z_r000189859_t000306680_gen.wav	S	Eygló óttast að ekki verði hægt að afgreiða frumvarpið sem lög fyrir mánaðamót.	0	model1	306680	216	empty

Tafla 1: Dæmi um þær upplýsingar sem þarf fyrir hvert hljóðdæmi í gagnagrunni MOSA.

Undirbúningur gagna
Markmið okkar er að meta hversu góðum árangri Parallel WaveGAN-líkönin okkar þrjú ná með gefnum taldæmum óséðra mælenda, og því völdum við 15 setningar úr talmálheildinni Talrómur 2. Hver setning sem við völdum var tekin upp af karlkyns og kvenkyns mælendum. Alls innihélt matssett okkar 90 taldæmi (30 taldæmi frá Parallel WaveGAN-líkani ótilgreinds kyns, 15 taldæmi frá hvoru líkani ákveðins kyns, og 30 grunnsannleikstaldæmi).

Öflun allra upplýsinga fyrir hvert taldæmi úr Talrómi 2, eins og sýnt var í töflu 1, getur verið tímafrek. Því var ákveðið að búa til bash-skriftu (make_index_csv.sh[59]) sem aflar sjálfkrafa öllum nauðsynlegum upplýsingum fyrir hvert taldæmi. Við sendum svo farsællega fyrstu AB-matsprófun okkar í MOSA, og frekari upplýsingar um AB-matsprófanirnar er að finna í fyrri hluta.

[59] https://github.com/cadia-lvl/IG-for-MOSI/tree/master

T9+T10 – Framendapípa talgervils

Skýrsla: Grammatek ehf, Háskólinn í Reykjavík
Aðilar: Grammatek ehf, Háskólinn í Reykjavík, Hljóðbókasafnið

Markmið

T9: Við lok annars árs munum við eiga mjög nákvæmt textanormunarkerfi fyrir íslensku. Verkefni þriðja árs eru tvíþætt: 1) þróun tóls fyrir áherslur (e. phrasing) og 2) notendaverk (e. use case task) sem felur í sér texta sérsviða. Til að merkja áherslur og hlé (e. pause) í texta sem inntak fyrir talgervilinn má nota þáttara til að sjá til þess að hlé birtist á skynsamlegum stöðum. Við munum nota þáttarana úr undirverkefni I5, en talið er að grunnþáttari sé nóg fyrir þetta verk. Ef aðrar lausnir virðast vænlegri verða þær greindar og gerð grein fyrir þeim. Enn fremur munum við skilgreina notkunardæmi í samstarfi við Hljóðbókasafnið. Kennslubækur eru mjög mikilvægt sérsvið hljóðbóka, t.d. í stærðfræði og efnafræði. Annað sérsvið eru opinberar skýrslur sem innihalda margar töflur, línurit o.s.frv. Við munum skilgreina sameiginlegt notkunardæmi sem er frábrugðið fréttatextum og koma á fót pípu úr lestrarefninu frá útgefanda til normaðs texta fyrir talgervilskerfið.

T10: Að eiga einingu sem umritar normaðan texta sjálfvirkt yfir í hljóðritunartákn. Annars vegar verður einingin tengd talgervilseiningu, til að hljóðrita sjálfkrafa óþekkt orð sem koma inn. Hins vegar má nota einingunа til að útbúa nýjar orðabækur, t.d. fyrir sérhæft sérsvið sem skortir of mörg orð í kjarnaorðabókinni.

Þessi verkþáttur er nátengdur G6, þróun framburðarorðabókarinnar.

Varða 7 – lýsing

T9: Áherslutól innleitt, með því að nota þáttara úr I5. Notkunarprófanir fyrir textanormun sérsviða skilgreindar og skilyrði tilbúin.
T10: Orðabyggð tungumálagreining og tungumála- (og mállýsku-) byggt g2p innleitt, bæði reglubyggt og LSTM. Samsetninga- og orðhlutafræðileg greining innleidd.

Afurðir

Öllum markmiðum hefur verið náð. Við höfum þó verið að bíða eftir því að tiltæka stofnhlutagreiningartólið *Kvistur* verði gert opið til að bæta samsetningagreiningu í g2p-tólinu. *Kvistur* var gerður aðgengilegur fyrir skiladagsetningu þessarar skýrslu og því erum við tilbúin að halda áfram þróun.

- Textahreinsari og þáttari hljóðbóka: https://github.com/grammatek/text-cleaner
- Áherslutól: https://github.com/cadia-lvl/phrasing-tool

- G2P með samsetninga- og tungumálagreiningu: https://github.com/grammatek/ice-g2p
- Uppfærð útgáfa Íslensku framburðarorðabókarinnar http://hdl.handle.net/20.500.12537/181

Skýrsla

Vegna þess hversu sterklega framendaeiningar talgervils eru háðar innbyrðis höfum við sameinað T9 (Forvinnsla texta, textanormun og áherslugreining) og T10 (Sjálfvirk hljóðritun) í eitt verkefni. Afurðir M7 eru allar einingar sem þarf fyrir forvinnslu texta fyrir talgervla ásamt þörfum fyrir framendastjórn talgervla sem skila á við M8.

Notkunardæmi fyrir forvinnslu sérsviðstexta: Þáttun hljóðbóka

Ásamt sérfræðingum frá Hljóðbókasafninu (hér eftir HBS) höfum við skilgreint notkunardæmi fyrir framleiðslu hljóðbóka með talgervilskerfi. Þetta notkunardæmi mun reyna á mörg atriði kerfisins í heild undir álagi. Helsta áhersla okkar verður á aðlögunarhæfni og þol textavinnslupípu: Þau sem búa til hljóðbækur munu þurfa að eiga bein samskipti við pípuna til að stilla mismunandi skref vinnslu fyrir hvert verk og þau verða að geta gert það án nokkurrar forritunarkunnáttu eða mikillar kunnáttu á talgervlum.

HBS hefur þá skyldu að anna eftirspurn við að búa til kennslubækur, gjarnan með stuttum fyrirvara. Því væri mikill kostur að einfalda og hraða því ferli með notkun talgervilskerfis í stað mannlegs lesanda. Sjálfvirk talgerving kennslubóka felur í sér ýmsar áskoranir (við lítum ekki á áskoranir við lestur skáldskaps eins og er): þar sem bækurnar eru kennslubækur er mikilvægt að efnið sé rétt merkt og uppsett, og skiljanleiki kerfisins eftir bestu gæðum. Útgefnar bækur á íslensku eru undirbúnar fyrir gerð hljóðbóka eftir samnorrænum viðmiðum, Nordic Guidelines for the Production of Accessible EPUB 3. Fyrsta skrefið í sjálfvirkri vinnslu fyrir talgervla er því þáttun þessara viðmiða. Eftirfarandi verk/þarfir voru skilgreindar fyrirfram fyrir þáttarann:

- Þagnir: hvar skal setja inn þagnir/hlé og hvernig skal ákvarða mismunandi lengd.
- Merkingar sem eru ekki til staðar í textanum: stundum bendir uppsetning bókar á sérstakt efni, t.d. box utan um texta. Í slíkum tilfellum gæti þurft aukamerkingar til að lesa textann.
- Eru tilfelli þar sem þarf að lesa greinarmerki eins og gæsalappir og sviga upphátt í stað þess að hunsa það, t.d. „svigi opnast“? Hvernig á að draga slíkt út sjálfvirkt?
- Röðun og hlé við lestur taflna, lista, og heimildaskráa.
- Breyting ákafa fyrir texta með áherslu, t.d. texta í tilvitnun eða skásettan texta.

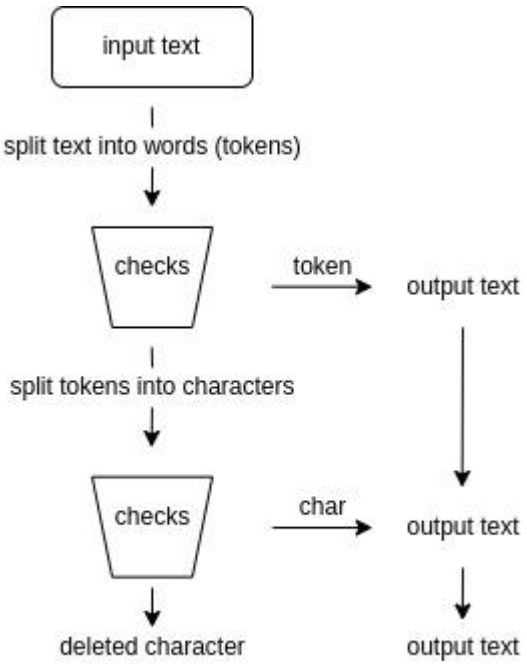
Þau sem búa til hljóðbækur munu einnig þurfa aðgang að kerfinu, til dæmis til að gera breytingar á orðabókinni eða meðhöndlun styttinga, og/eða breyta normuðum texta handvirkt milli vinnsluskrefa.

Fyrir núverandi notkunardæmi valdi HBS félagsvísindi sem aðalviðfangsefni. Það er grein sem margar bækur eru gefnar út innan jafnt og þétt, og gefur okkur tækifæri til að koma á fót traustu vinnsluferli hljóðbóka án þess að takast á við erfiðari verk eins og t.d. normun formúla í efnafræði- eða stærðfræðibókum. Við höfum útfært html-þáttara fyrir EPUB 3-staðalinn með dæmum af bókum úr viðfangsefninu. Hann fjarlægir tög og setur inn bið þar sem við á, og vinnur úr töflum, listum og efnisyfirlitum. Næstu skref eru ítruð þróun þáttarans með því að búa til dæmatexta og fá endurgjöf frá HBS. Í samstarfi við T13 (Stikaðir talgervlar), munum við skilgreina hvernig t.d. má búa til mislangar þagnir og mismunandi áherslustig sem fyrstu skref að sveigjanlegra hljómfalli.

Textahreinsari

Við höfum útfært textahreinsunartól tengt við eiginleika textaforvinnslu og ýmsar hreinsunarstillingar fyrir ákveðin verk. Textahreinsarinn notar margar athuganir til að meta hvaða hluta upprunalegs texta er haldið, hvaða hlutum er breytt og hvaða hlutar eru einfaldlega fjarlægðir.

Myndin að neðan lýsir flæði textahreinsarans í sinni einföldustu mynd. Hann tekur inntakstexta og framkvæmir nokkrar athuganir á hverju orði eftir því hvaða stillingar eru valdar. Þessar athuganir eru til þess að meta hvort orði er haldið, breytt eða sent áfram í hreinsarann. Ef orði er haldið eða breytt er því bætt aftan á „úttakstextann“. Ef það mistekst hins vegar heldur það áfram í hreinsaranum þar sem svipað ferli og áður fer fram. Við framkvæmum athuganir á hverjum staf í orðinu til að meta hvort halda eigi stafnum, breyta eða einfaldlega fjarlægja. Eins og áður, ef honum er haldið eða breytt, er honum bætt aftan á „úttakstextann“.



text cleaner flow diagram

Forvinnslustigið tryggir að engin samfelld greinarmerki séu til staðar, sem gerir notandanum kleift að bæta þeim að vild við enda taga án þess að hafa áhyggjur af samfelldum greinarmerkjum fyrir talgervilskerfið að lesa.

Þar sem textahreinsun fer eftir hverju og einu verki höfum við gefið textahreinsaranum ýmsar stillingar, eins og áður var nefnt. Meðal þessara stillinga eru, ásamt fleirum: að leyfa notendum að skilgreina eigin sett greinarmerkja og bókstafa, breyta hvaða setti strengja sem er í hvaða sett af streng sem er, og halda orðum merkt sem útlensk í textanum en „hreinsa“ önnur ómerkt tilvik, til að nefna nokkur dæmi. Að auki er gefinn kostur á nokkrum nálgunum við meðhöndlun tjámynda í texta. Umbreyting tjámyndar í punkt er

sjálfvalin nálgun en við getum einnig breytt þeim bæði í hvaða streng sem er og textalýsingu þeirra eins og hún er gefin upp í unicode.

Dæmi um texta fyrir og eftir hreinsun (fyrir fleiri dæmi og skjölun, sjá GitHub-hirslu):

1. Venjulegt notkunardæmi gæti litið svona út:

```
clean("Táknin σ og Σ eru notuð í tölfræði til að tákna summur.")
Táknin sigma og sigma eru notuð í tölfræði til að tákna summur.
```

2. Textahreinsarinn gefur sama úttak fyrir lágstafatákn og hástafatákn. En því má breyta fyrir ákveðin verk.

```
clean("Táknin σ og Σ eru notuð í tölfræði til að tákna summur.",
char_replacement={'σ': 'lástafur sigma', 'Σ': 'hástafur sigma'})
Táknin lágstafur sigma og hástafur sigma eru notuð í tölfræði til að tákna
summur.
```

3. Hvaða streng sem er má halda óbreyttum

```
clean("Í eðlisfræði er λ notað um öldulengd.", preserve_string=['λ'])
Í eðlisfræði er λ notað um öldulengd.
```

4. Við umbreytum tjámynd venjulega í punkt, en við getum einnig breytt þeim í lýsingu sína, sem getur verið nytsamlegt t.d. til að lesa tíst með talgervilskerfi.

```
clean("vá hvað er kalt úti! 🥶", clean_emoji=True)
Vá hvað er kalt úti! cold face
```

Þáttunartól
Þáttunartól var þróað með tiltækum markara og þáttara (https://github.com/hrafnl/icenlp/releases). Tólið tekur inn texta frá normaranum, með eina setningu í hverri línu, og notar málfræðileg mörk og málfræðilegar upplýsingar (e. part-of-speech tags and part-of-sentence grammatical information) til að bæta við eðlilegum þögnum í textann. Þagnirnar sem bætt er við eru m.a. þagnir fyrir samtengingar (þegar á við), aukasetningar, og sumar forsetningar. Öllum greinarmerkjum er breytt í viðeigandi langar þagnir.

Þrjár inntakssetningar eru því:

Niðurstaða dómsins var að í júlí tvö þúsund og fimmtán , hafi komist á munnlegur tímabundinn ráðningarsamningur , með gildistíma til loka september tvö þúsund og fimmtán , með fyrirvara af hálfu A um laun .
Af annarri grein laga númer hundrað þrjátíu og níu / tvö þúsund og þrjú verður ráðið að hlutlægar ástæður skuli ráða við ákvörðun um tímabundnar ráðningar .

Verður ekki séð að stefnandi hafi gert athugasemdir við að henni hafi þá verið boðinn nýr tímabundinn samningur , heldur hafi athugasemdir hennar snúið að launalið samningsins og hún neitað að skrifa undir vegna þess .

Og úttakssetningarnar lesast því:

Niðurstaða dómsins var <sil> að í júlí tvö þúsund og fimmtán <pau> hafi komist á munnlegur tímabundinn ráðningarsamningur <pau> með gildistíma til loka september tvö þúsund og fimmtán <pau> með fyrirvara af hálfu A um laun <sil>
Af annarri grein laga númer hundrað þrjátíu og níu <pau> tvö þúsund og þrjú verður ráðið <sil> að hlutlægar ástæður skuli ráða við ákvörðun um tímabundnar ráðningar <sil>
Verður ekki séð <sil> að stefnandi hafi gert athugasemdir við að henni hafi þá verið boðinn nýr tímabundinn samningur <pau> heldur hafi athugasemdir hennar snúið <sil> að launalið samningsins <pau> og hún neitað að skrifa undir vegna þess <sil>

Við þróun þáttunartólsins var keyrslutími hafður í lágmarki. Það tekur 31 millisekúndu að varpa setningunum að ofan frá normuðum til þáttaðra setninga.

Tungumálagreining byggð á orðum

Eftir samanburð við tiltæk söfn eins og Polyglot og TextBlob var tungumálagreining byggð á orðum lokst útfærð í g2p-einingu okkar, einfaldlega með íslenskri þrístæðutalningu á móti erlendri þrístæðutalningu orðanna í framburðarorðabókinni (verkþáttur G6). Inntaksorð eru nú send í g2p LSTM-líkan þjálfað á annaðhvort íslensku eða ensku (sjá M6), eftir því hvaða tungumál er talið líklegra með tungumálagreiningu þrístæða.

Þrístæðuaðferðin virðist henta betur þessu verki heldur en jafnvel bestu tungumálagreiningarsöfn eins og Polyglot. Þau tefjast við það að þurfa að velja frá tugum eða hundruðum mögulegra tungumála og lenda of oft á hvorki íslensku né ensku fyrir fullkomlega ásættanleg orð. Til samanburðar velur þrístæðuaðferð okkar einfaldlega annan tveggja kosta, með töluvert betri árangri.

Prófanir á íslensku fréttatextum um erlent efni bendir til þess að tungumálagreining og ensk LSTM bæti g2p-nákvæmni umtalsvert, bæði hvað varðar orðvillutíðni (WER) og fónemvillutíðni (PER), eins og sést í töflunni að neðan.

Aðferð	Orðvillutíðni (WER)	Fónemvillutíðni (PER)
ice-g2p-eining	26,93	8,61
+ Orðabókarleit	16,29	5,20
+ Tungumálagreining	15,46	3,98
+ Orðabókarleit, tungumálagreining	**10,56**	**2,93**

Í prófunartilviki okkar með einungis tungumálagreiningaraðferð okkar, með blönduðum texta tungumála með rétt yfir 1.000 orð, næst svipuð, en aðeins meiri lækkun í villutíðni miðað við að aðeins orðabókarleit sé notuð. Hvor aðferðin ein og sér skilaði þó greinilega einhverjum villum sem hin skilar ekki, og þar af leiðandi fást bestu niðurstöðurnar með því að nota báðar aðferðirnar.

Athugið að enska er eins og er eina erlenda tungumálið sem einingin okkar hefur eigið g2p-líkan fyrir. Það kann að vera að erlend orð úr öðrum tungumálum sem geta komið fyrir í íslenskum texta henti illa fyrir umritun af líkani sem þjálfað er á ensku og gætu jafnvel hljómað eðlilegar ef þau eru umrituð samkvæmt íslenska hljóðkerfinu. Væntanlega gildir þetta frekar fyrir skyld tungumál, eins og skandinavísk tungumál eða þýsku. Eins og er gerum við aðeins ráð fyrir ensku (sem við gerum ráð fyrir að sé algengasta erlenda tungumálið í íslenskum texta) og látum notandann velja hvort beita skal tungumálagreiningu við umritun texta.

T13 – Stikaðir talgervlar

Skýrsla: Háskólinn í Reykjavík
Aðilar: Háskólinn í Reykjavík, Grammatek ehf

Markmið

Að eiga fyrsta flokks stikaðan talgervil fyrir íslensku. Niðurstöðurnar eftir annað ár munu innihalda ítarlegar rannsóknir, tilraunir og undirbúningsvinnu fyrir hönnun hágæðakerfis. Niðurstöðurnar sýna að taugatalgervlar (e. neural TTS) hafa þróast á þá leið að fýsilegt væri að innleiða þá fyrir íslensku. Á þriðja ári verður áhersla lögð á að hraða þróun taugatalgervils, byggðum á FastSpeech2, sem og að kanna raddblöndun og mælendaaðlögun með því að nota gögn úr T2.

Seinkuð varða 6 – lýsing

Nýjar nálganir við SPSS afturvirkni útfærð fyrir íslensku og ensku. Nýjar nálganir við SPSS raddkótun útfærðar fyrir íslensku og ensku. Skýrsla um gæðamat fyrir þá bestu og nýjar nálganir við SPSS. Skýrsla og hönnun um hvernig skuli stjórna hljómfalli.

Varða 7 – lýsing

Frumgerð sérhæfðrar framendavinnslu fyrir SPSS. Skýrsla um gagnadrifna/eftirlitslausa möguleika fyrir hljómfall og áherslustjórnun (e. phrasing control) í taugatalgervlum. Skýrsla um raddblöndun og aðlögun.

Afurðir

Öllum markmiðum hefur verið náð.

Þeim markmiðum sem voru eftir frá M6 hefur verið náð. Við höfum skilað hágæða líkönum með notkun Tacotron2 og FastSpeech2 auk Parallel WaveGAN-raddkótara sem passar.

Við höfum þróað frumgerð af framenda, sem hluta af þjálfunarforskriftinni fyrir þessi líkön, sem framkvæmir nauðsynleg forvinnsluskref fyrir bæði þjálfun og ályktun. Þessi framendi verður stækkaður til að innihalda áherslugreiningu og normun.

Fræðigrein hefur verið unnin um gagnadrifið hljómfall og áherslustjórnun í tauganetstalgervli og um aðlögun og blöndun radda í tauganetstalgervli, og verður hún gerð aðgengileg máltæknisamfélaginu.

Eftirstandandi afurð frá M6 var þróun og útgáfa fyrsta flokks talgervilslausnar. Við tókum ákvörðun um að útfæra forskriftir okkar og þjálfa líkön okkar með ESPNet-tólasettinu,[60] sem er þróað og viðhaldið af alþjóðlegum hóp rannsakenda sem eru fulltrúar stofnana t.d. í Bandaríkjunum, Japan og Kína. Ákvörðunin um að nota ESPNet frekar en aðra verkvanga eins og Mozilla TTS eða Merlin er að mestu vegna mikillar virkni á ESPNet, og vegna mikils stuðnings við fjölda líkanaútfærslna.

Við notuðum útfærslu[61] af Parallel WaveGAN þróaða af Tomoki Hayashi, sem vinnur mikið í ESPNet. Þessi útfærsla er að fullu samhæfanleg líkönum þróuðum í ESPNet, sem nota mel-tíðnirófrit. Ákvörðunin um að þróa forskrift okkar innan þessa tólasetts mun vonandi gera rannsakendum í hinu alþjóðlega samfélagi kleift að nota hin frábæru talgervilsgögn sem við höfum safnað þegar við framkvæmum togbeiðnir í opinbera hirslu ESPNet, og bætum stuðningi við málheildir okkar við listann.

Þjálfunarferlið fyrir líkönin er eftirfarandi. Fyrst eru þagnir fyrir og eftir tal klipptar úr hljóðskránum með MFA-samröðun. Næst er umritunum varpað frá normuðum texta yfir á atkvæðagreint, hljóðvarpað inntak með áherslumerkingum á sérhljóðum. Orðskiptitókum var einnig bætt við til að bæta raddirnar. Fyrstu tilraunir höfðu sýnt að án atkvæðagreiningar og orðskiptinga kom út rangur framburður sumra samsettra orða og óeðlileg áhersla í einhverjum tilfellum.
Með þessu hljóðvarpaða inntaki voru Tacotron2-líkön þjálfuð fyrir 200 mynstrasöfn, með skoðun eftir hvert mynstrasafn. Tekið var meðaltal af stillingunum fyrir þær fimm skoðanir sem náðu bestum árangri á þróunarsettinu til að fá út endanlegt líkan fyrir hverja rödd.
Eftir að Tacotron2-líkönin voru þjálfuð var þeim beitt á öll þjálfunargögnin í kennaraþvingunarstillingu (e. teacher-forcing mode) til að fá út upplýsingar um lengd. Þessi gögn voru svo notuð til að þjálfa FastSpeech2-líkönin fyrir hverja rödd. Þau líkön voru þjálfuð fyrir 100 mynstrasöfn, og meðaltal fimm bestu skoðana tekin eins og nefnt var áður.

Raddkótarinn var þjálfaður í 800.000 skrefum, þar sem hvert skref samsvarar gróflega 7 sekúndum af gögnum. Við notuðum allar 8 raddirnar í Talrómi til að þjálfa þetta líkan, þar sem okkur grunar að frekari fínstilling líkansins á ákveðnum kynjum eða ákveðinni rödd skili stórlega hverfandi árangri. Hlustunarpróf sem prófar þessa tilgátu er á döfinni, en við munum opinberlega deila þjálfunarskriftum okkar til að gefa öðrum kost á að fínstilla raddkótarann ef viljinn er fyrir hendi.

[60] https://github.com/espnet/espnet
[61] https://github.com/kan-bayashi/ParallelWaveGAN

Tækniyfirfærsla

Skýrsla: Háskóli Íslands
Aðilar: Allir aðilar SÍM

Markmið

Að þróa eina eða fleiri afurð eða frumgerðir afurða sem byggja á afurðum kjarnaverkefnanna. Kostir slíkra tækniyfirfærsluverkefna (e. Technology Transfer Projects, TTPs) hvað varðar markmið máltækniáætlunarinnar eru: (a) færsla þekkingar frá núverandi meðlimum SÍM til annarra íslenskra fyrirtækja, helst stórra fyrirtækja með stóran hóp notenda; (b) prófanir og staðfestingar á afurðum kjarnaverkefna í atvinnuumhverfi; og (c) miðlun og kynning á máltækniáætlun í heild sinni.

Allir þættir þessarar nýju viðbótar við máltækniáætlun, að ósk Almannaróms, eru framkvæmdir í nánu samstarfi við Almannaróm.

Varða 7 – lýsing

M7: Ítarleg verkáætlun fyrir tækniyfirfærsluverkefnin. Fyrirtækin sem vinna þau eru nú meðlimir SÍM.

M8: Skilgreint síðar

M9: Skilgreint síðar

Afurðir

Markmið hafa ekki náðst og vörður 8 og 9 hafa ekki verið endurskoðaðar. Skýrsla sem fjallar um kjarnaverkefnin sex, verkþætti þeirra, stöðu og notagildi, var afhent í nóvember. Verkefni hafa ekki verið valin og ekki framkvæmdaraðilar heldur, en við höfum haft samband við nokkur fyrirtæki til að kynna fyrir þeim afurðir máltækniáætlunar og hugsanleg verkefni sem nota þær.