Eesti keele ajaväljendite tuvastaja uus versioon

Juhendaja: Siim Orasmaa (*siim . orasmaa @ ut . ee)*

Ajaväljendite tuvastaja on programm, mis leiab tekstist ajaväljendid (piiritleb ajaväljendi-fraasid) ning esitab nende semantika normaliseeritud kujul (seab vastavusse konkreetsed kalendrilised ajad, nt *"22. aprillil 2021."* => 2021-04-22, *'tänavu oktoobris'* => 2023-10).
Eesti keele ajaväljendite tuvastamiseks on loodud reeglipõhine programm (Orasmaa 2010, 2012), mis toetub morfoloogilisele analüüsile ning märgendab ajaväljendeid TIMEX3 märgendusskeemi alusel.

Käesoleva töö eesmärgiks on luua eesti keele ajaväljendite tuvastaja uus versioon, mis kasutab ära vana süsteemi teadmust ning laiendab seda uusimate võimalustega keele automaattöötluses, eesmärgiga saavutada parem kvaliteet ning läbipaistvam väljund. Ideed uuendusteks:

- Ajaväljendite eraldamisel saab sisendiks võtta automaatse süntaksianalüüsi väljundi, mis peaks oluliselt parendama fraaside tuvastamise kvaliteeti.
- Reeglite loomisel saab aluseks võtta vanad reeglid (reeglite formaadi kohta vt siit) ning esitada need EstNLTK grammatikate ja/või tuvastamisreeglite kujul. Kindlasti tuleks uurida võimalusi reeglite (pool-)automaatseks konverteerimiseks vanalt kujult uuele.
- Reeglites saab võtta kasutusele vahekuju ajaväljendi kontekstist sõltumatu semantika[1] esitamiseks, tuginedes näiteks LTIMEX (Mazur ja Dale 2011) või SCATE (Bethard ja Parker 2016) esituskujule. See peaks muutma nii reeglite ülesehituse kui süsteemi väljundi selgemaks ning võimaldaks vajadusel semantika üle arvutada.
- Saab katsetada Bert'i neuromudeleid keerukamate semantikaprobleemide lahendamisel (nt ajaväljendite ankurdamine ning konkreetse ja üldise tähenduse eristamine (Orasmaa 2012)).

Süsteemi arendamisel saab kasutada eesti keele ajaväljenditega märgendatud korpuseid, lisaks on uurimustöös võimalik kasutada Eesti keele koondkorpuse (ca 705 tuhat dokumenti) ajaväljendite märgendust, mis on loodud automaatselt vana süsteemi abil. Töö sisse kuulub uue süsteemi süstemaatiline hindamine ja võrdlus vanaga.
Lõpptulemusena peaks valmima (valdavalt) Pythonis implementeeritud ajaväljendite tuvastaja moodul, mis ühildub EstNLTK programmeerimisliidesega.

Kirjandus

Orasmaa, S. (2010). Ajaväljendite tuvastamine eestikeelses tekstis. Magistritöö. https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=1141
Orasmaa, S. (2012). Automaatne ajaväljendite tuvastamine eestikeelsetes tekstides. Eesti Rakenduslingvistika Ühingu aastaraamat, (8), 153-169. http://dx.doi.org/10.5128/ERYa8.10
Mazur, P., Dale, R. (2011). LTIMEX: representing the local semantics of temporal expressions. In 2011 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 201-208). IEEE.
Bethard, S., Parker, J. (2016). A semantically compositional annotation scheme for time normalization. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 3779–3786. European Language Resources Association (ELRA), Portorož, Slovenia.

1 Näiteks, ajaväljendi *„eile"* kontekstist sõltumatu semantika on *„1 päev enne tänast";* teades *tänast kuupäeva*, saame leida ka *„eile"* kontekstist sõltuva semantika ehk ajaväljendile vastava konkreetse kuupäeva.