Osavõtt statistlise masintõlke võistlusest

Juhendaja: Mark Fišel

Tase: tiim kahest-kolmest inimesest, kus igaüks teeb bakalaureuse- või magistritööd sarnasel teemal aga keskendudes erinevate aspektide peale / pädev üksik bakalaureuse- või magistritöö.

Kirjeldus: Masintõlge tegeleb tõlkimisega loomulike keelte vahel. Statististiline masintõlge on selle alamliik, kus tõlketeadmisi (sõna- ja fraasisõnastiku kirjeid, ümberpaigutamise reegleid, ja muid parameetreid) arvutatakse automaatselt suure tõlkenäidete hulga peal -- sellise lähenemise peal on ehitatud Google Translate, Bing Translator ning üldiselt suur hulk kaasaegseid tõlkesüsteeme. Suurt hulga tõlkenäiteid nimetatakse treenimiskorpuseks.

Et võrrelda erinevaid mudeleid ja lähenemisviise statistilisele masintõlkele, korraldatakse järgmisel 2015. aastal juba 10. korda statistilise masintõlke võistlust -- the WMT Shared Task on Machine Translation. Osalejatele antakse ühte ja sama treenimiskorpust, ning lõpuks hinnatakse süsteemi väljundi tõlkekvaliteeti eraldi tehtud test-korpuse peal, millest osalejad saavad alguses ainult sisend-osa. Võrdlemiseks vt. eelmiste aastate üritusi: http://statmt.org/wmt13/, http://statmt.org/wmt14/

Osavõtt võistlusest ja antud teemal töö kirjutamine tähendaks:

- tavalise statistilise masintõlke tehnoloogiast aru saamine
- baassüsteemide loomine tõlkimiseks võistluses antud keeltepaaride jaoks
- tõlkesüsteemile muutuste kavandamine ja rakendamine. Muutused peaksid olema keelepaaridest sõltumatud -- nii on nii lihtsam kui ka lõbusam ja üldisemini rakendatavam.

Enamik tarkvarast baassüsteemi loomiseks on olemas; nõutud oleks pigem skriptide kirjutamine ning terve protsessi käima saamine.

Ajakava antakse hiljem korraldajate poolt, aga orienteeruvalt saab treeningandmeid kätte detsembris 2014 - jaanuaris 2015. Test-hulka avaldatakse reeglina märtsis. Pärast saab rahulikult baka-/magistritööd kirjutada.