TARTU ÜLIKOOL

HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND

EESTI JA ÜLDKEELETEADUSE INSTITUUT

ÜLDKEELETEADUSE OSAKOND

Kristjan Poska

Nimeolemite tuvastamine 19. sajandi vallakohtu protokollides

Bakalaureusetöö

Juhendaja Siim Orasmaa, PhD

Tartu 2021

Sisukord

Sissejuhatus 3

1. Vallakohtutest 19. sajandil 5

1.1. Vallakohtu protokollid 6

1.2. Vallakohtu protokollide tähtsus 7

2. Nimeolemite tuvastamine 9

2.1. Nimeolemite tuvastamine eesti keeles 10

2.2. Lühidalt nimeolemite tuvastamise protsessist 11

2.3. Nimeolemite tuvastamise kasulikkus 12

3. Nimeolemite tuvastaja mudeli loomine 14

3.1. Kuldstandardi korpus ehk andmestik 16

3.2. Mudeli loomise eeltöö 17

3.3. Tuvastaja mudeli loomine 19

3.4. Mudeli kvaliteedi hindamine 19

3.5. Eksperimendid 21

3.5.1. Tunnuste komplektid 22

3.5.2. Eksperiment nimeolemite ühestamisega 27

3.5.3. Tulemuste sõltuvus treeningandmete hulgast 28

3.6. Lõpptulemus 29

Kokkuvõte 33

Kirjandus 35

Named Entity Recognition in 19th Century Parish Court Protocols. Summary 38

LISAD 39

Sissejuhatus

Nimeolemite tuvastamine (ingl *named entity recognition*) on üks infoeralduse ülesannetest, mis hõlmab endas tekstis infoelementide tuvastamist ja klassifitseerimist (Marrero jt 2013: 1). Nimeolemite hulka kuuluvad näiteks isikunimed (*Kersti Kaljulaid*), organisatsioonid (*Euroopa Liit*) ning asukohad (*Tartu*). Eesti keele nimeolemite tuvastaja väljatöötamisega ning kättesaadavaks muutmisega on peamiselt tegelenud kaks teaduslikku tööd: Alexander Tkatšenko magistritöö aastal 2010, mis tegeles nimeolemite tuvastaja väljatöötamisega ning Rasmus Maide bakalaureusetöö aastal 2020, mis tegeles nimeolemite tuvastaja EstNLTK[1] teeki integreerimisega.

Kuna nimeolemite tuvastamiseks loodud vahendid on kohandatud vaid tänapäeva kirjakeelele, siis uuritakse selles bakalaureusetöös, kuivõrd on võimalik neid vahendeid kohandada vana kirjakeele analüüsimiseks. Kui nimeolemite tuvastajat saaks kohandada vanale kirjakeelele, siis aitaks see kaasa eelkõige kirjandusteadlaste ning ajaloolaste, kuid ka keeleteadlaste, sh arvutilingvistide ja keeletehnoloogide tööle. Tuvastatud nimeolemid võimaldaksid lihtsustada näiteks nime, asukoha või organisatsiooni järgi tekstidest info otsimist. Vanale kirjakeelele kohandatud nimeolemite tuvastaja aitaks üldiselt tõsta eesti keele automaatse analüüsi taset.

Teadaolevalt ei ole keegi peale Maarja-Liisa Pilviku jt (Pilvik jt 2019) vana kirjakeele nimeolemite tuvastamisega tegelenud. Eelnimetatud artikliga võrreldes on selle bakalaureusetöö käigus analüüsitavad vallakohtu protokollid suurema varieeruvusega: artiklis kasutatud vallakohtu protokollid olid käsitsi parandatud ehk normaliseeritud, kuid siinses bakalaureusetöös kasutatavad protokollid on sisestatud võimalikult algsel kujul ehk tekste normaliseerimata. Samuti on oluline erinevus, et Maarja-Liisa Pilviku jt artiklis katsetati kirjakeele nimeolemite tuvastaja järelparandamist, kuid siin töös treenitakse süsteem käsitsi märgendatud korpusel ümber. Suurimaks probleemiks seoses nimeolemite tuvastaja loomisega võib osutuda vana kirjakeele morfoloogilise analüsaatori puudumine, sest suur osa nimeolemite tuvastusest põhineb analüsaatori tööl.

[1] https://github.com/estnltk/estnltk

Töö eesmärk on luua käsitsi märgendatud Rahvusarhiivi korpuse pealt masinõppel põhinev EstNLTK nimeolemite tuvastaja ning hinnata selle mudeli kvaliteeti. Eesmärgist tulenevad ka uurimisküsimused: kui hea on bakalaureusetöö käigus loodav nimeolemite tuvastaja mudel ning kuivõrd on see samal tasemel teiste omasugustega?

Eesmärgist tulenevalt saab töö autor uurida selle mudeli efektiivsust võrreldes tänapäeva nimeolemite tuvastaja mudelitega. Hetkeseisuga on tänapäeva kirjakeele masinõppel saavutatud kõrgeim tulemus EstNLTK nimeolemite tuvastaja abil 0,87 ehk 87% tõenäosus anda õige märgend (Tkatšenko jt 2013). Edasised, süvaõppel põhinevad tulemused, on jõudnud ka kõrgemale, 0,90 ehk 90% ringi (Tanvir jt 2021).

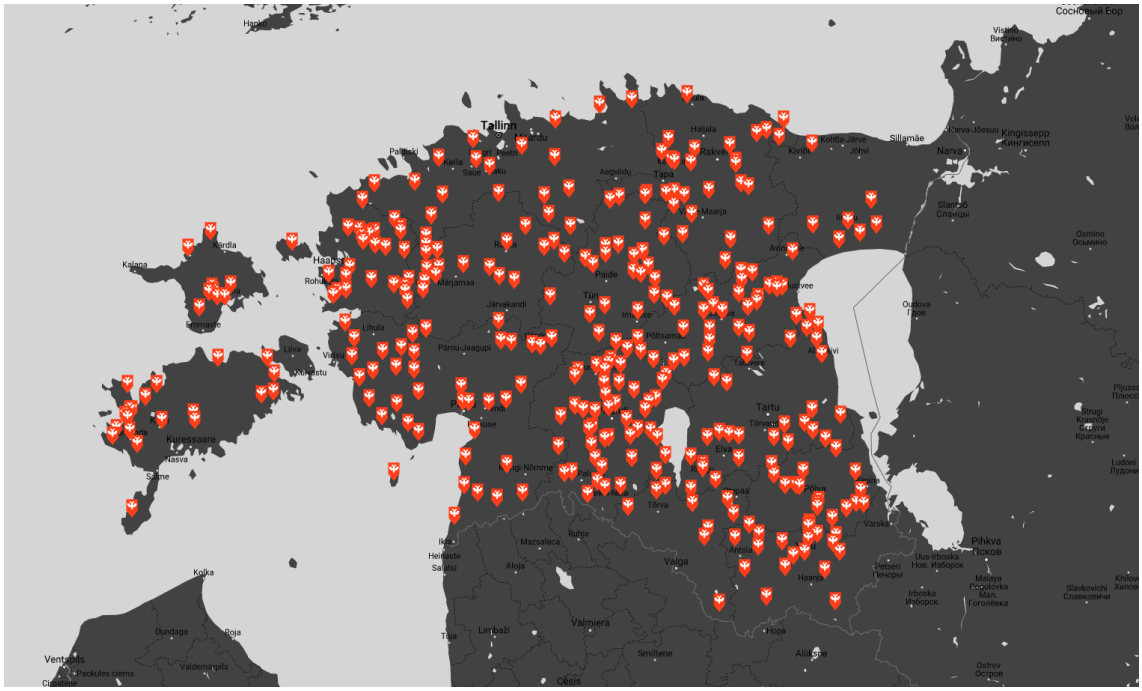
Töö jaguneb kaheks teooriaosaks ja üheks praktiliseks osaks: esimeses teooriaosas räägitakse vallakohtutest ja nende ajaloost ning vallakohtu protokollide olemusest ja tähtsusest. Teises teooriaosas räägitakse nimeolemite tuvastamisest, selle vajalikkusest ning protsessist. Kolmandas osas ehk praktilises osas räägitakse bakalaureusetöö käigus välja töötatud nimeolemite tuvastaja loomisest.

Bakalaureusetöö autor avaldab tänu töö juhendajale Siim Orasmaale tema kannatlikkuse, terase pilgu ning alati konstruktiivse ja vajadusel kriitilise tagasiside eest.

1. Vallakohtutest 19. sajandil

Vallakohus (sks *Gemeindegericht*) oli vallakogukonna kui kujuneva talurahva omavalitsusüksuse tuumik, sinna koondusid nii õiguslikud, politseilised kui ka halduslikud funktsioonid. Vallakohtute (tol ajal talurahvakohtute) ülesanne oli lahendada talupoegade omavahelisi tülisid ning muid väiksemaid süütegusid. (Linnus 1970: 231)

Valdade täpset arvu 19. sajandil pole teada, kuid külanõukogude moodustamisel 1945. aastal loeti valdu kokku 236 (231, kui jätta välja saarevallad, kus nõukogusid ei olnud) (EE 11). Joonisel 1 on näha vallakohtute asukohti tsaariaegses Eestis ehk Eestis enne 1917. aastat. Eesti Rahvusarhiivi on kogunenud ligi 450 protokolli- või lepinguraamatut ning 2247 säilikut, mille digiteerimiseks otsib Rahvusarhiiv hetkel vabatahtlikke.



Joonis 1. **Vallakohtute asukohad tsaariaegses Eestis**[2]

Oletatakse, et kõige esimese vallakohtu rajas krahv Karl Johann Mellin 1750. aastal Tuhala mõisas (Anepaio 2007: 353). Sellel ajal seda veel vallakohtuks ei nimetatud:

[2] https://www.ra.ee/vallakohtud/index.php

ametlik nimetus oli siiski mõisakohus. Eesti Entsüklopeedia ütleb, et vallakohtu ametlik nimetus oli mõisakohus kuni 1820. aastate lõpuni (EE 6: 475).

Kohtul oli palju ülesandeid, kuid need kõik olid seotud talupoegade eluolu ja vajadusel selle haldamisega. Näiteks lahendas vallakohus erinevaid varalisi probleeme, samuti probleeme maade ja talude vahel. Lisaks sellele lahendati ka üldiseid perekonna- ja moraaliprobleeme, selgitati talupoegadele õiguslikku korraldust ja hoiti üldist korda vallas. (Pilvik jt 2019: 141)

1.1. Vallakohtu protokollid

Protokollimine vallakohtutes seati sisse 1819. aastal, mil tuli välja Liivimaa talurahva vabastamisseadus (Traat 1980: 62–63). Sellest ajast pärinevad ka esimesed vallakohtute kirjalikud allikad. 1820.–1866. aastatest on säilinud ainult Liivimaa kubermangu Eesti osa vallakohtute arhivaalid: need on lünklikud ja kohati üsna abitus eesti keeles, sageli ka saksa keeles (Linnus 1970: 232). Säilikuid sellest perioodist on vähe, sest 1817. aastal vallakohtud kui sellised kaotati, kuid taastati 1866. aastal (Talving 2012: 3). Alates 1866. aastast läks olukord seaduse tõttu paremaks ning sellest ajast pärineb suur hulk kirjalikke allikaid Eesti ala kohta (Linnus 1970: 232). Kuna 1866. aasta seadus sätestas, et protokolle tuleb kirjutada kogukonna enamuse keeles, siis paljud neist protokollidest on eestikeelsed.

Vallakohtu protokollide seas on põhiliselt kassaraamatud, postikande raamatud. Kohtutoimikud on peamiselt seotud võlanõuete, süüteoasjade (sh pisivargused jmt) ning ka vähemalt määral pärandus- ja eestkosteasjadega. Hilisematest aastatest (19. sajandi lõpp ning 20. sajandi algus) pärineb ka protokolliraamatuid, ringkirju ja määruseid, kirjavahetusi kohtuotsuste täitmise, võlgade sissenõudmise, uute vallakohtunike valimise jmt kohta. (Linnus 1970)

Järgnev näide pärineb 1889. aasta 20. jaanuarist. Protokoll on kirja pandud Ahja vallakohtus:

Kusta Matsoon kaebas, et Mihkel Soe on mõtsas puulõikamise juures tema kallale tulnud ja kirwega talle läbi kübara haawa päha löönud, mille eest ta 300 rubla nõuab.

Mihkel Soe ütles kaibuse peale, et Kusta Matsoon on ise tema kallale tulnud, teda hullus löönud ja muidu peksnud, nii et tema seda sugugi ei tea, kuda ma talle lõi.

Tunnistaja Adam Saag ütles, et Kusta Matsoon on ühte puud mõtsast maha lõigata tahtnud, mida Mihkel Soe ei ole lasknud, selle peale läinud naad tõuklema pärast karwupidi kokku kus juures Mihkel Soe kirwega Kusta Matsoonile päha oli löönud, pärast näinud ka et weri mööda pääd maha jooksnud.

Tunnistaja Jaan Kusma ütles et tema kakelust pole näinud, aga werist Kusta Matsoni pääd on ta küll näinud pärast tapelust.

Tunnistaja Joosep Mägi ja Peeter Saag ei ole tulnud tehti

Otsus: Tunnistajad Joosep Mägi ja Peeter Saag 27al kohtu ette tallitada.

Protokollist on näha, et kohtumõistmise protsessi puhul on lähtutud üldtuntud tõest, et tuleb ära kuulata mõlemad osapooled ning tunnistajad. Protokolli otsusest saab täheldada, et lõppotsust ei tehta enne kui kõik tunnistajad on üle kuulatud.

19. sajandi esimeses kasutati ametliku keelena vallakohtu protokollide kirjapanekuks lisaks eesti keelele ka suures osas saksa ja vähemas osas vene keelt (Pilvik jt 2019: 140). 1866. aastal sätestas Aleksander II Baltimaade vallaseaduses, et protokollide keelena peab kasutama kogukonnas räägitavalt keelt – seetõttu on enne seda aastat säilinud väga palju protokolle ka saksa ja vene keeles, kuid pärast seda peamiselt eesti keeles. Siinse bakalaureusetöö andmestiku seas on näiteks venekeelseid protokolle Väätsa vallakohtust Järvamaalt ning Loona vallakohtust Saaremaalt, kuid ka saksakeelseid protokolle Abja vallakohtust.

1.2. Vallakohtu protokollide tähtsus

Vallakohtu protokollid kätkevad endas väärtuslikku informatsiooni 19. sajandi talurahva elust: nende keelekasutusest, sotsiaalsetest suhetest, majanduslikust seisust, varalistest tehingutest jpm (Linnus 1970). Lisaks filoloogidele pakuvad vallakohtu protokollid huvi

ka ajaloolastele, ajalooentusiastidele ja etnoloogidele. Samuti võib see huvi pakkuda inimestele, kes soovivad uurida oma sugupuu kohta või mingi kindla asukoha või nime kohta.

Filoloogidele pakub vallakohtu protokollide puhul huvi eelkõige keelelised nüansid: murded, keelenormile vastavus jne. Eestis kasutati tol ajal paralleelselt kaks kirjakeelt: lõunaeesti (tartu) ning põhjaeesti (tallinna) kirjakeelt. Seetõttu on näiteks huvitav uurida, kas ja kuidas võeti uus kirjaviis vastu pärast Eduard Ahrensi 1843. aasta grammatikat (Pilvik jt 2019: 146). Kuigi trükis hakati uut kirjaviisi kasutama alles 1870.–1880. aastatel, siis võib aimata, et 19. sajandi lõpus ilmunud protokollid on juba uues kirjaviisis (Pilvik jt 2019: 146).

2. Nimeolemite tuvastamine

Nimeolemite (ka nimeüksuste) tuvastamine (ingl *named entity recognition*) hõlmab endas tekstis informatsioonielementide tuvastamist (ehk identifitseerimist) ja klassifitseerimist (Marrero jt 2013: 1). Nimeolemite mõiste esitati esimest korda aastal 1996 kuuendat korda toimunud konverentsil Message Understanding (MUC-6). Mõiste esitajad olid arvutilingvistid Ralph Grishman ning Beth Sundheim (Grishman, Sundheim 1996: 467). Sellest ajast saati on täheldatud, et nimeolemite, näiteks isikunimede, organisatsioonide, asukohtade jmt tuvastamine on tähtis osa näiteks tekstikokkuvõtete loomisel (Prasad, Kantesaria 2020), informatsiooni filtreerimisel (Hidalgo jt 2005) ja suguluse uurimisel (Dai 2019).

Nimeolemite tuvastamine seisneb selles, et tuvastamiseks loodud mudelile antakse ette mingi tekst, näiteks

> Eesti Vabariik on riik Põhja-Euroopas. Eesti Vabariigi president on Kersti Kaljulaid.

ning kõikidele sõnedele selles tekstis antakse nimeolemi märgend

> LOC(Eesti Vabariik) O(on) O(riik) LOC(Põhja-Euroopas). ORG(Eesti Vabariigi) O(president) O(on) PER(Kersti Kaljulaid).

Tekstis on kasutatud märgendit LOC, mis tähistab asukohta, ORG, mis tähistab organisatsiooni, PER, mis tähistab isikut, ning O, mis tähistab nimeolemite välist teksti (ehk mitte-nimeolemit).

Kuigi parimad inglise keele mudelid saavutasid juba 2005.–2008. aastal nimeolemeid tulemusi tuvastades nimeolemeid 95% täpsusega, pole nende tuvastamise ülesanne kaugeltki mitte lahendatud (Marrero jt 2013: 9). Parimate nimeolemite tuvastamise mudelite saamiseks kasutatakse tihtipeale üsna kirjakeelseid tekste, näiteks ajaleheartikleid, raamatuid, Vikipeedia artikleid jpm, kuid keel ei koosne vaid kirjakeelest. Kuna maailmas on tõusuteel väga palju erinevaid nimeolemitega seotud tekstifiltreerimise protsesse (nt arvamuste kaevandamine (ingl *opinion mining*) või

emotsioonidetektorid), siis oleks vaja nimeolemite tuvastajaid ka nendele kohandada (Marrero jt 2013: 9–11), seega tegelikult ei ole ülesanne veel lahendatud.

Lisaks kirjakeelele on olemas ka suuline keel (kõnekeel, jututubade tekstid, vestlused sõprade vahel), mille keeleline varieeruvus ehk ebastandardsus on palju suurem kui kirjakeeles. Osa keelest on vana kirjakeel, mille nimeolemite tuvastamisele siinne töö keskendub. Vanal eesti kirjakeelel on palju tunnuseid, mis eristavad seda tänapäeva kirjakeelest, näiteks *v* asemel *w* või *n*-täht omastava käände tunnusena. Kuigi tänapäeva kirjakeele morfoloogiline analüsaator võib suuremast osast tekstist aru saada, peab ta eelnimetatud erisuste tõttu nii mõnelgi juhul kaotust tunnistama.

2.1. Nimeolemite tuvastamine eesti keeles

Eesti keele nimeolemite tuvastamisega tegi oma magistritöös 2010. aastal algust Aleksandr Tkatšenko. Tkatšenko kasutas oma töös uudisportaali Delfi[3] artikleid, mis pärinesid aastatest 1997–2009. Korpuse suurus oli 496 artiklit ning korpus koosnes 84 175 sõnest. Magistritöös kasutati tõenäosuspõhist mudelit ning selle abil saavutati selle korpuse parimaks tulemuseks 0,86 ehk 86% tõenäosus anda nimele õige märgend. (Tkatšenko 2010)

2013. aastal jätkas Tkatšenko nimeolemite tuvastamisega ning koos Timo Petmansoni ja Sven Lauriga kirjutasid nad artikli (Tkatšenko jt 2013), kus rääkisid lähemalt lahti selle protsessi sisu. Sel korral koosnes korpus lisaks Delfi artiklitele ka Postimehe[4] artiklitest: samast ajavahemikust (1997–2009), mis ka Tkatšenko magistritöös. Korpus kasvas 496 artiklilt 572 artikli suuruseks, korpus koosnes 184 638 sõnest. Artiklis räägitakse, milliseid tunnuseid tekstidest eraldati ning kuidas neid nimeolemite tuvastamisel kasutati (Tkatšenko jt 2013: 80). Korpuse parim tulemus oli 0,87 ehk 87%-ne tõenäosus anda õige märgend.

2021. aprillis tutvustasid Tanvir jt (Tanvir jt 2021) nimeolemite tuvastamiseks keelespetsiifilist mudelit BERT. Selle mudeli eeltreenimiseks kasutati tekstikorpust Estonian National Corpus 2017 (Kallas, Koppel 2018), mis oli tol hetkel suurim

[3] https://www.delfi.ee/
[4] https://www.postimees.ee/

saadavalolev korpus. Enne nimeolemite tuvastamist kasutati teksti eeltöötlemiseks EstNLTK teeki (Laur jt 2020), mille abil eemaldati XML/HTML märgendid ning eemaldati võõrkeelsed tekstid. Selle mudeliga saavutati nimeolemite tuvastamisel tulemus *ca* 0,90 ehk 90% ringis.

2.2. Lühidalt nimeolemite tuvastamise protsessist

Nimeolemite tuvastamise võib inimesele tunduda arusaadav ja isegi „lihtne“, aga selle automatiseerimiseks vajaminev aeg ning töömaht on väga suured. Nimeolemite tuvastamisel kasutatakse peamiselt kahte erinevat lähenemisviisi: reeglipõhiseid ning tõenäosuspõhiseid meetodeid. Tõenäosuspõhiste meetodite hulgast on viimasel ajal eraldi välja hakatud tooma ka sügavõpet (ingl *deep learning*), mis põhineb tehisnärvivõrkudel (Fišel 2021).

Reeglipõhiste meetodite puhul antakse arvutiprogrammile ette hulk reegleid (nii morfoloogilisi kui ka süntaktilisi), mille põhjal saab mudel õppida ning hakata nimeolemeid tuvastama. Üks kõige tähtsamaid reegleid nimeolemite tuvastamisel on see, kui sõne algab suurtähega (või läbiva suurtähega) (Riaz 2010). Eesti keeles saab lisanäitena välja tuua nimede eesliited, näiteks lühendile dr (doktor) järgneb üldjuhul isikunimi. Samuti esineb sarnane reegel ettevõtete nimedes: lühenditele AS, OÜ või MTÜ järgneb (või eelneb) tihtipeale ettevõtte nimi (samuti läheb see kokku suurtähereegliga).

Morfoloogilise tüpoloogia abil jaotatakse keeled isoleerivateks, aglutineerivateks ning flekteerivateks keelteks (Erelt 2005: 7). Uuritud on morfoloogia mõju nii nimeolemite tuvastamisele kui ka süntaktilisele analüüsile. Tulemused näitavad, et kuigi väga põhjalike aglutineerivate keelte puhul (nt araabia, mongoli, soome) on vaja teha ära suur eeltöö morfoloogilise informatsiooni ammutamisel (näiteks reeglipõhise märgendaja loomisel), siis annab see eeltöö lõpuks häid tulemusi (Güngör jt 2019).

Statistiliste mudelite põhist meetodit on hakatud kasutama, et vähendada kogenud arvutilingvistidel grammatilise mudeli väljatöötamiseks kuluvat aega (Padmanabhan 2013). Mudeli väljatöötamine põhineb treeningandmetel ning mudeli testandmetele rakendamisel. Statistiliste mudelite puhul on kaks peamist teed. Traditsioonilises masinõppes on arvutiprogrammile määratud tunnuste komplektid ning algoritmi

ülesandeks on õppida seaduspärasused tunnuste ning nimeüksuste esinemise vahel. Süvaõppes suudab programm juba ise tunnused välja selgitada (Fišel 2021). Lihtsa näitena tunnuse kohta võib tuua selle, et kui lause keskel on mõni suure algustähega sõne, siis on see suure tõenäosusega nimi.

Statistilise ehk tõenäosuspõhise mudeli väljatöötamine vajab piisavalt suurt käsitsi märgendatud korpust, mille põhjal mudelit treenida (Nothman jt 2013: 152). Seetõttu on kõige suurem sellise mudeli väljatöötamise osa eeltöö: kuldstandardi korpuse (ehk sajaprotsendiliselt õigete märgenditega käsitsi korpuse) loomine ja teksti eeltöötlus. Suure eeltöö tegemine annab seevastu ka parema tulemuse, sest mida rohkem on mudelil infot, mille pealt õppida, seda suurema tõenäosusega annab ta parema tulemuse (Nothman jt 2013).

Sügavõppest (ka süvaõpe) kui sellisest räägiti juba 1970. aastal, sellest kirjutas vene informaatik Mikhail Bongard oma teoses „Pattern Recognition“ (1970). Küll aga oli enne 20. sajandi lõppu neid meetodeid väga raske kasutada, sest arvutitel ei piisanud jõudlusest. 21. sajandil saab aga rääkida juba superarvutitest: teadaolevalt maailma parim arvuti, jaapani superarvuti Fugaku, suudab lahendada *ca* 440 000 matemaatilist tehet sekundis (Buchholz 2021). Seetõttu on hakatud sügavõppest rääkima palju optimistlikumalt.

Võrreldes sügavõpet traditsioonilise masinõppega, võib näha, et kuigi sel on mitmeid eeliseid (lihtne töövoog, toorsisendi piisavus, konteksti arvestamise automaatne õppimine), on ka sügavõppel mitmeid puudujääke. Peamised puudujäägid on näiteks 5–20 korda suurem andmete vajalikkus edukaks treenimiseks võrreldes masinõppega ning ka läbipaistmatus – mõned sügavõppe osad on tõlgendatavad, kuid enamasti puudub arendajatel arusaam, kuidas närvivõrgud täpselt ennustavad. (Fišel 2021)

2.3. Nimeolemite tuvastamise kasulikkus

Nimeolemite tuvastamist kasutatakse suuremahuliste andmete töötlemisel ning nendest vajaliku informatsiooni eraldamisel. Selle protsessi kasulikkust illustreerib näiteks Wei jt (2020) uurimus, kus kasutati nimeolemite tuvastamist biomeditsiinilistest tekstidest tarkvaranimede eraldamiseks. Selle protsessi abil, saavutades pealkirjadest nimede

eraldamisel tulemuse 91,79%, loodi kõrgkvaliteetne indeks tarkvaranimedest, mida biomeditsiinis kasutada saab.

Nimeolemite tuvastamine on samuti väga kasulik abivahend näiteks otsingumootorites: kellegi või millegi kohta nime põhjal info otsimine on mitmeid kordi efektiivsem kui lihtsalt tekste järjest läbi lugeda. 21. sajandi alguses hakati suurematele otsingumootoritele, sh Google[5], Bing[6] ja Yahoo![7], andma patente, mille otsingualgoritmides mängisid nimeolemid suurt rolli. (Alasiry 2015: 12–13) Lisaks sellele kasutatakse nimeolemite tuvastamist ka sisu soovitamise automatiseerimisel (näiteks internetist ajalehti lugedes) ja kasutajatuge pakkudes (Gupta 2018).

Soomlased Kimmo Kettunen ning Teemu Ruokalainen viisid 2017. aasta juunis läbi uurimuse, kus tuvastasid nimeolemeid Soome ajalooarhiivi ajalehekogus Digi[8]. Seda arhiivi kasutavad väga paljud inimesed: sugupuu-uurijad, pärandiseltsid, teadustöö tegijad ning ka ajalooentusiastid. Lisaks sellele soovitakse arhiiv teha kättesaadavamaks ka üldisele õppetööle (Kettunen, Ruokalainen 2017: 1).

[5] https://www.google.com/
[6] https://www.bing.com/
[7] https://www.yahoo.com/
[8] https://digi.kansalliskirjasto.fi/

3. Nimeolemite tuvastaja mudeli loomine

Bakalaureusetöö eesmärk on käsitsi märgendatud korpuse põhjalt luua masinõppel põhinev nimeolemite tuvastaja ning hinnata selle mudeli kvaliteeti.

Tõenäosuspõhise nimeolemite tuvastaja mudeli loomisel on tarvis tunda kuldstandardi korpuse mõistet. Kuldstandardi korpuse all mõeldakse üldjuhul käsitsi märgendatud korpust, mille märgendust saab võtta nii-öelda kullaproovina: märgendid on selles korpuses kõige õigemad.

Mudelit kontrollides saab mudeli märgendusi võrrelda kuldstandardi protokollide märgendustega – saab võrrelda nende märgenduste asukohta (indeksite järgi) ning seda, kas mudel on määranud tekstipositsioonile õige märgendi. Tabelis 1 on näidatud EstNLTK nimeolemite märgendaja mudeli märgendeid ning nende selgitusi.

Tabel 1. **Märgendid, nende vaste käsitsi märgendatud korpuses ning nende selgitus EstNLTK nimeolemite märgendaja mudelis**

Märgend EstNLTK mudelis	**Märgend käsitsi märgendatud korpuses**	**Selgitus**
PER	Isik	Isikunimi
LOC	Koht	Asukoht
ORG	Org	Organisatsioon
MISC	Muu, Teadmata, Ese	Mitmesugune (ingl *miscellaneous*)
LOC_ORG	KO_koht	Mitmene märgend asukoha ja organisatsiooni märkimiseks

Sõnatasandil märgendamisel lisandub eelnimetatud märgendite ette kas B- või I-, mis vastavalt tähendavad algust (ingl *beginning*) ning sisemist (ingl *inside*). Lisandub ka märgend O, mis tähendab nimeolemite hulga välist sõna.

Käsitsi märgendatud korpuse märgendid erinesid veidi standardsetest EstNLTK nimeolemite märgendaja märgenditest. Selle korpuse märgendusjuhendi[9] järgi olid nimeüksuste liigid järgnevad:

- Isik
- KO_koht
- Koht
- Org
- Ese
- Muu
- Teadmata

Märgend *Isik* tähistab isikute mainimisi tekstis. Tekstis mainitud isik võib olla nimetatud ees- ja perekonnanimega, eesnime esitähe ja perekonnanimega, ainult initsiaalidega ning eraldiseisva eesnime või perekonnanimega. *Isik*u märgendi on saanud näiteks „Maria Regendiini“, „H. Liivak“, „Jaani“ ja „Trei“, kuid ka „Märt Laas'le“, „J. M.“ ning omavahel põimunud isiku- ja isanimi „Peeter Kristjani p Petersoni“ või „Mihkel Tõnise poeg Reinberg“.

Märgend *KO_koht* on mõeldud nimeüksustele, mis mõnes kontekstis tähendavad kohti ning teises kontekstis selle kohaga seotud inimrühmi. Põhiliselt kasutati seda märgendit talunimede puhul, millele viidatakse ka sõnadega *krunt*, *maja*, *pere* ja ka *ase*. Samuti on selle märgendi saanud mõisate, valdade, kogukondade, külade ja kõrtside nimed. Selle märgendi on saanud näiteks väljendid „Poka talu“, „Ahjamoisa“, „Kiwwijerve koggokonna“, „Ninna kortsus“, „Peterburi Linnas“.

Märgendi *Koht* said sellised kohanimed, mille puhul pole protokolli alusel võimalik öelda, kas tegu on küla, valla vmt asutusüksusega. Samuti on see märgend kasutusel ka siis, kui ilma liigsõnata kohanimi on kohakäändes. Selle märgendi said näiteks väljendid „Herjanurmest“ ja „Puurmanni metsa“.

[9] Publitseerimata nimeolemite märgendusjuhend (Kadri Muischnek, Anna Edela, Siim Orasmaa).

Märgendi *Org* said protokollides organisatsiooni nimed ja nimetused. Näiteks on selle märgendi saanud „Kohhila mõisa politsei“ ning „Kaiu walla kohtusse“.

Märgendi *Ese* said nimelised esemed. Eelkõige said selle märgendi protokollides nimetatud laevad, näiteks „laew „Eduard““ või „Jaht Klara“, kuid ka nimelised tooted, näiteks „Bairich õlut“.

Märgendit *Muu* on kasutatud nimede puhul, mis ei mahu eelnimetatud kategooriate alla, aga mille puhul on siiski võimalik identifitseerida, millega on tegu. Selle märgendi on saanud näiteks sündmused („Nuustaku laadal“, „Paide Mardilaada“), aga ka kohtupidamisel kasutatud seaduseraamatute nimed („Liiwimaa talurwahwa seaduse raamatu“ ning „Tallorahwa seaduse ramato“).

Märgendi *Teadmata* said juhud, kus ei õnnestunud nime liiki määrata.

3.1. Kuldstandardi korpus ehk andmestik

Eksperimendi aluseks on 1500 vallakohtu protokolli, mis pärinevad peamiselt 19. sajandi lõpust ning 20. sajandi algusest. Protokollid pärinevad Rahvusarhiivi ühisloome projektist[10]. Protokollid on kuldstandardi korpuse loomise eesmärgil käsitsi märgendatud BRAT tööriista[11] abil ning nimeolemeid on selles kuldstandardi korpuses ümmarguselt 27 500. Protokollide seas on kirjeid lisaks tolle aja suurematele piirkondadele Harjumaale ja Tartumaale ka Järvamaa, Saaremaa, Võrumaa jne valdadest.

Tabelis 2 on välja toodud nimeolemite osakaal üle kõikide kuldstandardi korpuses olevate protokollide.

[10] www.ra.ee/vallakohtud
[11] https://brat.nlplab.org/index.html

Tabel 2. **Nimeolemite osakaal kuldstandardi korpuses %-des**

Märgend	**Osakaal**
PER	84,00
LOC_ORG	9,91
LOC	3,65
ORG	1,52
MISC	0,92

Tabelist on näha, et kõige rohkem oli nende protokollide seas PER märgendiga nimeolemeid ehk isikunimesid (84,00%). Kui mõelda vallakohtu protokollide teksti struktuuri peale, siis on see oodatav: keegi süüdistab kedagi, keegi on tunnistaja, keegi mõistab kellelegi midagi välja ning keegi kirjutab otsusele alla. Veel enam mõjutab seda asjaolu, et kohtumõistjaid oli erapooletuks jäämise eesmärgil tihti mitu. Lisaks sellele on suur osakaal ka LOC_ORG märgendiga nimeolemitel (9,91%), sest protokollides viidatakse tihti taludele, mille ümber kohtuasi keerleb, või valdadele, kus kohus toimub, ning sellele organisatsioonile, kes kohut mõistab (näiteks Rabiwere walla kohus).

3.2. Mudeli loomise eeltöö

Esimene ülesanne oli BRAT tööriista abil märgendatud protokollid teisendada EstNLTK jaoks sobivasse formaati ehk *Text* objektideks. Igal protokollil oli lisaks tekstifailile ka *ann*-laiendiga fail (märgendus ehk *annotation*), mis oli abiks nende objektide loomisel. Märgendusfailis oli kirja pandud andmed formaadis Tn ehk *trigger number* (põhimõtteliselt järjekorranumber), nimeolemi märgend, algusindeks, lõpuindeks ning nimi. Tabelis 3 on näide *ann*-laiendiga ehk märgendeid sisaldavast failist.

Tabel 3. **Näide BRAT-tööriista abil märgendatud protokollide .ann failist**

Tn	**Märgend, algus, lõpp**	**Nimi/nimeolem**
T1	Org 24 56	Rawila ja Palwere walla kohtusse
T2	Isik 94 104	Jaak Urgas
T3	Isik 120 134	Madis Weermann
T4	Isik 135 144	Hans Härg
T5	Isik 146 157	Jaan Leppik
T6	KO_koht 158 168	Albu walla
T7	Isik 215 229	Anop Prikülile
T8	Isik 466 477	Anop Prikül
T9	Isik 602 612	Jaak Urgas
T10	Isik 632 646	Madis Weermann
T11	Isik 651 660	Hans Härg

Tabelist 3 on näha, et BRAT-tööriista abil märgendatud nimeolemite seas on palju isikunimesid (*Isik* märgendiga nimed). Sellele viitab ka isikunimede (ehk PER-märgendite) 84-protsendiline osakaal korpuses. Samuti näeb tabelist 3, et BRAT-tööriista märgendite sildid erinevad kuldstandardi korpuses kasutatud märgenditest. Nende teisendust näeb tabelist 1.

BRAT-tööriista märgendid oli vaja teisendada süsteemi treenimisel ja hindamisel kasutatavateks märgenditeks, sest esimese bakalaureusetöö eksperimendi raames proovis autor tekstidel ka tänapäeva kirjakeele nimeolemite tuvastaja mudelit, mis kasutab teistsuguseid märgendeid (vt peatükk 3). Tabelis 1 on näidatud, kuidas BRAT-tööriista märgendid teisenduvad EstNLTK-le vajalikeks märgenditeks.

3.3. Tuvastaja mudeli loomine

Pärast seda, kui kuldstandardi korpus oli JSON-failide (ingl *JavaScript Object Notation*) kujul olemas, sai nende põhjal hakata treenima nimeolemite tuvastaja mudelit. Selleks oli kuldstandardi korpus vaja jaotada kuueks alamosaks, millest viit kasutati ristvalideerimise meetodi abil mudeli treenimiseks ning ühte mudeli lõplikuks testimiseks. Alamosade jaotamisel oli vaja jälgida, et kõikides alamosades oleks kuldstandardi korpuse märgendusi proportsionaalselt sarnases koguses, vastasel juhul oleks mudeli treenimine kallutatud.

EstNLTK teegist kasutati treenimiseks estner moodulit, mille mudeli treenijale oli võimalik anda tekstid, mille põhjal mudelit treenida, ning märgendajad / tunnuste komplektid, mille põhjal oli võimalik tekstile tunnuseid lisada. Mudeli treenimisel kasutati ristvalideerimist: üks alamosa viiest oli testimiseks ning ülejäänud treenimiseks. Niimoodi kasutati testimiseks ning treenimiseks võrdselt kõiki alamosasid.

3.4. Mudeli kvaliteedi hindamine

Mudeli kvaliteedi hindamiseks kasutatakse F-skoori, mis on täpsuse ja saagise harmooniline keskmine. Täpsuseks nimetatakse positiivsete tulemuste jagatist positiivsete ja negatiivsete (ehk kõikide) tulemuste koguarvuga. Saagis on positiivsete tulemuste jagatis positiivsete ja valenegatiivsete tulemuste koguarvuga.

Allpool välja toodud valemites on näidatud, kuidas täpsust, saagist ning f-skoori segadusmaatriksi (joonis 2) andmete abil arvutatakse.

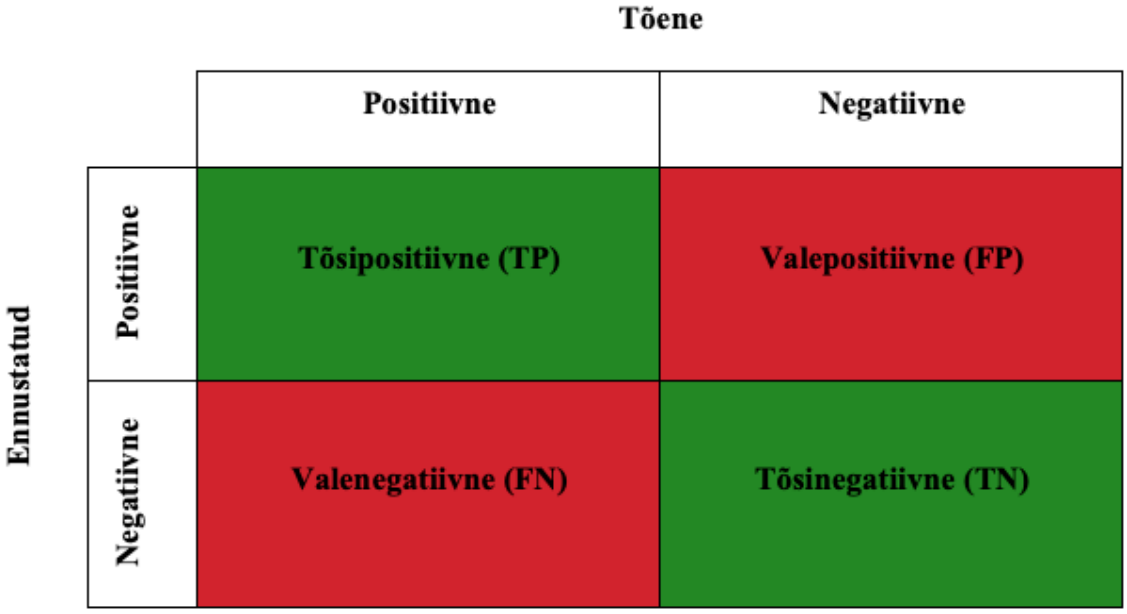
$$Täpsus = \frac{TP}{TP + TN}$$

$$Saagis = \frac{TP}{TP + VN}$$

$$F\text{-}skoor = 2 * \frac{täpsus * saagis}{täpsus + saagis} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Valemites on TP (tõsi)positiivne (ingl *true positive*), TN (tõsi)negatiivne (ingl *true negative*), VN väärnegatiivne (ingl *false negative*) ning VP valepositiivne (ingl *false positive*).

Joonisel 2 viidatud segadusmaatriksist on võimalik näha, kuidas valemites kasutatud väärtused illustreerituna välja näevad. Tabeli ülemine rida näitab tõest ehk õiget väärtust ning alumine osa mudeli poolt ennustatud väärtust. Kui näiteks tegelik väärtus on positiivne ning ka mudel ennustab, et see on positiivne, siis on tegemist tõsipositiivse väärtusega. Seevastu kui tegelik tulemus on positiivne, kuid mudel ennustab, et see on negatiivne, siis on tegu valenegatiivse tulemusega.



Joonis 2. **Segadusmaatriks (ingl *confusion matrix*) illustreerimaks positiivsete ja negatiivsete tulemuste vahekorda**

Täpsuse ja saagise arvutamise näitlikustamiseks võtame 100 pärisnime. Kui mudel tuvastab 40 nime, millest tegelikult 20 on pärisnimed, kuid 20 valesti tuvastatud, siis täpsus on 20/40 ehk 1/2 ~ 50%. Seevastu on saagis selle näite puhul 20/100 ehk 1/5 ~ 20%.

Kui täpsus on kõrge, kuid saagis on madal, on mudel oma tuvastatud kogumist suure osa õigesti määranud, kuid tuvastanud liiga vähe nimesid. Vastupidiselt, kui täpsus on madal ning saagis kõrge, on mudel tuvastanud suure osa kõikidest nimedest, kuid lisaks õigetele ka väga palju nimesid valesti tuvastanud. Seega on hea leida tasakaal mõlema tulemuse vahel ning sellest veel parem on, kui mõlemad tulemused on kõrged.

Nimeolemite tuvastamise koha pealt on täpsusel ning saagisel mõlemal väga suur roll: lisaks nime leidmisele ning märgendi õigesti määramisele on tähtsad ka sõnapiirid. Vallakohtu protokollide seas on ka väga pikki nimesid, näiteks tabelis 3 välja toodud „Rawila ja Palwere walla kohtusse“. Märgendaja peab olema võimeline ära tundma, et terve see viiesõnaline fraas on üks nimi, kuid samal ajal peab ka mõistma, et see nimi koosneb omakorda nimedest. Kui märgendaja tuvastab eelnimetatud nimest vaid „Rawila“, siis on nimi küll teoreetiliselt tuvastatud, kuid mitte terviklikult.

3.5. Eksperimendid

Nimeolemite tuvastamise mudeli loomisel on kaks peamist tähtsat komponenti: tunnused ning õppimisalgoritm. Tunnused määravad nimeolemitele erinevad erijooned, mille abil neid testimisandmetes tuvastada. Näited tunnustest on sõna lemma (algvorm), sõnaliik, kuulumine mõnesse (koha)nimeloendisse ehk geograafilisse leksikoni (ingl *gazetteer*) (Tkatšenko jt 2013: 80). Sarnaselt Tkatšenko jt 2013. aasta nimeolemite tuvastamisele kasutas bakalaureusetöö autor õppimisalgoritmina tingimuslikel juhuslikel väljadel põhinevat mudelit ehk CRF (*conditional random fields*) mudelit ning sarnaseid tunnuste komplekte.

Töö käigus loodav nimeolemite tuvastaja suudab kõigi eelduste kohaselt märgendada vanas kirjakeeles kirjutatud protokolle paremini kui tänapäeva kirjakeele nimeolemite tuvastaja. Selle võrdluse huvides märgendas töö autor treeningandmed (5 alamhulka, 1250 faili) kõigepealt tänapäeva kirjakeele mudeliga.

Tabel 4. **Tänapäeva kirjakeele nimeolemite tuvastaja tulemus treeningandmetel**

Mudel	**Täpsus**	**Saagis**	**F-skoor**
LOC_ORG	0,575	0,557	0,565
Ühestades LOC märgendiks	0,583	0,565	0,574
Ühestades ORG märgendiks	0,578	0,560	0,569

Tabelist 4 on näha, et tänapäeva kirjakeele nimeolemite tuvastaja mudel andis treeningandmetel (1.–5. alamhulgal) tulemuseks 0,565.

Kuna EstNLTK nimeolemite tuvastaja mudel ei kasuta märgendamisel LOC_ORG märgendit ning kuna LOC_ORG märgend on mitmetähenduslik, siis tegi töö autor ka väikse eksperimendi nii, et ühestas kõik LOC_ORG märgendiga nimed treeningandmete hulgas kas LOC märgendiks või ORG märgendiks, sest viimast kahte tunneb ka EstNLTK nimeolemite märgendaja.

Tulemustes (vt tabel 4) oli näha väikest paranemist: kui ühestada LOC_ORG märgendid LOC märgendiks, siis sai mudel skooriks 0,574, kui ühestada need ORG märgenditeks, siis sai mudel tulemuseks 0,569. See tähendab, et LOC_ORG kui mitmene märgend on pigem lähemal märgendile LOC kui märgendile ORG.

3.5.1. Tunnuste komplektid

Tunnuste komplektid viitavad sellele, et iga komplektiga lisatakse tekstile uusi tunnuseid või kihte, mille põhjal on võimalik nimeolemite tuvastamise kvaliteeti parendada. Tkatšenko (2013) eeskujul olid tunnuste komplektid mudeli treenimiseks järgnevad:

1. Baastunnused (näiteks näiteks kas sõne on suurtäheline või väiketäheline, kas sõne sisaldab numbreid või mitte, mis on sõne pikkus jne);
2. Baastunnused ja morfoloogilised tunnused (seejuures lemmasid kasutamata);
3. Baastunnused ja morfoloogilised tunnused (kasutades lemmasid);

4. Baastunnused, morfoloogilised tunnused ning lausepõhised tunnused (tunnused lause esimesele ja viimasele sõnele);
5. Baastunnused, morfoloogilised tunnused, lausepõhised tunnused ning nimeloend ehk leksikon (ingl *gazetteer*).
6. Baastunnused, morfoloogilised tunnused, lausepõhised tunnused, nimeloend ja globaalsed tunnused (aitavad tuvastada konteksti selle sõne samas dokumendis eelneva esinemise põhjal).

Tunnuste komplekte oli vaja mudelite treenimisel järk-järgult juurde lisada, sest see tõstab mudeli kvaliteeti (parandab täpsust/saagist ning vähendab tuvastusvigu). Samuti annavad head tunnused mudelile ka üldistusvõime: see peaks töötama mitte ainult treeningandmetel, vaid ka (sarnase keelekasutusega) tundmatutel tekstidel. Lõppkokkuvõttes osutus kõige paremaks mudeliks siiski kõige suurema tunnuste komplektide arvuga mudel ehk 6. mudel (vt tabel 5).

Tabel 5. **Mudelite tulemused tunnuste komplektide lõikes**

Mudel	**Täpsus**	**Saagis**	**F-skoor**
Baas (1)	0,866	0,832	0,849
Baas + morf (lemmadeta) (2)	0,873	0,849	0,861
Baas + morf (lemmadega) (3)	0,901	0,864	0,882
Baas + morf + lausepõhised tunnused (4)	0,898	0,863	0,880
Baas + morf + lausepõhised tunnused + nimeloend (5)	0,900	0,864	0,882
Baas + morf + lausepõhised tunnused + nimeloend + globaalsed tunnused (6)	0,906	0,875	0,890

Nendest tulemustest on näha, et suurimad hüpped toimuvad 1.–3. mudeli juures, 4. ja 5. mudel on üsna sama tulemusega ning 6. mudel tegi jällegi veidi suurema hüppe. Üllatav on asjaolu, et algse hüpoteesi kohaselt tänapäeva kirjakeele morfoloogiline analüsaator vana kirjakeele tekstidel väga hästi ei tööta. Seevastu on näha, et baastunnuste ning

täieliku morfoloogilise analüüsi lisamise vahel on peaaegu 0,4 ehk 4% suurune tulemuse paranemine (vt tabel 5). Sarnast paranemist nägid ka Tkatšenko jt (2013: 82) oma artiklis, kus kasutati treening- ning testandmetena ajaleheartikleid.

Kuna 4. ning 5. mudel jäid tulemuselt üsna statsionaarseks, otsustas töö autor uurida nende mudelite eripärasid. Kui 4. mudeli juures lausepõhiste tunnuste muutmist on raskem teha, siis 5. mudeli nimeloendi ehk leksikoniga sai teha lisaeksperimente.

Nimeloend koosneb kahest tulbast, mis on eraldatud tühikutega. Esimeses tulbas on mingi nimi, organisatsioon, asukoht jm, mis võib olla ükskõik millisel kujul (mitmest sõnast koosnev, ainult esitähed jm). Teises tulbas on selle nime märgend. Kuigi märgendite tulp on olemas, siis mudeli treenimisel ei ole kindlalt määratud, et nimi selle märgendi nimeloendist saab. Märgendite tulba märgend mõjutab tulemust koosmõjus teiste tunnustega.

Kuna EstNLTK vaikimisi nimeloend on koostatud tänapäeva kirjakeele nimedest, siis oli esimene idee luua uus nimeloend, kus on kasutatud ka vanematest tekstidest pärit nimesid. Allikaid leksikoni loomiseks oli kaks:

- Rahvusarhiivi vallakohtute ühisloome protokollid[12], kus olid erinevate inimeste poolt käsitsi märgendatud kohanimed ning isikunimed (mis olid ebaühtlase kvaliteediga – mõni märgendaja märgendas paremini, mõni halvemini;
- vana kirjakeele korpuse[13] vallakohtu protokollide morfoloogilised analüüsid, kus oli vaid info, et tegemist on nimega, kuid puudus nime liik.

Mõlemal allikal olid failid, kust oli võimalik nimed välja noppida. Ühisloome protokollide puhul tuli silmas pidada, et nimeloendi hulka ei satuks nimesid, mis pärinevad treeningandmete seas olevatel protokollidel, need välistati nimeloendi loomisel.

Nendest kahest algallikast loodi kokku kolm nimeloendit:

- gazetteer_vk.txt, milles olid vaid ühisloome projekti protokollidest pärit nimed,
- gazetteer_tsv.txt, milles olid vana kirjakeele korpusest pärit nimed,

[12] https://www.ra.ee/vallakohtud/
[13] https://vakk.ut.ee/

- gazetteer_both.txt, milles olid eelnimetatud nimeloendid kokkupanduna.

Kuna kõige parema tulemuse eelneval mudeli treenimisel andis 6. mudel (vt tabel 5), kasutas töö autor nimeloendi muutmist just sellel mudelil. Kõikide eelnimetatud nimeloenditega oli tulemus väiksem kui tänapäeva kirjakeele nimeloendiga (vt tabel 6).

Tabel 6. **Mudeli tulemused uue nimeloendiga**

Mudel	**Täpsus**	**Saagis**	**F-skoor**
Baas + morf + lausepõhised tunnused + nimeloend + globaalsed tunnused (algne 6. mudel)	0,906	0,875	0,890
Baas + morf + lausepõhised tunnused + globaalsed tunnused + gazetteer_vk	0,894	0,872	0,883
Baas + morf + lausepõhised tunnused + globaalsed tunnused + gazetteer_tsv	0,893	0,874	0,883
Baas + morf + lausepõhised tunnused + globaalsed tunnused + gazetteer_both	0,895	0,873	0,884

Töö autor koostas ka tabeli tulemustega nimeolemite lõikes ehk kõikide märgendite eraldi tulemused (vt tabel 7). Sellest tabelist oli näha, et uute nimeloenditega on kõige madalam tulemus LOC märgendil (~0,51) kusjuures kõikidel teistel märgenditel oli ca 0,15 võrra kõrgem tulemus. Selle märgendi väiksest tulemusest sündis ka järgmine idee – lisada kõige parema tulemuse saanud nimeloendile juurde LOC märgendiga nimesid (asukohti).

Tabel 7. **Parima mudeli (6. mudel) tulemused nimemärgendite lõikes**

Märgend	**Täpsus**	**Saagis**	**F-skoor**
PER	0,937	0,926	0,931
LOC_ORG	0,752	0,663	0,703
LOC	0,611	0,444	0,513

ORG	0,780	0,732	0,755
MISC	0,746	0,635	0,683

Sellest eksperimendist sündis kaks uut nimeloendit: lisatud LOC märgenditega gazetteer_both.txt ning samuti lisavariantidega versioon eelnimetatust. LOC-märgendid lisandusid nimeloendile gazetteer_both juurde EstNLTK tänapäeva kirjakeele nimeolemite tuvastaja nimeloendist – töö autor võttis sealt välja kõik LOC märgendiga nimed ning lisas need eelnimetatud nimeloendisse. Lisavariantidega nimeloend loodi seetõttu, et vallakohtu protokollides kasutati tihti *v*-tähe asemel *w*-tähte (*vald – wald* jt). Seega koosnes lisavariantidega versioon gazetteer_both.txt nimedest, LOC märgenditest ning lisavariantidest, kus oli kõikidest *v*-tähte sisaldavatest nimedest loodud ka *w*-tähega versioon. Parima tulemuse üle kõikide mudelite sai viimane, LOC märgendite ning lisavariantidega nimeloendit kasutav mudel (vt tabel 8).

Tabel 8. **Mudeli tulemused nimeloenditesse LOC märgendite ning lisavariantide lisamisega**

Mudel	**Täpsus**	**Saagis**	**F-skoor**
Baas + morf + lausepõhised tunnused + nimeloend + globaalsed tunnused (algne 6. mudel)	0,906	0,875	0,890
Baas + morf + lausepõhised tunnused + globaalsed tunnused + gazetteer_both + LOC	0,908	0,875	0,891
Baas + morf + lausepõhised tunnused + globaalsed tunnused + gazetteer_both + LOC ja lisavariandid	0,909	0,876	0,892

3.5.2. Eksperiment nimeolemite ühestamisega

Nimeolemite ühestamine oli üks lisaeksperimentidest, mida siinse töö autor tulemuse parandamiseks läbi viis. Ühestamise mõistet kasutatakse tihemini morfoloogilise analüüsi puhul, kus valitakse konteksti arvestades mitmete erinevate morfoloogiliste tõlgenduste vahel õige variant, kuid siinse bakalaureusetöö puhul toimus ühestamine nimeolemite märgendite puhul.

Rasmus Maide uuris oma bakalaureusetöös, kui mõni nimi on saanud 90% ulatuses ühe märgendi ning 10% ulatuses mõne muu märgendi, siis kas võime öelda, et mudeli tulemus paraneb, kui ka väiksema osa märgendid muuta enamuse märgenditeks (Maide 2020: 18–21). See tähendab, et kui nimi „Jaan“ on saanud 90% juhtudest B-PER märgendi ning 10% juhtudest I-PER märgendi, siis kas tulemus paraneb, kui muuta ka I-PER märgendid B-PER märgenditeks.

Selle eksperimendi läbiviimiseks võttis töö autor parima mudeliga märgendatud protokollid ning lõi neist kõigepealt sõnastiku, kus luges kokku kõik erinevad märgendid, mida üks nimi protokollides on saanud, näiteks:

„Jaan“ : {B-PER: 421, I-PER: 120},
„talu“ : {O: 610},
„mõis“ : {I-LOC: 130, O: 726},
„Peep“ : {B-PER: 212},
...

Pärast selle sõnastiku loomist lõi töö autor uue sõnastiku, kus andis nendele märgenditele protsentuaalsed väärtused, näiteks kui sõne „Jaan“ esines tekstis eelneva näite põhjal kokku 421+120 = 541 korda, siis märgend B-PER esines protsentuaalselt kokku (421/541)*100% ≈ 77,8% ning märgend I-PER (120/541)*100% ≈ 22,2%.

Seejärel oli vaja sõnastikust eemaldada kõik sõned, millel oli ainult üks märgend, kuna antud eksperiment tegeleb ainult mitu märgendit saanud nimedega. Seega eemaldati sõnastikust eelneva näite põhjal sõned „talu“ ning „Peep“, sest mõlemad sõned on saanud vaid ühe märgendi.

Eksperimenti oli vaja jooksutada mitmes erinevas konfiguratsioonis. Nimelt on tähtis, et kui eelneva näite põhjal sõna *mõis* on suuresti O (ehk nimeolemite-välise) märgendiga, siis ei tohiks neid väheseid I-LOC märgendeid hoobilt O märgendiks teha, sest see annaks valed tulemused. Kui meil on nimeolemid, mis on saanud harva mõne nimemärgendi (ehk mitte O), siis kui ka need teha O märgendiks, muutuks liialt suur osa sõnedest nimeolemite-väliseks. Seega kui nimeolemi märgendi, näiteks sõna *mõis* puhul märgendi I-LOC muuta O märgendiks, siis peaks piir olema kõrgem (O märgendit peaks olema näiteks vähemalt 95–99% ulatuses).

Tabel 9. **Nimeolemite märgendite ühestamise tulemused**

	Algne mudeli tulemus	**Uus tulemus**
Täpsus	0,909	0,908
Saagis	0,876	0,877
F1	0,892	0,892

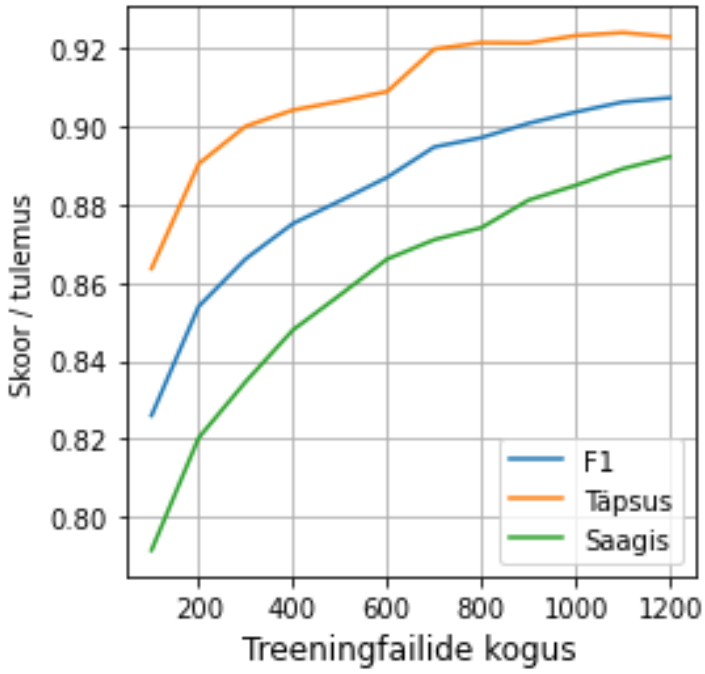
Eksperimendi tulemustest oli näha, et kõige parema tulemuse andis konfiguratsioon, kus O märgendid jäetakse täielikult välja ning nimeolemite märgendid ühestatakse juhtudel, kui üks märgenditest kaalub üle teis(t)e 95% ulatuses (ehk kui ühte märgendit on teiste märgendite seast rohkem kui 95%). F-skoor paranes sellise konfiguratsiooni puhul minimaalselt: 0,892105-lt 0,892179-le (vt tabel 9), seega statistiliselt ei ole see liialt tähelepanuväärne. See-eest annab eksperiment mõista, et suund võib olla õige. Tabelis 9 on algne mudeli tulemus on madalam kui lõplik tulemus, sest algne mudeli skoor põhineb ristvalideerimisel 1.–5. alamhulgal, kasutades neist nelja alamhulka treenimiseks ning ühte testimiseks.

3.5.3. Tulemuste sõltuvus treeningandmete hulgast

Tähtis eksperiment nimeolemite tuvastamise puhul on näha, kuidas treeningandmete hulk mõjutab mudeli kvaliteeti. Tkatšenko jt leidsid 2013. aastal, et nende artiklis kasutatud

treeningkorpuse puhul on kõige järsem tõus kuni 300 dokumendini 572-st, pärast seda hakkab skoor tasanema (Tkatšenko jt 2013: 81–82).

Siinse töö autor leidis eksperimendi läbiviimise käigus (joonis 3), et kasutades treeningkorpusena 1.–5. alamhulka ning testides seda 6. alamhulga peal, tõuseb skoor samuti alguses väga järsult (kuni 700 dokumendini 1250-st), kuid hiljem tasaneb. See on kooskõlas Tkatšenko jt eksperimendiga: tõus laugeb umbes poole pealt ehk 50–60% vahemikus. See-eest näitab selline graafik siiski tõusutrendi ning võib eeldada, et kui treeningandmeid oleks veelgi rohkem, oleks ka tulemused kõrgemad.



Joonis 3. **Tulemuste sõltuvus treeningandmete hulgast**

3.6. Lõpptulemus

Nimeolemite tuvastaja mudeli loomise aluseks oli 1500 käsitsi märgendatud korpuse faili. Mudeli treenimiseks jaotati need 1500 faili kuute alamhulka, millest treenimiseks kasutati esimest viit – kuues alamhulk jäi parima mudeli testimiseks. Kuna mudeli treenimine toimus algselt ristvalideerimise teel ainult esimese viie alamhulga peal (s.t, et treenimiseks oli 4 alamhulka ja testimiseks 1), siis on ristvalidatsiooni teel saadud tulemus veidi madalam. Esimese nelja alamhulga peal treenitud (ja viienda peal testitud)

mudeli parim keskmine tulemus oli 0,892 (jättes hetkel välja nimeolemite ühestamise eksperimendi, kuna see kasv oli niivõrd väike).

Parima mudeli ja lõpliku tulemuse saamiseks võttis töö autor treeningandmeteks kõik esimesed viis alamhulka ning treenis mudeli parima tulemuse saanud mudeli seadetega. Mudeli lõplik testimine toimus kuuenda alamhulga peal, mille puhul sai töö autor tulemuse, mis on nähtavad tabelist 10.

Tabel 10. **Parima nimeolemite tuvastaja mudeli tulemused**

Mudel	**Täpsus**	**Saagis**	**F-skoor**
Baas + morf + lausepõhised tunnused + globaalsed tunnused + uus *gazetteer* + LOC ja lisavariandid	0,925	0,892	0,909

Täpsus on ~0,92, mis tähendab, et kõikidest nimedest, mis mudel on tuvastanud, on umbes 92% õiged. Saagis on ~0,89, mis näitab, et kõikidest nimedest, mida mudel oleks kokku pidanud tuvastama, on mudel tuvastanud 89%. Kuna f-skoor on täpsuse ning saagise harmooniline keskmine, siis (0,925 + 0,892) / 2 ≈ 0,909. See tähendab, et mudel suudab õigesti tuvastada umbes 90% nimedest.

Eesti kirjakeele nimeolemite tuvastaja mudeli kõrgeim saavutatud tulemus (f-skoor) on 0,87, seega siinses bakalaureusetöös treenitud mudel saavutas üsna arvestatava ning võrreldava tulemuse.

Tähtis osa sellest tulemusest on ka välja tuua parima mudeli tulemused nimeolemite liikide (nt ORG, LOC jt) lõikes. Selleks väljastas töö autor ka tulemused liikide lõikes, mida näeb tabelist 11.

Tabel 11. **Parima mudeli tulemused nimeolemite lõikes**

Märgend	**Täpsus**	**Saagis**	**F-skoor**
PER	0,954	0,937	0,945
LOC_ORG	0,786	0,695	0,738
LOC	0,602	0,490	0,540
ORG	0,838	0,781	0,809
MISC	0,684	0,578	0,627

Eelnevas tabelis on märgendid pandud nimekirja, mis järgib nende esinemissagedust töö aluseks olevas korpuses (vt tabel 2). PER märgendi puhul on tulemus arusaadav: kuna korpus koosneb suuresti just isikunimedest (84%), siis on see saanud ka kõige parema tulemuse. Huvitav on aga see, et LOC märgend (mis on esinemissageduselt kolmas) on saanud kõige madalama tulemuse, sealjuures esinemissageduselt neljas märgend, ORG, on saanud isegi kõrgema tulemuse kui esinemissageduselt teine märgend, LOC_ORG.

Tabel 12. **Segadusmaatriks parima tulemuste puhul erinevate nimemärgendite lõikes**

Ennustatud / **Tõene**	**LOC**	**LOC_ORG**	**MISC**	**ORG**	**PER**
LOC	71	10	0	1	12
LOC_ORG	5	276	0	0	11
MISC	0	0	26	0	5
ORG	0	1	0	57	1

PER	2	3	0	0	3352

Töö autor koostas mudeli illustreerimiseks segadusmaatriksi (ingl *confusion matrix*) (vt tabel 12), mis näitab ennustatud ning tõeste märgendite vahekorda. Segadusmaatriksist on näha, et mudel üledefineerib märgendit PER ehk annab liiga tihti selle märgendi mingile teisele nimeolemile, näiteks LOC (12), LOC_ORG (11), MISC (5) või ORG (1).

Samuti saab segadusmaatriksi põhjal öelda, et LOC märgend aetakse väga tihti segi mõne teise märgendiga: mudel arvab, et see võib olla LOC_ORG (10), ORG (1) või PER (12). Kuna LOC märgend on korpuses osakaalult kolmas märgend (3,65%), siis võib eeldada, et see osakaal siiski niivõrd väike, et kui see aetakse mõne teise märgendiga sassi, siis langeb selle õige märgendi tulemus märgatavalt.

Üllatav on segadusmaatriksist näha ka seda, et märgendeid ORG ning MISC, mis on korpuses osakaalult väga madalad, vastavalt 1,52% ning 0,92%, pole peaaegu et kordagi mõne teise märgendiga sassi aetud. See tulemus näitab, et mudel saab väga hästi hakkama ORG-tüüpi nimedega (näiteks „Kohhila mõisa politsei“, „Röa-mõisa walitsusele“ või „Kaiu walla kohtusse“) ning MISC-tüüpi nimedega (näiteks „Tallorahwa seaduse ramato“ ja ka „T. S. R.“ ning „Laewa „Kalewi““ või „Jaht Klara“).

Kui võrrelda lõpliku mudeli tulemusi nimeolemite lõikes (tabel 11) ning mudeli tulemusi enne uute nimeloendite jmt eksperimentide tegemist (tabel 7), siis on näha, et kuigi PER märgendi tulemus lõplikus tulemuses väheneb, siis kõikide teiste märgendite tulemus suureneb. Lõpptulemus oli siiski parem, sest PER märgendit on korpuses ülekaalukalt kõige rohkem ning kui see mõne punkti võrra vähenes, siis toimus teistes selle arvelt tasakaalustumine.

Kokkuvõte

Käesoleva töö eesmärk oli luua käsitsi märgendatud Rahvusarhiivi korpuse pealt masinõppel põhinev EstNLTK nimeolemite tuvastaja ning hinnata selle mudeli kvaliteeti. Mudeli hindamiseks võttis töö autor eesmärgiks võrrelda siinses töös loodud nimeolemite tuvastajat EstNLTK tänapäeva kirjakeele nimeolemite tuvastajaga.

Töö tähtsus seisneb selles, et nimeolemite tuvastamine kui infoeralduse ülesanne on tähtis nii ajaloolastele, sugupuu-uurijatele, filoloogidele jpt aladele. Samuti annab nimeolemite tuvastaja mudel võimaluse uurida täpsemalt ka vanu tekste, mida siiani tänapäeva kirjakeele nimeolemite tuvastajaga täpselt uurida ei suudetud.

Töö käigus loodi nimeolemite tuvastaja mudel Rahvusarhiivi korpuse pealt, kus jagati algsed 1500 faili kuueks alamosaks, millest esimest viit kasutati ristvalideerimise meetodil treenimiseks ning kuuendat lõplikuks testimiseks. Kõige suurema tähelepanu said mudelid, kus oli võimalik muuta nimeloendit ehk leksikoni (ingl *gazetteer*) ja viia see vastavusse vana kirjakeelega. Lisaks eksperimenteeris töö autor nimeolemite ühestamisega ning treeningandmete järk-järgulise suurendamisega.

Töö käigus kasutati EstNLTK tänapäeva kirjakeele nimeolemite tuvastajat, mille seadeid muudeti vastavalt vajadusele. Huvitav oli näha, et eesti kirjakeele morfoloogiline analüsaator sai vähemalt nimeolemite tuvastamise seisukohast vanas kirjakeeles kirjutatud tekstide analüüsimisega üsna hästi hakkama.

EstNLTK tänapäeva kirjakeele nimeolemite tuvastaja parim saavutatud f-skoor on 0,87, mis saavutati kirjakeelsetel ajalehetekstidel. See-eest kasutades tänapäeva kirjakeele nimeolemite tuvastajat siinse töö treeningandmetel mõõdeti märksa väiksem tulemus vahemikus 0,565–0,569. Siinses töös loodud nimeolemite tuvastaja mudel saavutas parimaks skooriks tulemuse 0,90. See on võrreldes tänapäeva kirjakeele tulemusega üsna arvestatav saavutus.

Tulemuse edasiseks parandamiseks saab suurendada treeningandmete hulka, sest tulemustest oli näha, et mida rohkem on treeningandmeid, seda suurem on tõenäosus, et mudeli tulemus tuleb kõrge. Samuti on võimalik luua veel suurmaid nimeloendeid, kuhu

lisada juurde vanemaid nimesid. Lisaks eelnevatele võimalustele saaks edasi arendada ka nimeolemite ühestamise eksperimenti ning seda täpsemalt määratleda.

Kirjandus

Alasiry, Areej Mohammed 2015. Named Entity Recognition and Classification in Search Queries; https://bit.ly/2QPKLhL. Vaadatud 06.05.2021.

Anepaio, Toomas 2007. Vallakohus – kas ainult talurahva kohus?. Communal courts – peasant courts only? – Ajalooline Ajakiri. The Estonian Historical Journal 3–4, 343–368.

Bongard, Mikhail Moiseevich 1970. Pattern Recognition. New York: Spartan Books.

Buccholz, Katharina 2021. These are the World's Most Powerful Supercomputers. – World Economic Forum. https://bit.ly/3b3KpdZ. Vaadatud: 06.05.2021.

Dai, Hong-Jie 2019. Family member information extraction via neural sequence labeling models with different tag schemes. – BMC Medical Informatics and Decision Making 10–19, 257.

EE 6 = Eesti Entsüklopeedia 6. köide. 1992. Tallinn: Valgus.

EE 11 = Eesti Entsüklopeedia 11. köide. 2002. Eesti Entsüklopeediakirjastus. https://bit.ly/3w6vec7. Vaadatud: 15.05.2021.

Erelt, Mati 2005. Keeletüpoloogiast. – Oma Keel 1, 5–13. https://bit.ly/3xObz2q. Vaadatud: 25.04.2021

Fišel, Mark 2021. Meetodid – Tehisintellekti Algkursus. https://bit.ly/3xObYSu. Vaadatud: 06.05.2021.

Gupta, Shashank 2018. Named Entity Recognition: Apllications and Use Cases. – Medium. https://bit.ly/33nyK5z. Vaadatud: 06.05.2021.

Güngör jt = Güngör, Onur, Tunga Güngör, Suzan Üsküdarli 2019. The Effect of Morphology in Named Entity Recognition with Sequence Tagging. – Natural Language Engineering 25, 147–169.

Kallas, Jelena, Kristina Koppel 2018. Estonian National Corpus 2017. https://bit.ly/2RtP3ev. Vaadatud: 06.05.2021.

Kettunen, Kimmo, Teemu Ruokalainen 2017. Names, Right or Wrong: Named Entities in an OCRed Historical Finnish Newspaper Collection. – Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage. Göttingen Germany: ACM, 181–186.

Laur jt = Laur, Sven, Siim Orasmaa, Dage Särg, Paul Tammo 2020. EstNLTK 1.6: Remastered Estonian NLP Pipeline. – Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 7152–7160.

Linnus, Jüri 1970. 19. sajandi talurahvakohtute materjalid rahvakultuuri uurimise allikana. – Emakeele Seltsi aastaraamat 16. Tallinn: Teaduste Akadeemia Kirjastus, 231–242.

Maide, Rasmus 2020. Eesti keele nimeolemite märgendaja analüüs ja parandamine. Tartu Ülikool. https://bit.ly/3nU0VCK. Vaadatud: 06.05.2021.

Marrero jt = Marrero, Mónica, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, Juan Gómez-Berbís 2013. Named Entity Recognition: Fallacies, challenges and opportunities. – Computer Standards & Interfaces 35–5, 482–489.

Nothman jt = Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy, James R. Curran 2013. Learning multilingual named entity recognition from Wikipedia. – Artificial Intelligence 194, 151–175.

Padmanabhan, Pyari 2013. Named Entity Recognition using Statistical Model Approach. – International Journal of Computer Applications 73–14, 31–33.

Pilvik jt = Pilvik, Maarja-Liisa, Kadri Muischnek, Gerth Jaanimäe, Liina Lindström, Kersti Lust, Siim Orasmaa, Tõnis Türna 2019. Möistus sai kuulotedu: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine. – Eesti Rakenduslingvistika Ühingu aastaraamat 15, 139–158.

Riaz, Kashif 2010. Rule-Based Named Entity Recognition in Urdu. – Proceedings of the 2010 Named Entities Workshop. Uppsala, Sweden: Association for Computational Linguistics, 126–135.

Prasad, Sandhya, Mahek Laxmikant Kantesaria 2020. Named Entity Recognition in Document Summarization. https://bit.ly/3oike8R. Vaadatud: 15.05.2021.

Talving, Hanno 2012. Vallamajad. Tallinn. https://bit.ly/3tnhCr9. Vaadatud: 06.05.2021.

Tanvir jt = Tanvir, Hasan, Claudia Kittask, Sandra Eiche, Kairit Sirts 2021. EstBERT: A Pretrained Language-Specific BERT for Estonian. https://bit.ly/3xR1eCK. Vaadatud: 06.05.2021.

Tkatšenko, Aleksandr 2010. Nimega üksuste tuvastamine eestikeelsetes tekstides. Tartu Ülikool. https://bit.ly/3nUpjUX. Vaadatud: 06.05.2021.

Tkatšenko jt = Tkatšenko, Aleksandr, Timo Petmanson, Sven Laur 2013. Named Entity Recognition in Estonian. – Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing. Sofia, Bulgaria: Association for Computational Linguistics, 78–83.

Traat, August 1980. Vallakohus Eestis 18. sajandi keskpaigast kuni 1866. aasta reformini. Tallinn: Eesti Raamat.

Wei jt = Wei, Qiang, Yaoyun Zhang, Muhammad Amith, Rebecca Lin, Jenay Lapeyrolerie, Cui Tao, Hua Xu 2020. Recognizing software names in biomedical literature using machine learning. – Health Informatics Journal 26–1, 21–33.

Named Entity Recognition in 19th Century Parish Court Protocols. Summary

The aim of this bachelor's thesis was to train a named entity recognition model based on a corpus consisting of 1500 19th century parish court protocols. Named entity recognition means that the model should recognize people's names, locations, organisations and other miscellaneous names in the protocols. The protocols originate from The National Archives of Estonia. The Python library named EstNLTK (version 1.6) was used for training the model.

The 1500 protocols were divided into 6 subdistributions: 5 for training the model and testing it using cross-validation (by having 4 subdistributions for training and 1 for testing) and the 6th subdistribution for testing the best model, which would later be trained on all 5 initial training subdistributions.

The experiments were conducted based on Aleksandr Tkachenko's, Timo Petmanson's and Sven Laur's article on Named Entity Recognition in Estonian (2013). The experiments mostly consisted of using different sets of features when training the model. The biggest opportunity to achieve a higher score was seen in creating new gazetteers for the trainer (by adding old names instead of ones from literary Estonian). The author of this thesis also conducted experiments based on training data quantity and combining ambiguous named entities.

The highest score this trained model achieved on the corpus was an f-score of 0,90. In comparison, the highest achieved score on the named entity recognition model trained on newspaper articles written in the literary language of Estonian is 0,87.

Further improvements for this model include increasing the data quantity because the experiment on data quantity showed that increasing the training data increases the score of the model. While the experiment on combining ambiguous named entities returned a minimal increase in the score, further experiments on this should also be conducted.

LISAD

Bakalaureusetööga seotud koodirepositoorium on kättesaadaval lingil https://github.com/pxska/bakalaureus.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kristjan Poska,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Nimeolemite tuvastamine 19. sajandi vallakohtu protokollides“, mille juhendaja on Siim Orasmaa, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kristjan Poska

25.05.2021