Nýsköpunarsjóðs námsmanna Sumar 2010

Endurbættur þáttari fyrir íslenskan texta

Umsjónarmenn:
Hrafn Loftsson, Háskólinn í Reykjavík
Eiríkur Rögnvaldsson, Háskóli Íslands
Jón Eðvald Vignisson, CLARA

Námsmaður:
Ragnar Lárus Sigurðsson
Háskólinn í Reykjavík

Útdráttur

Markmið rannsókna í máltækni eru af ýmsum toga, einna helst að einfalda samskipti milli fólks og tölva. Að geta haft samskipti við tölvu með náttúrulegu tungumáli gæti einfaldað líf og störf margra til muna, því er þetta mjög áhugavert rannsóknarsvið. Forritið IceParser á sinn þátt í þessu ferli en það sér um að merkja setningarliði og setningafræðileg hlutverk í íslenskum texta. Þessar merkingar einfalda útdrátt upplýsinga af ýmsu tagi úr textanum sem hægt er að nýta til margvíslegra verkefna. Í þessu verkefni voru gerðar endurbætur á IceParser til þess hann nýttist enn betur í rannsóknarverkefnum hjá Máltæknisetri og einnig í raunverulegum máltækniverkefnum hjá íslenskum fyrirtækjum.

Inngangur

Máltækni er mjög lifandi vísindagrein, hún er lifandi vegna þess að viðfangsefni greinarinnar, tungumálin, taka sífelldum breytingum í takt við lífið. Það þýðir einnig að mjög flókið er að kenna tölvum að skilja mál. Í grunninn hefur tölva enga þekkingu á tungumáli, til þess að hún geti skilið tungumál í formi texta þarf að mata hana með öllum reglum sem málið varðar.

Til hliðsjónar er gott að nefna muninn á forritunarmálunum sem tölvur skilja og eru notuð til að smíða forrit til keyrslu á tölvunum og hins vegar náttúrulegum tungumálum. Forritunarmál er hægt að skilgreina fullkomlega, þ.e.a.s. hægt er að skrifa mengi af reglum sem skilgreina öll möguleg orð og setningar forritunarmálsins. Þetta er ekki hægt með náttúruleg mál vegna þess hve breytileg og flókin þau eru, ný orð bætast við ööru hvoru og gömul breytast. Eitt orð í náttúrulegu tungumáli getur einnig haft meira en eina merkingu sem eykur flækjustig málsins til muna. Þetta gerir það að verkum að í raun er smíði og viðhald máltækniforrits óendanlegt verkefni, því sífellt þarf að kenna því merkingu nýrra orða og orðatiltækja og þar með auka nákvæmni greiningarinnar.

Í þessu verkefni var tiltekið máltækniforrit, IceParser [1, 2], endurbætt með það í huga að forritið nýttist betur við rannsóknir á sviði máltækni en einnig til að það nýttist fyrirtækjum betur við greiningu á íslenskum texta. IceParser, sem er hluti af IceNLP[1] forritasafninu [3], þáttar íslenskan texta, þ.e. greinir hann setningafræðilega – bæði einstaka setningarliði (nafnlið, sagnlið o.s.frv.) og setningafræðileg hlutverk (frumlag, andlag o.s.frv.). Þessi hugbúnaður var þróaður fyrir fjórum árum og hefur þegar verið nýttur í ýmsum rannsóknarverkefnum á sviði máltækni.

Ýmsri nýrri virkni var bætt við IceParser í verkefninu og villur lagfærðir. Breytingum er lýst næsta kafla.

Meginmál

Eins og kom fram hér að ofan var verkefni mitt að gera endurbætur á þáttaranum IceParser. Hann samanstendur af röð stöðuferjalda sem leita í texta af fyrirfram skilgreindum mynstrum - ef einhver strengur finnst sem passar við ákveðið mynstur þá setur þáttarinn inn merki sem tilgreinir hvers konar mynstur fannst. Mikið getur farið úrskeiðis í svona stórri vél eins og IceParser er, jafnvel eitt auka bil í skilgreiningu á mynstri getur reynst vandkvæð villa.

Inntakið í IceParser er markaður texti en í honum hefur sérhvert orð verið greint í orðflokk og beygingarleg einkenni. Greiningarstrengurinn er kallaður mark og forrit sem skilar af sér þess konar greiningu er kallað markari. Dæmi um markara er málfræðilegi reglumarkarinn IceTagger [4]. Þáttarinn IceParser byggir síðan greiningu sína í setningarliði og setningafræðileg hlutverk á markaða textanum.

[1] IceNLP er opinn hugbúnaður, aðgengilegur á http://icenlp.sourceforge.net

Hér er dæmi um hvernig setning breytist þegar hún er send í gegnum ferjöld IceParser.

- Hér er setningin eins og IceParser tekur við henni (mörk birtast á eftir sérhverju orði):
  - Stóru lkfnvf strákarnir nkfng borðuðu sfg3fþ mikið aa . .
- Ferjaldið Phrase_Advp - merkir atviksliði:
  - Stóru ^lkfnvf$ strákarnir ^nkfng$ borðuðu ^sfg3fþ$ [AdvP mikið ^aa$ AdvP] . ^.$
- Ferjaldið Phrase_AP - merkir lýsingarorðsliði:
  - [AP Stóru ^lkfnvf$ AP] strákarnir ^nkfng$ borðuðu ^sfg3fþ$ [AdvP mikið ^aa$ AdvP] . ^.$
- Ferjaldið Phrase_NP - merkir nafnliði:
  - [NP [APn Stóru ^lkfnvf$ AP] strákarnir ^nkfng$ NP] borðuðu ^sfg3fþ$ [AdvP mikið ^aa$ AdvP] . ^.$
- Ferjaldið Func_SUBJ merkir frumlög:
  - {*SUBJ> [NPn [APn Stóru ^lkfnvf$ AP] strákarnir ^nkfng$ NP] *SUBJ>} [VP borðuðu ^sfg3fþ VP] [AdvP mikið ^aa$ AdvP] . ^.$
- Að lokum kemur setningin út, hreinsuð af öllum auka merkjum:
  - {*SUBJ> [NP [AP Stóru lkfnvf AP] strákarnir nkfng NP] *SUBJ>} [VP borðuðu sfg3fþ VP] [AdvP mikið aa AdvP] . . (*)

Um leið og hafist var handa við fyrstu breytingarnar var ákveðið að best væri að hagræða ræsingarferli IceParser. Áður hafði það verið mjög skorðað og tók ekki mið að því að IceParser væri ræsanlegur í mismunandi hömum. Breytingarnar fólust aðallega í samþáttun á kóða í þeim tilgangi að ekki þyrfti að gera breytingu nema á einum stað ef breytinga væri þörf. Þetta leiddi því af sér minnkun á kóða sem aftur leiðir af sér einföldun sem er ávalt góð.

Fyrsta breytingin sem ráðist var í var sú að hægt væri að ræsa IceParser í sérstökum ham sem tæki mið af beygingarlegu samræmi orða. Þessi ræsistilling gefur IceParser merki um það að hann eigi að bera saman kyn, tölu og fall orða sem eru sett saman í nafnliði og persónu og tölu orða þegar frumlög sagna eru merkt (fyrri útgáfa af IceParser byggði að mestu leyti eingöngu á orðaröð). Þetta gerir það að verkum að IceParser sleppir því t.d. að setja saman orð í nafnliði ef orðin sambeygjast ekki - þar að leiðandi verður hann varkárari og merkir síður eitthvað sem á ekki við. Einnig var sú breyting gerð að nú er hægt að láta hann merkja þessi tilfelli sérstaklega með sérstakri villumerkingu sem gæti mögulega gagnast ef nýta ætti IceParser í málfræðivilluleit. Hér er ekki átt við hefðbundna stafsetningavilluleit heldur væri hægt að benda á röð orða og segja að þar væri einhver villa, t.d að kyn, tala eða fall orðanna væri ekki eins.

Hér er dæmi um málfræðilega rangan texta sem IceParser hefur merkt.

- IceParser ræstur í venjulegum ham.
  - {*SUBJ> [NP Allir fokfn [AP stór lkensf AP] strákarnir nkfng NP] *SUBJ>}
  - [VP léku sfg3fþ VP]
  - {*OBJ< [NP sér fpkfþ NP] *OBJ<} . .
- Hér er IceParser síðan beðin um að merkja sérstaklega (með spurningamerki) þá liði sem ekki var samræmi í.
  - {*SUBJ> [NP? Allir fokfn [AP stór lkensf AP] strákarnir nkfng NP?] *SUBJ>}
  - [VP léku sfg3fþ VP]
  - {*OBJ< [NP sér fpkfþ NP] *OBJ<} . .

Næsta verkefni var að breyta því hvernig IceParser skilar af sér merktum texta. Fyrri útgáfa gat einungis skilað textanum í belg og biðu í textaskrá eða einum merktum lið (setningarlið eða setningafræðilegu hlutverki) í hverri línu. Þetta er ekki mjög hentugt fyrir þá sem vilja nýta sér IceParser í öðrum forritum því þeir gætu þurft að byrja á að búa sér til einhvers konar ferli til að lesa inn þessa skrá og smíða sér einhverja grind utan um greininguna. Ákveðið var að hægt yrði að biðja

um textann á sama sniði og áður, þ.e. texta í belg og biðu eða einn liður í hverri línu, en bæta jafnframt við þremur nýjum úttakssniðum.

Fyrsta nýja úttakssniðið er sameining setningarliðamerkinga og setningafræðilegra hlutverka. Hér fyrir neðan er setningin merkt með (*) að ofan þar sem merkingarnar hafa verið sameinaðar:

- [NP-SUBJ> [AP Stóru lkfnvf AP] strákarnir nkfng NP-SUBJ>] [VP borðuðu sfg3fþ VP] [AdvP mikið aa AdvP] . .

Hin tvö nýju úttakssniðin eru JSON[2] og XML[3] snið. Þessi tvö snið eru mjög mikið notuð í tölvuheiminum og því augljós kostur. Með því að styðjast við þessi skráarmynstur er textinn settur inn í ákveðnar gagnagrindur sem eru síðan auðveldar í innlestri fyrir önnur forrit. Nú til dags er jafnvel orðið algengt að stuðningur sé fyrir hendi til innlesturs úr svona skrám í mörgum forritunarmálum.

Á meðan á þessum breytingum stóð komu í ljós nokkrar vel leyndar villur í ýmsum hlutum stöðuferjaldanna. Sumar ollu því jafnvel að orð gátu horfið úr textanum eða færst úr stað. Þetta voru mjög vandfundnar villur því oft þurfti mjög sérstök tilfelli til að framkalla þær. Að lokum tókst þó að laga þær, í það minnsta þær sem að fundust. Eftirfarandi er lýsing á einni þessara villna.

Sá texti sem IceParser tekur við þarf að vera sérstaklega merktur, hverju orði þarf að fylgja merki sem inniheldur upplýsingar um orðið, s.s. kyn, tölu og fall. IceParser á að geta tekið við alls konar texta og hunsað það sem ekki á við, því er gert ráð fyrir því að erlend orð geti verið hluti af inntakstexta. Í sumum tilfellum voru þó erlend orð þau sömu og mörkin sem IceParser bjóst við að fá frá markaranum. Dæmi um þetta er franska orðið "au" en í markamengi IceTagger er "au" notað fyrir „atviksorð-upphrópun“. Þetta var í raun ekki mjög alvarleg villa en gat engu að síður valdið mönnum hugarangri og því var æskilegt að gera lagfæringar á því.

Lagfæringin fólst í því að bætt var inn auka stöðuferjaldi alveg fremst í röðinni í IceParser sem hafði það hlutverk að finna mörkin (greiningarstreng sérhvers orðs) og setja utan um þau einhvers konar merkingar til að koma í veg fyrir að líklegt væri að mörkin gætu komið fyrir í hefðbundnum texta. Þessum merkingum er síðan hægt að breyta auðveldlega. Einnig þurfti að bæta við stöðuferjaldi alveg aftast til þess að taka þessar merkingar síðan af textanum áður en honum er skilað út.

Hér er dæmi um inntakstexta í IceParser. Efri línan er fyrir þessa breytingu, en í neðri línunni er búið að setja merkingar utan um orðamörkin.

- Jæja **au** , , svo aa hún fpven er sfg3en orðin sþgven **au** e pair lkfnsf ? ?
- Jæja **^au$** , ^,$ svo ^aa$ hún ^fpven$ er ^sfg3en$ orðin ^sþgven$ **au** ^e$ pair ^lkfnsf$? ^?$

Niðurstöður

Í verkefni sem þessu er erfitt að segja að einhverri ákveðinni niðurstöðu hafi verið náð. Hins vegar er ljóst að virkni hlutaþáttarans IceParser hefur sannarlega verið endurbætt. Hann er nú auðveldari í notkun og býður upp á fleiri ræsimöguleika sem eykur notagildi hans til muna. Ég er sannfærður um að IceParser muni nú nýtast fleiri aðilum í greiningu á íslenskum texta, bæði rannsóknarfólki og fyrirtækjum.

Breytingarnar á IceParser sem gerðar voru í þessu verkefni eru nú komnar inn í nýjustu útgáfu af IceNLP á http://icenlp.sourceforge.net.

[2] http://en.wikipedia.org/wiki/JSON
[3] http://en.wikipedia.org/wiki/XML

Heimildir

[1] Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. IceParser: An Incremental Finite-State Parser for Icelandic. In J. Nivre, H-J. Kaalep, K. Muischnek and M. Koit (eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA-2007)*. Tartu, Estonia.

[2] Hrafn Loftsson and Eiríkur Rögnvaldsson. 2008. Linguistic richness and technical aspects of an incremental finite-state parser. In *Proceedings of "Partial Parsing 2008", workshop at the 6th International Conference on Language Resources and Evaluation, LREC 2008*. Marrakech, Morocco.

[3] Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. IceNLP: A Natural Language Processing Toolkit for Icelandic. In *Proceedings of InterSpeech 2007, Special session: "Speech and language technology for less-resourced languages"*. Antwerp, Belgium.

[4] Hrafn Loftsson. 2008. Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, **31(1)**, 47-72.

Nemandi:

Ragnar Lárus Sigurðsson

Umsjónarmaður:

Hrafn Loftsson