Hrafn Loftsson, lektor við tölvunarfræðideild HR

Nýtt íslenskt-enskt grófþýðingarkerfi frá Máltæknisetri

Inngangur

Máltækni 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 og tölfræði. Tölvunarfræðideild Háskólans í Reykjavík (HR) var stofnaðili að *Máltæknisetri* (áður nefnt Tungtæknisetur) (http://www.maltaeknisetur.is) árið 2005 ásamt Málvísindastofnun Háskóla Íslands (HÍ) og orðfræðisviði Stofnunar Árna Magnússonar í íslenskum fræðum. Máltæknisetrið er vettvangur fyrir samstarf þessara aðila um rannsóknir, þróun og kennslu í máltækni.

Í upphafi ársins 2009 fékk Máltæknisetur ásamt samstarfsmönnum öndvegisstyrk frá Rannís fyrir verkefnið „Hagkvæm máltækni utan ensku – íslenska tilraunin“ (sjá http://iceblark.wordpress.com). Hluti þessa styrks gekk til rannsóknarverkefnis á sviði vélrænna þýðinga sem hefur nú getið af sér nýtt kerfi til grófþýðinga úr íslensku yfir á ensku. Í þessari grein er fyrst fjallað almennt um vélrænar þýðingar og síðan fjallað nánar um þetta nýja grófþýðingarkerfi.

Vélrænar þýðingar

Í vélrænum þýðingum eru tölvur (hugbúnaður) notaðar til að þýða texta úr einu tungumáli, *frummáli*, yfir á annað tungumál, *markmál*. Vélrænar þýðingar eru eitt elsta rannsóknarsvið innan tölvunarfræði og máltækni – rannsóknir á sviðinu má rekja allt aftur til 1950 þegar vísindamenn við Georgetown University og hjá IBM gerðu tilraunir með þýðingar milli rússnesku og ensku (John Hutchins, 2005). Ýmsum aðferðum hefur verið beitt við þróun þýðingarkerfa í gegnum árin. Tvær þeirra helstu eru regluaðferðir og tölfræðiaðferðir.

Í regluaðferðum (e. *rule-based methods*) byggir þýðingarkerfið á málfræðilegum reglum og orðasöfnum. Kerfið þarf jafnframt aðgang að málvinnslutólum eins og markara (e. *part-of-speech tagger*) og þáttara (e. *parser*) því nauðsynlegt er að geta greint textann annars vegar í orðflokka og beygingarlegar myndir og hins vegar í einstaka

setningarliði og setningafræðileg hlutverk. Þekktasta dæmið um þýðingarkerfi sem byggir á regluaðferðum er SYSTRAN-kerfið (Mary Flanagan og McClure, 2002) sem notað er víða um heim.

Í kerfum sem byggja á tölfræðiaðferðum (e. *statistical methods*) eru þýðingar búnar til með hjálp tölfræðilíkans. Stikar (e. *parameters*) líkansins fást með sjálfvirkri greiningu á samhliða málheildum, þ.e. textum á tungumáli A og sömu (þýddum) textum á tungumáli B. Í „hreinum“ tölfræðikerfum er engin málfræðiþekking innbyggð, hvorki á frummálinu né markmálinu. Þýðingarvélin, sem fyrirtækið Google býður upp á, http://translate.google.com, byggir t.d. á tölfræðiaðferðum.

Áður en Máltæknisetur hóf þróun á þýðingarkerfi á milli íslensku og ensku þá voru í raun einungis tvö kerfi aðgengileg almenningi sem gátu þýtt íslenskan texta á sómasamlegan hátt yfir á annað tungumál. Hið fyrra er fyrrnefnd þýðingarvél Google og hið síðara er þýðingarkerfi Stefáns Briem sem byggir á regluaðferð og er aðgengilegt á vefsvæðinu http://www.tungutorg.is.

Grófþýðingarkerfi Máltækniseturs

Þýðingarverkefni Máltækniseturs nefnist „Þróun grófþýðingarkerfis með tiltækum opnum tólum“. Grófþýðingaraðferðin (e. *shallow-transfer machine translation*) er afbrigði af regluaðferð þar sem megin markmiðið er að koma merkingu til skila en minni áhersla er lögð á gæði þýðingar. Jafnframt er lögð áhersla á að þýðing gangi hratt fyrir sig í rauntíma.

Eins og heiti verkefnisins gefur til kynna þá er lögð áhersla á að nota opin tiltæk tól og gögn við þróun kerfisins. Hér er t.d. átt við að nýta málvinnslutól, sem þegar hafa verið þróuð, og opin gögn í þeim tilgangi að stytta þróunartímann. Mikil áhersla er lögð á að allur hugbúnaður og gögn sem notuð eru við þróunina séu opin því *i)* þannig geta fleiri tekið þátt í áframhaldandi þróun kerfisins og *ii)* þau gögn og forrit sem verða til geta nýst á auðveldan hátt í öðrum verkefnum.

Helstu tól og gögn sem notuð hafa verið hingað til í verkefninu eru þessi:

- *Apertium-kerfið* (Carme Armentano-Oller o.fl., 2005)
- Reglumarkarinn *IceTagger* (Hrafn Loftsson, 2008)
- Lemmarinn *Lemmald* (Anton K. Ingason o.fl., 2008)

- Íslensk-enskur orðalisti frá bókaútgáfunni *Forlagið*.

Apertium-kerfið myndar grunn (e. *platform*) sem er sameiginlegur þeim grófþýðingarkerfum er byggja á regluaðferðum og nýta sér þennan grunn. Þróunaraðilar sem nota grunninn þurfa eingöngu að þróa tól og setja saman gögn sem eiga við viðkomandi frum- og markmál.

IceTagger greinir sérhvert orð í orðflokka og beygingarleg einkenni og Lemmald varpar orðmyndum yfir í nefnimyndir (lemmur). IceTagger og Lemmald eru hluti af IceNLP forritunarsafninu (Hrafn Loftsson og Eiríkur Rögnvaldsson, 2007) sem nýlega var gert að opnum hugbúnaði (sjá http://icenlp.sourceforge.net/).

Orðalistinn frá Forlaginu, sem hefur að geyma um 18.000 nefnimyndir ásamt þýðingum, er notaður til að mynda grunn að tvímála (íslensk-ensku) orðasafni í Apertium-kerfinu. Ljóst er að gríðarlega mikil vinna sparast í verkefninu með því að geta nýtt þann orðalista.

Lítum nú á það ferli sem á sér stað við þýðingu á íslenskri setningu, *S*, yfir í enska setningu, *T*, í Apertium-kerfinu. Gerum ráð fyrir að þýða skuli setninguna *S* = `stóru strákarnir borðuðu góða súpu`. Fyrst markar IceTagger sérhvert orð í *S* og Lemmaldið skilar viðkomandi nefnimynd. Úttakið (eftir vörpun yfir á það snið sem Apertium-kerfið krefst) er þetta:

```
stór<adj><pst><m><pl><nom><vei>
strákur<n><m><pl><nom><def>
borða<vblex><act><past><p3><pl>
góður<adj><pst><f><sg><acc><sta>
súpa<n><f><sg><acc><ind>
```

Fyrst í hverri línu er nefnimynd orðsins en þar á eftir fylgja tókar (e. tokens) sem saman mynda málfræðilegt mark. Sem dæmi má nefna að markið `<n><m><pl><nom><def>` merkir nafnorð (n=noun), karlkyn (m=masculine), fleirtala (pl=plural), nefnifall (nom=nominative) og viðskeyttur greinir (def=definite).

Þessu næst eru svokallaðar tilfærslureglur (e. *transfer rules*) keyrðar til að framkvæma grunnþýðinguna. Úttakið að ofan er þá þáttað í einstaka setningarliði (t.d. nafnliði), tilfærslum beitt innan liða (t.d. er íslenskur viðskeyttur greinir gerður að enskum ákveðnum greini fremst í nafnlið) og íslenskum nefnimyndum varpað yfir í

samsvarandi enskar nefnimyndir með hjálp tvímála orðasafnsins. Eftir þetta lítur úttakið þannig út:

```
the<det><pl>
big<adj><pst><m><pl><nom><vei>
boy<n><m><pl><nom><def>
eat<vblex><act><past><p3><pl>
a<det><sg>
good<adj><pst><f><sg><acc><sta>
soup<n><f><sg><acc><ind>
```

Að lokum er orðhlutafræðileg myndun (e. *morphological generation*) keyrð til að mynda réttu ensku orðmyndirnar úr nefnimyndunum og viðkomandi mörkum. Hin endanlega þýðing er þá *T* = `the big boys ate a good soup`.

Samstarfsaðili Máltækniseturs í þessu rannsóknarverkefni er Universitat d'Alacant. Helstu rannsakendur frá þeim háskóla eru doktorsneminn Francis M. Tyers og Dr. Mikel L. Forcada, prófessor, sem leitt hefur þróun Apertium-kerfisins frá árinu 2004. Hjá Máltæknisetrinu hafa eftirtaldir komið að verkefninu: Martha Dís Brandt, meistaranemi í máltækni við HR, Hlynur Sigurþórsson, meistaranemi í tölvunarfræði við HR, Dr. Hrafn Loftsson, lektor við tölvunarfræðideild HR (verkefnisstjóri), og Eiríkur Rögnvaldsson, prófessor í íslenskri málfræði við HÍ.

Á vefsetrinu http://nlp.cs.ru.is hefur nú verið sett upp form til að gera notendum kleift að þýða íslenskan texta yfir á ensku. Jafnframt hefur verið þróuð vefþjónusta sem gerir öðrum vefsetrum kleift að senda setningar inn til þýðingar í rauntíma (sjá upplýsingar á http://nlp.cs.ru.is).

Lokaorð

Ef vel tekst til við þróun ofangreinds kerfis gæti það orðið sú þýðingarvél sem flestir reiða sig á fyrir grófþýðingar úr íslensku yfir í ensku. Í framtíðinni ráðgerir síðan Máltæknisetur að þróa kerfi sem þýðir úr ensku yfir á íslensku, ásamt því að þróa kerfi sem þýðir á milli íslensku og annarra tungumála, t.d. færeysku.

Heimildir

Anton K. Ingason, Sigrún Helgadóttir, Hrafn Loftsson og Eiríkur Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using Hierachy of Linguistic Identities (HOLI). In *B. Nordström og A. Rante, editors, Advances in Natural Language Processing, 6th International Conference on NLP, GoTAL 2008, Proceedings*. Gothenburg, Sweden.

Carme Armentano-Oller, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Boyan Bonev, Sergio Ortiz Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez og Felipe Sánchez-Martínez. 2005. 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.

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

Hrafn Loftsson og 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.

John Hutchins. 2005. The history of machine translation in a nutshell. http://www.hutchinsweb.me.uk/Nutshell-2005.pdf.

Mary Flanagan og Steve McClure. 2002. SYSTRAN and the Reinvention of MT. http://wwwv5.systransoft.com/IDC/26459.html.