MoLL
Marka og lemma leiðrétting

Atli Snær Ásmundsson
Háskóli Íslands / Trjábankar MLT302F
asa71@hi.is

Útdráttur

Þessi grein er yfirlit yfir notandaviðmótið MoLL (Marka og Lemmu Leiðrétting). Tilgangur þess er að auðvelda yfirferð á textum sem markaðir (og lemmaðir) hafa verið stafrænt með því að útvíkka markastrengi svo léttara sé að lesa upplýsingar úr þeim. Auk þess auðveldar MoLL leiðréttinguna með því að stinga upp á mögulegu marki.

1 Bakgrunnur

Mikilvægur þáttur í þróun máltæknihugbúnaða er skipting texta niður í orð og aðrar heildir, og enn nánar niður í orðflokka og formgerðir (*e. parts of speech*). Í beygigarmálum líkt og íslensku er þetta enn mikilvægara.

Á seinustu árum hefur þróun þess konar hugbúnaða, þ.e. til sjálfvirkrar mörkunar og lemmunar, sprottið upp hér á íslandi. Má þá helst nefna IceNLP (Loftsson og Rögnvaldsson, 2007) og nýlega ABLTagger (Steingrímsson et al., 2019) sem orðflokkagreinir rétt í um 95% tilvika. En þrátt fyrir þetta er handvirk yfirferð líka mikilvæg svo hægt sé að leiðrétta það sem tölvan merkir vitlaust og þá bæta nákvæmni stafrænnar úrvinnslu seinni tíma.

Þetta er þó hægara sagt en gert. Handvirk yfirferð getur tekið vikur, ef ekki mánuði. Stærð gagna og þ.a.l. fjöldi orða sem tölvur geta unnið með er gríðarlegur og engin leið að segja til um nákvæmlega hvaða orð það eru sem tölvunni yfirsést. Því þarf yfirfarari að fara yfir öll gögnin frá A til Ö, ráða úr markastrengjunum og lemmunum, og skrá niður breytingar ef þarfar eru. Þessi vinna fer oftast fram í töfluviðmóti, líkt og Excel.

Leiðréttingin sjálf getur svo verið vandmál í sjálfu sér. Nokkuð einfalt getur verið að leiðrétta lemmur þar sem skráning og leiðréttning er fullkomlega eftir höfði yfirfarara, og þó líklegast einhver hjálp frá orðabókum. Flækjan liggur oftast í markastrengjunum. Markastrengur er einfaldlega strengur stafa (bók- eða tölustafa) þar sem hver stafur táknar einhvern orðflokk (n = nafnorð), undirflokk (a = ábendingarfornafn) eða formgerð (k = karlkyn). Staða hvers stafs og samesetning fyrri stafa ráða því hvaða merkingu núverandi stafabil hefur og getur því verið að sami bókstafur tákni mismunandi formgerðir eftir umhverfi og stöðu.

Við úrvinnslu þarf yfirfarari því oftast að styðjast við einhvers konar lyklablað (*e. cheat-sheet*) sem gefur upp markamengið (upplýsngar um hvernig má lesa úr markastreng). Yfirfarari þarf því oft að taka augun af töfluviðmótinu til þess að lesa úr strengnum og getur það verið hægt yfirferðar þar sem auðvelt er að fara línuvillt eða einfaldlega týna staðsetningunni. Yfirfarari mun líklegast ná tökum á merkingu hvers stafs innan tíðar en vert er að benda á að verkefni sem þessi eru oft sett í hendur grunnnema og þarf því oft að kynna nýju fólki fyrir yfirferðaraðferðum.

Þessi vandamál voru höfð að leiðarljósi við þróun notandaviðmótsins MoLLs. Í næstu köflum verður farið yfir uppsetningu og virkni forritsins, því næst notkun forritsins og loks verður farið yfir möguleg næstu skref og vankanta viðmótsins.

2 Uppsetning og þróun

Allar hliðar viðmótsins eru unnar með forritunartungumálinu Python (útgáfa 3.8) en aukalega eru pakkar á borð við Pandas[1] og Pygubu[2] notaðir. Viðmótið sjálft er hannað í Pygubu Designer og notar það Tkinter eininguna sem fylgir með nýjustu útgáfum Python. Meðhöndlun skráa er svo í gegnum Pandas.

2.1 Viðmót

Líkt og áðan koma fram er hönnunin gerð með Tkinter einingunni innan Python. Með henni er

[1] https://pandas.pydata.org/
[2] https://github.com/alejandroautalan/pygubu

mögulegt að setja upp allskyns útlit og viðmót en það verður þó að teljast nokkuð seingert. Pygubu Designer nýtir sér Tkinter eininguna og gerir notendum kleift að hanna sín eigin viðmót í gegnum Pygubu Designer viðmótið. Notandaviðmótið er svo vistað sem UI (User Interface) skrá sem síðan er hægt að lesa inn í Python og vinna má með líkt og unnið er venjulega með Tkinter kóða sem hægt er að tengja við föll, breytur o.s.frv.

Viðmótið sjálft er svo sett upp sem klasi og allir hnappar og öll virkni, sem og utanaðkomandi föll, eru innifalin í viðmótinu sjálfu. Eina sýnilega fallið sem keyranlegt er í gegnum aðalforritið er svokallað keyrslufall (*e. run*) sem gangsetur klasann/viðmótið.

2.2 Gagnaskipan

Gögnin eru lesin inn af Pandas og geymd í klasa af týpunni DataFrame sem er nokkurs konar orðabókar tag (*e. dictionary type*). Með þessu eru gögnin sett upp í ákveðið Python-vænt töflusnið og létt er að skipta upp dálkum, bæta við dálkum og vista svo skjalið. Eins og er eru aðeins CSV skrár leyfðar og þurfa þær að vera rétt uppsettar svo forritið virki rétt. En nánar verður farið út í það í kafla 3.1.

Unnið er með hverja röð fyrir sig og dálkum skipt upp í viðeigandi breytur sem listar (*e. lists*) og haldið er utan um núverandi röð með heiltölubreytu. Þessum breytum er svo raðað inn í viðmótið og síðan uppfærðar eftir hentisemi og virkni viðmótsins.

2.2.1 Útvíkkun marka

Helsti eiginleiki viðmótsins er virkni þess til þess að útvíkka markastrengi og sýna notanda útvíkkað mögulegt næsta mark.

Meðfylgjandi viðmótinu er JSON skrá sem inniheldur allar þær upplýsingar sem nauðsynlegar eru til þess að lesa úr streng sem fylgir reglum markamengis eins og þær koma fram í Íslenskri orðtíðnibók (Magnússon et al., 1991). Eins og er er útvíkkun marka tvískipt. JSON skráin nýtist í uppástungur fyrir mögulega markastrengi en kóði sem fengin er frá Árnastofnun (og notaður er til þess að útvíkka mörk í málgreiningartóli þeirra[3]) er notaður til þess að fullvissa að mark sé rétt uppsett. Nánar verður farið út í þessi atriði í kafla 4.

[3]http://malvinnsla.arnastofnun.is/

3 Notkun viðmóts

Áður en forritið er notað er mikilvægt að viðkomandi hafi sótt nýjustu útgáfu Python (3.8, ekki er víst að kóðinn keyri á mikið eldri útgáfum) ásamt því að niðurhala þeim pökkum sem nauðsynlegir eru til keyrslu, þá má finna í skránni `requirements.txt` og er hægt að hala niður með skipuninni

```
pip3 install -r requirements.txt
```

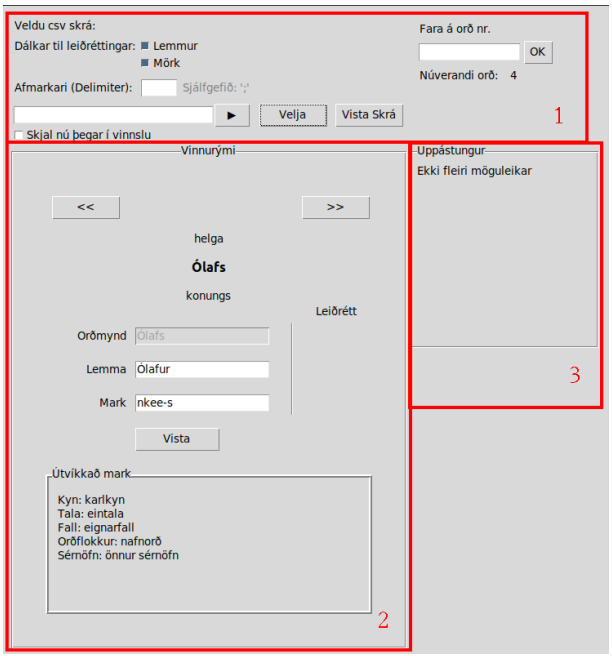
Athugið að skipuninn verður að vera keyrð innan forritsmöppunnar. Til þess að keyra forritið er svo skipunin

```
python3 moll_gui.py
```

keyrð. Forritið hefur aðeins verið prófað á Linux stýrikerfi og því ekki öruggt að það keyri rétt á öðrum kerfum.

Á mynd 1 má sjá útlit viðmótsins. Grunnurinn er nokkurn veginn þrískiptur:

1. Skráaruppsetning
2. Vinnurými
3. Uppástungur

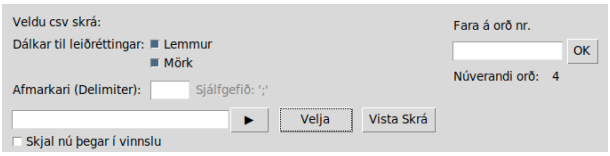


Mynd 1: Útlit viðmóts

Þessi skipting er ekki fullkomin því í raun má flokka svæði 3 undir svæði 2 en einnig er hægri hluti svæðis 1 partur af vinnusvæðinu. En megin vinnan fer fram á svæði 2. Hér að neðan verður reynt að fara ítarlega í gegnum alla virkni hvers svæðis fyrir sig.

3.1 Skráaruppsetning

Það fyrsta sem gera þarf er að velja skrá. Í þessari fyrstu útgáfu forritsins er aðeins leyfilegt að vinna með CSV skrár. Athuga þarf að þær séu rétt uppsettar svo forritið lesi þær rétt. Enginn haus má vera til staðar í skránni og dálkar þurfa að vera raðaðir rétt: Fyrsti dálkur skal vera orðmynd, annar dálkur markastrengur og sá þriðji lemma. Ekki er heimilt að hafa fleiri dálka, en leyfilegt er að skráin innihaldi bara orðmynd og markastreng, en ekki orðmynd og lemmu einungis (þ.e., orðmynd og mark eru ekki valkvæðir dálkar).



Mynd 2: Skráarsvæði

Áður en skránni er hlaðið inn verður að velja hvaða dálka á að leiðrétta, lemmur eða mörk (Mynd 2). Sé ekkert valið verður ómögulegt að bæta við leiðréttingum, en þó verður hægt að fletta í gegnum skránna. Áhrif þessara valkosta verður svo útskýrð nánar í kafla 3.2. Einnig þarf að velja afmarkara skráarinnar, en semíkomma er sjálfgefinn afmarkari ef ekkert er gefið upp.

Neðst er svo hak sem mikilvægt er að haka í ef skráin sem unnið er með er nú þegar í vinnslu, þ.e. hefur verið vistuð í gegnum viðmótið. Þetta er gert þar sem vistaðar skrár innihalda hausa og nýja dálka.

Rest svæðisins þarfnast í raun ekki nánari útskýringa þar sem hún lýsir sér sjálf. Orðaflakkarinn (hægra megin) mætti í raun flokka undir vinnurýmið en staðsetning hans ætti að breyta litlu þar sem lítið ætti að þurfa að nota hann. Loks er skráin svo vistuð með viðeigandi hnappi en einnig má vista hana með flýtileiðinni `Ctrl+S`.

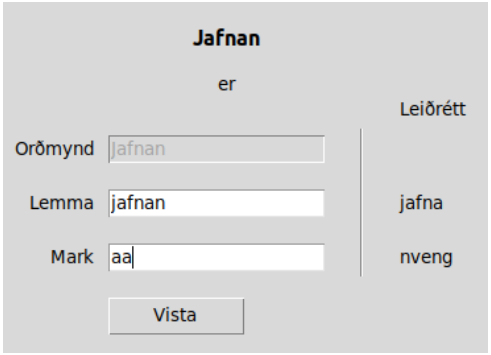
3.2 Vinnurými

Mesta vinna viðmótsins fer fram á svokölluðu vinnurými (Mynd 3). Efsti partur rýmisins sýnir orðið sem unnið er með ásamt orði á undan og eftir. Örvahnappana er svo hægt að nota til þess að flakka á milli en þó er frekar mælt með því að nota upp og niður örvar lyklaborðsins þar sem sú leið er fljótari og liggur í raun betur við þar sem orðin flakka upp og niður, en ekki til hliðar.

Miðju parturinn samanstendur af þremur innsláttargluggum: Orðmynd, Lemma og Mark.

Mynd 3: Vinnurými

Hægra megin við gluggana er svo rými sem sýnir leiðréttingar. Fyrsti glugginn er alltaf óvirkur en hugmyndin er að nýta hann frekar í næstu útgáfur þar sem möguleikinn stendur til að breyta orðmyndinni. En hér er það ekki talið mikilvægt.



Mynd 4: Dæmi um leiðrétta lemmu og mark

Á mynd 4 sést hvernig leiðrétt lemma og mark kemur fram þegar búið er að vista leiðréttingu. Aðeins er hægt að vista léiðrétt mark ef það er löglegt (þ.e. ef "Útvíkkað mark"glugginn sýnir niðurstöður), þetta er gert til þess að reyna að koma í veg fyrir villur. Ef engu er breytt er þó mikilvægt að vista svo hægt sé að reikna út hve langt viðkomandi var kominn þegar skjalið er aftur opnað. Þegar leiðréttur dálkur er sá sami og upprunalegur dálkur sýnir leiðréttingarhliðin aðeins bandstrik ". Sé ekki hægt að vista dofnar "Vista"hnappurinn. Hægt er að vista með því að smella á hnappinn en einnig með flýtileiðinni `Alt+S`. Við vistun líta

bæði innsláttarglugginn og leiðréttingarglugginn eins út en sé farið á annað orð og svo aftur til baka má sjá að upprunaleg lemma og mark haldast í innsláttarglugganum, sé notandi óviss með breytingu og vilji endurskoða upprunalegar myndir.

Þegar skrá er valin þarf að velja hvaða dálka á að leiðrétta, líkt og bent var á í kafla 3.1, sé ekkert valið verða allir dálkar gerðir óvirkir, líkt og orðmynda dálkurinn. Sé mark aðeins valið verður það eini virki dálkurinn, og sama sé aðeins lemma valin.

Seinasti hluti vinnurýmisins sýnir svo útvíkkað mark, sé markið löglegt. En þar sem sá hluti kallast á við uppástungurýmið hentar betur að tala um þetta saman.

3.3 Markamengi

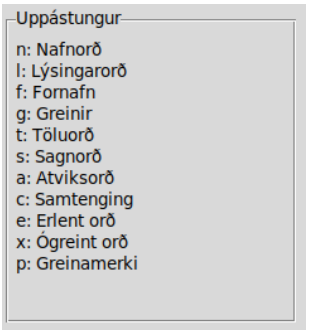
Í þessari fyrstu útgáfu MoLLs er aðeins eitt markamengi mögulegt, en það er mengið sem fyrr var talað um, úr Íslenskri orðtíðnibók (Magnússon et al., 1991). Markamengislykillinn gegnir tveimur hlutverkum í forritinu. Í glugganum sem kallast "Útvíkkað markǫg annars vegar í uppástungurýminu.

3.3.1 Úvikkaður markastrengur

Neðst á mynd 3 má sjá glugga merktan "Útvíkkað mark". Innan hans koma fram allar þær upplýsingar sem markastrengurinn stendur fyrir.

Útvíkkað mark er þó eingöngu sýnt ef markastrengur er löglegur. Sé hann hins vegar ólöglegur inniheldur glugginn skilaboðin `"Markastrengur ekki samþykktur"`. Þetta getur gerst ef vitlaust mark er slegið inn eða þá að ekkert sé slegið inn. Sé markastrengur "á réttri leiðǫn þó ekki kláraður (t.d. nve, en ekki nven) munu sömu villuskilaboð koma upp þó svo allir stafir séu á réttum stað. Þetta undirstrikar mikilvægi þessa að strengur sé réttur svo hægt sé að vista hann. Sé hann ekki fullkláraður verður ómögulegt að vista hann.

Ef fyrsti stafurinn í markastreng er ólöglegur koma aftur á móti upp villuskilaboðin `"Ólöglegur upphafsstafur"`. Þetta er gert til þess að greina á milli villna, þar sem fyrri villan þýðir í raun ekki endilega að notandi sé á rangri leið, aðeins það að strengurinn sé á einhvern hátt ófullkominn. Seinni villan bendir notanda þó á að viðkomandi sé á villigötum frá upphafi.



Mynd 5: Uppástungurými

3.3.2 Uppástungur

Á mynd 5 sést uppástungurýmið. Þetta er í raun ákveðinn hjálpargluggi sem nýtist einna helst þeim sem eru að feta sig áfram í nýju markamengi.

Þegar markastrengur er tómur kemur upp val um orðflokk (Mynd 5). Eftir það koma fram upplýsingar um næstu mögulegu skref (sé t.d. `n` valið ætti næst að koma val um kyn). Þegar markastrengur er fullkláraður og allir valmöguleikar hafa verið valdir sýnir glugginn skilaboðin `"Ekki fleiri möguleikar"`. Vert er að benda á að markastrengur þarf ekki að hafa nýtt alla möguleika til þess að markastrengur sé löglegur (líkt og útvíkkaða rýmið gefur til kynna)

Líkt og útvíkkaða rýmið sýnir villu þegar fyrsti stafur er rangur sýnir uppástungurýmið skilaboðin `"Óþekktur valmöguleiki"`við sömu villu.

4 Næstu skref

Þar sem segja mætti að forritið sé enn á tilraunastigi eru ýmis atriði sem má betrumbæta og virkni sem bæta má við.

Fyrst ber að nefna markamengið, en eins og er er ekki valmöguleiki um nema eitt mengi, sem getur varla talist valmöguleiki. Undirliggjandi virkni forritsins hefur þó verið gerð með það í huga að seinna megi bæta við virkninni að notandi geti notað eigið markamengi. Flestir bitar kóðans eru ekki háðir einu markamengi en þó eru einhverjir bitar sem gera innleiðslu nýs mengis erfiða en þó alls ekki óhugsandi. Hér má þá einnig benda á vandann sem talað var um í kafla 2.2.1, þ.e. að lestur markastrengja sé tvískiptur. Þetta mætti einnig bæta með því að sameina virknina.

Næst mætti svo bæta við fleiri skráartegundum sem vinna mætti með, þá helst Excel skrár. Jafnvel væri hægt að bæta við nánari stillingum um hvernig skráin sé uppsett svo ekki þurfi að breyta skránni, sé hún "rangt"uppsett, til þess að forritið

virki sem skyldi.

Nokkur lítilsháttar mál sem betur mættu fara eru einnig til umræðu. Í útvíkkaða rýminu er til að mynda engin ákveðin röð á flokkum sem sýndir eru. Þetta gæti verið nokkuð óhentugt þar sem best lægi við að flokkarnir væru í þeirri röð sem strengurinn er sleginn inn. Einnig mætti bæta við að sé vitlaust mark slegið inn í miðjum streng sýni uppástungurýmið að fyrri möguleiki hefði verið rangt sleginn inn. Þetta eru þó, líkt og nefnt var, lítilsháttar vandamál og ættu í raun ekki að hamla notandanum að miklu leyti.

Athugasemdir um galla eða betrumbætur sendist á `asa71@hi.is`.

Heimildir

Hrafn Loftsson og Eiríkur Rögnvaldsson. 2007. IceNLP: A natural language processing toolkit for icelandic. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 1, pages 1533–1536.

Friðrik Magnússon, Stefán Briem, og Jörgen Pind. 1991. *Íslensk orðtíðnibók*. Orðabók Háskólans, Reykjavík.

Steinþór Steingrímsson, Örvar Kárason, og Hrafn Loftsson. 2019. https://doi.org/10.26615/978-954-452-056-4_133 Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1161–1168, Varna, Bulgaria. INCOMA Ltd.