Keeletehnoloogia ja Bürokratt

Kristiina Vilbaste
Riigi Infosüsteemi Ameti Bürokrati keeletehnoloogia projektijuht

Eesti avalikus sektoris on juba mitu aastat tegeletud mitmesuguste tehisintellekti- ehk kratilahenduste loomisega. Praeguseks on selliste projektide arv ületanud juba saja piiri ning avaliku sektori teadlikkus ja huvi valdkonna vastu on ajas järjest kasvanud. Viimasel paaril aastal enam kõneainet pakkunud algatusena võib välja tuua riiklike virtuaalassistentide koostoimelise võrgustiku, mis kannab nime Bürokratt.

Bürokrati sünniajaks võib pidada 2020. aasta veebruarit, mil Majandus- ja Kommunikatsiooniministeerium avalikustas kontseptsiooni avalike teenuste toimimisest tehisintellekti- ehk kratiajastul. Kontseptsioon sai nimeks Bürokratt ning tegemist on visiooniga avalike teenuste tarbimisest ja riigiga suhtlemisest virtuaalsete assistentide kaudu.[1]

Bürokrati eesmärk on muuta riigiga suhtlemine nii füüsilistele kui ka juriidilistele isikutele praegusest oluliselt lihtsamaks ja kiiremaks. Tähtis on, et kui inimene soovib riigiga suhelda, siis ei pea ta teadma, millise asutuse poole täpsemalt pöörduda on vaja, vaid ta suunatakse automaatselt õige asutuse poole, kes tema pöördumisega tegeleb. Ka ei pea inimene enam lähtuma asutustega kontakteerumisel nende lahtiolekuaegadest, sest Bürokratt vastab inimese päringutele ööpäev ringi ning ka nädalavahetustel ja riigipühadel.[2]

Eesti riikliku virtuaalassistendi esimene versioon valmis 2022. aasta kevadel ning järjepidevalt lisatakse tootele uusi võimekusi ning parandatakse selle kasutajamugavust. Praeguseks on Bürokratt paigaldatud viie asutuse kodulehele ning peagi on liitumas veel asutusi. Inimestel on võimalik asutusega nende kodulehel oleva Bürokrati vestlusakna kaudu kindlatel aegadel ühendust võtta. Eeltreeninguga hõlmatud küsimustele oskab Bürokratt ise vastata. Teemad, mida ei ole treenitud, suunatakse klienditeenindajale.

Nagu näeme, siis on visioon ja praegune olukord üksteisest veel üsna kaugel, kuigi iga iteratsiooniga liigutakse ideaalile lähemale. Bürokratt ei ole lihtsalt klassikaline infosüsteem, mille eesmärk on andmete varundamine ning nende liigutamine punktist A punkti B, vaid keerulisem sümbioos IT ja tehisintellekti lahendustest. Kuna Bürokratt on oma olemuselt virtuaalassistent, siis kasutatakse või hakatakse selles tulevikus kasutama peamiselt keeletehnoloogilisi (aga ka mõningaid muid tehisintellektil baseeruvaid) lahendusi. Siinkohal ongi sobilik rääkida, milliseid keeletehnoloogilisi lahendusi Bürokratt tänasel päeval juba kasutab ning kuidas aitavad tulevikus lisanduvad keeletehnoloogilised võimekused visioonile lähemale jõuda.

[1] **O. Velsberg.** #Bürokratt – virtuaalne assistent kodanikule. 26.02.2021.
[2] Kratid Eesti heaks. Virtuaalne abiline Bürokratt.

Enne aga veel paari sõnaga sellest, mida keeletehnoloogia endast kujutab. Lihtsustatult öeldes on keeletehnoloogia eesmärk võimaldada inimestel suhelda masinatega, kasutades selleks loomulikku inimkeelt. Keeletehnoloogia jaguneb vormi põhjal kaheks suuremaks haruks: inimkõne töötlemisele keskenduvaks kõnetehnoloogiaks ja kirjalike tekstide töötlemisel põhinevaks tekstitehnoloogiaks.[3] Levinumad keeletehnoloogia rakendusvaldkonnad on näiteks kõnetuvastus, kõnesüntees (teksti automaatne ettelugemine tehishääle poolt), automaattõlge, tekstidest automaatne teadmuse ammutamine ning viimasel paaril kuul palju kõneainet pakkunud automaatne tekstiloome, milles OpenAI arendatud ChatGPT[4] on näidanud tõeliselt häid tulemusi.

Tänane Bürokratt

Tulles aga tagasi Bürokrati juurde, siis praegu saab Bürokratiga suhelda vaid kirjalikult Bürokrati vestlusakna kaudu. Seetõttu kasutatakse Bürokratis praegu ennekõike tekstitehnoloogia lahendusi. Nagu eespool mainitud, on Bürokratt praegu paigaldatud viie asutuse (Tarbijakaitse ja Tehnilise Järelevalve Amet, majandustegevuse register, Politsei- ja Piirivalveamet, Statistikaamet, Viimsi vallavalitsus) kodulehele ning nende asutuste lehtedel olevat Bürokratti on treenitud vastama vaid selle asutuse spetsiifilistele küsimustele. Selleks et Bürokratt oleks võimeline iseseisvalt küsimustele vastama, vajab ta suurt mahtu õppematerjali ehk küsimuste ja vastuste paare.

Bürokratti on integreeritud spetsiaalsed masinõppe algoritmid, mis töötlevad sissetulevate kliendipöördumiste tekste. Tekstid tokeniseeritakse ehk jagatakse sõnaüksusteks, lemmatiseeritakse ehk kõik sõnad viiakse nende grammatilisse algvormi ning eemaldatakse stoppsõnad. Stoppsõnad on näiteks sidesõnad, kaassõnad, määrsõnad, asesõnad. Töödeldud kliendipöördumiste tekste kasutatakse sisendina masinõppe mudelite treenimiseks. Mudelite treenimine on pidev protsess, kus asutuses Bürokratiga töötavad spetsialistid annavad mudelile pidevalt tagasisidet tema tehtud otsuste õigsuse kohta ning mudel korrigeerib end saadud tagasiside järgi, et tulevikus pöördumisele paremini vastata.

Ka on valmis arendatud, kuid Bürokratiga veel mitte integreeritud isikuandmete anonüümimise tööriist[5], mis võimaldab kliendipöördumistes sisalduvaid isikuandmeid minimaalse vaevaga anonüümida nii, et teksti tarbijal pole enam võimalik aru saada, millise isiku andmeid tekst sisaldab. Selle tööriista loomiseks on kasutatud nimeüksuste tuvastamise tehnoloogiat ehk NER-i (*named entity recognition*).

Nimeüksuste tuvastamine tähendab, et kasutades Tartu Ülikooli loodud EstNLTK[6] teeki ja regulaaravaldisi eraldatakse tekstidest automaatselt konkreetsetele isikutele viitavad andmed, näiteks isikute ja organisatsioonide nimed, kohanimed, isikukoodid, telefoninumbrid,

[3] **M. Mihkla, L. Piits.** Vaimult suureks keeletehnoloogia toel. – Pikksilm. Tallinn: Arenguseire Keskus, 2022, lk 1.
[4] https://openai.com/blog/chatgpt.
[5] https://github.com/buerokratt/Data-Anonymizer.
[6] https://github.com/estnltk/estnltk.

meiliaadressid jms. Ka see tehnoloogia baseerub masinõppe mudelitel, mida on võimalik tagasisidestamise abil täpsemini nimeüksusi tuvastama õpetada.

Kuidas edasi?

Bürokrati 2023.–2024. aasta teekaardil on välja toodud järgmised olulised keeletehnoloogiaga seotud verstapostid:

- Bürokratiga peab saama suhelda suulises vormis;
- Bürokratt peab olema võimeline inimesele suuliselt sünteeshääle abil vastama;
- Bürokrati vestlusaknas ja nn tagatoas peab saama pöördumisi ning vastuseid automaatselt tõlkida.

Muutmaks Bürokratiga suhtlemist eri kasutajarühmadele mugavamaks, lisatakse Bürokratile TalTechis välja töötatud kõnetuvastuse lahendus kiirkirjutaja[7]. Kõnetuvastus on keeletehnoloogia üks olulisemaid alamtehnoloogiaid, mis võimaldab inimkõne automaatselt tekstiks muuta ehk transkribeerida. Seejärel on võimalik kirja pandud teksti juba klassikaliste tekstianalüüsi vahenditega töödelda ning sealt Bürokratiga suhtleva inimese pöördumise põhjus välja selgitada. Kõnetuvastuse liidestamine Bürokratiga parandab oluliselt avalike teenuste kättesaadavust ka nende kasutajarühmade puhul, kel on see varem eri tegurite tõttu problemaatiline olnud (nt vaegnägijad).

Kui Bürokratt on võimeline suulist keelt mõistma, siis on järgmise sammuna oluline, et ta oleks võimeline ka suuliselt vastama. Selleks kasutatakse kõnesünteesi ehk tehnoloogiat, mis võimaldab teksti inimesele kohati äravahetamiseni sarnase sünteeshääle abil ette lugeda. Eestis on kõnesünteesi arendamisega tegeletud juba pikalt, seda nii Eesti Keele Instituudis[8] kui ka Tartu Ülikoolis[9]. Sarnaselt kõnetuvastusega parandab ka kõnesüntees eelkõige vaegnägijate ligipääsu Bürokrati kaudu pakutavatele avalikele teenustele.

Kuna riigiga suhtlemise vajadus on ka mitte eesti keelt kõnelevatel inimestel, siis on tähtis, et Bürokratti oleks lisatud võimekus tõlkida olulisemates võõrkeeltes (nt inglise, vene, ukraina) kirjutatud tekst automaatselt eesti keelde. Kuid ka vastupidi, Bürokrati või klienditeenindaja kirjutatud eestikeelseid tekste peab olema võimalik automaatselt võõrkeelde tõlkida. Praegu plaanitakse Bürokratt liidestada Eesti Keele Instituudi Tõlkevärava[10], Tartu Ülikooli neurotõlke[11], Euroopa Parlamendi eTranslationi[12] ning Microsoft Azure'i ja Google'i API-dega. Bürokratti kasutav asutus saab ise valida, millise neist kasutusele võtab.

Kui Bürokratiga liitunud asutuste arv on kasvanud piisavalt suureks, on aeg astuda oluline samm lähemale Bürokrati suurele visioonile. See tähendab, et inimene ei pea enam tulevikus ise teadma, millise asutuse poole ta täpsemalt oma probleemiga peaks pöörduma. Inimene saab

[7] https://koodivaramu.eesti.ee/taltechnlp/kiirkirjutaja.
[8] https://www.eki.ee/heli/.
[9] https://neurokone.ee/.
[10] https://tolkevarav.eki.ee/.
[11] https://neurotolge.ee/.
[12] https://commission.europa.eu/resources-partners/etranslation_en.

oma pöördumise esitada Bürokrati kaudu ja see suunab tema pöördumise edasi õige asutuse poole. Lihtsustatult öeldes on tegemist automaatse tekstide klassifitseerijaga ehk spetsiaalseid masinõppe mudeleid on treenitud kõigi liitunud asutuste spetsiifiliste sisendandmetega ehk neile laekunud kliendipöördumistega. Selle lahenduse kohta on juba läbi viidud pilootprojekt ning see kannab hellitavalt nime Siimuke[13].

See on järjepidev protsess, kuna Bürokratiga liituvate asutuste arv aja jooksul kasvab ning Bürokratt peab suutma ka lisanduvate asutuste pöördumisi klassifitseerida. Teine põhjus, miks mudeleid on vaja järjepidevalt edasi arendada, peitub selles, et asutused võivad tulevikus hakata pakkuma teistsuguseid teenuseid kui praegu ehk nendeni jõudvad kliendipöördumised kannavad teistsugust sisu. Ka peab arvestama sellega, et keel, eriti internetis kasutatav keel, muutub pidevalt ning mudelid vajavad pidevalt värsket informatsiooni inimeste tegeliku keelekasutuse kohta.

Lisaks sellele, et Bürokratt peab olema võimeline suunama automaatselt inimeste pöördumised vastava asutuse juurde ning neile ka nii palju ise vastama kui võimalik, on vaja, et Bürokratt oskaks vastata ka olulisematele üldist laadi küsimustele. Näiteks kui inimene soovib teada, kes on parasjagu Eesti president või kui tihti korraldatakse Eestis valimisi, siis saaks Bürokratt ka sellistele küsimustele vastata. Siinkohal oleks mõistlik ära kasutada juba olemasolevaid võimalusi, nagu eespool mainitud ChatGPT, mis on võimeline üllatavalt hästi vastama ka Eestit puudutavatele küsimustele ning seda eesti keeles.

Ehk kõige uuenduslikum ja põnevam keeletehnoloogiauuendus, mida tulevikus Bürokratis näha loodame, on see, et Bürokratiga saaksid mugavamalt suhelda ka vaegkuuljad. Praegu on vaegkuuljatel küll võimalik kasutada asutustega suhtlemisel viipekeele tõlgi abi, kuid kuna praegu pole üheski Eesti ülikoolis võimalik eesti viipekeele tõlgiks õppida, siis väheneb viipekeele tõlgi teenuse kättesaadavus Eestis üha enam. Just sel põhjusel peame uurima võimalusi, kuidas seda probleemi tehisintellekti abil lahendada.

Näeme, et tulevikus peaks Bürokratt olema võimeline mõistma viipekeelse kõneleja viibeldud kõnet. Viipekeele automaatset tuvastamist lahendatakse küll masinnägemise tehnoloogia abil, kuid kuna viipekeele puhul on siiski tegemist ühega paljudest inimkeeltest, võib seda tehnoloogiat käsitada ka keeletehnoloogilise lahendusena. Lisaks viipekeele automaatsele tuvastamisele on vaja luua ka lahendus, mis oleks võimeline ise etteantud tekstilise või kõnelise sisendi järgi viipeid genereerima.

Kokkuvõte

Keeletehnoloogial on digitaliseeruvas maailmas üha olulisem roll. Olgu selleks kõnetuvastus, kõnesüntees, tekstiandmete automaatne töötlus ning sealt teadmuse ammutamine, me kohtame neid lahendusi oma igapäevaelus kogu aeg. Üks valdkond, kus keeletehnoloogia lahendused on väga laialdaselt kasutusel, on just virtuaalassistendid. Nagu teisteski avalikkusele väga hästi

[13] **Majandus- ja Kommunikatsiooniministeerium**. Kodanike ja ettevõtjate pöördumised suunatakse tulevikus automaatselt õigele asutusele ja ametnikule. 17.09.2021.

tuntud virtuaalassistentides (nt Siri, Alexa), on ka Bürokratis praegu kasutusel või kasutusele tulemas väga erinevad keeletehnoloogilised lahendused, mis aitavad klienditeenindajate tööd oluliselt automatiseerida ning avalikke teenuseid ka inimestele igal ajal kättesaadavaks muuta.