Informaatikud raallingvistika globaalkonverentsil

Umbes aasta tagasi (25-29 jaanuar 2014) toimus Tartu Ülikoolis iga kahe aasta tagant erinevas riigis peetav rahvusvaheline wordneti konverents nimega „7th Conference on Global WordNet (GWC)“. Konverentsi ettekandjateks on tüüpiliselt teadlased, kelle teadustöö keskendub wordneti loomisele, täiendamisele ja rakendamisele erinevates loomuliku keele töötlemise ülesannetes nagu masintõlge, informatsiooni otsimine, sõnatähenduste ühestamine, korpuste annoteerimine jne. Peale selle ülesanded, mis vajavad semantilist analüüsi.

Konverentsi tarbeks valmis ühisartikkel emeriitprofessor Leo Võhanduga (juhendaja) ja Tartu ülikooli teaduri Heili Oraviga, mis tuli ka ettekandele.

Artiklis käsitletav uurimisteema keskendus universaalsetele (st keelest sõltumatutele) testmustritele, mis on sobilikud wordnet-tüüpi sõnastike hierarhiliste struktuuride hindamiseks. Kuigi doktoritöö raames on testmustreid peaaegu kümme, siis konverentsi artiklis kajastasime vaid kolme: asümmeetriline ring, suurim suletud alamhulk ja väikesed kuni kolme tasemega hierarhiad.

Kuna wordnet kvaliteedist sõltub loomuliku keele töötlemis ülesande kvaliteet on kriitilise tähtsusega ka wordneti enda kvaliteet. Meie poolt pakutud testmustrid aitavad valdkonna spetsialistidel (keeleteadlane, leksikograaf) leida wordneti hierarhilises struktuuris võimalikke ebakohti.

Ettekanne tekitas erilist huvi antud valdkonna tipptegijate nagu Piek Vossen, Christiane Fellbaum, Maciej Piasecki ja Alexandre Rademaker hulgas. Taoline tähelepanu on ka põhjendatud, kuna pakutud testmustrid wordneti kvaliteedi hindamiseks on genereeritavad automaatselt ja kasutatavad suvalise keele wordneti jaoks.

Wordnetist

Wordnetti on vaadeldud/kasutatud taksonoomiana, tesaurusena ja mõnikord ka nö kergekaalulise leksikaalse ontoloogiana. Esimene wordnet loodi Princetoni Ülikoolis 1990ndate alguses. Tänapäeval on üle 70 erineva keele wordneti, milliseid lisandub iga aastaga veelgi juurde.

Wordneti peamiseks ehituskiviks on sünonüümide hulk e sünohulk, millesse koondatakse kõik sama tähendusväljaga sõnad. Nt {ämber, pang}. Sünohulgad on üksteisega semantilistes seostes, millest osa (nt hüponüümia ja meronüümia) tekitavad hierarhilisi struktuure ja teised (nt lähisünonüümia ja antonüümia) mitte. Tekkivad struktuurid jaotuvad grammatilise kategooriate nagu *substantiiv*, *verb*, *adjektiiv* ja *adverb* alusel neljaks. Wordneti kasutamise seisukohalt on kõige olulisemaks semantiliseks seoseks *hüponüümia* koos *substantiivide* ja *verbide* sünohulkadega. Nimetatud grammatilised kategooriad koos hüponüümia seosega teevadki wordnet-tüüpi sõnastikest artaktiivse leksikaalse ressursi erinevate NLP ülesannete lahendamisel.

Konverentsist osavõttu toetas IT Akadeemia, mille eest oleme tänulikud.

TTÜ Informaatikainstituudi doktorant Ahti Lohk ja
emeriitprofessor Leo Võhandu