Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Informaatika õppekava

Sander Saska

Morfoloogilise muuttüübi automaatne tuvastamine

Bakalaureusetöö (9 EAP)

Juhendaja(d) Siim Orasmaa, PhD

Tartu 2024

Morfoloogilise muuttüübi automaatne tuvastamine

Lühikokkuvõte: Eesti keel on pidevas arenemises – uusi sõnu tekib juurde erinevate võimalustega. Keelekasutajad oskavad uusi sõnu käänata või pöörata sageli intuitiivselt, kuid keeleteaduses on see intuitsioon formaliseeritud muuttüüpide näol. Käesolev töö uurib, kuidas automatiseerida sõna muuttüübi tuvastamist. Selleks on loodud kaks LSTM-põhist mudelit, mis ennustavad sõnadele muuttüüpe. Mudelite algandmed on võetud Vabamorfi põhileksikonist, mis koosneb peaaegu 74 000 algvormist. Algvormidele on sünteesitud kõik võimalikud sõnavormid ja on ümber töödeldud tehisnärvivõrgule sobivaks kujuks. Üks mudel on treenitud ainult sõna kirjapildil, mille täpsuseks saavutati 95.8%, teisel on lisaks antud sõnaliik, täpsuseks 97.8%.

Võtmesõnad:
LSTM, muuttüübid, Vabamorf, tehisnärvivõrgud, tehisintellekt, masinõpe, klassifitseerimine

CERCS: P176 – Tehisintellekt

Automatic detection of morphological inflection types

Abstract: Estonian language is constantly evolving, as new words are created in different ways. Language users often know intuitively how to inflect new words, but in linguistics this intuition is formalized in the form of inflection types. This work researches how to automate the identification of inflection types. To this end, two LSTM-based models have been created to detect and predict inflection types. The initial data for the models are taken from Vabamorf's morphology lexicon, which consists of almost 74 000 lemmas. All possible word forms are synthesized for the lemmas and the result is transformed into a suitable form for the LSTM-based models. One model is trained on only words, with an accuracy of 95.8%, and the other model is trained on words and parts of speech, with an accuracy of 97.8%.

Keywords:
LSTM, inflection types, Vabamorf, artifical neural networks, artificial intelligence, machine learning, classification

CERCS: P176 – Artificial intelligence

Sisukord

1 Sissejuhatus

Eesti keel on pidevas arenemises – uusi sõnu tekib juurde uute mõistete nimetamiseks ja selleks on erinevaid võimalusi: sõnade liitmisel („suunamudija“); tuletamisel („guugeldamine“); laenamisel („droon“); tehisloomel („mupo“); murdetüvede taaselustamisel („kilb“) või isegi olemasolevale kirjakeele sõnale uue tähenduse andmisel („heitunu“ tähenduses „hirmunu“ kui ka „inimene, kes on töö otsimisest loobunud“) [Ees22]. Varasemalt on tehtud sõnavõistlusi[1] ja sõnakorjeid, et saada uusi sõnu ja ergutada inimeste keelehuvi. Näiteks 2023. aasta Sõnaus, kuhu laekus kokku 3350 sõna, millest rahvahääletusega ja hindamiskomisjoniga valiti võidusõnadeks järgmised sõnad: „asejäse“, „digikelts“, „iilitiivik“, „kõdustama/kõdusti“, „lamar“, „põimsus“, „rohepaplus“ ja „töökaim“ [Raa23].

Inimestena suudame sageli varasema keelekogemuse põhjal oletada, kuidas uusi sõnu tuleks käänata või pöörata. Teisisõnu me teame „muuttüüpi“ intuitiivselt. Lingvistid on meie keelekogemuse põhjal pakkunud välja teoreetilised muuttüübid, mis grupeerivad samamoodi käänduvaid või pöörduvaid sõnu. Käesolev töö uurib, kuidas muuttüübi määramist automatiseerida – tuvastada etteantud sõna muuttüüp automaatselt.

Bakalaureuse töö on rakenduslik töö ning selle eesmärk on uurida:

- kui täpselt suudab tehisnärvivõrgul põhinev mudel tuvastada muuttüüpe toetudes „Väikese vormisõnastiku”[2] (VVS) tüübikirjeldustele;
- millisest infost piisab tehisnärvivõrgule muuttüübi tuvastamisel.

Luuakse kaks LSTM-põhist mudelit, millest üks on treenitud sõna kirjapildil, teine kirjapildil ja sõnaliigil. Mudelitest on hiljem kasu automaatse morfoloogilise analüüsi kohandamisel uutele sõnade, aga ka murdekeele ja vana kirjakeele sõnade analüüsimisel.

Töö jaguneb viieks peatükiks. Esimeses peatükis antakse teoreetiline ülevaade muuttüüpidest, kirjeldatakse algandmeid ja tehisnärvivõrke. Teises peatükis kirjeldatakse metoodikat, kuidas leksikonist andmeid koguti, algvormidele muutevorme sünteesiti ja tehisnärvivõrgu alusmudel üles ehitati. Kolmandas peatükis eksperimenteeritakse alusmudelite peal hüperparameetrite häälestamist. Neljandas peatükis analüüsitakse tulemusi häälestatud mudelitel ja viimases peatükis erinevaid edasiarendusvõimalusi sellele tööle.

Valminud töö kood koos graafikute ja mudelitega on kättesaadav järgmiselt veebilehelt[3].

[1] Eesti Keele Instituut. Sõnavõistluste tava Eestis. 2021. `https://eki.ee/teatmik/sona/uued-sonad/sonavoistlused/` (14.05.2024)

[2] Ü. Viks. Väike vormisõnastik. I: Sissejuhatus & grammatika, II: Sõnastik & lisad. Tallinn: Eesti Teaduste Akadeemia. 1992.

[3] S. Saska. Morfoloogilise muuttüübi automaatne tuvastamine. 2024. `https://github.com/SanderSaska/MorfoloogiliseMuuttyybiAutomaatneTuvastaja` (14.05.2024)

2 Teoreetiline ülevaade

Järgmises peatükis selgitatakse muuttüüpide mõistet ja kuidas neid kasutatakse 2018. aasta õigekeelsussõnaraamatus (ÕS). Ajaloo alampeatükis tutvustatakse varasemaid muuttüüpide liigitamise viise, millele järgnevad varasemalt loodud muuttüüpe ennustavad mudelid. Lisaks selgitatakse viimases kahes alampeatükis töös kasutatavat sõnastikku ja tehisnärvivõrke.

2.1 Muuttüübid ÕS-is

Morfoloogiline[4] muuttüüp on rühm, kus sõnu käänatakse või pööratakse samamoodi, see tähendab kasutatakse samasuguseid grammatilisi tunnuseid ja tüvesid samades vormides [EER20] [Ere+18] [Vik15]. Näiteks sõnad „punane“ ja „kollane“ kuuluvad samasse rühma, sest neid käänatakse analoogselt: *punane, punase, punast* ja *kollane, kollase, kollast*.

Sõnu on keeles palju, mistõttu nende rühmitamiseks on rohkelt võimalusi, näiteks on Elmar Muugi „Väikeses õigekeelsus-sõnaraamatus“[5] muuttüüpe kokku 1166, kuid aja jooksul on efektiivsema grupeerimisega saadud ainult 38 [Vik15]. Lugejasõbraliku süsteemi loomiseks lisatakse igale muuttüübile tähistuseks seda esindav tüüpsõna, number ja/või mõni muu tunnus.

Näiteks on 2018. aasta ÕS-is tüübid numereeritud 1–38, kus esimesed 26 muuttüübi kirjeldavad käändsõnu ja viimased 12 pöördsõnu. ÕS-i sõnastikus on käändsõnadel ja pöördsõnadel muuttüübi number pandud sõna kõrvale noolsulgude vahele, millele järgneb, kuid mitte alati, käänete järelliited alates omastavast. Muutumatutel sõnadel (määrsõnad, kaassõnad, hüüdsõnad ja sidesõnad) sellised noolsulud puuduvad [Ere+18]. Näiteks sõnade „punane“ ja „kollane“ noolsulgudest on võimalik välja lugeda, et sõnadele vastab muuttüüp numbriga 10 ja omastava käände järelliide -se (joonis 1). Järelliitest -se saab tuletada sõnade omastav käändevorm „punase“ ja „kollase“.

punane <10: -se>, **kollane** <10: -se>,

Joonis 1. Sõnade „punane“ ja „kollane“ muuttüübi number 2018. aasta ÕS-is [Ere+18]

Kõik muuttüübid on ÕS-is eraldi peatükis järjestikuliselt esitatud, kus lisaks kirjeldatakse nende ülesehitust ja erinevaid liike. Iga muuttüüp algab tüüpsõnaga ja sellele järgnevad põhivormid (joonis 2). Tüüpsõna on sõna algvorm ehk lemma, mis on kujutatud paksus kirjastiilis. Põhivormid on käänded või pöörded, millest on võimalik tuletada kõiki teisi vorme.

[4] sõnavormide ülesehituse või moodustamisega seotud. EKI ühendsõnastik 2024. Eesti Keele Instituut, Sõnaveeb 2024. `https://sonaveeb.ee/search/unif/dlall/dsall/morfoloogiline/1` (14.05.2024)

[5] E. Muuk. Väike õigekeelsus-sõnaraamat (1. trükk). Tartu: Eesti Kirjanduse Selts. 1933. (5. trükk 1936)

Käändsõnadel on põhivormideks kolm järgmist käänet ainsuse ja mitmuse vormides: omastav, osastav ja sisseütlev. Pöördsõnadel on põhivormideks *ma*-tegevusnimi, *da*-tegevusnimi, kindla kõneviisi oleviku ainsuse 1. pööre, lihtmineviku ainsuse 1. ja 3. pööre, möönva kõneviisi olevik, *v*-kesksõna, *nud*-kesksõna, umbisikulise tegumoe kindla kõneviisi olevik (*-kse*) ja *tud*-kesksõna. Tüüpsõna ja põhivormide loetelule järgneb all tabel häälikumuutustega, mis võivad esineda nii tüve lõpus kui ka tüve sees. Läbi nende on lihtsam käänata või pöörata otsitud sõna, kui tüüpsõnast jääb puudu [Ere+18].

1

ohutu, ohutu, ohutut, ohutusse, ohutute, ohutuid, ohututesse *ja* ohutuisse

av-ta	lm-ta	`aasta, kasvataja, psühhol`oogia, matemaatika, k`õrge, `õhtu, elamu, k`ohkumatu

Joonis 2. Muuttüüp 1 – **tüüpsõna**, põhivormid ja häälikumuutused [Ere+18]

3 ja 5

armas, `armsa, `armsat, `armsasse, `armsate, `armsaid, `armsatesse *ja* `armsaisse

vv, lv	as → sa	ergas : `erksa
	is → sa	õilis : `õilsa, õnnis : `õndsa
	us → sa	ainus : `ainsa
	es → sa	tõrges : t`õrksa

5 ja 3

armas, `armsa, armast, `armsasse, `armsate, `armsaid, `armsatesse *ja* `armsaisse

vv, lv	as → sa	ergas : `erksa
	is → sa	õilis : `õilsa, õnnis : `õndsa
	us → sa	ainus : `ainsa
	es → sa	tõrges : t`õrksa

Joonis 3. Põhimuuttüüpide 3 ja 5 rööptüübid ÕS-is [Ere+18]

Lisaks esineb ÕS-is ka rööptüüpe, kus on põhitüübile antud viitena heledamas kirjas teine põhitüübinumber (joonis 3). Esmapilgul võivad need paista samad, kuid tegelikkuses eristab neid mingi muutevorm. Joonise 3 näitel on erinev osastav kääne: „armast“ ja „armsat“. Rööptüübid tulenevad sõna mitmetähenduslikkusest või kui sõna tohib hääldada mitmes (II ja III) vältes, näiteks on sõna „armas“ nii omadussõna kui ka nimisõna, omadussõnana kui „väga armastatud“, nimisõnana kui „inimene, keda keegi (väga) armastab“. Sõna „ahne“ on mitme vältega *ahne* ja *‘ahne* [Ere+18].

2.2 Muuttüüpide ajalugu

Eesti keeles on muuttüüpide arv ja järjekord ajaloos muutunud. Järgnev toetub Ülle Viksi artiklile [Vik15]. Tema sõnul on tüüpide arvu mõjutanud tüübistiku liigitusalusena arvestatud tüvemuutuste detailsus. Esimene tänapäevase kujuga muuttüüpide süsteem oli Elmar Muugi „Väikeses õigekeelsus-sõnaraamatus“. Muugi sõnastikus oli esialgu muuttüüpe kokku 1166, kuid järgmiste trükkidega (kokku viis) taandus see lõpuks 845-le. Suurem koondamine tehti järgmisena „Suures õigekeelsus-sõnaraamatus“[6], kus saadi tulemuseks ainult 115 tüüpi. Pool sajandit hiljem, 1999 ÕS-ga[7] ja 2006 ÕS-ga[8], langes muuttüüpide arv 69-le, kuniks võeti kasutusele uute põhimõtetega tüübistik, mille järgi on muuttüüpe tänaseni 38. Selle loomise ajendiks oli vajadus luua järjekindel ja formaalne klassifikatsioon, millega saaks genereerida uutele sõnadele muutevorme. Samuti aitas see teist keelt kõnelevatel isikutel lihtsamini eestikeelseid sõnu käänata ja pöörata. Uue tüübistiku esimene rakendus oli VVS, mis leidis kasutust andmestikuna mitmes automaatselt tehtavas morfoloogilises süsteemis. Hiljem võeti sama tüübistik kasutusele ka ÕS-ides 2013[9] ja 2018[10].

Väike arv muuttüüpe on kasulik masinõppemudeli loomisel, sest väheste klasside arv tähendab, et igasse klassi langeb rohkem sõnu, mistõttu on mudel palju kindlam oma vastuse andmisel. Samas on raskem kategoriseerida mitmetähenduslike või mitme vältega sõnu, sest enam ainult sõna kirjapildist ei piisa, vaid tuleb võtta arvesse eristavaid omadusi, näiteks sõnaliik või sõnavälde.

2.3 Muuttüüpide ennustamine

Morfoloogilist muuttüüpi on eesti keeles varemgi ennustatud. Eesti Keele Instituudi tarkvaraga Tüübituvastaja [Eesa] on võimalik väiketähtedega algvormile leida sellele vastav muuttüüp ja sõnaliik ning see kasutab klassifitseerimiseks VVS-i muuttüübistikku. Algoritm on reeglipõhine ja kasutab muuttüübi määramiseks sõna fonoloogilist[11] struktuuri [Eesb], näiteks nimisõnareeglis *2 Clik 25_A* esimene number *2* tähistab algvormi silpide arvu, tähed *Clik* viimaseid häälikuid (konsonant ehk kaashäälik + *lik* lõpp) ning allkriipsuga ühendatud number-tähe paar *25_A* muuttüübi numbrit ja sõnaliiki (adjektiiv ehk

[6] A. Kask, E. Raiet, J. V. Veski. Suur õigekeelsus-sõnaraamat (I, II). Tartu: Teaduslik Kirjandus. 1948–1951

[7] T. Erelt, T. Leemets, S. Mäearu ja M. Raadik. Eesti keele sõnaraamat. ÕS 1999. Tallinn: Eesti Keele Sihtasutus. 1999

[8] T. Erelt, T. Leemets, S. Mäearu ja M. Raadik. Eesti õigekeelsussõnaraamat. ÕS 2006. Kirjakeele normi alus. Tallinn: Eesti Keele Sihtasutus. 2006

[9] T. Erelt, T. Leemets, S. Mäearu ja M. Raadik. Eesti õigekeelsussõnaraamat ÕS 2013. Tallinn: Eesti Keele Sihtasutus. 2013. `http://www.eki.ee/dict/qs2013/` (14.05.2024)

[10] Vaata [Ere+18]

[11] keele häälikusüsteemiga seotud. EKI ühendsõnastik 2024. Eesti Keele Instituut, Sõnaveeb 2024. `https://sonaveeb.ee/search/unif/dlall/dsall/fonoloogiline/1` (14.05.2024)

omadussõna). Selle reegli alla kuuluvad sõnad nagu „petlik“ ja „piinlik“. Tegusõnadel kasutatakse samasuguse ülesehitusega reegleid. Algoritmi sisendsõnana nõutakse eestikeelse sõna väiketähtedega algvormi. Kui anda sisendiks sõna mõnes muus käändes või pöördes, siis tagastatakse vale muuttüüp või #, mis tähistab keelavat reeglit (sõna peab vastama algvormi tingimustele). Liitsõnade korral tuleb Tüübituvastajale anda ette liitsõna viimane komponent algvormis, vastasel korral tagastatakse samuti vale muuttüüp või #. Programmi graafiline kasutaliides võimaldab sõnu ainult ühekaupa sisestada, mis teeb võimatuks programmi masstestimise suurel sõnade hulgal.

Muuttüüpide ennustamisi on tehtud ka teistes keeltes. Uurijad Nikola Ljubešić, Filip Klubička, Miquel Esplà-Gomis ja Nives Mikelić Preradović on kasutanud pooljuhendatud masinõpet, et leida tundmatutele horvaatia keelte sõnadele vastavat muuttüüpi ja algvormi. Uue sõna korral leitakse selle sõna kõik võimalikud algvormi-muuttüübi paarid. Mudel kuvab kasutajale kandidaat-paarid, mis on järjestatud sõnatüve tunnuste, korpuse statistika ja olemasolevast muutkondade leksikonist saadud info alusel. Lõpliku otsuse teeb kasutaja [Lju+15]. See on üks võimalus, kuidas muuttüüpi ennustada. Mudel aitab lihtsustada uute sõnade lisamist leksikoni, sest ei pea enam ise üles leidma, milline muuttüübi number vastaks sõnale kõige paremini.

Leidub uurijaid, kes on lähenenud sellele ülesandele teisiti. Uurijad Grégoire Détrez ja Aarne Ranta tutvustavad tarku muuttüüpe, mis ei hoia endas kõiki vorme, vaid regulaaravaldistel ja sõneoperatsioonidel põhinevaid tingimusi ning argumente selliste tingimuste loomiseks. Kui sõna vastab targale muuttüübile, siis tagastatakse traditsiooniline lingvistikas kasutatav muuttüüp. Selliste tarkade muuttüüpidega saab muuttüüpide arvu vähendada kuni sõnaliikide arvuni. Igale sõnaliigile luuakse rida tarku muuttüüpe ning eesmärk on leida kõige paremini muuttüüpe eristatavad sõnavormide jadad, mis sobivad tarkadele muuttüüpidele argumentideks. Tulemuseks on erinevate argumentide jadadega tark muuttüüp, mis suudab määrata müüttüüpe kõikidele teatud sõnaliigiga sõnadele [DR12].

Käesolevas töös sarnaneb eesti keele morfoloogilise muuttüübi automaatne tuvastaja rohkem esimese uurimisrühma lähenemisega. Ei leita kõik võimalikud algvormi-muuttüübi paarid, mille vahel peab kasutaja tegema otsuse, vaid sõna tükeldatakse tähejärjendiks ja antakse sisendina tehisnärvivõrgule, mis leiab sellele kõige sobivama muuttüübi numbri. Närvivõrk ise leiab tähejärjenditele iseloomulikud omadused ning loob erinevate kaaludega seosed muuttüüpidele. Lisaks töötab tuvastaja sama tüübistikuga (VVS) nagu Tüübituvastaja. Liitsõnad on võimalik anda ette täielikul kujul ning see oskab arvestada kõikide sõnedega ükskõik mis käändes või pöördes.

2.4 Sõnastik

Algandmetena kasutatakse morfoloogilise analüsaatori Vabamorfi põhisõnastikku, et muuttüübi tuvastaja suudaks tuvastada vähemalt niipaljude sõnade muuttüüpe, kui oskab Vabamorf analüüsida/sünteesida ilma oletamiseta. Järgnev sõnastiku ülevaade tugineb Vabamorfi autorite Heiki-Jaan Kaalep, Rene Prillop ja Tarmo Vaino loodud leksikoni kirjeldusele [KPV22]. Selle järgi on põhisõnastikus 73831 kirjet, millest suurem osa moodustavad VVS-i sõnad (40%) ja Filosofti tesaurus[12] (26%). Peaaegu viiendik (17%) on lisatud jooksvalt analüsaatori arendamise käigus.

```
allikas-liigitus:aeg|sisu
```

Joonis 4. Sõnastikukirje formaat

Sõnastikukirje (joonis 4) koosneb järgmistest elementidest:

- `allikas-liigitus` – näitab, kust või millega seoses on kirje tulnud;
- `aeg` – kirje lisamise aeg, kus kaks esimest numbrit tähistavad kuud (`00` kui pole teada) ja kaks viimast aastat;
- `|` ja `:` – info eraldajad.

Näiteks kirjes `vvs:0093|aku!\S\.16!` on allikaks VVS ja ajaks aasta 1993 ilma kuuta.

Püstkriipsust (`|`) paremal pool olev element `sisu` jaguneb järgmisteks komponentideks: tüvi, morfoloogiline kood ja erandid. Tüvi hoiab endas sõna algvormi ja lisaks võib olla põhivorme, mille arv sõltub morfoloogilisest koodist. Tüvi on eraldatud sümbolite `|` ja `!` vahele.

```
tüvi!\sõnaliik\.muuttüüpˆvokaalˆ.SG_või_PL
vokaaltüveliste_vormide_arv!%vana_lõpp>uus_lõpp[I%**erandid
```

Joonis 5. Morfoloogilise koodi formaat

Morfoloogiline kood (joonis 5) koosneb järgmistest elementidest:

- `!` ja `%` – info eraldajad;
- `sõnaliik` – tüve sõnaliik;
- `muuttüüp` – muuttüübi number;
- `vokaal` – vokaalmitmuse moodustamise viis;
- `SG` – tüvel on ainult ainsusevormid;

[12] Eesti keele tesaurus. Filosoft. 2007. `https://www.filosoft.ee/thes_et/` (14.05.2024)

- `PL` – tüvel on ainult mitmusevormid;
- `vokaaltüveliste_vormide_arv` – määrab, milliste käänete vokaalmitmuselised vormid moodustatakse;
- `vana_lõpp>uus_lõpp[I` – vokaallõpulise tüve teisendus, kui mitmuse moodustamisel lisandub *i*-ga algav tunnus.

```
lisa:0094|üli_k<ool] üli_k<ool]i üli_kool]i!\S\.22^E^!
 vvs:0093|ülemus ülemuse!\S\.11^I^+7!
 vvs:0093|ürask üraski!\S\.2+9!%i>e[I%
lisa:0094|T<artu_m<aa!\H\.26.SG!
lisa:0094|teksa[D!\S\.16.PL!
vvsv:0093|t<arni[MA tarni[B!\V\.28!
```

Joonis 6. Näited sõnastikukirjetest

Joonis 6 näitab, et iga kirje ei sisalda kõiki infokilde ja tüvede arv on varieeruv vastavalt morfoloogilises koodis määratud reeglitele, millega saadakse tüve kõik teised vormid.

Viimasena on erandid, mis sisaldavad reeglipäratuid käände- või pöördevorme. Valdav hulk ($\frac{6}{7}$) kirjeid sõnastikus seda ei sisalda. Erandite plokk järgneb morfoloogiakoodile ja algab kahe tärniga `**` (joonis 7).

```
lisa:0094|T<artu!\H\.1.SG! **$ADT$: T<artu
 vvs:0093|t<asku!\S\.1+9! **$ADT$: t<asku
```

Joonis 7. Erandid sõnastikukirjetes tähistatud sümbolitega `**`

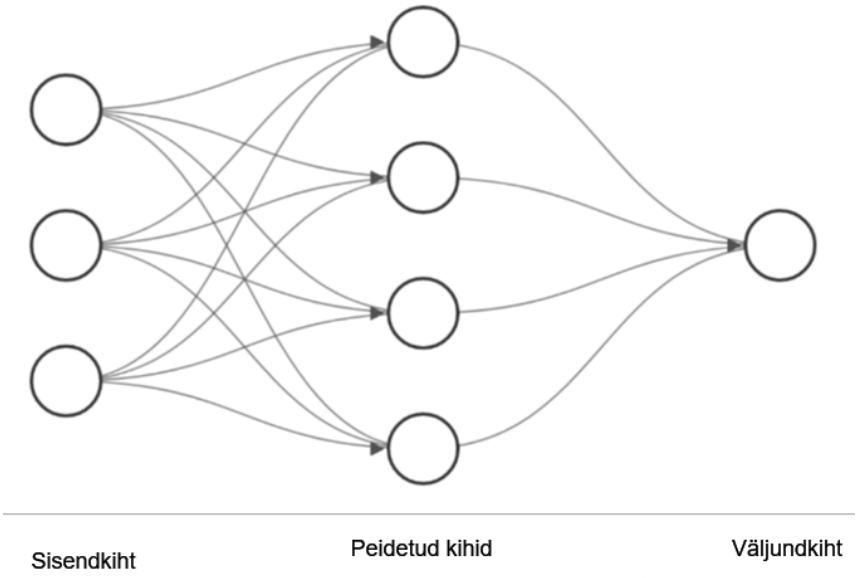
Iga erand algab vorminimega, mis on `$`-märkide vahel ja lõppeb `:` sümboliga. Pärast seda järgnevad erandsõnad, mis eraldatakse `;` märgiga. Näidetes on vorminimeks ADT, mis tähistab aditiiv suunduv ehk lühike sisseütlev. Erandlikumatel tegusõnadel on eraldi välja toodud 13 tüve ja väga erandlikult käänduvad sõnad (näiteks liitsõnad, mille kõik osised käänduvad) on määratud muuttüüpi 0 ning neil on sõnastikukirjes kõik vormid üles loetud.

2.5 Tehisnärvivõrgud

Järgnevad närvivõrkude kirjeldused tuginevad Ian Goodfellow, Yoshua Bengio ja Aaron Courville raamatule „Deep Learning“ [GBC16]. Selle järgi nimetatakse tehisnärvivõrke

võrkudeks, sest need koosnevad tavaliselt paljudest erinevatest kokkuliidetud funktsioonidest ehk tehislikest neuronitest. Tehisnärvivõrgu eesmärk on lähendada mõnda funktsiooni f^*. Selleks defineerib tehisnärvivõrk funktsiooni $\mathbf{y} = f(\mathbf{x}; \theta)$ ja andmete pealt treenitakse välja parameetrite θ väärtused, et funktsioon f oleks võimalikult lähendane funktsioonile f^*.

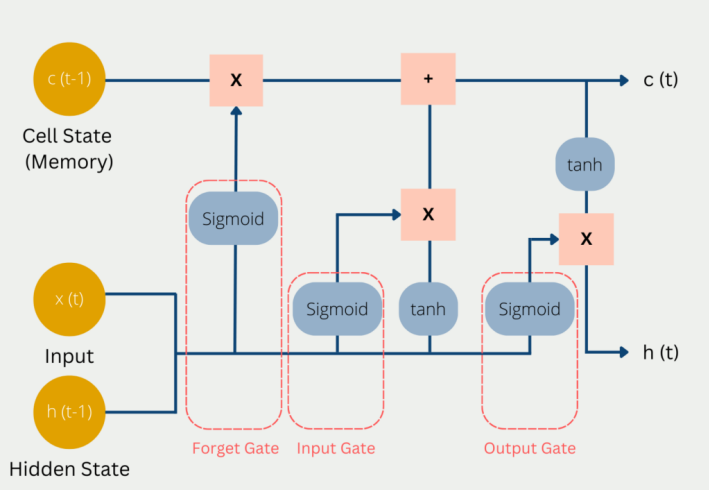
Mudelit on seostatud suunatud graafiga, mis kirjeldab, kuidas funktsioonid on kokku pandud. Sisendkiht võtab endasse sisestusi, mis vastab muutujale x funktsioonis $f(\mathbf{x}; \theta)$. Närvivõrgu treenimise ajal juhitakse funktsiooni $f(x)$ vastavusse funktsiooniga $f^*(x)$. Iga sisendi x juurde kuulub klass $y = f^*(x)$. Treeningandmed määravad, mida väljundkiht peab igas punktis lähteandmega x tegema – andma väärtuse, mis on lähedane ennustatavale klassile y. Õppimisalgoritm peab otsustama, kuidas neid kihte kasutada, et kõige paremini rakendada funktsiooni f lähendamist funktsioonile f^*. Kuna treeningandmed ei näita iga sellise kihi jaoks soovitud väljundit, nimetatakse neid peidetud või varjatud kihtideks. Viimast kihti nimetatakse väljundkihiks, kus arvutatakse viimase funktsiooniga võimalikult lähedane väärtus ennustatavale klassile y.



Joonis 8. Pärilevivõrgu ülesehitus[LeN19]

Pärilevivõrgud[13] on üks perekond tehisnärvivõrkudest. Pärilevivõrke nimetatakse pärileviks, sest info liigub ühes suunas: sisendkiht $\rightarrow$ peidetud kihid $\rightarrow$ väljundkiht (joonis 8). Tagasilevi, mil mudeli väljundid suunatakse iseendasse tagasi, ei ole, aga kui pärilevivõrke nendega laiendata, siis on tulemuseks rekurrentsed närvivõrgud ehk RNN-d. Need on järjestikuste andmete töötlemiseks mõeldud närvivõrkude perekond, mis jagavad samu θ parameetrite väärtusi mitme epohhi ehk ajasammu vahel.

[13] teisisõnu mitmekihilised pertseptronid/närvivõrgud (MLP)



Joonis 9. LSTM neuroni ülesehitus[14]

Rekurrentsete närvivõrkude perekonda kuulub ka muuttüübi tuvastajas kasutatav LSTM[15]. Tuginedes Aston Zhangi, Zack Liptoni, Mu Li ja Alex Smola sügavõppe raamatule [Zha+23], LSTM-i neuronid on RNN-i neuronite edasiarendus – nad suudavad talletada olulist (ja kauget) informatsiooni mitme epohhi vahel või treenimiskäigu algusest lõpuni välja. Lisaks ei lasta igal uuel mitteolulisel infol talletatud infot mõjutada.

LSTM-i neuron (joonis 9) koosneb

- neuroni olekust (*cell state*), mis hoiab endas talletatud infot mitmel epohhil ja mille väärtust väljastatakse igal epohhil;
- varjatud olekust (*hidden state*), mis hoiab endas eelmise epohhi infot, kirjutatakse üle pidevalt, ja mille väärtust väljastatakse igal epohhil;
- sisendväravast (*input gate*), mis otsustab, kas praeguse epohhi sisend (*input*) peaks mõjutama neuroni olekus olevat infot;
- unustamisväravast (*forget gate*), mis otsustab, kas neuroni olekus tuleks unustada kogu info;
- väljundväravast (*output gate*), mis otsustab, kas neuroni olek peaks mõjutama varjatud oleku väljundit.

Esmalt liidetakse sisendi ja eelmise epohhi varjatud väärtused, seejärel liigub info läbi funktsioonide ja siis teostatakse ülejäänud binaarsed tehted. Väravad koosnevad sigmoid funktsioonidest, mille väärtused jäävad vahemiku $(0, 1)$. Mida lähemale väärtusele 1, seda rohkem on väravad lahti ehk seda rohkem lastakse infot mõjutada. Tanh funktsioon normaliseerib väärtused vahemikku $(-1, 1)$.

[15] *Long Short Term Memory*, pikk lühimälu
[15] LSTM neuroni ülesehituse joonis. 2022. `https://databasecamp.de/en/ml/lstms` (14.05.2024)

3 Metoodika

Selles peatükis kirjeldatakse meetodeid morfoloogilise muuttüübi automaatse tuvastaja ülesehitamiseks. Alustatakse vajalike andmete kogumisega sõnastiku igast kirjest. Andmete kogumisele järgneb kõikide võimalike muutevormide sünteesimine. Pärast sünteesimist töödeldakse andmed tehisnärvivõrgule sobivaks vormiks ja viimases alampeatükis kirjeldatakse muuttüüpe tuvastava tehisnärvivõrgu loomist.

3.1 Leksikonist andmete kogumine

Vabamorfi põhisõnastikus on iga kirje inforohke, alates allikast kuni muutevormide sünteesimiseks vajalike reegliteni välja. Miinimumsisend muuttüübi ennustamisel on sõnavorm, aga on võimalik, et läheb vaja ka sõnaliigi infot. Samas robustse süsteemi loomiseks peaks mudel kasutama sisendina sõnavormi ja sisend ei tohiks nõuda täiendavat lingvistilist analüüsi.

Seega on leksikonist vaja koguda igast kirjest sõnavorm, muuttüüp ja sõnaliik. Saaks ka ilma sõnaliigita, sest muuttüübid on jagatud käändsõnadesse ja pöördsõnadesse ning ülejäänutel muuttüüp lihtsalt puudub. Samas tuleb sõnaliigi kogumine kasuks leksikoni analüüsimisel ja on võimalik, et mudel ei suuda järeldada ainult sõnavormi kujust, kas sõnavorm on üldse käänatav või pööratav.

Kogu töö toimus Google Colabi[16] keskkonnas, sest seal on vajalikud teegid ja riistvara kiirendajad olemas. Sellest tulenevalt ei pidanud lokaalsesse arvutisse lisatarkvara alla laadima ja paigaldama.

Sõnastikust vajaliku info kogumiseks käidi tsükliga rida haaval kirjeid läbi ning sõna, muuttüübi ja sõnaliigi eraldamiseks kasutati regulaaravaldisi[17]. On teada, et sõnastikukirje on kujul `allikas-liigitus:aeg|sisu` (joonis 4), kus kõik vajalik info asub elemendi `sisu` all. Igast kirjest algvormi kogumiseks on vaja kätte saada sõnavormide jada `tüvi`, mis on eraldatud sümbolite `|` ja `!` vahele. Selle töö teeb kõik ära regulaaravaldis `\|(.*?)!`. Eraldajaga `|` jagatakse kirje järjendiks `[allikas-liigitus:aeg, sisu]`, millest on vaja viimast elementi `sisu`. Eraldajaga `!` jagatakse element `sisu` järjendiks `[tüvi, [morfoloogiline kood, erandid]]`, kust on vaja esimest elementi `tüvi`. Selle saab kätte regulaaravaldises kirjeldatud sulgudega. Kuna sõnavormide arv jadas on kirjeti erinev, siis tuleb omakorda jada `tüvi` jagada tühikuga järjendiks $[\texttt{tüvi}_0, \texttt{tüvi}_1, \texttt{tüvi}_2, \ldots]$, kus iga element $\texttt{tüvi}_i$ vastab sõnavormile indeksil `i`. Järjendist on võimalik kätte saada algvorm, mis asub esimesel indeksil ehk $\texttt{tüvi}_0$.

[16] Sissejuhatus *Google Colabi*. `https://colab.research.google.com/` (14.05.2024)

[17] Pythoni standardteegi *re* dokumentatsioon. `https://docs.python.org/3/library/re.html` (14.05.2024)

Muuttüübi eraldamiseks kasutati regulaaravaldist `\.(\d{1,2})`, sest on teada, et muuttüüp asub kohe pärast sümbolit . ja sellele järgnevad üks või kaks numbrit. Regulaaravaldist `\\(\\(\w)\w{0,5}\\|\(\w)\w{0,5}\.` kasutati sõnaliigi eraldamiseks. Sõnaliiki oli keerulisem kätte saada, sest morfoloogilise koodi kuju järgi on sõnaliik tavaliselt eraldatud sümbolitega \, näiteks kirje `vvsm:0093|<all!**\D\!**`. Samas esineb leksikonis kirjeid, kus neid sümboleid pole ning sõnaliik järgneb kohe info eraldajale !, näiteks kirje `lisav:0919|**<eel_k<eet[MA. . . **!V.34!**`. Lisaks võib sõnaliigile juurde olla pandud mingi tähega tähistatud tunnus, näiteks tabusõna (t) või liitsõna koosseisu mittesobiv sõna (m), mistõttu tuleb võtta ainult esimene täht, millega tähistatakse sõnaliiki ennast.

Kui vajalik info oli kirjest korjatud, tuli järgmisena statistika, täpsemalt uuriti eraldi sõnaliikide ja muuttüüpide sagedusi leksikonis. Nendest sagedustabelitest sai järeldada, kui ühtlaselt on klassid jaotunud ja andsid aimu, kuidas tuleks jagada andmestik treenimis-, valideerimis- ja testhulgaks.

Lisades olev tabel 4 kirjeldab sõnaliikide sagedusi Vabamorfi põhisõnastikus. Tuleb arvestada, et mitme sõnaliigiga sõnade korral on esimene neist sisestatud sagedustabelisse, teised arvestamata. Esikohal on nimisõnad ($\sim 55\%$), järgnevad tegusõnad ($\sim 14.5\%$) ja omadussõnad ($\sim 14.5\%$), mis on kõik muutuvad sõnad. Järgmisena tuleb esimene muutumatute sõnade rühm, määrsõna (8.75%). Neid tuleb alles hoida, et mudel oskaks eristada käänduvaid ja käändumatuid ning tagastada vastusena sümbol „-“, millega tähistatakse olematut muuttüüpi, ehk tegu on muutumatu sõnaga.

Lisades olevad tabelid 5, 6 ja 7 kirjeldavad muuttüüpide sagedusi Vabamorfi põhisõnastikus. Mitme muuttüübiga kirjed olid klassifitseeritud eraldiseisvana, et oleks võimalik tuvastada rööptüüpe. Sagedustabelis on näha, et muuttüübid on leksikonis ebaühtlaselt jaotatud, mistõttu tuli hulkadeks jagamisel veenduda, et vähemalt üks kirje igast klassist oleks esindatud treeninghulgas. Vastasel korral oleks võinud närvivõrgule sattuda ette tundmatu muuttüüp, millega polnud teda varem treenitud.

Kogutud andmete kontrollimine oli keeruline. Sõnaliigi ja muuttüübi saab kontrollida koostatud statistika kaudu, kui sagedustabelis ei esine soovimatuid klasse, siis on nende eraldamine olnud korrektne. Sõnu otseselt kontrollida, see tähendab vaadelda igat sõna eraldi, on mahukas, mistõttu veenduti põhjalikult loodud algoritmi korrektsuses, millega saab eeldada, et sisu on ka õige.

Selle alampeatüki kood on kättesaadav järgmiselt veebilehekülgelt[18].

[18]Morfoloogilise muuttüübi automaatne tuvastamine - I osa. `https://github.com/SanderSaska/MorfoloogiliseMuuttyybiAutomaatneTuvastaja/blob/main/Morfoloogilise_muutt%C3%BC%C3%BCbi_automaatne_tuvastamine_I_osa.ipynb` (14.05.2024)

3.2 Süntees

Kui algvormid koos muuttüüpidega ja sõnaliikidega oli kogutud, siis sünteesiti sõnadele kõik teised muutevormid kasutades EstNLTK süntesaatorit. Tegu on Vabamorfi algoritmiga, mistõttu see tagab sünteesitud vormide korrektsuse, sest algvormide leksikon on samuti Vabamorfi oma. Iga muutevorm on tähistatud lühendiga[19], näiteks kääne nimetav on `n` (lugemine) ja da-tegevusnimi jaatavas kõnes on `da` (lugeda). Käändsõnadel on lisaks käänetele olemas ka ainsuse ja mitmuse tähistus. Pöördsõnadel on arvestatud kõneviisi, tegumoe, aja ning ainsuse ja mitmuse isikutega [Lau+20].

Süntesaatori funktsioon vajab algvormi (`lemma`) ning teatud muutevormi tähistust (`vorm`), et sünteesida sõnale muutevorm. Kõikide vormide sünteesimiseks koguti kokku kõik tähistused, sõltuvalt sõnast kas käänded või pöörded, võeti tsüklina ükshaaval ette ja anti koos algvormiga funktsiooni `synthesize(lemma, algvorm)`. See protsess tehti omakorda läbi kõikide algvormidega, et saada kätte terviklik andemstik, mida närvivõrgule hiljem treenimiseks sisestati.

Ilma muuttüübita sõnad ehk muutumatud sõnad lisati andmestikku koos „-“ muuttüübiga. Ühe muuttüübiga algvormidel polnud probleeme teiste vormide sünteesimisega, sest tagastati järjend muutevormide järjenditest. Mõni muutevormide järjend võis olla tühi, sest sünteesitakse kõik võimalikud vormid, aga sõna käändub ainult ainsuses, näiteks nimi „Tartu“, või mitmuses, näiteks sõna „andmed“. Järjenditest loeti kõik sõnavormid kokku, tühjad järjendid jäeti vahele, ja andmestikku lisati sõna koos kõikide võimalike muutevormidega.

Mitme muuttüübiga ehk rööptüüpidega algvormidel tuli arvestada, millised muutevormid kuuluvad mõlema alla ja millised eristavad teineteist. Õigekeelsussõnaraamatu muuttüübistikust on võimalik uurida läbi tüüpsõnade, millised muutevormid kattuvad põhimuuttüüpidega ja millised on eristatavad. Näiteks sõna „armas“ vastab rööptüübile 3 ja 5 ja esimesed kolm põhivormi on järgmised: muuttüübil 3 – *armas, ‘armsa,* ***‘armsat***, muuttüübil 5 – *armas, ‘armsa,* ***armast***. Järjendis oleksid sõna „armas“ kolm põhivormi järgmiselt: `[[armas], [armsa], [armast, armsat]]`. Paksus kirjastiilis on tähistatud eristatavad muutevormid, teised muutevormid vastavad mõlemale muuttüübile. Sellest lähtuvalt kirjutati funktsioon, mis arvestab süntesaatori poolt loodud järjestusega muutevormide järjendis ja paneks muutevormide järjendis igat eristatavat muutevormi vastama teatud muuttüübile (sh rööptüübile). Sünteesimine ise oli sarnane ühe muuttüübiga sõnadel, sünteesiti sõnavormide järjendid, loeti kõik sõnavormid kokku, tühjad järjendid jäeti vahele, ja andmestikku lisati sõna koos kõikide võimalike muutevormidega.

[19] Lühendite nimekiri. `https://github.com/estnltk/estnltk/blob/main/tutorials/nlp_pipeline/B_morphology/00_tables_of_morphological_categories.ipynb` (15.04.2024)

Leidus käändsõnu ja pöördsõnu, millele süntesaator ei sünteesinud muutevorme. Enamiku moodustasid tegusõnaühendid, mida oli kokku 2844, näiteks „kiiremaks minema“ ja „takistusi tegema“. Nendel pöördus ainult viimane sõna, mistõttu lisati need andmestikku koos muutevormidega juhul, kui viimast sõna polnud varem kohatud. Lisaks oli andmestikust välja jäetud veel 142 sõna, mille alla kuulusid sõnaühendiga pärisnimed, näiteks „Abu Dhabi“ ja „Baltic News Service“, sõnaühendiga nimisõnad, näiteks „ema vend“, „isa õde“ ja „nimetu sõrm“, ning veel teised tavapäratud sõnad, „m-s“, „n-s“ ja „ignoma“.

Selle alampeatüki kood on kättesaadav järgmiselt veebilehekülgelt[20].

3.3 Tehisnärvivõrk

Terviklik andmestik koostatud, viidi see järgmisena arvkujule, sest närvivõrgud kasutavad erinevaid matemaatilisi tehteid, näiteks maatrikstehteid, mida on võimalik teostada ainult arvudega [GBC16]. See tähendab, et sõnad tuli teisendada tähejadaks, kus iga täht tuli omakorda märgistada mingi täisarvuga, ning muuttüübid ja sõnaliigid kahendarvude jadaks, kus iga kahendarv tähistab kindla muuttüübi olemasolu (1 – on selle muuttüübiga sõna, 0 – ei ole). Kahendarvuks sellepärast, et tegu oli klassifitseerimisülesandega, kus muuttüübid on kindlad klassid, mida mudel tuvastab. Andmete teisendamiseks ja mudeli loomiseks kasutati avaliku lähtekoodiga teekidekogumiku Tensorflow[21] ja scikit-learn[22].

Tensorflow kihiga `TextVectorization`[23] teisendati tekstid numbrite jadaks. Parameetritega täpsustati kui pikk on väljundiks tekkiva jada pikkus ehk vektori suurus ja sõnastiku koostamisel eraldaks funktsioon tekstis üksusi tähthaaval. Teised parameetrid jäeti vaikeväärtusteks. Kiht kohandati üle andmestiku sõnade kasutades kihi meetodit `adapt`, mis koostab sõnastiku ja paneb iga tähe vastavusse mingi täisarvuga. Kihis tekkiva sõnastiku mahtu polnud vaja piirata, sest see koosnes eesti tähestikust ja lisaks võõrtähtedest, mida polnud palju. Kohandatud kihiga sai hiljem teisendada hulkadesse jaotatud sõnad ümber numbrite jadaks ehk numbervektoriteks.

Scikit-learni funktsioon `MultiLabelBinarizer`[24] teisendab muuttüübid ja sõnaliigid kahendarvude jadaks. Meetodiga `fit` kohandati funktsioon üle muuttüüpide/sõnaliikide jada. Meetodis loetakse kõik unikaalsed väärtused kokku ühte arvu ja igale sisendina antud muuttüübile/sõnaliigile koostatakse selle arvu pikkusega kahendjada, kus igal indeksil asuv

[20] Morfoloogilise muuttüübi automaatne tuvastamine - II osa. https://github.com/SanderSaska/MorfoloogiliseMuuttyybiAutomaatneTuvastaja/blob/main/Morfoloogilise_muutt%C3%BC%C3%BCbi_automaatne_tuvastamine_II_osa.ipynb (14.05.2024)

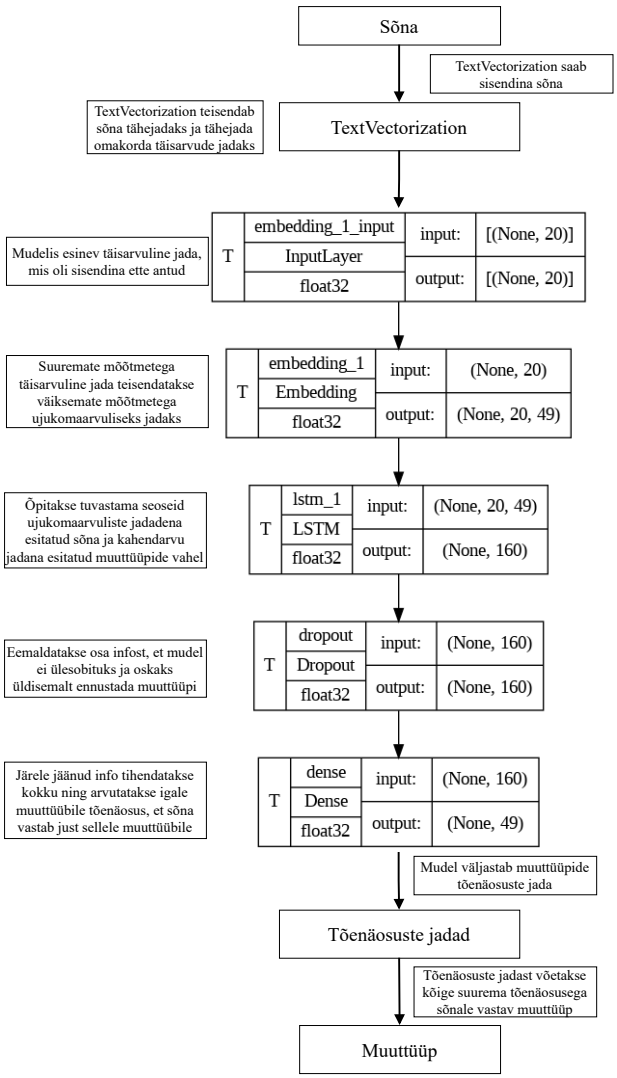
[21] Tensorflow. https://www.tensorflow.org/ (14.05.2024)

[22] scikit-learn. https://scikit-learn.org/stable/ (14.05.2024)

[23] Tensorflow kihi TextVectorization API. https://www.tensorflow.org/api_docs/python/tf/keras/layers/TextVectorization (14.05.2024)

[24] Scikit-learni funktsioon MultiLabelBinarizer. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html (14.05.2024)

kahendarv vastab unikaalse muuttüübi/sõnaliigi olemasolule. Kohandatud funktsiooniga sai hiljem teisendada hulkadesse jaotatud muuttüübid/sõnaliigid ümber kahendarvude jadaks.



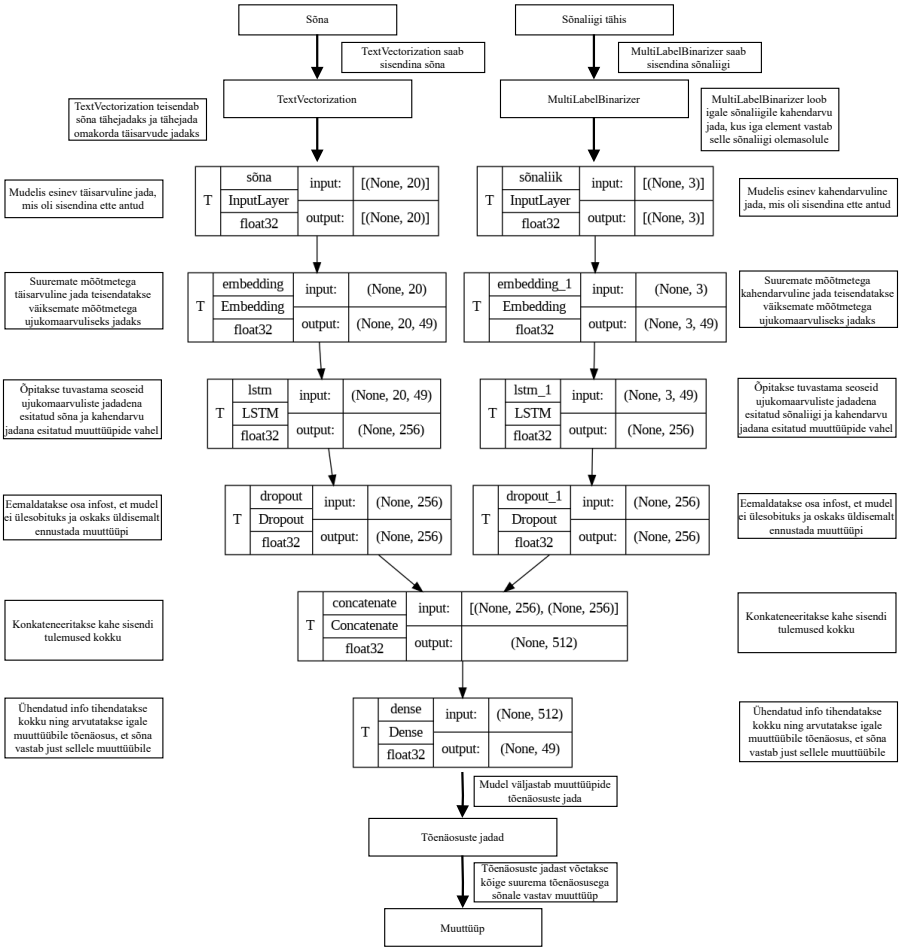
Joonis 10. Sõnamudeli ülesehitus

Eksperimenteeriti kahe närvivõrgu alusmudeliga: sõnavormide pealt muuttüüpi ennustav alusmudel ehk sõnamudel ning sõnavormide ja sõnaliikide pealt muuttüüpi ennustav mudel ehk sõnamudel sõnaliigiga. Mõlemal juhul tugineti LSTM-ile, et tuvastada seoseid tähejadade/sõnaliigitähiste ja muuttüüpide vahel.

Sõnamudel (joonis 10) koosneb viiest kihist[25]: InputLayer, Embedding, LSTM, Dropout ja Dense. Joonisel on iga kiht tähistatud vasakult T kastikesega ja lisaks on kõrval kastides täiendavat informatsiooni. Kihi kasti vasak osa kirjeldab kihile antud nime, kihi

[25] Kihtide kirjeldused. `https://www.tensorflow.org/api_docs/python/tf/keras/layers` (14.05.2024)

API nime ja andmetüüpi. Kihis InputLayer pole andmetüüp määratud, mistõttu pannakse sinna vaikeväärtusena `float32`. Kihi kasti parem osa kirjeldab ploki suurust (*batch size*) koos sisendandme ja väljundandme mõõtmetega. Mudeli ülesehitamisel ploki suurust pole määratud, sellepärast on kõigil mõõtmete alguses `None` väärtus. Ploki suurust määrati mudeli treenimisel.



Joonis 11. Sõnamudel sõnaliigiga ülesehitus

Peale selle loodi mudel, mis võtab sisendiks lisaks sõnale ka sõnaliigi (joonis 11) uurimaks, kas lisasisendiga õpib mudel täpsemalt ennustama muuttüüpe. Mudeli ülesehitus on erinev, mõlemale sisendile lisatakse kihid InputLayer, Embedding, LSTM ja Dropout. Dropout kihi väljundid ühendatakse Concatenate kihiga ja Dense kihiga väljastatakse tulemus.

Selle alampeatüki kood on kättesaadav järgmiselt veebileheküljelt[26] defineeritud funktsioonide alampeatükist.

[26]Morfoloogilise muuttüübi automaatne tuvastamine - III osa. `https://github.com/SanderSaska/MorfoloogiliseMuuttyybiAutomaatneTuvastaja/blob/main/Morfoloogilise_muutt%C3%BC%C3%BCbi_automaatne_tuvastamine_III_osa.ipynb` (14.05.2024)

4 Masinõppe eksperimendid

Selles peatükis kirjeldatakse alusmudeli hüperparameetrite häälestamist eesmärgiga leida mudel, mis ennustab muuttüüpi kõige täpsemalt. Esimeses alampeatükis õpib mudel ennustama ainult sõnade põhjal, teises alampeatükis lisatakse juurde sõnaliik, et uurida, kuidas uue sisendi lisamine mõjutab mudeli täpsust.

4.1 Sõnamudeli katsetamine

Andmed jagati treening-, test- ja valideerimishulkadesse kasutades scikit-learni funktsiooni `train_test_split`. Protsentuaalselt jagati kogu andmestik järgmiselt: treeninghulka 80%, testimishulka 10% ja valideerimishulka 10%. Muuttüüpide sagedustabelitest on näha, et muuttüüpidele vastavaid sõnu on andmestikus ebaproportsionaalselt. Seetõttu väärtustati `stratify` parameetris andmestiku muuttüüpide jada, mis jagab andmed kihiliselt hulkadesse. Kihiliselt selles mõttes, et igas hulgas on peaaegu sama protsent muuttüüpi esindavaid sõnu nagu protsent terves andmestikus. Näiteks kui terves andmestikus oli muuttüüpi 22 esindavaid sõnu umbes 18%, siis igas hulgas on muuttüüpi 22 esindavaid sõnu samuti 18%.

Hüperparameetrite häälestamiseks on esmalt kasutatud võrkotsingu (*grid search*) meetodit. Selle jaoks määratakse igale hüperparameetrile, mida soovitakse muuta, väärtuste jada. Seejärel võetakse kõik võimalikud väärtuste kombinatsioonid ja iga kombinatsiooniga treenitakse uut mudelit. Näiteks on kaks hüperparameetrit A ja B, väärtuste jadad on järgmised $A = [5, 10, 15, 20, 25]$ ja $B = [0, 0.1]$. Esimene väärtuste kombinatsioon on $\{5, 0\}$. Seega tuleb treenida $|A| \cdot |B| = 5 \cdot 2 = 10$ erinevat mudelit, et kõik võimalikud väärtuste kombinatsioonid mõlemast jadast oleksid kaetud.

Selle meetodi suurim kitsaskoht on ajakulu, sest hüperparameetrite ja väärtuste arv mõjutab treenitavate mudelite arvu ning kui andmestik on suur, peaaegu kaks ja veerand miljonit sõna, siis seda enam tuleb kaaluda, milliseid väärtusi on mõistlik katsetada.

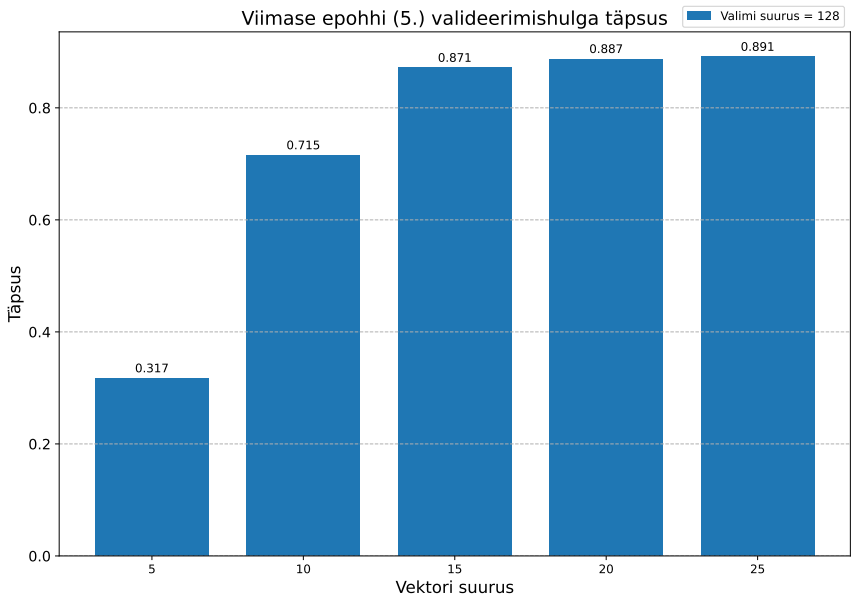
Esmalt katsetati vektori suuruse muutmist väärtustega 5, 10, 15, 20 ja 25. Vektori suurus muudab sõna tähejärjendi pikkust ja on mudeli väline hüperparameeter. Suurem vektori suurus annab suurema ülevaate sõnast, kuid see tähendab, et on ka rohkem infot, mille pealt mudel peab õppima. Näiteks vektori suurusega 5 näeb mudel sõnast „raudteejaam“ ainult tähejärjendit `[e, j, a, a, m]`, aga suurusega 10 `[a, u, d, t, e, e, j, a, a, m]`. Vektori suurus muudab otseselt infohulka ja mõjutab ajakulu (tabel 1). Hüperparameetri väärtuse tõstmine 5 võrra kasvatas ajakulu CPU-ga umbes 7 minuti võrra, aga GPU-ga ainult ligikaudu 20 sekundit, välja arvatud suurusega 25.

Täpsuste vahe oli märkimisväärne esimese kolme vektori suurustega 5, 10 ja 15 (joonis 12). Keskmine sõna pikkus andmestikus on 12 tähte. Alates suurusest 15 hakkas

Vektori suurused	5	10	15	20	25	Kokku
CPU ajakulu (s)	851	1373	1756	2193	2527	8700
GPU ajakulu (s)	664	681	698	724	677	3444

Tabel 1. Sõnamudeli treenimise ajakulud

täpsuse kasv taanduma, vektori suuruste 20 ja 25 täpsuste vahe oli ainult 0.4%. Mudeil sisemiste hüperparameetrite katsetamiseks võeti vektori suurus 20, mis oli piisavalt täpne ja väiksema ajakuluga kui riistvaraga CPU vektori suurusega 25. Täpsused valideerimishulga ja testhulga vahel erinesid väga vähe, tavaliselt 0.1%. Põhjus võib olla sõna ja tema muutevormide laialijagamine hulkadesse. Kui mudelis on muuttüüpi õpetatud mingite sõnavormide peal, siis peaks olema lihtsam ennustada teistel hulkadel samat muuttüüpi sama sõnatüve tõttu.



Joonis 12. Sõna sisendiga mudeli täpsused

Järgmisena katsetati mudeli sisemisi hüperparameetreid, täpsemalt A väljajätumeetodit (*dropout*), B õpisammu (*learning rate*) ja C LSTM kihi neuronite arvu. Hüperparameetrite väärtuste jadad olid järgmised $A = \{32, 64, 92, 128, 160, 192, 224, 256\}$, $B = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ ja $C = \{0.0001, 0.001, 0.01\}$. Võrkotsingu meetodit niipaljude väärtuste katsetamiseks ei sobi, sest treenida tuleks $|A| \cdot |B| \cdot |C| = 8 \cdot 6 \cdot 3 = 144$ erinevat mudelit viiel epohhil, mis oleks väga ajakulukas. Seetõttu võeti kasutusele hüperparameetrite häälestaja Hyperband [Li+18]. Algoritm treenib mõne epohhi jooksul

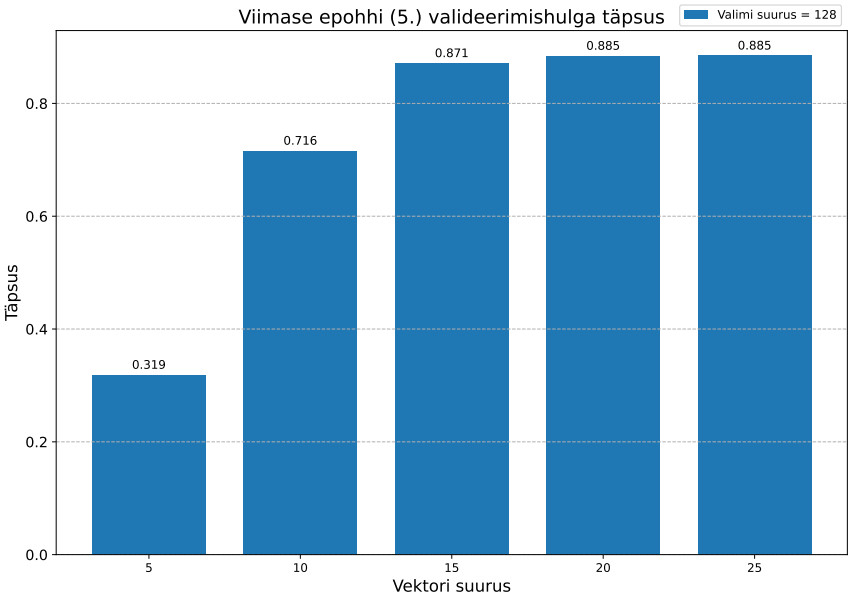
arvukaid mudeleid ja kannab järgmisesse vooru edasi ainult poole mudelitest, mis olid paremate tulemustega. Esimeses voorus treenitakse ainult kahe või kolme epohhiga, igas järgmises voorus tõstetakse epohhide arvu. Hüperparameetrite häälestajaga Hyperband kasvas mudeli täpsus $\sim 7\%$ võrra ja saadi sõnamudeli täpsuseks 95.8%.

4.2 Sõnamudel sõnaliigiga katsetamine

Sarnaselt tavalisele sõnamudelile toimus andmestiku jaotamine samamoodi ning kasutati võrkotsingu meetodit vektori suurusega ja samade väärtustega. Sõnaliikideks võeti kõik sõnastikus esinevad sõnaliigid (tabel 4 lisas), seal hulgas mitme sõnaliigiga klassid, näiteks sõna issand on nii nimisõna (muuttüüp 2) kui ka hüüdsõna (muuttüüp -). Uue sisendi lisamine tõstis ajakulu märgatavalt (tabel 2) nii riistvara CPU kui ka GPU peal. Kui sõnamudelil riistvaraga GPU tõusis ajakulu ligikaudu 20 sekundi võrra vektori suuruse tõstmisega, siis sõnaliigi lisamisega tõusis ajakulu umbes 6 minuti ja 20 sekundi võrra.

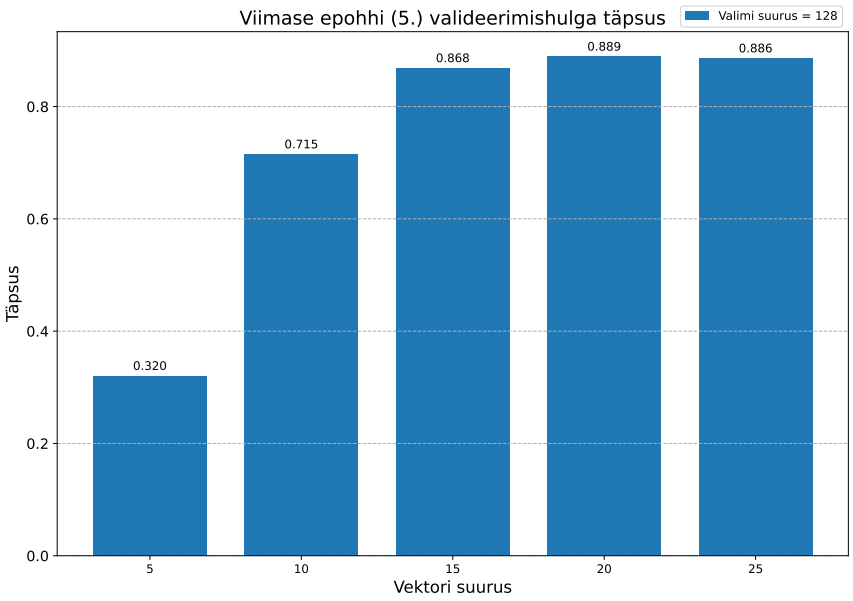
Vektori suurused	5	10	15	20	25	Kokku
CPU ajakulu (s)	1642	1825	2072	2449	2839	10827
GPU ajakulu (s)	859	1261	1640	1992	2378	8130

Tabel 2. Sõnamudeli sõnaliigiga treenimise ajakulud



Joonis 13. Sõna ja sõnaliigi sisendiga mudeli täpsused

Sõnaliigi lisamine mudeli täpsust paremaks ei teinud. Põhjus võib seisneda liiga täpses sõnaliigi klassifitseerimises ja sõnaliigi valesti määramises sünteesimise käigus. Sõnad, millel on üks muuttüüp ja mitu sõnaliiki ning üks sõnaliikidest on muutumatute rühmas, sünteesiti muutevormid, kuid muutevormidele määrati mõlemad sõnaliigid, mida ei tohiks. Seetõttu katsetati üldisemat klassifitseerimist: käändsõnad, pöördsõnad ja muutumatud sõnad, sest sedasi on muuttüüpe klassifitseeritud ka ÕS-is.



Joonis 14. Sõna ja üldisema sõnaliigi sisendiga mudeli täpsused

Üldisem sõnaliigi klassifitseerimine parandas täpsust vähe, mõne vektori suurusega isegi vähendas. Sellegi poolest sai sõnaliigiga sõnamudelitest kõige parema tulemuse mudel vektori suurusega 20 ja üldisema sõnaliigi sisendiga. Sarnaselt sõnamudelile häälestati samu mudeli sisemisi hüperparameetreid häälestajaga Hyperband, mis tõstis mudeli täpsust märgatavalt, peaaegu 9%. Lõpptulemuseks saavutati sõnaliigiga sõnamudel täpsusega 97.8%.

5 Tulemuste analüüs

Valminud sõnamudelitega (*.keras* failid) uuriti, kui hästi suudavad mudelid ennustada muuttüüpi sõnaliigiti ja muuttüübiti (tabelid 8, 9, 10, 11). Sõnamudel on võimeline ennustama muuttüüpe käändsõnadele peamiselt üle 95% täpsusega, pöördsõnadele umbes 95% täpsusega, kuid muutumatute sõnadega jääb hätta. Näiteks määrsõnu, mida on andmestikus 9.15%, ennustas sõnamudel õiget muuttüüpi ainult 14.3%. Pöördsõnadel on täpsus kehvem, kuna pöördvorme on rohkem kui käändevorme – pöördevorme on maksimaalselt 48, käändevorme maksimaalselt 29 (lühike sisseütlev kaasa arvatud). Mudelitel on vaja ära tunda palju vorme, mistõttu tekib raskusi sõnaliigi tuvastamisega, et tagastada õige muuttüüp.

Samas sõnaliigi lisamisega ennustas sõnamudel kõikidele sõnadele muuttüüpi keskmiselt 95% täpsusega. Vahe tuleneb mudelite ülesannetes – tavaline sõnamudel peab lisaks muuttüübi ennustamisele tuvastama ka sõnaliiki, ent sõnaliigi lisamisega saab mudel keskenduda peamiselt muuttüübi ennustamisele. Tuleb arvestada, et sõnaliik on nagu oraakel mudelile ja see on inimese poolt määratud. Kui tahetakse sõnaliigiga sõnamudelil ennustada muuttüüpi, siis tuleb anda lisaks sõnale kaasa ka üldisem sõnaliik – kas tegu on käänduva, pöörduva või muutumatu sõnaga. Sõnamudelil tuleb arvestada, et ennustatav sõna oleks kindlasti käänduv või pöörduv, sest muutumatutele sõnadele tuvastab mudel enamasti vale muuttüübi.

Sõnaliikide tabelist 8 (asub lisades) võib välja lugeda, et ühe sõnaliigiga muutumatudele sõnadele ennustati 100% täpsusega muuttüüpe, ent kui arvata kaasa ka mitme sõnaliigiga, mis langesid kokku käänd- või pöördsõnadega, siis täpsus langes 95.2%-ni. Erinevus võib tuleneda mitme sõnaliigiga sõnadele sünteesitud muutevormidest, millele on pandud mõlemad sõnaliigid. See pole aga alati õige, eriti sõnadel nagu „issand“, millel on ainult üks muuttüüp, aga mitu sõnaliiki – võib mõelda kui hüüdsõna, mis on muutumatu ehk muutevorme pole, ja nimisõna, mis on käänduv ehk muutevormid on olemas.

	Sõnade arv	Sõnamudeli täpsus	Sõnaliigiga sõnamudeli täpsus
Algvormid	6988	75.94%	91.7%
Teised vormid	217063	96.41%	97.88%

Tabel 3. Algvormide ja teiste muutevormide ennustamistäpsus mudelitel

Algvormidele ja teistele muutevormidele muuttüübi ennustamise täpsus erines märgatavalt (tabel 3). Sõnamudel ennustas muuttüüpi algvormile $\sim 20\%$ kehvemini kui teistele muutevormidele, sõnaliigiga sõnamudel $\sim 6\%$ kehvemini. On võimalik välja lugeda, et kui algvormile pole sõnaliiki määratud, siis mitme erineva sõnaliigiga sõnade puhul ei

oska mudel määrata, kas sisendsõna all mõeldakse käänduvat, pöörduvat või muutumatut sõna. Sarnane olukord tekib ka teistel muutevormidel. Teatud muuttüübiga käändsõnade kirjapilt võib kattuda teatud muuttüübiga pöördsõnade kirjapildiga või lausa muutumatu kirjapildiga. Näiteks sõna „targalt“ võib olla muutumatu kui ka muutuv sõna – sõna „tark“ alaltütlevas käändes. Teine näide on sõna „sulge“, mis võib olla osastav kääne käändsõnast „sulg“ või käskiva kõneviisi ainsuse 2. pööre pöördsõnast „sulgema“. Ilma kontekstita on raske tuvastada sõnale sõnaliiki.

Sellegi poolest, kuna mudelid on üsna täpsed, siis võiks olla neist kasu keeleteadlastel uute või murdesõnade analüüsimisel – abivahend muuttüüpide määramisel, mis vähendab ajakulu, kuid lõppotsuse teeb ikkagi kasutaja ise. Samuti saavad mudeleid kasutada ka keelehuvilised õppimiseks, ent see vajab programmeerimisoskusi ja täielikult neid mudeleid usaldada ei saa – pole 100% täpsusega mudelid.

6 Edasiarendusvõimalused

Edasiarendamise võimalusi on mitmeid. Saaks testida mudeleid eri liiki tekstidel, näiteks veebikeelega, kus on sagedased kirjavead ja släng, ja vana eesti kirjakeelega, mille sõnavara erineb tänapäevasest ning õigekiri on ebaühtlane. Peamine mõte on uurida, kui hästi suudab mudel ennustada muuttüüpe täiesti tundmatudele sõnadele ja kas mudelid on ülesobitatud. Selleks tuleb luua tundmatute sõnade andmestik, kus on sõnadele käsitsi määratud muuttüübid, mis nõuab aga lingvistilist ekspertteadmist ja jääb selle töö skoobist välja.

Testimisele lisaks saaks parandada või muuta lähenemist andmestiku koostamisele – on võimalik otsida lahendust maha jäetud sõnadele ja sõnaliigi valesti määramisele sünteesimise käigus. Sõnaliigi valesti määramist saab parandada, kui ühe muuttüübiga muutevormide sünteesimisel määrata muutevormidele ainult see sõnaliik, mille sõnad on muutuvad. Näiteks sõna „issand“ on nimisõna ja hüüdsõna ning ühe muuttüübiga. Sünteesimise käigus andmestikku lisamisel antakse kaasa mõlemad sõnaliigid kõikidele vormidele, kuid tegelikult peaks ainult algvormile andma mõlemad sõnaliigid ja muutevormidele sõltuvalt sõnaliigist, kas üks või mõlemad. Sõna „issand“ puhul peab andma ainult nimisõna sünteesitud muutevormidele.

Samuti saab uurida, kuidas mõjutab täpsust teistsugune jagamismeetod hulkadesse – kihilise jagamise asemel saaks jagada algvormi gruppide kaupa, kus ühes grupis on esindatud algvorm koos oma kõikide võimalike muutevormidega. Üks sõna koos selle kõigi vormidega oleks rangelt kas treening-, valideerimis- või testhulgas, mitte laiali erinevate hulkade vahel. Hiljem saab uurida, kas selline jagamismeetod parandas täpsust või mitte.

Ambitsioonikam edasiarendus on luua mudel, mis suudaks genereerida sõnale teisi muutevorme. Saaks uurida, kas praegune mudel, millest saab kätte muuttüübi, hõlbustaks tundmatule sõnale automaatset käände- või pöördevormide genereerimist.

7 Kokkuvõte

Inimestena suudame sageli varasema keelekogemuse põhjal oletada, kuidas uusi sõnu käänata või pöörata. Lingvistid on meie keelekogemuse põhjal pakkunud välja teoreetilised muuttüübid, mis grupeerivad samamoodi käänduvaid või pöörduvaid sõnu. Seetõttu uuris käesolev töö, kui täpselt suudab tehisnärvivõrgul põhinev mudel tuvastada muuttüüpe toetudes VVS tüübikirjeldustele ja millisest infost piisab tehisnärvivõrgule muuttüübi tuvastamisel.

Mudeleid treeniti ja testiti Vabamorfi põhileksikoni kirjetega. Igast kirjest eraldadi algvorm, muuttüüp ja sõnaliik. Algvormidele sünteesiti kõik võimalikud muutevormid ja muutevormidele määrati muutüüp ja sõnaliik vastavalt leksikoni kirjetele.

Andmete eeltöötlemisel teisendati sõnavormid tähejadadeks ja tähejadad omakorda täisarvulisteks jadadeks, kus igale arvule vastab kindle täht. Muuttüüp ja sõnaliik teisendati kahendarvude jadadeks, kus number 1 jadas tähendab kindla muuttüübi või sõnaliigi olemasolu. Jadad jaotati kihiliselt treening, test- ja valideerimishulkadesse protsentidega $80\% : 10\% : 10\%$. Kihiliselt sellepärast, et mudel oleks treenimisel teadlik kõikidest muuuttüüpidest.

Võimalikult täpsete mudelite saavutamiseks loodi alusmudelid, mille peal häälestati mudeli sisemisi kui ka välimisi hüperparameetreid. Esmalt kasutati võrkotsingu meetodit mudeli välimise hüperparameetriga vektori suurus ehk tähejada pikkus, mis andis $\sim 88\%$ täpsuseid mudeleid. Hiljem kasutati Hyperband häälestajat mudeli sisemiste hüperparameetrite väljajätumeetodi, õpisammu ja LSTM kihi neuronite arvu häälestamiseks, mis andis täpsuseks $\sim 95\%$.

Töös valmisid kaks muuttüüpi ennustavat mudelit, millest üks tugineb ainult sõnasisendile ning teine sõnasisendile ja inimese poolt määratud sõnaliigisisendile – käändsõna, pöördsõna või muutumatu sõna. Sõnasisendiga mudeli täpsuseks saadi 95.8% ja koos sõnaliigisisendiga mudeli täpsuseks 97.8%. Sõnasisendiga mudel oskab ennustada muuttüüpi käändsõnadele üle 95% täpsusega, pöördsõnadele umbes 95%, kuid muutumatute sõnadega jääb hätta. Sõna- ja sõnaliigisisendiga mudel on võimeline ennustama muuttüüpe kõikidele sõnadele täpsusega $\sim 95\%$.

Mudelist võiks olla kasu keeleteadlastel uute või murdesõnade analüüsimisel ja keelehuvilistel sõnade õppimisel. Lõppotsuse muuttüübi määramisel teeb ikkagi kasutaja ise.

Viidatud kirjandus

[Ees22] Eesti Keele Instituut. Kuidas uusi sõnu teha? 2022. https://eki.ee/teatmik/kuidas-uusi-sonu-teha/ (14.05.2024).

[Raa23] M. Raadik. Uudisrohesõnade korje 2023. 2023. https://eki.ee/teatmik/uudisrohesonade-korje-2023/ (14.05.2024).

[EER20] M. Erelt, T. Erelt ja K. Ross. Eesti keele käsiraamat. Uuendatud väljaanne. Tallinn: EKSA, 2020. http://www.eki.ee/books/ekkr20/ekkr20.pdf (14.05.2024).

[Ere+18] T. Erelt, T. Leemets, S. Mäearu ja M. Raadik. Eesti õigekeelsussõnaraamat: ÕS 2018. Tallinn: EKSA, 2018. http://www.eki.ee/dict/qs2018/qs2018.html (14.05.2024).

[Vik15] Ü. Viks. Muuttüübid ja ÕS 2013. *Oma Keel 2*, 2015, nr 31, lk 44–54. https://www.emakeeleselts.ee/omakeel/2015_2/06.pdf (14.05.2024).

[Eesa] Eesti Keele Instituut. Tüübituvastus. s.a. https://www.eki.ee/tarkvara/tyybituvastus/ (14.05.2024).

[Eesb] Eesti Keele Instituut. Morfoloogiamoodulite töö üldpõhimõtted. s.a. https://www.eki.ee/tarkvara/morf_yld.html (14.05.2024).

[Lju+15] N. Ljubešić, M. Esplà-Gomis, F. Klubička, and N. M. Preradović. Predicting inflectional paradigms and lemmata of unknown words for semi-automatic expansion of morphological lexicons. *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, 09/2015, pp. 379–387. https://aclanthology.org/R15-1050 (14.05.2024).

[DR12] G. Détrez and A. Ranta. Smart paradigms and the predictability and complexity of inflectional morphology. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, 04/2012, pp. 645–653. https://aclanthology.org/E12-1066.pdf (14.05.2024).

[KPV22] H.-J. Kaalep, R. Prillop ja T. Vaino. Vabamorfi morfoloogia-leksikon. Filosoft. 2022. https://github.com/Filosoft/vabamorf/blob/master/doc/morfi_leksikoni_kirjeldus.md (14.05.2024).

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. MIT Press, 2016. Chap. 2, 6, 10, pp. 29, 164, 367. http://www.deeplearningbook.org (14.05.2024).

[LeN19] A. LeNail. NN-SVG: Publication-Ready Neural Network Architecture Schematics. *Journal of Open Source Software*, 2019, 33, p. 747. https://doi.org/10.21105/joss.00747 (14.05.2024).

[Zha+23] A. Zhang, Z. C. Lipton, M. Li ja A. J. Smola. Dive into Deep Learning. Cambridge University Press, 2023. https://D2L.ai (14.05.2024).

[Lau+20] S. Laur, S. Orasmaa, D. Särg ja P. Tammo. EstNLTK 1.6: Remastered Estonian NLP Pipeline. Teoses: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 05.2020, lk 7154–7162. https://www.aclweb.org/anthology/2020.lrec-1.884 (14.05.2024).

[Li+18] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh ja A. Talwalkar. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 2018, 185, lk 1–52. http://jmlr.org/papers/v18/16-558.html (14.05.2024).

Lisad

Tabelid

Tabel 4. Algvormide sagedustabel, Vabamorfi sõnastiku sõnaliigid

Sõnaliik	**Sagedus**	%	**Näide**
Nimisõnad	40625	55.03%	ülikool
Tegusõnad	10724	14.53%	õppima
Omadussõna algvõrre	10336	14.0%	tark
Määrsõna	6460	8.75%	targasti
Pärisnimi	3777	5.12%	Tartu
Käändumatu omadussõna	584	0.79%	eesti
Omadussõna keskvõrre	414	0.56%	targem
Hüüdsõna	336	0.45%	hei
Kaassõna	204	0.28%	ääres
Põhiarvsõna	105	0.14%	üks
Asesõna	102	0.14%	teie
Omadussõna, ülivõrre	70	0.09%	kõige targem, targim
Järgarvsõna	48	0.06%	esimene
Sidesõna	23	0.03%	ja
Verbi juurde kuuluv sõna	21	0.03%	plehku
Kokku	73829	100.0%	

Märkus: mitme sõnaliigiga sõnadest on ainult esimene sõnaliik esindatud.

Tabel 5. Algvormide sagedustabel, Vabamorfi sõnastiku käändsõnade muuttüübid

Muuttüüp	**Sagedus**	%	**Näide**
22	13941	18.883%	riik
2	11260	15.252%	õpik
1	4640	6.285%	ohutu
12	4286	5.805%	oluline
11	3890	5.269%	harjutus
16	2722	3.687%	pere
11, 9	2462	3.335%	raskus
17	1721	2.331%	elu
25	1514	2.051%	õnnelik
26	1256	1.701%	idee
10	1085	1.47%	üldine
6	1063	1.44%	mõte
12, 10	903	1.223%	üldine
9	586	0.794%	katus
19	552	0.748%	seminar
13	531	0.719%	suur
7	455	0.616%	kallas
24	399	0.54%	puri
18	396	0.536%	nägu
5	227	0.308%	liige
23, 22	199	0.27%	poiss

Tabel jätkub järgmisel lehel

Tabeli 5 järg

Muuttüüp	**Sagedus**	%	**Näide**
19, 2	187	0.253%	akvaarium
14	184	0.249%	uus
20	108	0.146%	süli
4	106	0.144%	ase
21	99	0.134%	jõgi
8	79	0.107%	küünal
15	70	0.095%	käsi
3, 5	25	0.034%	armas
23	19	0.026%	poiss
10, 12	14	0.019%	üldine
3	8	0.011%	vaher
5, 7	5	0.007%	kallis
Kokku	54993	74.488%	

Tabel 6. Algvormide sagedustabel, Vabamorfi sõnastiku pöördsõnade muuttüübid

Muuttüüp	**Sagedus**	%	**Näide**
27	5581	7.559%	elama
28	2639	3.575%	õppima
29	722	0.978%	hüppama
36	551	0.746%	tulema
34	463	0.627%	saatma
30	214	0.29%	vaidlema
38	169	0.229%	käima
37	153	0.207%	võima
35	112	0.152%	petma
31, 27	38	0.051%	rabelema
32	37	0.05%	seisma
28, 27	31	0.042%	sulgema
33	11	0.015%	naerma
Kokku	10721	14.521%	

Tabel 7. Algvormide sagedustabel, Vabamorfi sõnastiku muud muuttüübid

Muuttüüp	**Sagedus**	%	**Näide**
-	8061	10.918%	targasti
0	54	0.073%	ise

Märkus: muuttüüpide näited, v.a muud muuttüübid, on võetud 2018 ÕS-i raamatust [Ere+18].

Tabel 8. Algvormide sagedustabel koos mudelite täpsustega, sõnaliigid andmestikus pärast sünteesimist

Algvormid koguandmestikus			**Täpsus testandmestikul**	
Sõnaliik	**Sagedus**	%	**Sõnamudel**	**Sõnaliigiga sõnamudel**
Nimisõnad	40394	57.21%	96.9%	97.8%
Omadussõna algvõrre	10333	14.64%	95.9%	97.5%
Tegusõnad	7875	11.15%	95.0%	98.5%
Määrsõna	6457	9.15%	14.3%	100.0%
Pärisnimi	3639	5.15%	89.4%	93.1%
Käändumatu omadussõna	584	0.83%	0.53%	100.0%
Omadussõna keskvõrre	414	0.59%	96.8%	99.5%
Hüüdsõna	336	0.48%	25.0%	100.0%
Kaassõna	204	0.29%	0.0%	100.0%
Põhiarvsõna	102	0.14%	94.8%	95.4%
Asesõna	100	0.14%	83.1%	85.0%
Omadussõna, ülivõrre	70	0.1%	97.8%	99.6%
Järgarvsõna	48	0.07%	95.9%	97.7%
Sidesõna	23	0.03%	100.0%	100.0%
Verbi juurde kuuluv sõna	21	0.03%	0.0%	100.0%
Kokku / mikrotäpsus	70600	100%	95.8%	97.8%

Märkus: mitme sõnaliigiga sõnadest on ainult esimene sõnaliik esindatud.

Tabel 9. Algvormide sagedustabel koos mudelite täpsustega, käändsõnade muuttüübid andmestikus pärast sünteesimist

Algvormid koguandmestikus			**Täpsus testandmestikul**	
Muuttüüp	**Sagedus**	%	**Sõnamudel**	**Sõnaliigiga sõnamudel**
22	13708	19.416%	97.9%	98.8%
2	11219	15.891%	96.5%	98.1%
1	4629	6.557%	93.1%	94.9%
12	4281	6.064%	99.2%	99.3%
11	3889	5.508%	99.1%	99.5%
16	2685	3.803%	89.6%	92.9%
11, 9	2462	3.487%	96.5%	98.4%
17	1721	2.438%	94.1%	96.5%
25	1514	2.144%	99.4%	99.7%
26	1240	1.756%	95.7%	97.1%
10	1085	1.537%	96.0%	95.6%
6	1055	1.494%	92.0%	92.3%
12, 10	902	1.278%	95.9%	96.2%
9	582	0.824%	94.1%	95.5%
19	545	0.772%	96.1%	97.2%
13	531	0.752%	92.5%	96.6%
7	453	0.642%	93.7%	95.5%
24	397	0.562%	91.7%	95.3%
18	393	0.557%	92.9%	96.5%
5	226	0.320%	78.9%	87.4%

Tabel jätkub järgmisel lehel

Tabeli 9 järg

Algvormid koguandmestikus			**Täpsus testandmestikul**	
Muuttüüp	**Sagedus**	%	**Sõnamudel**	**Sõnaliigiga sõnamudel**
23, 22	199	0.282%	87.8%	91.5%
19, 2	187	0.265%	96.1%	97.2%
14	184	0.261%	87.7%	92.9%
20	107	0.152%	90.6%	95.0%
4	105	0.149%	93.5%	93.7%
21	99	0.14%	93.0%	93.7%
8	79	0.112%	94.2%	92.2%
15	70	0.099%	95.4%	98.3%
3, 5	22	0.031%	81.5%	92.4%
23	19	0.027%	96.0%	98.2%
10, 12	14	0.02%	80.0%	70.0%
3*	11	0.016%	65.7%	71.4%
5, 7	5	0.007%	66.7%	83.3%
Kokku	54618	77.363%		

* Kirjeid on tekkinud juurde, kuna süntesaator on osadele algvormidele sünteesinud kaks sama kirjapildiga sõnavormi ja andmestiku on pandud samad sõnavormid erinevate sõnaliikidega.

Tabel 10. Algvormide sagedustabel koos mudelite täpsustega, pöördsõnade muuttüübid andmestikus pärast sünteesimist

Algvormid koguandmestikus			**Täpsus testandmestikul**	
Muuttüüp	**Sagedus**	%	**Sõnamudel**	**Sõnaliigiga sõnamudel**
27	5029	7.123%	97.3%	98.9%
28	1957	2.772%	92.6%	98.6%
29	518	0.734%	88.5%	97.4%
30	186	0.263%	94.7%	99.2%
34	70	0.099%	76.6%	90.9%
31, 27	37	0.052%	93.3%	100.0%
36	16	0.023%	73.7%	90.8%
28, 27	14	0.02%	61.8%	61.8%
35	11	0.016%	53.8%	76.9%
38	11	0.016%	46.2%	65.4%
32	10	0.014%	60.9%	78.3%
33	7	0.01%	71.9%	75.0%
37	6	0.008%	35.7%	75.0%
Kokku	7872	11.15%		

Tabel 11. Algvormide sagedustabel koos mudelite täpsustega, Vabamorfi sõnastiku muud muuttüübid

Algvormid koguandmestikus			**Täpsus testandmestikul**	
Muuttüüp	**Sagedus**	%	**Sõnamudel**	**Sõnaliigiga sõnamudel**
-	8058	11.413%	13.3%	95.2%
0	52	0.074%	65.5%	67.6%

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Sander Saska**,

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

   Morfoloogilise muuttüübi automaatne tuvastamine,

   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 4.0, 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.

Sander Saska
15.05.2024