Vélrænar grófþýðingar

Hrafn Loftsson
Tölvunarfræðideild
Háskólinn í Reykjavík
hrafn@ru.is

Útdráttur

Vélrænar þýðingar, sem rekja má allt aftur til 1950, er eitt elsta rannsóknarsvið innan tölvunarfræði og máltækni. Í þessu erindi er saga þeirra rifjuð upp og stuttlega gerð grein fyrir helstu aðferðum við vélrænar þýðingar. Aðal áherslan er lögð á svokallaðar grófþýðingar, þ.e. þýðingar sem hafa það að megin markmiði að koma merkingu til skila. Grófþýðingarkerfið *Apertium* er kynnt og fyrirhuguð notkun þess í þróun íslensks-ensks þýðingarkerfis. Megin tilgangur verkefnisins er að þróa kerfi með því að nota opinn hugbúnað og tiltæk tól, eins og þau sem þegar hafa verið þróuð til að greina íslensku. Þetta verkefni er hluti rannsóknarinnar "Hagkvæm máltækni utan ensku – íslenska tilraunin", sem hlaut öndvegisstyrk RANNÍS í upphafi árs 2009.

1 Inngangur

Í vélrænum þýðingum eru tölvur (hugbúnaður) notaðar til að þýða texta af 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 Georgetown University og IBM gerðu tilraunir með þýðingar á milli rússnesku og ensku (Hutchins, 2005). Í kjölfar þessara tilrauna fengu rannsóknaverkefni í vélrænum þýðingum víða styrki og þá sérstaklega í Bandaríkjunum.

Árið 1966 kom hins vegar út hin vel þekkta *ALPAC* (Automatic Language Processing Advisory Committee) skýrsla (Pierce and Carroll, 1966) sem var mjög gagnrýnin á rannsóknir í vélrænum þýðingum. Höfundar skýrslunnar færðu rök fyrir því að lítill sem enginn árangur hefði náðst í vélrænum þýðingum og ýmsar setningar í skýrslunni eru mjög afdráttarlausar, eins og:

> "[...] it might be simpler and more economical for heavy users of Russian translations to learn to read the documents in the original language."

> "[...] there is no immediate or predictable prospect of useful machine translation."

Niðurstaða höfunda var í raun sú að vélrænar þýðingar væru vonlausar og að í stað þess að halda áfram að styrkja verkefni á þessu sviði væri nær að styðja grunnrannsóknir í máltækni og þróun á hjálpartólum fyrir þýðendur.

Umrædd skýrsla hafði mikil áhrif á fjárveitingar til verkefna á sviði vélrænna þýðinga og það var í rauninni ekki fyrr en í kringum 1990 að sviðið gekk í endurnýjun lífdaga (sjá kafla 2.2).

2 Helstu aðferðir

Aðferðum við vélrænar þýðingar má skipta í þrjá flokka: regluaðferðir, tölfræðiaðferðir og hliðstæðuaðferðir.

2.1 Regluaðferðir

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 setningaliði og setningarfræðileg hlutverk. Þekktasta dæmið um þýðingarkerfi sem byggir á regluaðferðum er *SYSTRAN* kerfið (Flanagan and McClure, 2002) sem notað er víða um heim.

Svokölluð tilfærsluaðferð (e. transfer-based method) er afbrigði af regluaðferð. Í tilfærsluaðferð er frummálið greint orðhlutafræðilega og

setningarfræðilega (og stundum merkingarfræðilega). Síðan er innra form (e. internal representation) frummálsins búið til og þýðingin að lokum mynduð úr innra forminu með notkun orðalista og reglna um tilfærslur á orðum og setningaliðum. Sérstaklega er fjallað um þessa aðferð í kafla 4.

2.2 Tölfræðiaðferðir

Í kerfum sem byggja á tölfræðiaðferðum (e. statistical methods) eru þýðingar búnar til með hjálp tölfræðilíkans. Kennistærðir (e. parameters) tölfræðilíkansins fást með sjálfvirkri greiningu á samhliða málheildum, þ.e. textum á tungumáli A og sömu (þýddum) textum í tungumáli B. Í "hreinum" tölfræðikerfum er engin málfræðiþekking innbyggð, hvorki á frummálinu né markmálinu.

Þessar aðferðir hafa náð mikilli útbreiðslu síðan að tímamótagrein rannsóknarhóps IBM um aðferðina kom út árið 1990 (Brown et al., 1990). Ástæður útbreiðslunnar eru helstar þær að: i) samhliða málheildum hefur fjölgað ört sl. 20 ár; ii) aðferðirnar eru óháðar tungumálum; og iii) regluaðferðir krefjast oft mikillar vinnu við sametningu orðasafna, reglna og tóla.

Þessum aðferðum má í raun þakka að verkefni á sviði vélrænna þýðinga fengu á ný náð hjá styrkveitendum.

2.3 Hliðstæðuaðferðir

Hliðstæðuaðferðir (e. example-based methods) (Nagao, 1984) byggja, eins og tölfræðiaðferðir, á samhliða málheildum en geta einnig nýtt sér tól eins og markara og þáttara á sama hátt og gert er í regluaðferðum. Hlutar setningar eru þýddir sérstaklega með því að nota hliðstæður úr áður þýddum setningahlutum. Síðan eru hlutirnir settir saman til að mynda þýdda útgáfu af upphaflegu setningunni í frummálinu.

Þýðingarminni (e. translation memory), er afbrigði af hliðstæðuaðferð. Munurinn er sá að í þýðingarminni eru nýjar þýðingar, sem byggja á hliðstæðum í gagnagrunninum, ekki búnar til sjálfvirkt af kerfinu heldur er það þýðingarpar (<frummál, markmál>) í grunninum fundið sem talið er vera líkast því pari sem ætti að koma út úr þýðingunni. Þýðingarparið er síðan sýnt þýðanda til yfirferðar og leiðréttingar.

Kerfi sem byggja á þýðingarminni eru t.d. notuð hjá Evrópusambandinu (`http://langtech.jrc.it/DGT-TM.html`).

3 Íslensk þýðingarkerfi

Nokkur þýðingarkerfi á netinu bjóða upp á þýðingu úr íslensku yfir á ensku en flest þeirra skila slakri þýðingu. Þegar t.d. setningin "Hann er mjög góður kennari" er slegin inn hjá `http://www.translation-guide.com/free_online_translators.php` þá fæst þýðingin "Hún is mjög góður teacher."

Annað gildir um Tungutorg Stefáns Briem, `http://www.tungutorg.is`. Kerfið, sem byggir á regluaðferð, var opnað almenningi 29. mars 2008 eftir að hafa verið í þróun í mörg ár. Þegar ofangreind setning er slegin inn í Tungutorg þá fæst þýðingin "She is superfine a teacher". Þetta dæmi sýnir ágætlega mikilvægi þess að beita tilfærslum í kerfum sem byggja á regluaðferðum því betri þýðing er auðvitað "She is a superfine teacher."

Annað íslenskt þýðingarkerfi sem hefur verið í umræðunni er kerfi frá Alnet hf. Fyrirtækið fékk styrk frá Tækniþróunarsjóði 2005 til að þróa íslenskt-enskt þýðingarkerfi. Ekki er þó ljóst hvers konar kerfi er um að ræða og þegar þetta er ritað hefur kerfið ekki verið sett á markað.

4 Grófþýðingar

Grófþýðingar (e. shallow-tranfer machine translation) er tegund tilfærsluaðferðar (regluaðferðar). Megin markmið grófþýðingar er að koma merkingu til skila en minni áhersla er lögð á gæði þýðingar. Jafnframt er áhersla á að þýðing gangi hratt fyrir sig (í rauntíma). Full þáttun (e. full parsing) á texta frummáls er ekki framkvæmd heldur vinna reglur um tilfærslur á úttaki úr hlutaþáttun (e. shallow parsing). Grófþýðingarkerfi hafa yfirleitt verið notuð til að þýða á milli skyldra tungumála.

4.1 Apertium kerfið

Apertium kerfið (`http://wiki.apertium.org`) er þýðingarkerfi sem byggir á opnum gögnum og opnum hugbúnaði. Hugmyndin með kerfinu er að mynda grunn (e. platform) sem er nauðsynlegur öllum þýðingarkerfum er byggja á tilfærsluaðferðum. Þróunaraðilar sem nota grunninn þurfa þá eingöngu að þróa þær auðlindir og tól sem eiga við viðkomandi frum- og markmál (sjá að neðan). Apertium er þróað hjá Universitat d'Alacant undir stjórn Dr. Mikel L. Forcada. Forcada hefur fært ýmis rök fyrir því að hentugt sé að

beita regluaðferðum við þróun þýðingarkerfa, m.a. vegna:

> "[...] the amounts of sentence-aligned parallel text (of the order of hundreds of thousands or millions of words) required to get reasonable results in pure corpus-based MT, such as statistical MT" (Forcada, 2006).

Forcada er jafnframt eindreginn stuðingsmaður opins hugbúnaðar og telur að þýðingarkerfi sem byggja á opnum hugbúnaði séu miklu hentugri "litlum" tungumálum heldur en kerfi sem byggja á lokuðum lausnum (Forcada, 2006).

Apertium samanstendur af nokkrum megin einingum:

1. **Orðhlutafræðileg greining** (e. morphological analysis)
   - Sérhvert orð í setningu frummáls er greint í orðhluta og möguleg málfræðileg mörk (e. tags).
2. **Mörkun** (e. PoS tagging)
   - Sérhvert orð í setningu frummáls er einrætt (e. disambiguated), þ.e. aðeins eitt mark er valið m.t.t. til samhengis.
3. **Tilfærslur** (e. transfer modules)
   - Hlutaþáttun er framkvæmd og tilfærsla á orðum innan liða eða á setningaliðum.
   - Lemmu (flettimynd) orðs í frummáli er varpað yfir í tilsvarandi lemmu í markmáli.
4. **Orðhlutafræðileg myndun** (e. morphological generation)
   - Setning í markmálinu er mynduð út frá lemmum, tilfærslum og mörkum.

Til að geta notað Apertium grunninn fyrir þýðingu á tilteknu frummáli yfir í tiltekið markmál er nauðsynlegt að hafa aðgang að eða útfæra eftirfarandi auðlindir og tól:

1. Orðasafn með orðhlutafræðilegri greiningu fyrir orð í frummálinu.
2. Orðasafn með orðhlutafræðilegri greiningu fyrir orð í markmálinu.
3. Tvímála orðasafn (e. bilingual dictionary).
4. Markaða málheild til að þjálfa tölfræðimarkara.
5. Reglur um tilfærslur á orðum eða setningaliðum.

4.2 Íslenska Apertium-verkefnið

Í febrúar 2009 hófst samstarf Tungutækniseturs og Universitat d'Alicant um þróun á grófþýðingarkerfi frá íslensku yfir í ensku. Verkefnið er hluti rannsóknarinnar "Hagkvæm máltækni utan ensku – íslenska tilraunin", sem hlaut öndvegisstyrk RANNÍS í upphafi árs 2009. Markmiðið með verkefninu er:

1. Að finna hagkvæmar leiðir til að búa til gögn og reglur sem nauðsynleg eru fyrir grófþýðingarkerfi.
2. Að finna leiðir til að samþætta tiltæk tól (eins og íslenskan markara, lemmara og þáttara) inn í Apertium kerfið.
3. Nota þessar aðferðir til að búa til frumgerð af grófþýðingakerfi á milli íslensku og ensku.

Fyrstu tveir liðirnir að ofan gefa til kynna að lögð er áhersla á að þurfa ekki að búa til auðlindir og tól frá grunni heldur finna hagkvæmar leiðir til að nýta gögn og tól sem þegar eru til í þeim tilgangi að stytta þróunartímann. Í þessu sambandi má nefna að meðlimir Tungutæknisetur hafa þegar þróað lemmarann *Lemmald* (Ingason et al., 2008) (sem skilar uppflettimynd orða), reglumarkarann *IceTagger* (sem skilar málfræðilegu marki orða) (Loftsson, 2006; Loftsson, 2008) og hlutaþáttarann *IceParser* (sem merkir setningaliði og setningarfræðileg hlutverk) (Loftsson and Rögnvaldsson, 2006). Þessi tól, eða hluta þeirra, ætti að vera hægt að aðlaga að virkni Apertium og flýta þannig fyrir þróun. Vonast er til að reynslan af þessu verkefni geti nýst við þróun grófþýðingakerfa fyrir önnur tungumál.

Áður hefur verið nefnt að Apertium verkefnið byggir á opnum gögnum og opnum hugbúnaði. Þau tiltæku tól sem þegar hafa verið þróuð fyrir íslensku og nýtt verða í verkefninum munu jafnframt verða gerð opin. Tilgangurinn með því er sá að þar með geti fleiri aðilar komið að áframhaldandi þróun kerfisins og jafnframt að þau gögn og forrit sem til verða geti nýst á auðveldan hátt í öðrum verkefnum.

Að lokum má benda á að íslenska Apertium-verkefnið passar vel við hluta þeirra aðgerða

sem felast í þingsályktunartillögu um íslenska málstefnu sem samþykkt var á Alþingi þann 12. mars 2009 (`http://www.althingi.is/altext/136/s/0248.html`), þ.e.:

- Að málleg gagnasöfn og hugbúnaður til að vinna með íslenskt mál verði gerð opin og frjáls eftir því sem kostur er.
- Að nothæf þýðingarforrit milli íslensku og valinna erlendra mála, a.m.k. ensku, verði gerð innan fimm ára.

4.2.1 Dæmi um ferli í þýðingu

Í þessum kafla verður sýnt dæmi um það ferli sem gert er ráð fyrir að eigi sér stað við þýðingu í íslenska-enska Apertium-verkefninu.

Gerum nú ráð fyrir að þýða eigi íslensku setninguna "stóru strákarnir borðuðu góða súpu" yfir á ensku. Með því að nota tól sem þegar hafa verið þróuð fyrir íslensku (tiltæk tól), þ.e. *IceTagger* og *Lemmald* þá fæst eftirfarandi úttak:

```
stóru lkfnvf stór
strákarnir nkfng strákur
borðuðu sfg3fþ borða
góða lveosf góður
súpu nveo súpa
```

Fyrsti tókinn í hverri línu er orðmyndin sjálf, annar tókinn er málfræðilega markið og þriðji tókinn er lemman. Málfræðilega markið gefur upplýsingar um orðflokk og beygingarleg einkenni. Sem dæmi má nefna að markið "nkfng" merkir nafnorð (n), karlkyn (k), fleirtala (f), nefnifall (n), viðskeyttur greinir (g) og markið "sfg3fþ" merkir sagnorð (s), framsöguháttur (f), germynd (g), þriðja persóna (3), fleirtala (f), þátíð (þ).

Ofangreint úttak er hins vegar ekki á því sniði sem Apertium krefst og því þarf að varpa úttakinu yfir á eftirfarandi snið:

```
^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>$
```

Munurinn er sá að Apertium nægir lemman sjálf og mörkin eru á öðru sniði en mörkin í íslenska markamenginu. Í Apertium stendur t.d. "<n><m><pl><nom><def>" fyrir "noun" (nafnorð), "masculine" (karlkyn), "plural" (fleirtala), "nominative" (nefnifall), "definite" (ákveðinn greinir). Þegar hér er komið við sögu er þá búið að keyra einingar nr. 1 og 2 (orðhlutafræðileg greining og mörkun) í listanum yfir Apertium-einingar í kafla 4.1 en athugið þó að tiltæk tól hafa verið notuð í stað þess að þróa samsvarandi Apertium-einingar.

Næst þarf að keyra tilfærslur og byrjað er á því að þátta úttakið í einstaka setningahluta. Niðurstaðan gæti orðið eitthvað á þessa leið:

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

Hér hafa lemmurnar og mörkin verið afmörkuð með upplýsingum um hvar tilteknir setningaliðir byrja og enda (skrifa þarf Apertium-reglur sem byggja á orðaröð til að afmarka setningaliði). T.d. merkja "[NP", og "[VP" upphaf nafnliðs og sagnliðs og "NP]" og "VP]" merkja endir viðkomandi liða.

Eftir tilfærslur innan liða og vörpun á íslenskum lemmum yfir í enskar (með notkun íslensks-ensks orðasafns) þá fæst síðan úttakið:

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

En hvernig þýðast lemmurnar "stór strákur" ásamt tilsvarandi mörkum yfir í "the big boy"? Í Apertium-kerfinu þarf að skrifa reglu fyrir tilfærslu innan liðar sem á við nákvæmlega svona tilvik. Reglan, sem útfærð er með reglulegum segðum, hefur það hlutverk að breyta íslenska mynstrinu <lýsingarorð, nafnorð með viðskeyttum greini> yfir í enska mynstrið <determiner, adjective, noun>.

Að lokum þarf að keyra síðustu Apertium-eininguna, þ.e. orðhlutafræðilega myndun. Sú eining tekur úttakið úr tilfærslunum og notar lemmurnar ásamt mörkunum til að mynda réttu orðmyndirnar “the big boys ate a good soup”. Þessi eining þarf á orðhlutafræðilegum greini fyrir ensku að halda.

4.2.2 IceParser

Ein af þeim rannsóknarspurningum sem leitast verður við að svara í þessu verkefni er sú hvort hægt er að nota úttakið úr *IceParser* í þýðingunni. Eins og áður hefur komið fram er *IceParser* hlutaþáttari sem skilar af sér upplýsingum um setningaliði og setningarfræðileg hlutverk. Fyrir setninguna í dæminu að ofan þá skilar *IceParser* eftirfarandi úttaki:

{*SUBJ> [NP [AP stóru lkfnvf AP] strákarnir nkfng NP] *SUBJ>} [VP borðuðu sfg3fþ VP] {*OBJ< [NP [AP góða lveosf AP] súpu nveo NP] *OBJ<}

Ofangreint úttak sýnir ekki eingöngu skiptingu í setningaliði heldur einnig upplýsingar um setningarfræðileg hlutverk eins og frumlag (*SUBJ) og andlag (*OBJ). Í stað þess að skrifa sérstakar Apertium-reglur til að framkvæma hlutaþáttun þá væri fýsilegt að geta notað úttakið úr *IceParser* beint inn í Apertium.

4.3 Mat á virkni

Þegar frumgerð af þýðingarkerfinu er tilbúin þá verður reynt að meta hvernig til tókst. Útbúin verða fjölda prófunartilvika sem metin verða á tvo vegu. Í fyrsta lagi verður hópur notenda beðinn um að meta þýðingarnar með því að svara eftirfarandi spurningu: “Býr þýðingarkerfið til úttak sem er nægjanleg gott þannig að minni tími fari í að leiðrétta úttakið í stað þess að búa þýðinguna til í höndunum frá grunni?”. Í öðru lagi mun kerfið verða metið m.t.t. villuhlutfalls sem mælt er í fjölda orða (per 100 orð) sem þarf að setja inn, eyða og skipta út til að gera þýðinguna ásættanlega.

Heimildir

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

M. Flanagan and S. McClure. 2002. SYSTRAN and the Reinvention of MT. `http://wwwv5.systransoft.com/IDC/26459.html`.

M. L. Forcada. 2006. Open-source machine translation: an opportunity for minor languages. In *Strategies for developing machine translation for minority languages (5^{th} SALTMIL workshop on Minority Languages, organized in conjunction with LREC 2006)*, Genoa, Italy.

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

A. K. Ingason, S. Helgadóttir, H. Loftsson, and E. Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using Hierachy of Linguistic Identities (HOLI). In B. Nordström and A. Rante, editors, *Advances in Natural Language Processing, 6^{th} International Conference on NLP, GoTAL 2008, Proceedings*, Gothenburg, Sweden.

H. Loftsson and E. Rögnvaldsson. 2006. A shallow syntactic annotation scheme for Icelandic text. Technical Report RUTR-SSE06004, Department of Computer Science, Reykjavik University.

H. Loftsson. 2006. Tagging a Morphologically Complex Language Using Heuristics. In T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala, editors, *Advances in Natural Language Processing, 5^{th} International Conference on NLP, FinTAL 2006, Proceedings*, Turku, Finland.

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

M. Nagao. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In *Proceedings of the international NATO symposium on Artificial and human intelligence*, Lyon, France.

J. R. Pierce and J. B. Carroll. 1966. Language and machines – computers in translation and linguistics. ALPAC report, National Academy of Sciences, National Research Council (Publication 1416). `http://www.nap.edu/openbook.php?isbn=ARC000005`.