



keele tehnoloogiast

Google teatas 31. jaanuaril, et täiendab oma keskkonda seitsme keele automaattõlke võimalusega. Albaania, galeegi, ungari, malta, tai ja türgi keele kõrval tehti võimalikuks ka eesti keele masintõlge. Kokku toetab Google'i keskkond nüüd 41 keelt. Ehkki pakutav tõlge eesti keelde toob kohati muige suule, on tegu ikkagi omamoodi verstepostiga. Kuidas hinnata seda sammu eesti keele ja ka keeletehnoloogia seisukohast?

Google'i automaatne masintõlge, mis oskab nüüd ka eesti keelt paljudesse teistesse keeltesse ja sealt tagasi tõlkida, on loomulikult suurepärase uudis ja verstepost edasiseks.

Kuidas nad seda tegid?

Google'i kui ettevõtte omapära on ennekõike globaalne väga suure infohulga kogumine ja sellega ümberkäimise võime. Veebi indekseerimise käigus kogunevad kokku teksti-andmebaasid (korpused) kõikides keeltes. Neid tekste analüüsid saab tuvastada sisult sama teksti erikeelseid versioone ehk moodustada n-ö paralleelkorpuse. Kui korpused on piisavalt palju näiteid, siis hakkavad seal korduma ka sõnad, fraasid ja terved laused. Tekstide automaatse statistilise analüüsi käigus saab tuvastada, kuidas erinevaid sõnu, fraase ja lauseid tüüpiliselt samas kontekstis on tõlgitud. Mida rohkem on näiteid, seda usaldusväärsemalt saab teha automaatselt „tõlget“ võrreldes lihtsalt andmebaasis varem olnud tõlkenäiteid. Piisavalt suurte tekstihulkade puhul võib põhimõtteliselt „tõlkida“ isegi nii, et vastava tõlkeprogrammi tegijate tiimis ei ole ühtegi selle keele oskajat. Nii võib öelda Google'i eesti keele tõlke kohta, et ilmselt kasvas nende paralleelkorpus nii suureks, et nad julgesid pidada selle põhjal tehtavat tõlget piisavalt kvaliteetseks.

Kuhu liigub eesti keele automaattõlge siinmail? Põhimõtteliselt lähenevad Eesti teadlased, näiteks Heiki-Jaan Kaalep, tõlkele samamoodi: automaattõlke jaoks on vaja koguda piisavalt kvaliteetsed tekstikorpused ning nende pealt trennida, trennida ja trennida tõlkesüsteeme. Ennekõike keskendutakse inglise-eesti-inglise tõlkele. Kindlasti on tõlget võimalik kõige kiiremini paremaks teha spetsiifilistes valdkondades, näiteks õigustõlke või tehnilise tõlke vallas – eeldusel, et kasutada on piisavalt palju kvaliteetseid tõlkenäiteid. Seal peaks mehaaniline toortõlge küll lihtsustama käsitööd.

Eriti huvitav on küsimus, kuidas peaks automaattõlkele kaasa aitama need jõupingutused, mida on tehtud ja tehakse eesti keele sõnamoodustuse, lauseehituse ja muu analüüsi

Google'is tõlkimiseks saadaolevad keeled:

albaania, arabia, bulgaaria, eesti, galeeni, heebrea, hiina, hindi, hispaania, hollandi, horvaadi, indoneesia, inglise, itaalia, jaapani, katalaani, korea, kreeka, leedu, läti, malta, norra, pilipino, poola, portugali, prantsuse, rootsi, rumeenia, saksa, serbia, slovaki, sloveeni, soome, taani, tai, tšehhi, türgi, ukraina, ungari, vene, vietnami

vallas ning kuidas neid meetodeid statistilise tõlkega kõige paremini siduda. Tööd peab seega jätkama nii andmete kogumise, süstematiseerimise ja puhastamise vallas (korpuste koostamine) kui eesti keele spetsiifiliste aspektide sidumise statistilisse tõlkesse.

Teksti sisust „aru saamise“ osas on tööd samuti väga palju. Selle lihtsam näide peaks olema mõistete süsteemide ja ontoloogiate ning tõlkesõnastike koostamine. Kui need ressursid on olemas, saab ka vabateksti kõigepealt proovida siduda vastavate täpsemate mõistetega, ning vastavalt suunata tõlget õiges suunas.

Tagasi Google'i juurde – tehnilised vahendid, mis võimaldavad veebis kiiresti enamvähem suvalises keeles lehekülje sisust umbkaudugi aimdust saada oma emakeeles, on väga teretulnud. Selliste süsteemide taga on osavad insenerid ja arvutiteadlaste tiimid, kelle kasutada on teatud mõttes piiramatud ressursid. Meie kõige piiratum ressurss on inimesed – huvitavaid ja olulisi ülesandeid jaguks meil kümnetele uutele tõsistele tegijatele. Kuid doktorantide pealekasv ei ole just väga kiire protsess.

Ilukirjanduse ja luule masintõlge ei allu aga loodetavasti veel nii pea esteetiliselt nauditava loomingulise tõlke nõuetele. Oleks ju väga kurb, kui suleseppade loomingulisus oleks arvutitele lihtsalt saavutatav pelgalt kõiki varasemaid inimkonna tekste analüüsides. ●

LAURI KULPSOO



Vastas Tartu ülikooli arvutiteaduse instituudi bioinformaatika professor

JAAK VILO

LOE VEEL

■ Google Translate: <http://translate.google.com/>