Máltækni í HR

Inngangur

Máltækni (tungutækni) er rannsóknar- og þróunarsvið sem hefur það að markmiði að smíða kerfi sem geta unnið með og skilið náttúruleg tungumál og stuðlað að notkun þeirra í samskiptum manns og tölvu. Máltækni er þverfaglegt svið sem spannar t.d. tölvunarfræði, málvísindi, gervigreind, tölfræði og sálfræði. Sviðið er tiltölulega nýtt hér á landi en segja má að skýrsla, sem unnin var fyrir menntamálaráðuneytið á árunum 1998–1999, hafi markað upphafið [1].

Í skýrslunni var lagt til að átak yrði gert til að efla máltækni á Íslandi enda væru Íslendingar að dragast verulega aftur úr öðrum þjóðum á þessu sviði. Sérstaklega var hvatt til að útbúnar yrðu ýmiss konar máltæknieiningar í þeim tilgangi að auðvelda notkun íslensku í upplýsingatækniþjóðfélaginu. Með máltæknieiningum er t.d. átt við málheildir (safn fjölbreyttra texta sem geymdir eru á stöðluðu sniði á rafrænu formi) og tól sem greina texta af ýmsu tagi. Safn máltæknieininga, þ.e. BLARK (e. Basic Language Resource Kit), er nauðsynlegur grunnur undir frekari rannsóknir og þróun í máltækni fyrir sérhvert tungumál [2].

Tölvunarfræðideild HR var stofnaðili að *Tungtæknisetri* (http://www.tungutaekni.is/info/verkefnid.html) árið 2005 ásamt Málvísindastofnun HÍ og orðfræðisviði Stofnunar Árna Magnússonar í íslenskum fræðum. Tungutæknisetur er vettvangur fyrir samstarf þessara aðila um rannsóknir, þróun og kennslu í máltækni.

Meistaranám

Haustið 2007 settu HR og HÍ á laggirnar sameiginlegt meistaranám í máltækni. Markmið með náminu er tvíþætt: annars vegar að útskrifa nemendur með þekkingu til að stjórna verkefnum og útfæra lausnir á sviði máltækni; hins vegar að undirbúa nemendur undir doktorsnám á sviðinu. Nám í máltækni er bæði fyrir nemendur með BA-próf í hugvísindagreinum (almennum málvísindum, íslensku og erlendum tungumálum o.fl.) og BS-próf í raungreinum (tölvunarfræði, rafmagns- og tölvuverkfræði o.fl.).

Skipulag námsins er á þann veg að nemandi tekur námskeið í íslensku, tölvunarfræði og máltækni, í mismiklum mæli eftir því hver bakgrunnur nemandans er. Námskeið í íslensku eru kennd í HÍ, námskeið í tölvunarfræði í HR og námskeið í máltækni eru kennd í báðum háskólunum. Forsvarsmenn námsins eru Hrafn Loftsson lektor, hjá HR, og Eiríkur Rögnvaldsson prófessor, hjá HÍ. Nánari upplýsingar um námið má finna á http://nlp.ru.is/meistaranam.htm.

Rannsóknarverkefni

Hér verður nokkrum rannsóknarverkefnum starfsmanna og meistaranemenda í máltækni við HR lýst stuttlega en nokkur þessara verkefna hafa verið unnin fyrir tilstilli Tungutæknisetur með styrk frá Rannís. Upplýsingar um verkefnin og vísindagreinar þeim tengd má finna á http://nlp.ru.is.

Hrafn Loftsson hefur þróað markara, þ.e. forrit sem greinir (markar) sérhvert orð í íslenskum texta í orðflokk og beygingarleg einkenni, eins og kyn, tölu og fall fyrir fallorð, og hátt, mynd, persónu, tölu og tíð fyrir sagnorð. Markarinn, sem nefnist *IceTagger*, byggir á málfræðilegum reglum og mörkunarnákvæmni hans er um 91,5% [3].

Ida Kramarczyk, meistaranemi í máltækni við HR, vinnur nú að verkefninu "Aukin mörkunarnákvæmni íslensks texta". Verkefnið, sem styrkt er af Rannís árin 2007-2009, miðar að því að auka nákvæmnina með því að nota stærra orðasafn, þróa einfaldara markamengi (mengi greiningarstrengja), endurbæta útgáfu af IceTagger og setja saman nokkrar tegundir af mörkurum.

Í verkefninu "Hlutaþáttun íslensks texta", sem styrkt var af Rannís árið 2006, þróaði Hrafn Loftsson (í samvinnu við Eirík Rögnvaldsson) svokallaðan hlutaþáttara sem tekur inn markaða setningu (úr markara eins og t.d. IceTagger) og greinir setninguna í einstaka setningarliði (nafnliði, sagnliði o.s.frv.) og setningafræðileg hlutverk (frumlög, andlög o.s.frv.). Hlutaþáttarinn, sem nefnist *IceParser*, samanstendur af röð af stöðuferjöldum sem gerir hann mjög hraðvirkan [4]. IceTagger og IceParser er hægt að prófa á síðunni http://nlp.ru.is/icenlp_isl.htm.

Martha Dís Brandt, annar meistaranemi í máltækni við HR, vinnur nú að verkefninu "Grófþýðing íslensks texta með tiltækum tólum". Með grófþýðingu er átt við að áherslan er lögð á að koma merkingu til skila en ekki að koma fram með "fullkomna" þýðingu. Árið 2009 fékk Tungutæknisetur ásamt samstarfsmönnum öndvegisstyrk frá Rannís fyrir verkefni á sviði máltækni. Hluti þessa styrks gengur til umrætts meistaraverkefnis sem felst í því að þróa þýðingarkerfi á milli íslensku og ensku með því að nota máltæknieiningar sem þegar hafa verið búnar til (eins og IceTagger og IceParser). Verkefnið er unnið í samstarfi við University of Alicante sem þróað hefur grófþýðingarkerfið *Apertium* [5].

Skiptinemarnir Verena Henrich og Timo Reuter frá University of Applied Sciences, Darmstadt, vinna nú að meistaraverkefni sínu "A Statistically Enhanced Grammar Checker". Í verkefninu er þróuð aðferð, sem er óháð tungumálum, til að leita að málfræðivillum í textum. Aðferðin byggist á því að safna saman miklu magni texta á viðkomandi tungumáli og vinna síðan tölfræðiupplýsingar úr gögnunum. Þessar upplýsingar notar hugbúnaður síðan til að gefa vísbendingar um hvort setningar í texta notanda innihalda málfræðivillur.

Heimildir

[1] Rögnvaldur Ólafsson, Eiríkur Rögnvaldsson og Þorgeir Sigurðsson (1999). *Tungutækni: Skýrsla starfshóps*. Menntamálaráðuneytið, Reykjavík.

[2] S. Krauwer (2003). The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of SPECOM 2003*. Moskva.

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

[4] Hrafn Loftsson og Eiríkur Rögnvaldsson (2007). IceParser: An Incremental Finite-State Parser for Icelandic. In *Proceedings of the 16th Nordic Conference of Computational Linguistics, NODALIDA-2007*. Tartu.

[5] C. Armentano-Oller, A. M. Corbí-Bellot, M. L. Forcada, M. Ginestí-Rosell, B. Bonev, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez, and F. Sánchez-Martínez. An open-source shallow-transfer machine translation toolbox: consequences of its release and availability. In *Proceedings of OSMaTran: Open-Source Machine Translation, A workshop at Machine Translation Summit X*, Phuket, Thailand, 2005.