TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Rasmus Maide

Eesti keele nimeolemite märgendaja analüüs ja parandamine

Bakalaureusetöö (9 EAP)

Juhendaja: Sven Laur, PhD

Tartu 2020

Eesti keele nimeolemite märgendaja analüüs ja parandamine

Lühikokkuvõte: Nimeolemite märgendamine on infoeralduse ülesanne, mis seisneb tekstist pärisnimede leidmises ja kategooriatesse jagamises. Eesti keele kohta on avaldatud üks uurimus, mille tulemusena valmis eesti keele nimeolemite märgendaja ning see on kättesaadav EstNLTK projekti raames. Antud bakalaureusetöö eesmärk on märgendaja viia üle EstNLTK projekti uusimasse versiooni ning analüüsida selle tööd. Analüüsi tulemuste põhjal pakutakse välja reeglipõhiseid parandusi olemasoleva märgendaja kvaliteedi parandamiseks. Edukate paranduste põhjal uuendatakse olemasolevat märgendajat.

Võtmesõnad: nimeolemite märgendamine, loomuliku keele töötlus, statistika, reeglipõhised mudelid, masinõpe

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

Analysis and Improvement of Named Entity Recognition for Estonian

Abstract: Named entity recognition is a task in information extraction that aims to find proper names from text and categorizing them. There is one previous published research on named entity recognition for Estonian and as a result of that research, a named entity recognizer for Estonian was developed which is accessible through the EstNLTK project. The purpose of this thesis is to port the recognizer to the newest version of EstNLTK and analyse its performance. As a result of that analysis, rule-based improvements are proposed for the named entity recognizer. The improvements that have a positive effect on the performance of the named entity recognizer are implemented.

Keywords: named entity recognition, natural language processing, statistics, rule-based models, machine learning

CERCS: P170 Computer science, numerical analysis, systems, control

Sisukord

Sissejuhatus

Nimeolemite märgendamine (*named entity recognition)* on infoeralduse ülesanne loomuliku keele töötluses, mis seisneb toortekstist nimeolemite leidmises ja nende kategooriatesse jaotamises [1].

Nimeolemite märgendamine on tihti esimene samm teiste infoeralduse ülesannete jaoks [1]. Näiteks saab seda kasutada meelestatusanalüüsis, kui on vaja teada saada kliendi meelestatust mingi nimeolemi suhtes [1]. Nimeolemid võivad olla kasulikud ka küsimustele vastamises ja informatsiooni tekstiga sidumisel näiteks Vikipeedias [1].

Nimeolemite märgendamise kohta eesti keeles on avaldatud üks uurimus, mille tulemusena valmis vabalt kättesaadav eesti keele nimeolemite märgendaja [2]. Antud töö eesmärk on see märgendaja muuta kättesaadavaks EstNLTK kõige uuemas versioonis, analüüsida selle kvaliteeti ja pakkuda välja reeglipõhiseid võimalusi selle parandamiseks.

Esimeses peatükis antakse ülevaade nimeolemite märgendamise ülesandest, levinud probleemidest selle juures ja lähenemistest, mille abil on märgendamist parandatud. Samuti sisaldab see ülevaadet olemasoleva eestikeelse märgendaja omadustest.

Töö teises peatükis on kirjeldatud olemasoleva eesti keele nimeolemite märgendaja üle viimist EstNLTK versiooni 1.6. Kuna see versioon on veel arendusjärgus, siis selles versioonis olevat märgendajat on võimalik töös leitud meetoditega parandada.

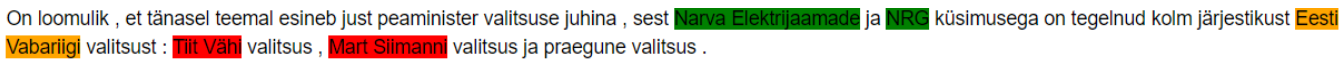
Kolmandas peatükis analüüsitakse praeguse nimeolemite märgendaja kvaliteeti ning leitakse märgendaja tüüpvigasid, mida oleks võimalik reeglipõhiselt parandada.

Neljandas peatükis lisatakse analüüsi põhjal leitud reeglid nimeolemite märgendajasse ning hinnatakse, kas märgendaja kvaliteet nende muutuste järel paranes.

Lisades on viidatud koodirepositooriumile, kus on bakalaureusetöö jaoks tehtud analüüsi kood.

1. Nimeolemite märgendamine

Enim on uuritud pärisnimede märgendamist, mis jaotatakse tihti omakorda kolme kategooriasse: isikunimed, asukohtade nimed ja organisatsioonide nimed [3]. Antud uurimistöö eesmärgiks on olemasoleva eesti keele nimeolemite märgendaja parandamine. See tuvastabki vaid pärisnimesid, seega antud uurimistöö tegeleb just sellega. Joonisel 1 on näide kõigi kolme kategooria kohta nimeolemite märgendamises.



Joonis 1. Nimeolemite tuvastaja töö näide: organisatsioonid on märgendataud rohelise taustaga, asukohad oranži taustaga ja inimesed punase taustaga.

Pärisnimede märgendamise peamised probleemid on nimeolemi piiride määramine ja olemitevaheline mitmesus [1].

Nimeolemi piiride määramisel ei piisa vahel lauses olevast infost ning otsuse tegemisel tuleb kasutada taustinfot [4]. Näiteks lausetes „Pidu toimus kusagil Eesti linnas.“ ja „Pidu toimus kusagil Tartu linnas.“ on esimeses lauses nimeolem vaid „Eesti“, teises lauses aga „Tartu linnas“ [4]. Kuna lausete erinevus seisneb vaid selles, millise pärisnimega on tegu, peab märgendaja õigete piiride märgendamiseks teadma, et „Eesti“ viitab riigile ja „Tartu“ viitab linnale [4].

Olemitevaheline mitmesus tuleneb sellest, et üks nimeolem võib kuuluda sõltuvalt kontekstist erinevatesse kategooriatesse [1]. Näiteks sõne „Washington“ võib olla nimeolem, mis viitab nii isikule, organisatsioonile kui asukohale [4]. Tabelis 1 on toodud näiteid kõigist kolmest variandist. Õige kategooria määramiseks ei piisa vaid nimeolemi leidmisest, märgendaja peab ka arvestama ümbritsevat konteksti. Tabelis ja töös edaspidi kasutatakse isikuid tähistava märgendi kohta lühendit „PER“, organisatsiooni tähistava märgendi kohta lühendit „ORG“ ja asukohta tähistava märgendi kohta lühendit „LOC“.

Tabel 1. Sõne „Washington“ erinevad võimalikud kategooriad

Kategooria	**Lause**
PER	*Washington* oli Ameerika Ühendriikide esimene president
ORG	Eileõhtuse mängu võitis *Washington* 121-109.
LOC	*Washington* on osariik Ameerika Ühendriikide loodeosas.

Nimeolemeid, millel on erinevaid võimalikke kategooriad nimetatakse mitmesteks ja ainult ühe võimaliku kategooriaga nimetatakse ühesteks [4]. Tkachenko jt. avaldamata uurimuse kohaslt on eesti keele koondkorpuses automaatse märgendamise järel 64% esinevatest nimeolemitest ühesed ja 36% mitmesed, kusjuures üheseid on üle 800 tuhande unikaalse, samas kui mitmeseid on üle 120 tuhande unikaalse [4]. Tabelis 2 on toodud üheste ja mitmeste nimeolemite täpsem jaotus Tkachenko jt. uurimuse järgi.

Tabel 2. Olemite sagedused eesti keele nimeolemite korpuses [4]. Arvud kolme tüvenumbri täpsusega.

Kategooria	Ülempiir		Ühesed nimeolemid			Mitmesed nimeolemid		
	Sagedus	Arvukus	Sagedus	Arvukus	Unikaalne	Sagedus	Arvukus	Unikaalne
Väga sage	$1*10^{-0}$	-	13.1%	1.42 M	56	3.0%	324 K	14
Sage	$1*10^{-3}$	13.5 K	35.6%	3.89 M	18.8 K	27.0%	2.90 M	16.4 K
Haruldane	$2.6*10^{-6}$	28	7.3%	793 K	89.8 K	5.5%	602 K	63.8 K
Väga haruldane	$3*10^{-7}$	5	8.0%	876 K	696 K	0.9%	103 K	42.8 K
Kokku			64.0%	6.98 M	805 K	36.0%	3.93 M	123 K

1.1 Järjestikuline märgendamine

Tavapärane nimeolemite märgendamise lähenemine on sõnahaaval järjestikuline märgendamine, kus loodud märgendid näitavad nii nimeolemi piire kui kategooriat [1]. Kuna nimeolemid võivad koosneda mitmest sõnast, aga märgenduse saab iga sõna eraldi, kasutatakse IOB märgendamist, et nimeolemite piire määrata [1]. IOB märgendamisel on võimalike märgendite arv 2n+1 kui kategooriaid on n. Igale kategooriale vastab kaks märgendit: B ehk *begin* – algus ja I ehk *inside* – sisemine [1]. Nendele sõnadele, mis ei kuulu ühegi nimeolemi alla, antakse märgend O ehk *outside* [1]. Ühe lause peal võib mudeli väljund välja näha näiteks selline: B-PER Oskar I-PER Luts O sündis B-LOC Järvepera I-LOC külas B-LOC Palamuse I-LOC kihelkonnas. Lõppkasutajatele kuvatavas väljundis teisendatakse märgendus olemitasandile ehk see näeks välja selline: [PER Oskar Luts] sündis [LOC Järvepera külas] [LOC Palamuse kihelkonnas]. Kui sõnatasandil ei oleks B ja I märgendusi, poleks võimalik teada kust läheb Järvepera küla ja Palamuse kihelkonna nimeolemite piir. Järjestikuline märgendamine tähendab, et nimeolemid ei saa lõikuda [1]. Näiteks nimeolem „Hugo Treffneri gümnaasium“ on organisatsioon ja „Hugo Treffneri“ selle sees ei saa eraldi isikunimeks märgendada.

1.2 Õppemeetodid

1.2.1 Reeglipõhine lähenemine

Nadeau ja Sekine ülevaate põhjal käsitleti nimeolemite tuvastamist eraldiseisva ülesandena esimest korda 1991. aastal Lisa F. Rau uurimuses, mille eesmärk oli eraldada ja tuvastada (firmade) nimesid. Tema metoodika oli reeglipõhine ja heuristiline [3]. Reeglipõhise lähenemise eelis masinõppe ees on, et puudub vajadus suure treeningkorpuse järele, mille loomine on ajakulukas ning seetõttu kasutatakse seda tänapäevalgi [3]. Chiticariu jt. 2013. aasta uurimuse järgi on suurem osa kommertslahendusi nimeolemite tuvastamise jaoks reeglipõhised, kuigi akadeemilised uurimused keskenduvad masinõppele [5]. Hetkel on levinud hübriidlahendused, mis kasutavad reegleid koos masinõppega, näiteks käiakse tekst läbi mitu korda reeglipõhise lähenemise järgi ning märgendatakse ära kõik ühesed nimeolemid ning hiljem lahendatakse masinõppe abil mitmeste nimeolemite probleem [1].

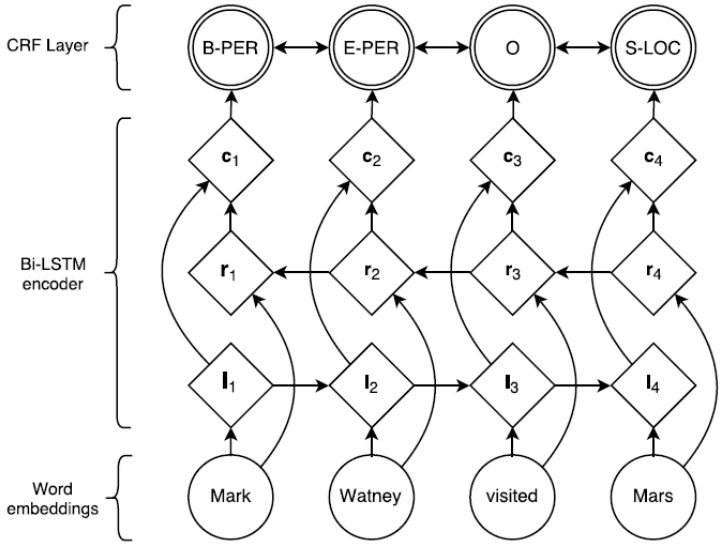
1.2.2 Masinõppepõhine lähenemine

Kui esialgu kasutati nimeolemite märgendamiseks reeglipõhiseid meetodeid, siis tänapäeval kasutatakse enam masinõpet [3][6]. Masinõppe kasutamine nõuab eelnevalt märgendatud treeningkorpust [3]. Nadeau ja Sekine artiklis jagatakse masinõppega nimeolemite märgendamine kolmeks, millest levinuim lähenemine on juhendatud masinõpe. Juhendatud masinõppe õppimisalgoritmidena kasutatakse näiteks peidetud Markovi mudeleid (*hidden Markov models)*, otsustuspuid (*decision trees)*, maksimaalse entroopia mudeleid (*maximal entropy models)*, tugivektor-masinaid (*support vector machines)* ja tingimuslikke juhuslikke välju *(conditional random fields* e. CRF*)*. Vähese juhendamisega masinõpe kasutab väikest hulka näiteid, mille konteksti uurides luuakse reeglid, mille järgi ülejäänud tekstist näiteid leida ja seeläbi nimeolemite leksikoni suurendada. Mõnes uurimuses on see lähenemine saavutanud juhendatud masinõppega samaväärseid tulemusi. Juhendamata masinõppe populaarseim lähenemine on klasterdamine, ent see ei anna iseseisvalt sama häid tulemusi kui juhendamata masinõpe [3].

Derczynski jt. analüüsisid 2015. aastal erinevaid populaarseid lähenemisi nimeolemite tuvastamise lahendamiseks, võttes testhulgaks kolm korpust Twitteri säutsudest. Parimaid tulemusi saavutasid NERD-ML tuvastaja, mis kasutab tugivektor-masinaid, K-lähima naabri algoritmi ja naiivse Bayesi algoritmi ning Stanfordi nimeolemite tuvastaja, mis kasutab tingimuslikke juhuslikke välju. Reeglipõhised tuvastajad saavutasid masinõppest selgelt nõrgemaid tulemusi [6].

1.2.3 Tehisnärvivõrkudel põhinev lähenemine

Alates 2011. a Collobert jt. tööst on nimeolemite tuvastamisel kasutatud tihti tehisnärvivõrke [1][7]. Praeguseks on populaarseim lähenemine tehisnärvivõrkude osas kahesuunaline pika lühiajalise mälu kiht (*bidirectional long short-term memory* e. bi-LSTM) [1][8][9]. Nagu eelnevalt mainitud, siis nimeolemite märgendamises sõnade märgendused järgivad reegleid, näiteks I-PER märgend ei saa järgneda märgendile O, mille saab mudelile ette anda graafilise mudeli, näiteks CRF abil. Tehisnärvivõrkudele pole võimalik seda otse ette anda, nii et levinud lähenemine on bi-LSTM-kihi järel kasutada CRF-kihti, et tagada reeglitest kinni pidamist [1]. Joonisel 2 on visualiseeritud sellise tehisnärvivõrgu ehitust.



Joonis 2. Diagramm Bi-LSTM kihist koos CRF-kihiga [9]. Bi-LSTM kiht töötleb sisendit mõlemas (l ja r kihid) ning selle väljund on kiht c, mis on sisendiks CRF-kihile, kus kontrollitakse märgendite reeglitele vastavust.

Sellised tehisnärvivõrkudel põhinevad mudelid on saavutanud paremaid tulemusi kui kõik ülejäänud lähenemised [9].

1.3 Nimeolemite märgendaja hindamine

Nimeolemite märgendaja hindamiseks kasutatakse üldiselt F1-skoori, mis on täpsuse[1] ja saagise harmooniline keskmine [1]. Hindamist sooritatakse olemite tasandil, näiteks lauses "Tartu linnas elab Mati Murakas" on kolm olemit: asukoht "Tartu linnas", mittenimeolem "elab" ja isik "Mati Murakas" [1]. Täpsus on õigete nimeolemite osakaal kõikide leitud nimeolemite seas, saagis on aga õigesti leitud nimeolemite osakaal kõikide tekstis olevate nimeolemite seas [10]. Üks põhjuseid, miks õigsuse[1] kasutamine on vähem populaarsem kui F1-skoor, on see, et F1-skoori puhul ei kasutata tõsinegatiivseid tulemusi, mille loendamine on nimeolemite tuvastamisel keeruline ning nende hulk on enamasti selgelt suurem kui tõsipositiivsete tulemuste puhul [11]. F1-skoori kasutamine tähendab, et tekstist ei ole vaja eraldada kõiki olemeid, mis on korretkselt märgendatud mittenimeolemiteks [11].

[1] Inglisekeelseid termineid *precision* ja *accuracy* tõlgitakse tihti mõlemat kui "täpsus". Segaduse vältimiseks on kasutatud *accuracy* tõlkena "õigsus" ning "täpsus" on kasutatud tähenduses *precision.*

Nimeolemite piiride leidmise ülesanne põhjustab aga probleeme ka F1-skoori abil hindamisel [11]. Kui märgendaja märgendab nimeolemi piirid korrektsest märgendusest erinevalt, siis F1-skoori silmis teeb märgendaja nii valepositiivse kui valenegatiivse vea, samas kui üldse mitte märgendades tehtaks vaid valenegatiivne viga [11]. Näiteks olgu õige nimeolem "Tallinna keskkonnaameti keskkonnahoiu osakond" märgendiga "ORG". Kui mudel ennustab, et "ORG" märgendi peaks saama "Tallinna keskkonnaameti", siis on see teinud kaks viga: valepositiivse vea, sest õigete nimeolemite hulgas ei olnud "Tallina keskkonnaameti" ja valenegatiivse vea, sest see ei lugenud nimeolemiks kogu korrektset olemit. Mudel, mis ei leiaks sellest tekstist ühtegi nimeolemit, teeks ainult valenegatiivse vea. Veel üks problem selle lähenemise juures on, et kui nimeolemi piirid on leitud korrektselt, aga mudel valib nimeolemile vale kategooria, on see samamoodi sama suur viga kui täiesti valesti leitud nimeolem [11]. Hoolimata nendest probleemidest on seda hindamismeetodit kasutatud nii *Conference on Computational Natural Language Learning (CoNLL)* nimeolemite tuvastamise võistlusel aastal 2003 kui ka antud töös aluseks võetud eesti keele nimeolemite märgendaja kvaliteedi hindamisel [2][10]. Esuli ja Sebastiani pakuvad oma töös lahendusena välja sõne (*token)* tasemel mudeli hindamist. Nende sõnul ei lahenda see vale kategooria määramise probleemi ning teeb selle tegelikult vaid hullemaks (kolmesõnaline LOC, mis on valesti märgendatud kui kolmesõnaline ORG, põhjustaks nüüd nii kolm valepositiivset kui ka kolm valenegatiivset viga), ent see probleem pole nii oluline. Lisaks sellele soovitavad nad F1-skoori makrokeskmistamist (*macroaveraging)*, mis tähendab esmalt kategooriaspetsiifilise F1-skoori arvutamist ja seejärel nende keskmise leidmist. Eelmainitud CoNLL 2003 võistlusel tõusis 10 tiimi konkurentsis selle hindamismeetodi järgi üks tiim üheksandalt kohalt teiseks ja üks kukkus kolmandalt viimaseks kümnendaks, mis näitab, kui suur olulisus on hindamismeetodi valikul [11].

1.4 Nimeolemite märgendamine eesti keeles

Hetkel ainus vabalt kättesaadav eesti keele nimeolemite märgendaja on kirjeldatud Tkachenko jt. 2013. aasta artiklis. Artikkel kirjeldab nimeolemite märgendusi raskusi eesti keeles, mis tulenevad eesti keele morfoloogilisest rikkusest ja vabast sõnajärjest. Loodud nimeolemite märgendaja kasutab juhendatud masinõpet, mis treeniti käsitsi märgendatud eestikeelsete uudiste korpusel. Ristvalideerimisel saavutas see märgendaja F1-skooriks 87%, mis on võrreldav tulemustega teistel sarnastel keeltel [2].

Kuna juhendatud masinõppe kasutamine nõuab suure märgendatud korpuse olemasolu, aga enne Tkachenko jt. artiklit polnud eesti keeles midagi sellist olemas, loodi see selleks tarbeks [2]. Korpus on loodud 572 uudisest, mis avaldati Postimehe ja Delfi portaalides vahemikus 1997-2009 [2]. See sisaldab 184638 sõna, millest 15411 on märgendatud kui nimeolemid ja jaotuvad kolme kategooriasse: isikunimed (5762 nimeolemit, 3588 unikaalset), asukoha nimed (5711 nimeolemit, 1589 unikaalset) ja organisatsioonide nimed (3938 nimeolemit, 1987 unikaalset) [2]. Sama korpust kasutab käsitsi märgendatud korpusena ka käesolev uurimus.

Masinõppe õppimisalgoritmina kasutasid Tkachenko jt. tingimuslikke juhuslikke välju, sest see meetod on nimeolemite tuvastamise ülesandes edukas tänu oma võimele kasutada suurt hulka tunnuseid. Tunnused jaotusid viite kategooriasse: põhitunnused, kuhu kuuluvad näiteks sõna väiketähestatud kuju, sõna algus ja lõpp ning kas sõna on lauses esimene; morfoloogilised tunnused, kuhu kuuluvad näiteks sõna lemma, kääne, lõpp, sõnaliik ja morfeemid, millest sõna koosneb; sõnastikupõhised tunnused, mis näitavad kas sõna või osa sõnast kuulub sõnastikku, mis koostati erinevatest eestikeelsetest avalikest ressursidest ning inglisekeelse nimeolemite märgendaja jaoks loodud sõnastikust; WordNeti tunnused, kus kasutatakse sõna semantilist suhet teiste sõnadega ja globaalsed tunnused, mis näitavad, kas eelnev või järgnev sõna on pärisnimi või kas see sõna esineb mujal tekstis suure algustähega lause keskel [2].

2. Nimeolemite märgendaja ettevalmistus

2.1 Nimeolemite märgendaja üle viimine EstNLTK uude versiooni

Olemasolev märgendaja on vabalt kättesaadav EstNLTK projekti versioonis 1.4. Hetkel arendatakse projekti versiooni 1.6 ning selle töö esimene eesmärk on märgendaja sellesse versiooni üle viia.

EstNLTK versioonid 1.4 ja 1.6 on mõlemad kirjutatud programmeermiskeeles Python. EstNLTK kasutab alusena Text klassi, mis on andmestruktuur teksti hoidmiseks. Versioonis 1.4 on see Pythoni sõnastiku (*dict)* alamklass, kuid versioonis 1.6 enam mitte.

Versioonis 1.4 põhineb nimeolemite märgendaja samuti Pythoni sõnastikel, mille abil on kirjeldatud kõik nimeolemite tunnused ning ka lõplik märgendamise tulemus. Nimeolemite märgendaja kood versioonis 1.4 on saadaval siin: https://github.com/estnltk/estnltk/tree/version_1.4/estnltk/estner

Töö tulemusena valminud nimeolemite märgendaja versioonis 1.6 kasutab sõnastike asemel nimeolemite kuvamiseks Layer klassi, mille abil saab kujutada teksti erinevaid atribuute. Kogu ülejäänud loogika nimeolemite märgendamisel jäi samaks kui versioonis 1.4. Testimine näitas, et kõik märgendatud nimeolemid ei ole identsed versioonides 1.4 ja 1.6, kuigi tunnused luuakse samamoodi. Selle põhjuseks on tõenäoliselt see, et versioon 1.6 kasutab uut morfoloogilist analüsaatorit, mille abil leitakse nimeolemite määramisel kasutatavaid tunnuseid, näiteks sõna lemma. Erinevused märgendajate vahel on siiski väga väikesed. Versiooni 1.6 märgendaja on hetkel kättesaadav aadressil https://github.com/estnltk/estnltk/tree/devel_1.6/estnltk/taggers/estner, selle kasutusjuhend on aadressil https://github.com/estnltk/estnltk/blob/devel_1.6/tutorials/taggers/ner_tagger.ipynb .

3. Nimeolemite märgendajate analüüs

Nimeolemite märgendaja analüüsi eesmärgiks on teada saada, mis tüüpi vigu olemasolev märgendaja teeb, et neid oleks võimalik parandada. Antud töö raames uut piisavalt suurt korpust pole võimalik käsitsi märgendada, seega tuleb kasutada olemasolevat korpust [12]. Kuna seda korpust kasutati ka olemasoleva märgendaja treenimiseks, on märgendaja täpsus selle korpuse peal ilmselt kõrgem kui tundmatu teksti peal. Sellegipoolest on võimalik märgendaja tehtavate vigade järgi aru saada, mis põhjustab märgendamisel suurimaid probleeme. Korpuses on iga sõna viidud vastavusse ühe IOB-märgendiga. Korpuses on 195957 märgendit O, 5762 korda märgendit B-PER, 5711 korda B-LOC, 3936 korda B-ORG, 2728 korda I-PER, 2207 korda I-ORG ja 600 korda I-LOC.

Manuaalseks analüüsiks kasutas autor 10000 esimest nimeolemit korpusest. Automaatne märgendaja märgendas teksti ning selle tulemust võrreldi etteantud korpuse märgendustega. Vigade loendamisel arvestati, et üks antud märgendaja omadusi on kaasata nimeolemisse ka nimeolemile järgnev punkt, mis tegelikult ei peaks olema nimeolemi osa. Tegemist ei ole sisulise veaga ning kui eeldada, et punkt ei peaks olema kunagi nimeolemi osa, on seda ka lihtne parandada, seega seda ei loetud veaks. Vigu loeti käsitsi, sest ka käsitsi märgendatud korpus sisaldab vigu, aga neid ei tohiks lugeda automaatse märgendaja vigadeks. Seega selliseid kohti ignoreeriti. Selliste vigade alla kuuluvad suuremad vead, näiteks Gruusia presidendi Saakašvili asukohaks märgendamine ning ka väiksemad vead, näiteks sõnele järgneva tühiku nimeolemi osaks lugemine. Selliseid vigu oli 10000 nimeolemi seas umbes 30. Kuna antud märgendaja on treenitud just selle korpuse peal, siis on korpuses kindlasti veel vigu, mida märgendaja on õppinud valesti märgendama ja mis ei tulnud seetõttu vigadena välja. Käsitsi kogu korpuse üle käimine ja vigade parandamine oleks liiga mahukas ning pole töö eesmärk.

Märgendaja tegi 13 nimeolemi piiri viga ehk leidis liiga lühikese või pika nimeolemi, 14 korral jagas ühe nimeolemi mitmeks osaks või pani mitu nimeolemit üheks kokku, 59 vale kategooria viga, 16 korda ei leidnud nimeolemit, 16 korral märgendas nimeolemi kohas, kus seda pole ja 15 korral oli viga segu piiri ja kategooria veast, näiteks nimeolemis Toomas Hendrik Ilves oli märgendatud Toomas Hendrik isikuks ja Ilves asukohaks. Kokku tegi märgendaja seega 10000 nimeolemi peale 133 viga. Nagu eelnevalt öeldud, ei tähenda see, et märgendaja täpsus oleks üle 98%, sest tegemist on selle märgendaja treeninghulgaga ning leidmata jäid vead, kus korpus ja märgendaja korraga eksisid.

Osa vigadest olid korduvad ja kõige sagedasem viga oli Euroopa Liidu kategooria viga, mida esines 27 korda ehk peaaegu pooled kategooria vead. 19 korral oli õigeks märgendiks LOC, aga märgendaja valis ORG ning 8 korral oli vastupidi. Käsitsi märgendatud korpuses ei ole Euroopa Liit süstemaatiliselt märgendatud ORG- või LOC-märgendiga. Selle asemel on mõnes lõigus see ORG ja mõnes LOC, kuid kogu korpuse peale järjepidevus puudub. See selgitab, miks märgendaja valib vahel Euroopa Liidu märgendiks ORG ja vahel LOC. Isegi Euroopa Liitu arvestamata oli aga kõige sagedasem viga kategooria valiku viga, nii et edasise töö käigus keskendutakse sellele.

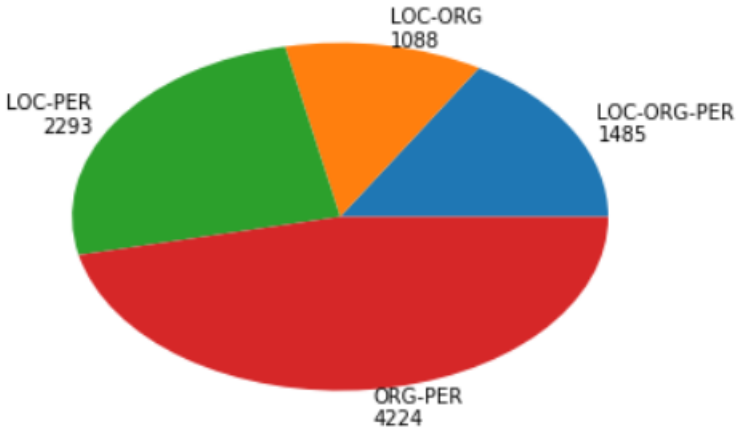
3.1 Nimeolemite mitmesus

Esmasest analüüsist oli näha, et üks peamistest probleemidest nimeolemite märgendaja jaoks on nimeolemite mitmesus. Selleks, et analüüsida, kuidas märgendaja valib nimeolemi kategooriat, märgendas autor Eesti keele koondkorpuse [13] tekste.

Valitud tekstideks osutusid Eesti Päevalehe tekstid aastast 1996 ja Eesti Ekspressi tekstid aastatest 1996-2001. Märgendaja leidis nendest tekstidest kokku 770 048 nimeolemit. Peale nende nimeolemite lemmatiseerimist oli unikaalseid nimeolemeid 152 692 ja nendest 52436 esinesid andmestikus enam kui korra ning seetõttu on võimalik öelda midagi nende mitmesuse kohta. 9090 nimeolemit ehk 17.3% olid sellised, mis olid märgendaja jaoks mitmesed ehk neid esines erinevates kategooriates.

Töö esimeses peatükis viidatud uurimuses koondkorpuses märgendatud tekstides olid mitmesed 36.0% leitud nimeolemitest. Kuna puuduvad täpsemad andmed eelneva uurimuse tulemuste kohta, ei saa kindlalt öelda, millest see erinevus tekkis, kuid on võimalik, et antud uurimistöö käigus märgendatud tekste oli vähem, mistõttu nendes ei esinenud teatud nimeolemeid rohkem kui korra ja seetõttu need ei saanudki jääda mitmesteks. Selle efekti võimendab ka asjaolu, et mida haruldasem on nimeolem, seda suurema tõenäosusega ei esine see nimeolemite märgendaja sõnastikus, mis teeb kategooria määramise keerukamaks. Teine võimalus seda erinevust põhjendada on, et kui korpuses on rohkem sõnu, siis on rohkem võimalusi märgendajal vähemalt korra kategooria valimisel eksida ning see tähendab, et nimeolem läheb ekslikult mitmeste nimeolemite alla. Näiteks juba antud uurimistöö mitmeste nimeolemite hulgas on “Eesti”, mis on saanud LOC-märgendi 44,244 korda ja ORG-märgendi 6 korda. Kuna märgendite sagedusel on enam kui seitsme tuhande kordne vahe,

siis tegelikkuses ei peaks tõenäoliselt tegemist olema mitmese nimeolemiga, kuid nii läheb see statistikasse kirja.



Joonis 3. Diagramm mitmesuse tüüpidest

Analüüsides tulemusi lähemalt on näha, et 46.47% mitmestest nimeolemitest on mitmesed ORG ja PER märgendite vahel. Teine kõige populaarsem mitmesus on 25.23% mitmestest nimeolemitest tekitav LOC-PER mitmesus.

Võrduluseks käsitsi märgendatud nimeolemite korpuses on mitmeseid nimeolemeid vaid 46, kuid enam kui korra esinevaid nimeolemeid on 1827 ehk vaid 2.52% nimeolemitest on mitmesed. Ka neid 46 lähemalt vaadates paistab, et seal sisaldub märgendamisvigu, näiteks on “Saakašvili” saanud 8 korral PER-märgendi, aga ühel korral LOC-märgendi, mis tõenäoliselt on ekslik. Nimeolemid, mis regulaarselt jäid erinevatesse kategooriatesse, olid reeglina kohanimed nagu „Tallinn“, „Pärnu“ või „Euroopa Liit“, mis sõltuvalt kontekstist võisid saada kas LOC-märgendi või ORG-märgendi. Käsitsi märgendatud korpuse mitmestest nimeolemitest on 56.52% LOC-ORG mitmesed, samas kui koondkorpuse märgendamisel on LOC-ORG mitmesust aga vähem kui ühtegi teist tüüpi mitmesust. Seega saab oletada, et suur osa koondkorpuse märgendamisel välja tulnud mitmestest nimeolemitest viitavad vigadele märgendaja töös ning kõige raskem on märgendaja jaoks just ORG- ja PER- märgendite eristamine.

3.2 Sõnastiku kasutus

Kategooria määramisel võib üks olulisemaid osasid nimeolemite märgendaja juures olla sõnastik, kus on kirjas, mis kategooria millisele nimeolemile peaks kuuluma. 9090 nimeolemist, mis jäid mitmeseks, olid aga 1750 (19.25%) eelnevalt sõnastikus olemas.

Kokku oli erinevaid nimeolemeid 152692, aga nendest valid 52436 esinesid andmestikus enam kui korra ning seetõttu on võimalik öelda midagi nende mitmesuse kohta. Neist omakorda 10831 (20.66%) nimeolemit esinesid samal kujul ka sõnastikus. Seega sõnastikul on küll statistiliselt oluline ($p<0.002$) roll kategooriate määramises, kuid vahe ei ole nii suur, et võiks eeldada, et märgendaja valib kategooria peamiselt sõnastiku abiga. Sõnastiku abil saab selgitada aga mitmeid vale kategooria vigu treeningkorpuses, näiteks « Austria » oli kolmel korral märgendatud LOC asemel PER märgendiga. Kuigi treeningandmetes oli Austria alati LOC, siis sõnastikus puudus Austria eraldi kui LOC, kuid Austria esines mitmes isikunimes, näiteks „archduke franz karl of austria“ ja „isabella of austria“. Inglisekeelsed nimeolemid on sõnastikus, kuna sõnastiku loomisel võeti aluseks inglisekeelse nimeolemite märgendaja jaoks loodud sõnastik.

4. Nimeolemite märgendaja parandamine

Reeglipõhise parandamise eesmärk on analüüsi käigus leitud märgendaja süstemaatilisi vigu lahendada reeglitega, mida mudelile lisada. Antud töös proovitakse parandamiseks kahte erinevat varianti.

Esimese lahendusena leitakse sõnu, mis eelnevad või järgnevad tihti ühe konkreetse kategooria nimeolemitele. Nende sõnade kohta luuakse reegel, et kui nimeolem vastavalt kas eelneb või järgneb sellele sõnale, siis see on alati just sellest konkreetsest kategooriast. Näiteks kui sõnale „osakond“ eelnev nimeolem on enamasti organisatsioon, siis kõik sellele sõnale eelnevad nimeolemid märgendatakse organisatsioonideks.

Teine lahendus on leida nimeolemeid, mis on korpuses jäänud mitmeseks, kuid tegelikult peaksid olema ühesed. Juba varem toodi näide, kus sõne „Eesti“ sai enam kui seitse tuhat korda suurema tõenäosusega märgendi „LOC“ kui märgendi „ORG“. Kui eeldada, et see nimeolem on ühene, tuleks kõik „ORG“-märgendid muuta „LOC“-märgenditeks.

4.1 Eelnevatel ja järgnevatel sõnadel põhinevad reeglid

Selleks, et leida, millised nimeolemile eelnevad või järgnevad sõnad nimeolemi kategooriat näitavad, kasutatakse koondkorpuse märgendatud tekste ning vaadatakse, millised sõned eelnevad ja järgnevad kõige enam mingist kategooriast nimeolemitele. Hetkel kasutatakse vaid Eesti Ekspressi märgendatud tekste, kus on 457114 nimeolemit. Sõnesid, mis eelnevad teatud kategooriast nimeolemile enam kui 100 korda ja eelnevad just sellele kategooriale enam kui 90% kordadest, on kokku 58. Neist 53 eelnevad enim PER-märgendile ja on enamasti ametid („kunstnik“, „ärimees“) või tiitlid („dr“,“hr“). 4 eelnevad enim ORG-märgendile, nendeks on „ajakiri“, „ajaleht“, „ajakirja“, „ajalehe“ ning üks, „väljaspool“, eelneb enim LOC-märgendile.

Paranduse idee on leida kõik kohad, kus leitud sõned eelnevad nimeolemile, mis märgendaja arvates ei ole sellest kategooriast, mis enim antud sõnele järgneb, näiteks kui sõnele „kunstnik“ järgneb nimeolem kategooriast „ORG“, ja muuta see kategooriaks, mis sellele sõnele tavaliselt järgneb. Kõik muudatused vaadatakse käsitsi üle ning hinnatakse, kumb kategooria on õigem. Kui muudatuste tulemusel märgendaja täpsus paraneb, on muudatus kasulik ning parandab märgendaja kvaliteeti.

Selle paranduse järel muutus 492 nimeolemi kategooria, millest 458 said PER-märgendi. Käsitsi neist 157 üle vaatamisel selgus, et 59 korral oli muutus korrektne, 79 korral muudeti õige kategooria valeks ning 33 puhul ei läinud märgendaja täpsus ei paremaks ega halvemaks. Nende 33 hulka kuuluvad nimeolemid, mis ei kuulu kumbagi kategooriasse, mis võivad kuuluda mõlemasse kategooriasse, on valede piiridega või pole üldse nimeolemid. Kategooriate lõikes toimunud muutused on esitatud tabelis 3.

Tabel 3. Muutused märgenduse õigsuses kategooriate kaupa

	PER	ORG	LOC
Vale -> õige	46	10	3
Õige -> vale	63	14	2
Jäi valeks	14	0	5

Sama metoodikaga leiti kõik sõned, mis järgnevad teatud nimeolemile kindlast kategooriast enam kui 100 korda ning see moodustab enam kui 90% kõikidest nimeolemitest, millele see sõne järgneb. Selliseid sõnesid on 38, neist 18 järgnevad enamasti PER märgendile, 15 LOC märgendile ja 6 ORG märgendile.

Selle paranduse abil muudeti kokku 569 nimeolemi kategooriat. Käsitsi neist 150 üle vaatamisel oli 39 nimeolemit, mille kategooria muutus õigeks ning 91, mille kategooria muutus valeks. 20 nimeolemi puhul olid piirid valed või pole kumbki kategooria õige. Seega see lähenemine on veel halvem kui nimeolemile eelneva sõna kasutamine.

Nii enne kui ka peale nimeolemit olevate sõnade järgi nimeolemi kategooria määramisel tõid kaasa rohkem vigu kui parandasid valesid märgendusi. Seega võib järeldada, et see meetod nimeolemite märgendaja parandamiseks ei sobi.

4.2 Mitmeste nimeolemite üheseks lugemine

Teine reeglipõhine lähenemine märgenduse parandamiseks on leida nimeolemeid, mis tegelikkuses ei ole mitmesed, kuid on saanud erinevates kohtades erinevate kategooriate märgendeid. Sarnaselt eelmisele lähenemisele leitakse kõik nimeolemid, mis esinevad korpuses vähemalt 100 korda ning mis on vähemalt 90% kordadest saanud ühe kategooria märgendi,

kuid vähemalt ühel korral mõne teise kategooria märgendi. Selliseid nimeolemeid on 267, kusjuures selliseid, mis said mõne teise kategooria enam kui 10% kordadest on 53.

267 leitud nimeolemi hulka kuulub 99 PER, 112 LOC ja 56 ORG olemit. Eesti Ekspressi tekstidest neid nimeolemeid otsides leidub 10135 korda neid sõnesid nii, et nende märgendus ei klapi tavapärase märgendusega. Sealjuures 2204 korral on need sõned saanud mõne teise märgenduse, aga 7931 pole need saanud ühtegi märgendust. Selle lähenemise abil on potentsiaalselt võimalik lisaks kategooria valikule parandada ka tuvastada muidu leidmata jäänud nimeolemeid. Kokku oli Eesti Ekspressi tekstidest leitud 457114 nimeolemit, seega kui kõik parandused on õiged, paraneks märgendaja täpsus enam kui 2%.

10135 potentsiaalsest parandusest 250 käsitsi üle vaatamisel selgus, et 127 parandust olid korrektsed, 76 parandust muutsid õige märgenduse valeks ning 47 puhul olid nii algne kui pakutud märgendus valed. Valitud 250 nimeolemit ei olnud juhuslikud, rohkem valiti neid, mille algne märgend ei olnud „0“, et saaks seda tabelit võrrelda allpool tabelitega 5 ja 6. Kui välja arvata nimeolemid, mis eelnevalt olid saanud „0“ märgenduse ehk ei olnud märgendaja poolt leitud, siis 66 parandust muutsid vale variandi õigeks ning 41 parandust õige variandi valeks. Kategooriate lõikes on selle analüüsi tulemused tabelites 4 ja 5.

Tabel 4. Tabel õigeks muudetud märgenduste kategooriate kohta. Veerud tähistavad automaatset märgendust, read tähistavad parandatud märgendust.

Uus/eelnev	0	ORG	PER	LOC
ORG	38		2	11
PER	8	17		8
LOC	15	8	20	

Tabel 5. Tabel valeks muudetud märgenduste kategooriate kohta. Veerud tähistavad automaatset märgendust, read tähistavad parandatud märgendust.

Uus/eelnev	0	ORG	PER	LOC
ORG	13		12	6
PER	0	8		0
LOC	23	7	8	

Tekkinud vigadest suure osa põhjustasid mõned sõned, mis tegelikult ei peaks olema üheste nimeolemite all, näiteks „Nõukogu“, mis on enam kui 90% korral LOC Nõukogude Liidu tõttu ja „I“, mis oli enam kui 90% kordadest ORG. Kuigi paistab, et automaatselt leitud üheste nimeolemite kogu parandab nimeolemite märgendaja tööd, võib parem lähenemine olla ühesed nimeolemid leida käsitsi märgendatud korpusest.

Selle teooria testimiseks võeti käsitsi märgendatud korpusest välja kõik nimeolemid, mis esinesid vähemalt 10 korda ja olid alati märgendatud sama kategooria alla. Selleks, et oleks võimalik leida ka uusi nimeolemeid, ei tohtinud ka need nimeolemid esineda kordagi korpuses märgendiga O. Vastasel juhul oleks näiteks PER kategooriasse läinud sõne „ligi“, mis ei esine küll üheski teises kategoorias, aga kui kõik sõne „ligi“ esinemised märgendada PER-märgendusega, siis tekiks valepositiivseid tulemusi ilmselt liiga palju. Niimoodi leiti 68 LOC-märgendiga, 35 ORG-märgendiga ja 46 PER-märgendiga nimeolemit. Ekspressi tekstides on 654 korral need nimeolemid märgendatud mõne muu kategooria nimeolemiks. Neist 100 nimeolemi käsitsi üle vaatamisel muutus 61 nimeolemi kategooria õigeks, 28 kategooria valeks ning 11 kategooria jäi valeks. Õigeks ja valeks muutumisi kategooriate kaupa näitavad tabelid 6 ja 7.

Tabel 6. Õigeks muudetud nimeolemite kategooriad. Veerud tähistavad eelnevaid kategooriaid, read tähistavad uusi kategooriaid.

Eelnev/Uus	ORG	PER	LOC
ORG		2	4
PER	2		6
LOC	19	26	

Tabel 7. Valeks muudetud nimeolemite kategooriad. Veerud tähistavad eelnevaid kategooriaid, read tähistavad uusi kategooriaid.

Eelnev/Uus	ORG	PER	LOC
ORG		2	0
PER	0		0
LOC	6	20	

Kategooriate kaupa andmeid vaadates on näha, et suur osa vigadest tuli PER-märgendust LOC-märgendiks muutes. Selle põhjustasid näiteks Katrin Saksa, Harald Rootsi ja Dave Hollandi nimelised inimesed. Sellest hoolimata tõi see meetod kaasa rohkem positiivseid kui negatiivseid muudatusi.

Siit tabelist on puudu tulp, kus on algselt „O“ kategooriasse määratud sõned, mis paranduse tulemusena muutuks nimeolemiks. Selle tulba lisamine oleks toonud kaasa väga palju valepositiivseid tulemusi, kuna see oleks andnud „LOC“-märgendi kõikidele rahvustele tekstis, näiteks „vene“, „prantsuse“ ja „briti“. 100 esimese näite põhjal oleks umbes kaks kolmandikku muudatustest oleks olnud nende valepositiivsete nimeolemite lisamine, seega see oleks kindlasti märgendaja kvaliteeti kehvemaks teinud.

4.3 Paranduste analüüs

Esimene pakutud reegel nimeolemite märgendaja parandamiseks oli eelnevate ja järgnevate sõnade põhjal nimeolemi kategooria valimine, mis ei töötanud, sest isegi kui eelnevad ja

järgnevad sõnad viitavad tihti sellele, mis kategooriast nimeolem võiks olla, siis see ei erista nimeolemite kategooriaid piisavalt, et selle abil saaks märgendajat parandada.

Teine reeglipõhine põhines analüüsi tulemusel, mis ütles, et automaatsel märgendamisel jäävad liiga paljud nimeolemid mitmeseks, kuid reaalsuses on mitmeseid nimeolemeid vähe. Enamasti ühest kategooriast olevate nimeolemite täielikult üheseks muutmine tõepoolest parandas märgenduse täpsust. Probleemiks osutus nimeolemite valik, mida üheseks lugeda. Automaatselt märgendatud korpuse abil leiab nimeolemeid, mis on just nendele tekstidele omased, näiteks EE on Eesti Ekspressi tekstides tihti organisatsioon, sest just nii viidatakse Eesti Ekspressile endale. Automaatselt märgendatud korpuse miinuseks on, et kui märgendaja süstemaatiliselt mingis kohas eksib, siis see eksimus võib kanduda edasi teistesse kohtadesse, mis muidu olnuks õigesti märgendatud.

Kui jätta välja nimeolemid, mis algselt said märgendi „0“, siis automaatselt leitud nimeolemite abil muudetaks üle kahe tuhande nimeolemi ning ligikaudu 62% muutustest teeksid vale variandi õigeks, samas kui 38% muudaksid õige variandi valeks. Käsitsi märgendatud korpusest saadud nimeolemite põhjal muutuks vaid üle kuuesaja nimeolemi kategooria, samas 69% parandustest muudaks vale kategooria õigeks ning 31% õige valeks. Juba nende arvude põhjal oleks märgendaja kogutäpsuse parandamiseks mõistlikum kasutada automaatse märgenduse põhjal saadud nimeolemite nimekirja. Lisaks sellele saab arvestada ka, et 62% algselt „O“-märgendiga olnud nimeolemitest said õige kategooria, samas kui 38% algselt „O“-märgendi saanud sõnedest märgendati ekslikult nimeolemiks. Kui eeldada, et selle lähenemise põhjal muudetakse 2% märgendatud nimeolemitest ning muudatustest 51% on õiged ja 30% valed, siis võib selle reegli põhjal paraneda nimeolemite märgendaja täpsus (51%-30%)/2% ≈ 0.4%.

Lisaks reeglile, mis märgendaja täpsust parandab, leidis antud uurimus ka probleemse koha märgendaja jaoks, mida saab kasutada tulevase treeningkorpuse loomiseks. Edasine uurimus sel teemal võiks sisaldada uue korpuse käsitsi märgendamist, kus oleks võimalikult palju levinud nimeolemitega lauseid, sest nagu antud uurimistöö näitas, võib nende nimeolemite kategooria ära õppimine parandada mudeli kvaliteeti.

Kokkuvõte

Töö esimeses osas viidi eesti keele nimeolemite märgendaja EstNLTK versioonist 1.4 versiooni 1.6. Samuti loodi EstNLTK versiooni 1.6 visualiseerimimoodul, mis lihtsustab automaatselt märgendatud teksti kontrollimist. Märgendaja üle viimisel ei olnud kõik märgendused identsed, kuid need olid väga sarnased.

Töö teises osas analüüsiti nimeolemite märgendaja omadusi. Analüüsist selgus, et nimeolemite märgendaja kõige levinum viga treeningkorpusel on vale kategooria valimise viga. Eesti keele käsitsi märgendatud nimeolemite korpuses esineb mitmeseid nimeolemeid 2.52%, kuid automaatselt märgendatud koondkorpuse tekstides on 17.34% nimeolemitest mitmesed. Sealjuures olid kategooriate lõikes mitmesuste osakaal korpustes selgelt erinev. See viitab sellele, et märgendaja eksib tihti kategooria valikul.

Töö kolmanda osa eesmärk oli parandada nimeolemite märgendajat reeglipõhise lähenemise abil. Esimese lähenemisena prooviti leida sõnu, mis eelnevad või järgnevad peamiselt ühe kindla kategooria nimeolemitele. Selle idee oli, et kui need sõnad vastavalt eelnevad või järgnevad mõne teise kategooria nimeolemitele, siis see nimeolem võib olla valesti märgendatud. See lähenemine ei osutunud edukaks, sest selle tagajärjel muutus rohkem õigeid märgendusi valeks kui valesid õigeks.

Teine lähenemine oli leida nimeolemeid, mis enamasti on saanud ühe kategooria märgenduse, kuid mõnel korral on ka teise kategooria alla märgendatud. Idee oli, et suurem osa nimeolemeid peaks kuuluma vaid ühte kategooriasse, seega teise kategooria alla märgendatud juhud võivad olla vead. Selle lähenemise järel märgendaja täpsus paranes ning analüüsi põhjal tähendab see muudatus ligikaudu 0.4% täpsuse paranemist.

Viidatud kirjandus

[1] Jurafsky, D., Martin, J. H. „Speech and Language Processing“ (3rd ed. draft), 2019, chapter 18

[2] Tkachenko A., Petmanson T., Laur S. „Named Entity Recognition in Estonian.“ *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, 2013, pp. 78 – 83.

[3] Nadeau D., Sekine S. „A survey of named entity recognition and classification“. *Lingvisticæ Investigationes*, 2007, vol. 30, no. 1, pp. 3-26

[4] Tkachenko A., Särg, D., Laur, S. Tammo, P. „Ner million“. Unpublished paper, 2020.

[5] Chicatriu L., Li, Y., Reiss, F. R. „Rule-based Information Extraction is Dead!

Long Live Rule-based Information Extraction Systems!“ *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 827-832

[6] Derczynski L., Maynard D., Rizzo G., van Erp M., Gorrell G., Troncy R., Petrak J., Bontcheva K. „Analysis of named entity recognition and linking for tweets“. *Information Processing & Management*, 2015, vol. 51, no. 2, pp. 32-49

[7] Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P. „Natural Language Processing (Almost) from Scratch“. *Journal of Machine Learning Research*, 2011, vol. 12

[8] Huang Z., Xu W. & Yu K. „Bidirectional LSTM-CRF Models for Sequence Tagging“ 2015

[9] Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C. „Neural Architectures for Named Entity Recognition.“ 2016. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* pp. 260-270

[10] Sang, E., De Meulder, F. „Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition“ 2003. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003.* pp. 142-147

[11] Esuli, A. Sebastiani, F. „Evaluating Information Extraction“. 2010. *Multilingual and Multimodal Information Access Evaluation.* pp. 100-111.

[12] Laur, S. „Nimeüksuste korpus“ Center of Estonian Language Resources. https://doi.org/10.15155/1-00-0000-0000-0000-00073L (08.05.2020)

[13] Eesti keele koondkorpus. https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et (08.05.2020)

Lisad

Töö käigus kirjutatud kood on kättesaadav lehel https://github.com/rasmusmaide/ner-analysis

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Rasmus Maide**

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

Eesti keele nimeolemite märgendaja analüüs ja parandamine,

mille juhendaja on **Sven Laur**

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.

Rasmus Maide

08.05.2020