TARTU ÜLIKOOL
Arvutiteaduse instituut
Andmeteaduse õppekava

Kristin Petersel

Finantsteemaliste artiklite lihtsustamine

Magistritöö (15 EAP)

Juhendajad: Mark Fišel, PhD
Lehar Oha

Tartu 2022

Finantsteemaliste artiklite lihtsustamine

Lühikokkuvõte:

Paljud finantsteemalised artiklid sisaldavad valdkonnaspetsiifilist sõnavara, mistõttu võivad need tavalugejale olla keerulised mõista. Magistritöö raames uuritakse tihedalt tarbitavate allikate artiklite põhjal, kui palju keerulist sõnavara neis sisaldub ning kui edukalt on keeletehnoloogia vahendite abil võimalik teksti lugejale lihtsamaks teha. Artiklites märgendati väliste lugejate poolt nende silmis keerulised sõnad, mida lihtsustatakse kahe meetodiga: teksti tõlkimine Tartu Ülikooli Neurotõlke abil ning keeruliste sõnade sünonüümimine, kasutades EstNLTK Wordneti sünohulkasid ning eeltreenitud skip-gram mudelit. Analüüs näitas, et vaadeldavates artiklites moodustasid keerulised sõnad ligi 2% kõikidest sõnadest. Nende lihtsustamisel olid tõlkemeetodi tulemused üldiselt edukamad, kuna sünonüümimise puhul ei olnud keerulise sõna sünonüüm alati lihtsamini mõistetav või siis ei leidunud Wordnetis vaadeldavale sõnale ühtegi sünonüümi.

Võtmesõnad: Teksti lihtsustamine, Sünonüümimine, Tõlkimine, Keeleteadus

CERCS: P176 Tehisintellekt

Simplification of Finance Articles

Abstract:

Many financial articles contain field-specific vocabulary, which can be difficult for the ordinary reader to understand. Based on articles from frequently consumed sources, the Master's thesis examines, how much complex vocabulary they contain and how successfully can language technology make the text easier for the reader. External readers marked complex words in the articles, which are simplified by two methods: translation of the text using Neurotõlge by the University of Tartu and synonymization of complicated words using the EstNLTK Wordnet synonym set and the pre-trained skip-gram model. The analysis showed that complex words accounted for nearly 2% of all words in the articles under consideration. In order to simplify them, the results of the translation method were generally more successful, as the synonym for a complex word was not always easier to understand or there was no synonym for the word in Wordnet.

Keywords: Text Simplification, Synonymization, Translation, Linguistics

CERCS: P176 Artificial intelligence

Sisukord

Sissejuhatus 4

1 Andmete kogumine 6

1.1 Alamandmestike moodustamine (märgendamiseks) 6

1.2 Keeruliste sõnade märgendamine 7

2 Metoodika 8

2.1 Tõlkimismeetod 8

2.1.1 Neurotõlge 8

2.1.2 Tõlkimine 9

2.2 Sünonüümimise meetod 10

2.2.1 EstNLTK 10

2.2.2 Sünonüümimine 11

2.2.3 Eeltreenitud skip-gram mudel 11

3 Tulemused 14

3.1 Tõlkimismeetod 15

3.2 Sünonüümimismeetod 18

4 Tuleviku töö 22

Kokkuvõte 23

Viidatud kirjandus 25

Lisad 27

I. Märgendamiseks koostatud juhend 27

II. Vähem õnnestunud paragrahvi tõlkimise tulemus 28

III. Algsõnade ja sünonüümide kaardistus 29

IV. Litsents 30

Sissejuhatus

Igal nooremal ja vanemal inimesel on oluline osata enda rahaga osavalt ümber käia. Seepärast on äärmiselt vajalik, et artiklitest, mida väljastavad lugematud allikad, aru saadakse. Paljud finantsteemalised artiklid on aga üsna spetsiifilise sõnavaraga, mistõttu ei pruugi sellest valdkonnast kaugemal olevatele isikutele vajalik sõnum või nõuanne olla arusaadav. Sel eesmärgil uurime väga levinud ja aktuaalsete artiklite põhjal, kui palju keerulist sõnavara kasutatakse erinevates finantsteemalistes artiklites ning kui edukalt on keeletehnoloogia vahendeid kasutades võimalik teksti lugejale selgemaks teha.

Tekstide lihtsustamist on varasemalt käsitletud ingliskeelsete tekstide põhjal. Lausete lihtsustamiseks on välja töötatud mitmeid juhendatud närvivõrkudel põhinevaid süsteeme [1], [2], mis on andnud märkimisväärselt paremaid tulemusi võrreldes varasemate mudelitega. Ideeliselt töötavad need kui tõlkesüsteemid, tõlkides originaaltekste lihtsasse keelde. Samal ideel tugineb ka juhendamata närvivõrkude süsteem, mis suudab teksti lihtsustada mitmel tasemel, konkureerides olemasolevate juhendatud meetoditega [3]. Magistritöös üritame sarnaseid tulemusi saavutada Tartu Ülikooli Neurotõlke abil.

Erinevate tekstide lihtsustamisel on sõnade asendamist sünonüümidega rakendatud üsna vähe, kuid rootsikeelsetele tekstidele leidub mitu lahendust. Ühes kirjutises on alternatiivsete sünonüümide kasutust uuritud kolme strateegia põhjal: vaadeldes sõna sagedust, pikkust ja sünonüümide rohkust [4]. Kõige edukamalt toimis meetod, kus kombineeriti sõna sagedus ning sünonüümide rohkus. Samuti on sünonüüme kasutatud meditsiinitekstide puhul, asendades sõnu juhul, kui sünonüüm hinnati algsest sõnast lihtsamaks [5]. Tuginedes mitmele loetavuse mõõdule tõi sünonüümidega asendamine kaasa nii mõnevõrra raskema teksti kui ka loetavuse kasvu. Magistritöös on sünonüümidega asendamise ajendiks aga lugeja märgendus, mis näitab, et sõna on peetud keeruliseks.

Magistritöös on finantsteemaliste artiklite lihtsustamisel esmalt suur osa tööst andmete kogumine. Selle moodustab andmete allikate tuvastamine, teemakohaste artiklite filtreerimine, nende kogumine ning töötlemine. Esimese puhul on oluline, et artiklid pärineksid varieeruvatest allikatest, kuid esindaksid platvorme, mida tavalugejad tarbivad. Kuna tavaliselt kõige rohkem tarbitavad allikad koosnevad eri liiki ja eri teemadel kirjutatud artiklitest, siis tuleb valida nende alamhulk, milleks on finantsteemasid käsitlevad artiklid. Valitud teemadega

tekstid kogutakse artiklite HTML dokumentidest kokku ühtseks tekstiks, mis koosneb artikli pealkirjast, alapealkirjadest, sisutekstist ning tsitaatidest. Seejärel on koostatud ka tööprotsess, mille käigus märgendatatakse artiklites kõik lugeja silmis keerulised sõnad. Artiklid koos keeruliste sõnade märgendustega on sisendiks lihtsustamismeetoditele.

Artiklite lihtsustamise väljakutset saab piiritleda üsna laialt, kuid käesolevas töös võtame detailsemalt vaatluse alla kaks meetodit. Esmalt rakendame parafraseerimist, mis baseerub Tartu Ülikooli tõlkevahendil Neurotõlge. See on üks vähestest masintõlke vahenditest, mis oskab tõlkida keelesiseselt, näiteks eesti keelest eesti keelde, ning töötab seeläbi ümbersõnastajana. Lisaks sellele kasutame võrdluseks ka tavapärasemat tõlkimist, et uurida, kas pöördkeele olemasolu parendab parafraseerimise tulemust.

Teiseks meetodiks on spetsiifiliste tekstis olevate keeruliste sõnade asendamine sünonüümidega ehk sünonüümimine (ingl *synonymization*). Igas finantsteemalises artiklis on märgendatud Tartu Ülikooli tudengite poolt nende silmis keeruliseks arvatud sõnad, milledele leiame eesti loomuliku keele töötluse tööriista EstNLTK abil sobivaimad sünonüümid. Viimase puhul on aga kitsenduseks see, et EstNLTK-s on vaid piiratud hulk sünonüüme, mistõttu mõne sõna jaoks ei pruugi leiduda ühtegi vastet. Sel juhul jääb keeruline sõna muutumatuks.

Peamised uurimisküsimused, millele magistritöö käigus vastuseid leitakse, on järgmised:

1. Kui palju keerulist sõnavara finantsteemalistes artiklites kasutatakse?
2. Kui edukas on lihtsustamine tõlkemeetodi abil?
3. Kui edukas on lihtsustamine sünonüümimise abil?

Magistritöö koosneb neljast peatükist. Esimeses peatükis kirjeldatakse finantsteemaliste artiklite kogumist, nendest alamandmestike moodustamist ning artiklites keeruliste sõnade märgendamise protsessi. Teine peatükk koosneb artiklite lihtsustamise metoodika kirjeldustest. Kasutatavateks meetoditeks on tekstide tõlkimine Neurotõlke abil ning keeruliste sõnade sünonümiseerimine. Kolmandas peatükis esitatakse keeruliste sõnade märgendamise tulemusi ning analüüsitakse finantsteemaliste artiklite lihtsustamise tulemusi rakendatud meetodite alusel. Saadud tulemuste põhjal kirjeldatakse neljandas peatükis edasiarenduse võimalusi tulevikus töö jätkamiseks. Viimaseks on magistritöö lisad, kus on välja toodud keeruliste sõnade märgendamiseks loodud juhend ning üks näide tõlkemeetodiga lihtsustamise edutumast tulemusest.

1 Andmete kogumine

Magistritöös kasutatavateks andmeteks on erinevad eestikeelsed finantsteemalised artiklid, mis on kogutud kolmest allikast nimetatud teemadel:

- Äripäev[1]: finantsjuhtimise, raamatupidamise ja makromajanduse teemadel,
- ERR[2]: majanduse kategooriast rahandus- ja finantsteemadel,
- RUP[3]: majanduse, äri, maksude ja raamatupidamise teemadel.

Artiklid on kokku kogutud veebikraapimise teel, tehes päringuid eelnevalt välja toodud teemade otsingutulemuste aadressidele. Selle tulemuseks on iga artikli kohta HTML dokument, millest on ühtseks tekstiks sisse loetud kõik võimalikud artikli osad: pealkiri, alapealkiri, sisutekst, tsitaadid. Allikasiseselt oli HTML-i struktuur enamasti ühesugune, kuid esines ka portaalide uuendusi, mille puhul oli vaja andmete kogumise protsessi kohandada. Kokku on kogutud 5269 finantsteemalist artiklit, mille hulgas on:

- 2727 (52%) Äripäeva,
- 515 (10%) ERR-i,
- 2027 (38%) RUP-i artiklit.

Kõik artiklid on avalikustatud 2019. aasta algusest kuni 2022. aasta märtsi lõpuni ning on varieeruvates pikkustes.

1.1 Alamandmestike moodustamine (märgendamiseks)

Artiklite lihtsustamise järgmise sammu teostamiseks on vaja nendest moodustada üheselt struktureeritud alamandmestikud. Andmestikud moodustatakse expertvaliku ja tagasipanekuta (TTA) lihtsa juhuvaliku põhimõtetel [6]. Selleks, et artiklitest moodustada üldkogumile sarnased alamandmestikud, valitakse igasse andmestikku 20 Äripäeva, 10 ERR-i ja 20 RUP-i artiklit. Allikate üldkogumi osakaalude poolest oleks kõige täpsem valida vastavalt 25, 5 ja 20 artiklit, kuid kuna Äripäeva ning ERR-i artiklite hulgad on väga erinevad, siis on otsustatud alamandmestikes neid veidi tasakaalustada. See tähendab, et expertvaliku tulemuseks on artiklite valimine kogustes 20:10:20. Järgneva sammuna rakendatakse allikate siseselt tagasipanekuta lihtsat juhuvalikut, mille puhul on kõikidel artiklitel võrdne valimisse sattumise

[1] Äripäev: https://www.aripaev.ee/
[2] Eesti Rahvusringhääling: https://www.err.ee/
[3] Raamatupidamisuudised: raamatupidamis- ja õigusuudisteveeb: https://www.rup.ee/

tõenäosus ja iga artikkel saab olla valitud maksimaalselt üks kord [6]. Tulemuseks on andmestikud, mis sisaldavad 50 erinevat artiklit kolmest erinevast allikast ning ei sisalda kattuvaid artikleid. Lisaks artikli sisutekstile, mis hõlmab ka pealkirja, alapealkirju ning tsitaate, on esitatud ka eraldi artikli pealkiri ja link. Igas andmestikus on üks näide märgendatud artiklist ning eraldi lehele on koostatud juhend keeruliste sõnade märgendamiseks (Lisa I).

1.2 Keeruliste sõnade märgendamine

Finantsteemaliste artiklite edukaks lihtsustamiseks peab olema kindel mõõdik, mille muutust on võimalik hiljem võrrelda. Selles töös on selleks artiklites olevad keerulised sõnad, mille olemasolu on võimalik peale lihtsustamist hinnata ning mille baasil saame läbi viia sünonümiseerimise. Kuna sõnade keerulisus on üsna subjektiivne, siis nende märgendamiseks on ette antud mõned suunised. Keeruliseks võib näiteks pidada sõnu, mille puhul:

- Sa ei tea sõna tähendust,
- Sul oleks vaja sõna tähendust meelde tuletada,
- Sa ei tea, mida sõna antud kontekstis tähendab,
- Sulle on antud sõna kasutus tekstis arusaamatu,
- midagi muud.

Lisaks sellele oleme esitanud kitsenduse, et institutsiooni nime ei saa antud kontekstis pidada keeruliseks, akronüümi aga saab küll. Juhul, kui institutsiooni nimi on akronüüm ning seetõttu tundmatu, siis selle saab lugeda keeruliseks. Siinkohal tulevad abiks märgendajad, kellel tuleb lugeda alamandmestikke kogutud finantsteemalisi artikleid ning juhendi abil märgendada ära kõik sõnad, mis on nende silmis keerulised. Märgendajateks olid Tartu Ülikooli keeleteaduste tudengid, kes kokku märgendasid 250 artiklit. Mõnede artiklite puhul oli märgendajaid kaks, mis tähendab, et ühekordselt märgendatud artiklite arv on 100 ning kahekordselt 150. Tulenevalt märgendajate töö subjektiivsusest käsitleme kõiki artikleid aga eraldiseisvatena. Töös kasutatavateks andmeteks on seega 400 märgendatud finantsteemalist artiklit.

2 Metoodika

Selles sektsioonis esitame ülevaate magistritöös kasutatavast metoodikast. Finantstekstide lihtsustamisel võtame vaatluse alla kaks erinevat lähenemist: parafraseerimine tõlkimismeetodi abil, kasutades Neurotõlget ning keeruliste sõnade sünonümiseerimine, kasutades EstNLTK-d ning eeltreenitud mudelit, et leida sobivaim sünonüüm.

2.1 Tõlkimismeetod

Finantsteemaliste artiklite lihtsustamiseks on võimalik teksti ümber sõnastada ehk parafraseerida. Selle jaoks kasutame Tartu Ülikooli poolt välja töötatud masintõlget Neurotõlge.ee. Neurotõlge on üks vähestest välja töötatud mudelitest, mida on treenitud samasse keelde tõlkima. Seda küll mitte kõikide keelte jaoks, vaid alamhulga näitel, kuhu kuulub ka eesti keel. Sel põhjusel saame väga kasuliku omadusega tõlkemudeli abil tõlkida otse eesti keelest eesti keelde ehk parafraseerida artiklite tekste. Lisaks sellele kasutame ka pöördkeeli, et uurida suuremate korpuste kasulikkust tekstide lihtsustamisel.

Kasutataval tõlkevahendil on ka alternatiive, mida on lähiaastatel uute aktuaalsete lahenduste abil aina paremaks arendatud. Näiteks M2M100 mitmekeelne kodeerija-dekooder (*seq-to-seq*) mudel, mis on koolitatud mitu-mitmele (ingl *many-to-many*) mitmekeelseks tõlkeks. M2M100 suudab otse tõlkida 100 keele 9900 suuna vahel ning selles mudelis on samuti võimaldatud ükskeelne tõlkimine [7].

2.1.1 Neurotõlge

Neurotõlge on Tartu Ülikooli poolt välja töötatud masintõlke mudel, mis töötab keelepõhise kooder-dekooder mudeli põhimõttel. Sellest on täpsemalt kirjutanud Purason oma bakalaureusetöös [8]. Keelespetsiifilised kodeerijad ja dekoodrid on üks lähenemisviis üks-ühele (ingl *one-to-one*) mudelite ja universaalsete mudelite puuduste kõrvaldamiseks. Sellise lähenemisviisi puhul ei ole iga keelepaari jaoks vaja eraldi mudelit, selle asemel on iga keele jaoks kodeerija ja dekooder. Näiteks, kui tõlgime eesti keelest inglise keelde, siis kasutatakse tõlkimisel eesti keele kodeerijat ja inglise keele dekoodrit. Neurotõlge on aastatega ka omajagu arenenud. Eelnevalt viidatud Purasoni bakalaureusetöö raames välja töötatud mudel ongi praegu produktsioonis olev Neurotõlge.

Neurotõlke parendamisele on aidanud kaasa ka 2018. aasta rahvusvahelisel masintõlkekonverentsil toimunud tõlkemootorite võistlus, kus võtsid osa mitmed innovaatilised süsteemid [9]. Kõige usaldusväärsemaks hinnati keeletehnoloogia ettevõtte Tilde tõlkesüsteem. Nemad tõid oma edu otseseks tulemuseks tehisintellekti valdkonnast tuntud neurovõrkude-põhise lähenemise, tänu millele saavutati läbimurre masintõlke kvaliteedis. Teisele kohale tuli Jaapanis loodud süsteem NICT, mille edu seisnes eelkõige suurte teadusarvutuste keskuste kasutuses. Kõikide süsteemide kohta on avalike teadusartiklite näol jagatud ka täpsemaid kirjeldusi, mille abil oli võimalik parandada ka Tartu Ülikoolis loodud tõlkemootorit Neurotõlge.

Neurotõlge oskab ülesandele sobivalt tõlkida sama keele siseselt ehk ükskeelselt. Regulaarsete lausetega ei tundu ükskeelne tõlge kasulik, sest selle käitumine seisneb peamiselt kopeerimises. Kui aga sisendil on omadusi, mida on mudeli poolt treeningu ajal harva või mitte kunagi nähtud, genereerib dekooder ikkagi lause regulaarsema või lihtsama versiooni [10]. Finantsteemaliste tekstide lihtsustamiseks on antud omadus väga asjakohane.

2.1.2 Tõlkimine

Finantsteemaliste artiklite parafraseerimiseks kasutame mitut tõlkimissuunda, milledest üks on ükskeelne ning teised on kakskeelsed. Nimetatud tõlkimisahelateks on:

- eesti – eesti,
- eesti – inglise – eesti,
- eesti – soome – eesti,
- eesti – saksa – eesti.

Kõige otsesemalt teostab parafraseerimise rolli ükskeelne tõlkimine. See tagastab enamjaolt küll sarnase teksti, kuid asendab treenimisel vähem esinenud sõnad sagedasematega, mis tavaliselt annavadki lihtsustatud sõnastamise tulemuse. Alternatiivselt uurime ka kakskeelseid tõlkesuundi, et näha, kas lihtsustamise edukamale tulemusele aitab kaasa pöördkeelte olemasolu. Pöördkeel on vahendajakeel või sillakeel, mida kasutatakse tõlkimise hõlbustamiseks kahe või enama teise keele vahel ning selleks on üldjuhul sagedaselt levinud keel, näiteks inglise keel [11]. Selles töös on peamine pöördkeel niisamuti inglise keel, kuna see on kõige levinum teine keel ja ülemaailmse äri keel, mis tähendab, et suurem hulk koolitatud tõlkijaid tunneb inglise ja teise keele kombinatsiooni kui teised keelekombinatsioonid.

Lisaks inglise keelele kasutame pöördkeelena ka soome ja saksa keelt. Soome keelt, kuna sellel on palju rohkem kasutajaid ning teisalt on see olemuselt ka üsna sarnane eesti keelele. Saksa keel on samas võrdlus inglise keelele, olles samuti suurema kasutajaskonnaga keel. Lisaks eelnevalt mainitule ühendame üheks variandiks nii ükskeelse kui ka kakskeelse tõlkesuuna:

- eesti – eesti – inglise – eesti.

Ühendades ükskeelse ja kakskeelse tõlkimise, kasutame sisuliselt lihtkeelset tõlget, mida on propageerinud Neubauer [12]. Lihtkeelne tõlge nõuab protsessi, mida nimetatakse intralingvistiliseks tõlkeks, keelesiseseks tõlkeks või lihtsalt ümbersõnastamiseks, mis on sõnaliste märkide tõlgendamine teiste samakeelsete märkide abil. Keelesisese tõlke lisamine keeltevahelise tõlke käigus või eraldi sammuna on oluline osa selleks, et tõlkijal õnnestuks lähteteksti sõnum analüüsifaasis taandada selle lihtsaimale kujule (mõtteks või tähenduseks). Seejärel kasutatakse seda tähendust sihtrühmale sobiva sihtteksti loomise alusena, andes tagasi sobivalt lihtsustatud sõnumi.

2.2 Sünonüümimise meetod

Finantsteemaliste artiklite lihtsustamist on võimalik ka detailsemalt lahendada. Nimelt, me saame artiklitest üles otsida keerulisemad sõnad ning asendada need sünonüümidega. Selleks, et eraldada keerulised sõnad terviklikust tekstist, kasutame EstNLTK-d. Seejärel leiame Wordnet abiga nendele sõnadele võimalikud sünonüümid ning eeltreenitud mudeli abil selgitame vastavalt kontekstile välja kõige sobivama sünonüümi. Viimaks asendame artiklis leitud keerulise sõna leitud sünonüümiga.

2.2.1 EstNLTK

EstNLTK on eesti keele töötlemiseks mõeldud Pythoni teek. EstNLTK pakub eesti keele jaoks loomuliku keele töötlemise funktsionaalsusi, nagu näiteks lõigu-, lause- ja sõna tokeniseerimine, morfoloogiline analüüs, nimeliste üksuste tuvastamine. Selles töös kasutame peamiselt morfoloogilist analüüsi, mille abil on võimalik finantsteemalise artikli teksti lausestada ning tekstis keeruliseks märgendatud sõna puhul teha kindlaks sõna algvorm.

Sõna algvormile saab seejärel leida sünonüümi, milleks kasutame EstNLTK Wordnet moodulit. Järgnev materjal tugineb H. Orava projektil [13]. Wordnet on leksikaal-semantiline andmebaas, mis sisaldab informatsiooni mõistete ja nendevaheliste semantiliste suhete kohta.

Wordneti põhiühik on sünohulk (või sünonüümihulk), mis koosneb kõigist ühte ja sama mõistet väljendavatest sõnadest (või sõnaühenditest). Juhul, kui sõnal ei ole ühtegi sünonüümi, siis sisaldab sünohulk ainult ühte sõna ning see tähendab, et antud mõistet saab väljendada ainult ühe sõnaga. Kusjuures, kõik sünohulga liikmed on ühest sõnaliigist. Sünohulgad on ühendatud ka viidetega, mis vastavad mõistetevahelistele suhetele, kuid selles töös ei ole antud suhetele vaja tähelepanu pöörata, sest meid huvitavad keeruliste sõnade kõik võimalikud sünonüümid.

2.2.2 Sünonüümimine

Finantsteemalistes artiklites keeruliste sõnade sünonümiseerimise jaoks tuleb esmalt artiklite tekstid lausestada. EstNLTK-s on lausete jaoks eraldi kiht, mis moodustatakse sisendtekstist ning see lõikab tervikliku teksti eraldi lauseteks. Seejärel saame morfoloogilise analüüsi abil lausetest välja võtta keeruliseks märgendatud sõnad ning leiame nende algvormi. Asendatava sünonüümi leidmiseks kogume kokku kõik võimalikud sünonüümid Wordneti sünohulkadest. Siinkohal toimime veidi erinevalt, olenevalt sõnade spetsiifikast, nimelt

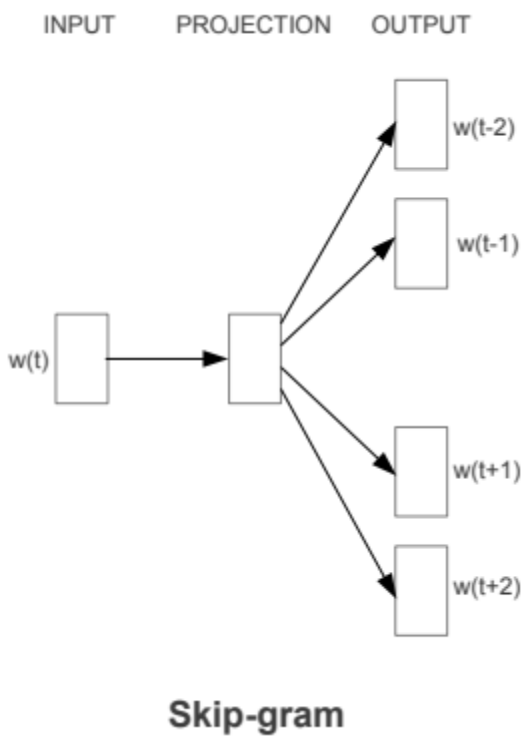
- kui tegemist on akronüümiga, on oluline hoida sõna läbivalt suurte tähtedega, kuna muidu ei ole võimalik sellele sünohulkasid leida;
- kui vaadeldav sõna ei ole üldse sünohulkades, siis jääb see puutumata;
- kui tegemist on sõnaga, millel on sõnavorm puudu, siis määratakse sellele vaikimisi ainsuse nimetav kääne, seega sünonüümi vorm jääb samaks.

Sõnavormide kogumine on oluline selleks, et sünonüümid saab leida vaid sõna algvormidele ning sealt saab tagasi samuti sünonüümi, mis on algvormis. Lausetesse asendamisel tuleb aga leitud sünonüümid viia korrektsesse vormi, et need algteksti sinna asemele sobituksid. Kokku kogutud sünonüümidest sobivaima leidmiseks kasutame Eesti Keeleressursside Keskuse eeltreenitud skip-gram mudelit. Kõrgeima skooriga sõna, mis ei ole sõna ise, viiakse seejärel korrektsesse vormi ning asendatakse lausesse keerulise sõna asemele.

2.2.3 Eeltreenitud skip-gram mudel

Sünonüümide hulgast keerulisele sõnale semantiliselt sobivaima leidmiseks kasutatakse eeltreenitud skip-gram mudelit. Algselt oli sünonüümimise käigus kavas kasutada üht

Huggingface Google eeltreenitud mudelit[4], kuid kasutusjuhendi puudulikkuse tõttu ei olnud võimalik seda implementeerida. Seetõttu võtsime kasutusele hoopis skip-gram mudeli, mis ennustab sisendiks antud keskse sõna baasil konteksti sõnu, püüdes maksimeerida sõna klassifitseerimist teise sõna põhjal samas lauses (Joonis 1). Täpsemalt, iga keskne sõna antakse sisendiks log-lineaarsele klassifikaatorile, millel on pidev projektsioonikiht, ning ennustatakse sõnu teatud vahemikus enne ja pärast keskset sõna [14].



Joonis 1. Skip-gram mudeli arhitektuur [15]

Sammhaaval toimib see mudeli loojate sõnul [14] järgmiselt:

- Sisendkihis kodeeritakse N eelnevat sõna vektoriaalse (ingl *1-of-V, one-hot*) kodeeringuga, kus V on sõnavara suurus. Iga sisendsümboli puhul saadakse V-suurune vektor, kus kõik on nullid, välja arvatud element, mis vastab sümboli järjestusele sõnavaras, mis saab väärtuseks 1.
- Kasutades jagatud projektsioonimaatriksit, projekteeritakse sisendkiht projektsioonikihile P, mis on mõõtmetega $N \times D$, kus D koosneb sõnade esitustest.

[4] Mudeli allikas: https://huggingface.co/google/t5-efficient-base-dm512

– Väljundiks on vektor, mis sisaldab iga antud sõnavara sõna kohta tõenäosust, et juhuslikult valitud lähedal asuv sõna on see just see sõnavara sõna [16].

Seega andes skip-gram mudelile ette keerulise sõna koos võimalike sünonüümidega, saame igale sünonüümile tõenäosuse, millest kõige suurema tõenäosusega on loetud kui kontekstilt kõige sarnasem sõna.

Mudel on treenitud eesti keele koondkorpuse põhjal, mis koosneb ajalehe, kirjanduse, teaduse ja õigusaktide tekstidest [17]. Korpus koosneb 16 miljonist lausest, 55 miljonist sõnast ja 2 miljonist sõnatüübist. Kuna sünonüümid leitakse sõnade algvormidele ehk lemmadele, siis kasutame mudelit, mis on treenitud samuti korpuse lemmatiseeritud versioonil.

3 Tulemused

Finantsteemaliste artiklite lihtsustamise eesmärgil märgendasid Tartu Ülikooli keeleteaduse tudengid 400 artiklis sõnu, mida nad pidasid keerulisteks. Mõnede artiklite puhul on keeruliste sõnade märgendamine sooritatud kahekordselt, see tähendab, et kaks inimest märgendasid samu artikleid. Hetkel käsitleme neid erinevate artiklitena, kuna märgendajate hinnangud olid subjektiivsed ning antud eraldiseisvalt.

Kokku märgendati 1467 sõna, mis moodustavad tekstide sõnade koguarvust 1,69%. Seega oli ligi 2% analüüsitud artiklite tekstist sõnastatud liiga keeruliselt või mõistmatult. Pelgalt protsendile peale vaadates ei tundu see kuigi suur osa tekstist, kuid reaalsuses mõjutab see 2% ka ümbritsevat teksti. Seetõttu võib tegelikult lugejale arusaamatuks jäänud teksti maht olla palju laialdasem ning keeruliste sõnade olulisus samuti märgatavalt suurem.

Detailsema analüüsi jaoks uurime sõnade lemmasid ehk algvorme. Unikaalsete lemmadega märgendatud keerulisi sõnu oli kokku 835. Nendest kõige tihedamini esinenud 50 sõna on esitatud sõnapilvena (Joonis 2). Jooniselt on selgesti näha, et kõige keerulisemaks või korduvalt keeruliseks on peetud sõna saneerimine. Sellele järgnevad struktuurne, keskpank, skp, inflatsioon ja käive.

Joonis 2. Sõnapilv märgendatud keeruliste sõnade lemmadest

3.1 Tõlkimismeetod

Finantsteemaliste artiklite lihtsustamiseks teostati ühe variandina tekstide ümbersõnastamist. Sel eesmärgil rakendasime tõlkemeetodit, mille käigus tõlkisime artikleid Tartu Ülikooli poolt loodud Neurotõlke abil. Uurides nii otsese ümbersõnastamise, pöördkeelte lisamise kui ka nende kombinatsiooni tulemuslikkust, kasutasime kokku viit erinevat tõlkesuunda:

- eesti – eesti,
- eesti – inglise – eesti,
- eesti – soome – eesti,
- eesti – saksa – eesti,
- eesti – eesti – inglise – eesti.

Tekstide tõlkimiseks tegime Neurotõlke API suunas päringuid, edastades tekste lõikude kaupa. Terviklike tekstide tõlkimisest hoidumine tulenes põhjusest, et üks päring võib sisaldada kuni 10 000 tähemärki, kuid mitmed artiklid on tervikuna palju pikemad. Mainitud faktor ei mõjutanud aga tõlke tulemusi, kuna mõndade lühemate lõikude ühisel tõlkimisel tuli välja, et igat lõiku vaadatakse eraldi ning ümber olevad paragrahvid ei mõjuta neid kuidagi. Tekstide lihtsustamise vaatepunktist ei erinenud tõlkesuundade tulemused omavahel kuigi palju (Tabel 1).

Tabel 1. Tõlkemeetodi tulemused

Tõlkesuund	**Teksti muutus (%)**	**Keerulisi sõnu (%)**	**Lihtsustatud sõnu (%)**	**Levenshteini kaugus**
eesti – eesti	12,6	63,6	36,4	317 830
eesti – inglise – eesti	12,7	67,1	32,9	366 747
eesti – soome – eesti	20,7	39	61	533 559
eesti – saksa – eesti	14,6	60,4	39,6	416 481
eesti – eesti – inglise – eesti	15,2	57,3	42,7	424 938

Tabelist on näha, et keerulisteks märgendatud sõnu jäi tekstidesse alles enamasti 60% ringis. Seega kuni 40% nendest sõnastati ümber ühel või teisel moel. Kõige parema tulemuse andis

aga tõlkesuund, kus oli kasutatud soome keelt pöördkeelena – keerulisi sõnu jäi teksti alles vaid alla 40%. Keeruliste sõnade parafraseerimise osakaalud olid võrreldes teiste tõlkesuundadega sisuliselt vastupidi. Edukaima lihtsustuse puhul on näha, et tekst on samuti selgelt kõige enam muutunud. Protsentuaalselt on tekstis sõnade erinevus 20,7% ning Levenshteini kaugus on 533 559.

Levenshteini kaugus (ka teisenduskaugus) on väikseim arv redigeerimistoiminguid, mis on vajalikud ühe stringi teisendamiseks. See näitab, kui mitu korda tuleb ühes sõnes tähte lisada, eemaldada või vahetada, et ühest sõnest saaks teine sõne [18]. Parima tulemusega tõlkemeetodi puhul on seega vaja algses korpusetekstis teha 533 559 muudatust, et see oleks võrdne tõlke tulemusega ning see on märkimisväärselt suurem kui teiste tõlkesuundade puhul.

Samas paneme tähele, et keelesisese tõlke (ehk otsese ümbersõnastamise) ühendamisel pöördkeele kasutusega on tulemused samuti üsna edukad, olles mõnevõrra tulemuslikumad kui ülejäänud kolm alternatiivi. Hübriidse tõlkesuuna korral on lihtsustatud pea 43% keerulistest sõnadest ning teksti muutus on alanenud 15,2%-ni. Üllatuslikult madalad näitajad on tekstisisese tõlke puhul eesti keelest otse eesti keelde ning kõige levinuma pöördkeele ehk inglise keele kasutamisel. Tekstid on muutunud kõige vähem ning ligikaudu kolmandik keerulistest sõnadest on saanud lihtsustatud versiooni või sõnastatud veidi ümber.

Tabelis kajastatud tulemuste põhjal uurime täpsemalt ka üht lõigu näidet. Toome välja ühe lõigu algteksti ning selle viis tõlget. Kas paneme ka siin näites tähele tugevat eelist, kui kasutada tõlkimisel soome keele abi?

- **Algne tekst**

  Kõrgeim brutokuupalk oli riigile kuuluvates asutustes ja ettevõtetes - 1817 eurot ning välismaa **eraõiguslikele isikutele** kuuluvates ettevõtetes - 1685 eurot. **Brutokuupalga aastakasv** oli kiireim riigile kuuluvates asutustes ja ettevõtetes, kus see oli 11,2 protsenti ning kõige aeglasem välismaa **eraõiguslikele isikutele** kuuluvates ettevõtetes - 5 protsenti.

- **eesti – eesti**

  Kõrgeim brutokuupalk riigiasutustes ja ettevõtetes - 1817 eurot ja välisriigi eraettevõtetes - oli 1685 eurot. Aastane brutokuupalga kasv oli kõige kiirem riigiasutustes ja ettevõtetes, kus see oli 11,2% ja kõige aeglasem välismaiste eraettevõtete puhul - 5%.

- **eesti – inglise – eesti**

  Kõrgeim brutokuupalk oli riigile kuuluvates asutustes ja ettevõtetes 1817 eurot ning välismaistele eraisikutele kuuluvates ettevõtetes 1685 eurot. Kõige kiiremini kasvas brutokuupalk riigile kuuluvates asutustes ja ettevõtetes, kus see oli 11,2 protsenti ja kõige aeglasemalt välismaistele eraisikutele kuuluvates ettevõtetes - 5 protsenti.
- **eesti – soome – eesti**

  Kõrgeim brutokuupalk oli avalik-õiguslikes asutustes ja ettevõtetes - 1817 eurot ja välismaistes eraettevõtetes - 1685 eurot. Aastane brutokuupalga kasv oli kõige kiirem avalik-õiguslikes asutustes ja ettevõtetes, kus see oli 11,2% ja kõige aeglasem välismaistes eraettevõtetes - 5%.
- **eesti – saksa – eesti**

  Kõrgeim brutokuupalk oli 1817 eurot riigiasutustes ja ettevõtetes ning 1685 eurot välismaal asuvatele eraisikutele kuuluvates ettevõtetes. Brutokuupalga aastane kasv oli kõige kiirem 11,2 protsendiga riigiasutustes ja ettevõtetes ning kõige aeglasem 5 protsendiga eraettevõtetes välismaal.
- **eesti – eesti – inglise – eesti**

  Kõrgeim brutokuupalk riigiasutustes ja ettevõtetes - 1817 eurot ja välismaistes eraettevõtetes - oli 1685. Kõige kiirem oli aasta brutokuupalga kasv riigiasutustes ja ettevõtetes, kus see oli 11,2% ja kõige aeglasem välismaistes eraettevõtetes - 5%.

Erinevaid tulemusi lugedes tundub kõige lugejasõbralikum olevat tõlge saksa keelega. Samas on tekstid siiski enamasti üsna sarnased, mistõttu ei paista, et tõlkimisel soome keele kasutamine palju lisandväärtust annab. Kõige märgatavam tundub olevat aga eesti keelesisese tõlke pikkus võrreldes algtekstiga. Mainitud tekst on lühenenud kõige rohkem ehk võib tunduda, et teksti sisu on mingil määral kompaktsemaks ja selgemaks muutunud, kuid samas lausete arusaadavus ei ole enam nii hea kui algse teksti puhul. Segadust lisavad ka mõttekriipsude kasutamine lausetes, mis mõjub kohati grammatiliselt ebakorrektsena.

Märgendatud keerulisi sõnu on selle näite puhul kaotatud või ümber sõnastatud üpris edukalt. Suur osa keerulistest sõnadest on saanud uue sõnastuse ning teinud seetõttu teksti lihtsamini loetavaks. Leidub aga ka negatiivseid näiteid, kus keerulisi sõnu pole pea üldse lihtsustatud ning parafraseerimise tulemusena on laused ka kohati keerulisemaks muutunud (Lisa II). Üldpilt Neurotõlke parafraseerimise tulemustest on aga üsna positiivne ning näeme, et Tartu Ülikooli keeleteadlased on välja töötanud väga kasuliku masintõlke mudeli.

3.2 Sünonüümimismeetod

Rahandusartiklite detailsema lihtsustamise eesmärgil asendasime keeruliseks märgendatud sõnu sobilike sünonüümidega. Kasutatud sünonüümisõnastik pärineb EstNLTK Wordneti moodulist, kus on suur hulk erinevaid sünonüüme eestikeelsete sõnade jaoks. Kuna neid on moodulis fikseeritud hulk, siis osadel sõnadel ei leidunud ühtegi algsõnast erinevat sünonüümi. Sel puhul jäid antud sõnad selle magistritöö raames muutumata.

Kokku oli muutumata sõnu 892, mis moodustasid kõikidest märgendatud sõnadest 61%. Nendest oli omakorda unikaalse lemmaga 518 sõna ehk 58%. Kõige tihedamalt märgendatud 25 sõna on allpool välja toodud sõnapilvena (Joonis 3). Kõige segasemaks jäi lugejate silmis keskpanga tähendus. Peale selle puudusid sünonüümid ka inflatsioonil, brutokuupalgal ning audiitorteenusel. Kui me mõtleme nende sõnade olemusele, siis me märkame, et nende selgitamiseks ei tundu piisav sellest, kui üritada lihtsalt mõne teise sõnaga terminit selgitada.

Esiteks, nendel sõnadel ei leidu kõige sobivamaid sünonüüme. Teiseks on tegemist taoliste spetsiifiliste terminitega, mille puhul aitaks kõige paremini või isegi eranditult hoopis see, kui sünonüümi asemel kasutada definitsiooni.

Joonis 3. Muutmata lemmad

Ligi 39% keerulistest sõnadest said asendatud sünonüümidega. Täpsemalt oli neid kokku 575 ja unikaalse lemmaga oli 63% ehk 360 sõna. Selleks, et mõista sünonüümidega asendamise tulemusi, vaatame nii muudetud sõnade algseid olekuid kui ka asendatud sünonüüme (Joonis 4, Joonis 5). Sagedasemalt esinenud sõnade puhul on tabelina välja toodud ka algsõnade ja

sünonüümide kaardistus (Lisa III). Kõige tugevamalt paistab jooniselt 4 välja saneerimine, mis näitab, et lihtsustatavate sõnade hulgast on see tekitanud tekstis kõige rohkem mõistmatust. Lisaks sellele on keerulisemaks märgitud ka sõna struktuurne ning akronüüm SKP.

Joonis 4. Sünonüümseks muudetud lemmad: algsed sõnad

Kõik eelnevas sõnapilves välja toodud ning ka väiksemate sagedusega keerulised sõnad said asendatud sünonüümidega. Vastavate sünonüümide visualisatsioon on üsna sarnane keeruliste omaga ning sõnapilvede vahel on võimalik näha väga otseseid seoseid (Joonis 5).

Joonis 5. Sünonüümseks muudetud lemmad: sünonüümid

Eelmises sõnapilves kõige enam esinenud saneerimine on asendatud sünonüümiga sanatsioon, funktsionaalse vasteks on ehituslik ning SKP puhul on akronüüm lahti kirjutatud sisemajanduse koguproduktiks. Viimaste puhul toimetab sünonüümimine üsna mõistlikult. Saneerimise ja sanatsiooni vahel ei ole aga sisuliselt üldse erinevust, sest neil on sama sõnatüvi. See võib olla ka selline sõna, mis vajab parema arusaadavuse jaoks definitsiooni välja toomist.

Sõnapilvede abil nähtud tulemuste põhjal uurime täpsemalt ka mõnda spetsiifilist artikli lõiku. Esmalt vaatame juba tõlkemeetodi paragrahvi all nähtud teksti ning võrdleme sünonüümimise meetodi tulemusi.

- **Algne tekst**

  Kõrgeim brutokuupalk oli riigile kuuluvates asutustes ja ettevõtetes - 1817 eurot ning välismaa **eraõiguslikele isikutele** kuuluvates ettevõtetes - 1685 eurot. **Brutokuupalga aastakasv** oli kiireim riigile kuuluvates asutustes ja ettevõtetes, kus see oli 11,2 protsenti ning kõige aeglasem välismaa **eraõiguslikele isikutele** kuuluvates ettevõtetes - 5 protsenti.

- **Sünonüümidega asendatud**

  Kõrgeim brutokuupalk oli riigile kuuluvates asutustes ja ettevõtetes - 1817 eurot ning välismaa eraõiguslikele **inimesile** kuuluvates ettevõtetes - 1685 eurot. Brutokuupalga aastakasv oli kiireim riigile kuuluvates asutustes ja ettevõtetes, kus see oli 11,2 protsenti ning kõige aeglasem välismaa eraõiguslikele **inimesile** kuuluvates ettevõtetes - 5 protsenti.

Antud näite puhul ei ole teksti lihtsustamine sünonüümide abil teinud tulemusliku soorituse. Enamus märgendatud sõnadest on endiselt tekstis, kuid üks sõna on siiski asendatud sünonüümiga. Sõna isikutele asenduseks on inimesile, mis on küll igati korrektne vorm, kuid lihtsuse mõttes oleme lugejatena harjunud pigem teistsuguse kirjapildiga, milleks on inimestele.

Sõnapilvede analüüsimisel panime tähele, et üsna palju esines saneerimist, mis asendati sünonüümiga sanatsioon. Idee poolest ei tundu see kuidagi lihtsustav, kuid vaatame ka spetsiifilisemat tekstilõiku.

- **Algne tekst**

  Saates sai veel täpsemini teada, millal tuleb **saneerimisega** alustada ning milliseid meetmeid sagedamini kasutatakse, millised on **saneerimise** õnnestumise võtmetegurid, mida tuleb arvestada **saneerimisega** seoses koroona kontekstis, milline on **saneerimisnõustaja** roll ja mis on suuremad probleemid, miks **saneerimine** ebaõnnestub.

- **Sünonüümidega asendatud**

  Saates sai veel täpsemini teada, millal tuleb sanatsiooniga alustada ning milliseid meetmeid sagedamini kasutatakse, millised on sanatsiooni õnnestumise võtmetegurid,

mida tuleb arvestada sanatsiooniga seoses koroona kontekstis, milline on saneerimisnõustaja roll ja mis on suuremad probleemid, miks sanatsioon ebaõnnestub.

Tulemuse põhjal näeme, et sünonüümimine on iseenesest toiminud pigem edukalt, kuid tekstide lihtsustamisele ei aita see kuigi palju kaasa. Siinkohal võib olla üsna kindel, et märgendaja jaoks on need lõigud võrdsed ning keerulisteks märgendataks nii saneerimine kui ka sanatsioon. Sünonüümidega ümbersõnastamine on aga andnud ka väga edukaid tulemusi.

- **Algne tekst**

  Eesti valitsussektori struktuurne puudujääk oli 2018. aastal rahandusministeeriumi hinnangul 1,4% **SKPst**, mis on suurem, kui riigi eelarvestrateegias või riigieelarve koostamisel eesmärgiks seati.
- **Sünonüümidega asendatud**

  Eesti valitsussektori struktuurne puudujääk oli 2018. aastal rahandusministeeriumi hinnangul 1,4% sisemajanduse koguproduktist, mis on suurem, kui riigi eelarvestrateegias või riigieelarve koostamisel eesmärgiks seati

Ülemise näite põhjal on näha, et asendatud sünonüüm on ka päriselt abiks ning teeb teksti lugejale lihtsamaks. Selliseid tulemusi on ka muude lausete näitel ning tekstidest on nende abil osa saanud kergemini loetavaks. Võrreldes sünonüümimist tõlkemeetodiga ning hinnates neid finantsteemaliste tekstide lihtsustamise aspektist, tundub, et tõlkimine oli selles siiski edukam. Tõlkemeetodi puhul oli boonuseks ka ümbritseva teksti ümbersõnastamine, mis võis kohati keerulist sõna muutmata aidata mõista selle sõna tähendust antud kontekstis. Sünonüümimise idee on küll hea, kuid selle täiustamiseks ja korrektselt üles ehitamiseks on vaja teha veel lisatööd ning uurimusi.

4 Tuleviku töö

Finantsteemaliste artiklite lihtsustamine on üsna tänuväärne tegevus nendele, kes on sellest valdkonnast kaugemal ning ei mõista koheselt kõiki kasutatavaid termineid. Magistritöös kasutatud lahendusmeetodid ei andnud eraldiseisvalt kumbki ideaalseid tulemusi. Neid meetodeid oleks edasises töös aga võimalik kombineerida. See näitab, et on palju lisafaktoreid, millele võiks tulevikus tähelepanu pöörata, kui sarnast ülesannet lahendada. Järgmiseks sammuks võiks veel olla tõlkimise ja sünonüümimise meetoditega lihtsustatud tekstide kuvamine erinevatele lugejatele. Selle eesmärgiks oleks lugejate poolne valideerimine:

- kas meetodid tegelikult ka tekste lihtsustasid,
- kas need püsisid umbes sama keerulised või
- kas läksid sootuks keerulisemaks.

Sünonüümimise parema töötamise jaoks saaks uurida finantssõnastike lisamise kasulikkust – kas need abistaksid sünonüümimata sõnadega või parendaksid üldiselt sünonüümide leidmist. Valdkonna sõnastike kaudu võib olla võimalik ka sünohulki laiendada. Sõnastike leidmisel saaks kindlasti abi küsida erinevatelt pankadelt ning muudelt asjaosalistelt, kellele töö tulemus kasulikuna võiks mõjuda.

Jätkates sünonüümimise lahendusega tundub lõpp-punktiks välja töötatud mudel, mis oskab teema-spetsiifilisi tekste lihtsustada kõige sobivamalt ning korrektsemalt. Tekstide lihtsustamine on üsna subjektiivne, mistõttu võib lausel olla mitmeid sobilikke lihtsustatud versioone. Mudelit ennast on aga selle tõttu keerulisem ehitada, et lihtsustamisel ei ole otsest tõde (ingl *ground truth*), millega tulemusi valideerida.

Ühe variandina näitas kasulikkust keeruliste sõnade sünonüümimise asemel hoopis nende lahti seletamine – näiteks lisades teksti allmärkuseid või kuvades sõna kohale automaatselt selle selgitusega lingi või seletuse pilvekesena (kui näiteks sõnale vajutada või selle kohal hiirega hõljuda). Sel juhul oleks ka vähem kartust, et teksti mõte võidakse muuta valeks, kuna originaalne sõna oleks olemas ning selgitus vaid täiendaks. Kui lisada sõnaseletusi aga tekstide sisse, oleks kindlasti tähtis, et sellised seletused on pigem napisõnalisemad ning annavad üldise mõiste tähenduse.

Kokkuvõte

Magistritöö raames tegelesime finantsteemaliste artiklite automaatse lihtsustamise ülesandega. Selle jaoks koguti Äripäevast, ERR-ist ja RUP-ist teemakohased artiklid, mille tekstides märgendasid Tartu Ülikooli tudengid enda silmis keerulised sõnad. Märgendatud finantstekstide lihtsustamiseks kasutasime kaht meetodit: tõlkimine ja sünonüümimine. Töö käigus leidsime vastused järgmistele uurimisküsimustele:

1. Kui palju keerulist sõnavara finantsteemalistes artiklites kasutatakse?
2. Kui edukas on lihtsustamine tõlkemeetodi abil?
3. Kui edukas on lihtsustamine sünonüümimise abil?

Artiklites märgendatud keerulisi sõnu uurides saime esimesele küsimusele vastuseks, et kokku märgendati 1467 sõna, mis moodustasid tekstide sõnade koguarvust 1,69%. Seega keerulist sõnavara oli finantsteemalistes artiklites ligi 2%. Protsent ise ei tundu väga suurena, kuid reaalsuses mõjutab see 2% ka ümbritsevat teksti, mistõttu võib lugejale arusaamatuks jäänud teksti maht olla palju laialdasem. Kuna aga sõna keerulisus on subjektiivne, siis ei ole see tulemus konstantne ning võib muutuda sõltuvalt lugejate teadmistest.

Tõlkemeetodi puhul kasutasime Tartu Ülikooli poolt välja töötatud Neurotõlget, mille tulemusena saime tekste ümber sõnastada. Lisaks sellele kasutasime võrdluseks tavapärast tõlkimist, et uurida, kas pöördkeele olemasolu parandab parafraseerimise tulemust. Tulemused näitasid, et kõige edukamalt kaotas tekstist keerulisi sõnu just pöördkeele kasutus, kui pöördkeeleks oli soome keel. Selle puhul muudeti keerulisi sõnu 60% ulatuses.Otsene keelesisene ümbersõnastus ei andnud aga oodatuid tulemusi ning selle sooritus oli üks halvimatest. Üllataval kombel ei andnud kõige populaarsem pöördkeel, inglise keel, pea üldse lisandväärtust. Keelesisese ja keeltevahelise tõlke ühendamisel (eesti – eesti – inglise – eesti) saime aga paremuselt teise tulemuse, milleks oli ligi 43%, vähendades ka teksti muutumist. Üldpilt Neurotõlke parafraseerimise tulemustest oli aga üsna positiivne ning näeme, et Tartu Ülikooli keeleteadlased on välja töötanud väga kasuliku masintõlke mudeli.

Teiseks meetodiks oli asendada tekstis sisalduvad keerulised sõnad sünonüümidega. Keerulised sõnad läbisid sünonüümimise protseduuri, kasutades EstNLTK-d, mis on eesti loomuliku keele töötlemiseks mõeldud Pythoni teek ning selles olevat Wordneti moodulit. Wordneti mooduli abil üritasime igale keerulisele sõnale leida sünonüümi ning asendasime

selle õiges vormis tagasi teksti. Wordnetis on vaid piiratud sünonüümide hulk, mistõttu mõnede sõnade jaoks ei leidunud ühtegi ning keeruline sõna jäi muutumatuks. Sünonüümimise tulemusel sai lihtsustatud või täpsemini muudetud ligi 40% keerulistest sõnadest. Sünonüümide asendamine ei töötanud aga alati kõige paremini, lihtsustades sõna pea identse sõnaga, millel on ka sama algvorm või leides sünonüümi, mis on võrreldes algse tekstiga hoopis arusaamatum. Teisalt oli ka juhtumeid, kus sünonüümija sai finantsteemalise teksti lihtsustamisega ideaalselt hakkama.

Võrreldes sünonüümimist tõlkimismeetodiga ja hinnates neid meetodeid finantsküsimusi käsitlevate tekstide lihtsustamise seisukohast, tundub, et tõlkimine oli siiski edukam. Tõlkemeetodi puhul oli lisaväärtuseks ka ümbritseva teksti ümbersõnastamine, mis võis aidata mõista selle sõna tähendust antud kontekstis, muutmata mõnikord keerulist sõna. Sünonüümimise idee ise on küll hea, kuid selle edendamiseks, korrektseks ülesehitamiseks ja täiustamiseks on vaja teha kindlasti lisatööd ja uuringuid.

Viidatud kirjandus

[1] S. Nisioi, S. Štajner, S. P. Ponzetto, ja L. P. Dinu, „Exploring Neural Text Simplification Models“, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada, 2017, lk 85–91. doi: 10.18653/v1/P17-2014.

[2] M. Cooper ja M. Shardlow, „CombiNMT: An Exploration into Neural Text Simplification Models“, lk 7.

[3] S. Surya, A. Mishra, A. Laha, P. Jain, ja K. Sankaranarayanan, „Unsupervised Neural Text Simplification“. arXiv, 21. august 2019. Vaadatud: 17. mai 2022. [Online]. Available at: http://arxiv.org/abs/1810.07931

[4] R. Keskisarkka, „Automatic Text Simplification via Synonym Replacement“, lk 83.

[5] E. Abrahamsson, T. Forni, M. Skeppstedt, ja M. Kvist, „Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language“, *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, Gothenburg, Sweden, 2014, lk 57–65. doi: 10.3115/v1/W14-1207.

[6] Natalja Lepik, Imbi Traat, „Tõenäosuslik valikuuring I“. Tartu Ülikool, 2016.

[7] A. Fan *et al.*, „Beyond English-Centric Multilingual Machine Translation“. arXiv, 21. oktoober 2020. Vaadatud: 17. mai 2022. [Online]. Available at: http://arxiv.org/abs/2010.11125

[8] Taido Purason, „Modular Septilingual Neural Machine Translation“, bakalaureusetöö, TÜ arvutiteaduse instituut, Tartu, 2021.

[9] Mark Fišel, „Selgusid parimad eesti-inglise-eesti tõlkemootorid“, nr ERR, nov 2018, Vaadatud: 13. mai 2022. [Online]. Available at: https://novaator.err.ee/874855/selgusid-parimad-eesti-inglise-eesti-tolkemootorid

[10] E. Korotkova, A. Luhtaru, M. Del, K. Liin, D. Deksne, ja M. Fishel, „Grammatical Error Correction and Style Transfer via Zero-shot Monolingual Translation“. arXiv, 22. oktoober 2019. Vaadatud: 17. mai 2022. [Online]. Available at: http://arxiv.org/abs/1903.11283

[11] „Pivot Language | The Content Wrangler“. https://thecontentwrangler.com/glossary/pivot-language/, https://thecontentwrangler.com/glossary/pivot-language/ (vaadatud 17. mai 2022).

[12] A. Neubauer, „Euro-murkiness: Plain language and "fighting the fog" in European Union translation“, University of Geneva, 2011. Vaadatud: 17. mai 2022. [Online]. Available at: https://archive-ouverte.unige.ch/unige:18416

[13] H. Orav, „Estonian Wordnet (2.4.0)Eesti Wordnet (2.4.0)“, sept 2021, doi: 10.15155/1-00-0000-0000-0000-001B3L.

[14] T. Mikolov, K. Chen, G. s Corrado, ja J. Dean, „Efficient Estimation of Word Representations in Vector Space“, *Proc. Workshop ICLR*, kd 2013, jaan 2013.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, ja J. Dean, „Distributed Representations of Words and Phrases and their Compositionality“, lk 9.

[16] Chris McCormick, „Word2Vec Tutorial - The Skip-Gram Model · Chris McCormick“, *http://www.mccormickml.com*, 19. aprill 2016. http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/ (vaadatud 17. mai 2022).

[17] „Eesti keele koondkorpus“. https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et (vaadatud 17. mai 2022).

[18] L. Yujian ja L. Bo, „A Normalized Levenshtein Distance Metric“, *IEEE Trans. Pattern Anal. Mach. Intell.*, kd 29, nr 6, lk 1091–1095, 2007, doi: 10.1109/TPAMI.2007.1078.

Lisad

I. Märgendamiseks koostatud juhend

Tere! Mina olen Kristin ja teen andmeteaduse magistritööd, kus üritan lihtsustada keerulisemaid termineid finantsteemalistes artiklites. Ma palun selleks Sinult abi, et need keerulisemad sõnad ära märgendada. Siinkohal on sõna "keeruline" üsna individuaalselt hinnatav. Keeruliseks võib näiteks pidada sõnu, mille puhul:

- Sa ei tea sõna tähendust
- Sul oleks vaja sõna tähendust meelde tuletada
- Sa ei tea, mida sõna antud kontekstis tähendab
- Sulle on antud sõna kasutus tekstis arusaamatu
- midagi muud.

Märkus: Institutsiooni nime **ei saa** antud kontekstis pidada keeruliseks, akronüümi aga võib küll.

Märgendamine

1. Palun märgista artikli teksti sees Sinu jaoks keerulisemad sõnad **BOLD**-iga (klaviatuuri otsetee **Ctrl+b**). Märgista kindlasti terve sõna, mitte ainult tüvi. Sõna saab teha aktiivseks hiirega vasaku topeltklõpsu abil ja siis märgendada Ctrl+b.

2. Kui terve artikkel on läbi vaadatud ja märgendatud, muuda palun "Märgendatud?" veeru väärtus **"ei"** → **"jah"**.

Suur-suur aitäh Sulle, et oled mind abistanud!

II. Vähem õnnestunud paragrahvi tõlkimise tulemus

- **Algne tekst**

  Aastavõrdluses kasvasid euroalal tööstusvaldkondade lõikes **tootjahinnad** samuti enim **energiasektoris**, koguni 25,4 p2rotsenti. **Vahekaupade** hinnad tõusid 10,6 protsenti, kestvate ja mitte-kestvate tarbekaupade hinnad 2,3 protsent ja **kapitalikaupade** hinnad 2 protsenti. **Energiasektorit** arvestamata kasvasid **tootjahinnad** 5,6 protsenti.

- **eesti – eesti**

  Aasta võrdluseks kasvasid ka tootjahinnad euroala tööstussektorite lõikes kõige rohkem, isegi 25,4 % võrra. Vahetoodete hinnad tõusid 10,6%, püsivate ja mittekestvate tarbekaupade hinnad 2,3% ja kapitalikaupade hinnad 2%. Välja arvatud energiasektor, kasvasid tootjahinnad 5,6 %.

- **eesti – inglise – eesti**

  Aastataguse ajaga võrreldes tõusid tootjahinnad euroalal tööstuse lõikes enim ka energiasektoris, koguni 25,4 protsendipunkti võrra. Vahekaupade hinnad tõusid 10,6 protsenti, püsivate ja püsivate tarbekaupade hinnad 2,3 protsenti ning kapitalikaupade hinnad 2 protsenti. Ilma energiasektorita tõusid tootjahinnad 5,6%.

- **eesti – soome – eesti**

  Aastasel tasandil tõusid ka tootjahinnad energiasektoris kõige rohkem, kuni 25,4 protsendipunkti. Vahetoodete hinnad tõusid 10,6 %, jätkusuutlike ja jätkusuutmatute tarbekaupade hinnad 2,3 % ja kapitalikaupade hinnad 2 %. Välja arvatud energiasektor, tõusid tootjahinnad 5,6 %.

- **eesti – saksa – eesti**

  Aastases võrdluses tõusid euroala tootjahinnad ka energiasektoris kõige enam tööstusharude kaupa, ulatudes 25,4 p2roseni. Vahetoodete hinnad tõusid 10,6 protsenti, püsivate ja mittepüsivate tarbekaupade hinnad 2,3 protsenti ja kapitalikaupade hinnad 2 protsenti. Energiasektorit arvesse võtmata tõusid tootjahinnad 5,6 protsenti.

- **eesti – eesti – inglise – eesti**

  Aastataguse ajaga võrreldes tõusid ka tootjahinnad euroala tööstussektorite lõikes kõige enam, isegi 25,4%. Vahetoodete hinnad tõusid 10,6%, vastupidavate ja mittekestvate tarbekaupade hinnad 2,3% ja kapitalikaupade hinnad 2%. Kui energiasektor välja arvata, tõusid tootjahinnad 5,6%.

III. Algsõnade ja sünonüümide kaardistus

Tabel 1. 25 kõige sagedasemat sünonüümivahetust

Algne sõna	**Sünonüüm**
saneerimine	sanatsioon
struktuurne	ehituslik
skp	sisemajandus koguprodukt
käive	ringlus
intress	kasvik
likviidsus	maksevõime
rahavoog	rahakäive
võlakiri	võlatäht
võõrandamine	võõrandus
emissioon	väljalaskmine
teine	2.
indeks	sihtnumber
tariif	tasumäär
diskontomäär	soodusemäär
varu	tagavara
äriühing	majandusühing
audiitor	revident
majanduslangus	surutis
efektiivne	tõhus
avalik	varjamatu
hulgikaubandus	hulgimüük
krunt	maavaldus
bilanss	bilansiaruanne
alampalk	miinimumpalk

IV. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kristin Petersel

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose **Finantsteemaliste artiklite lihtsustamine**, mille juhendajad on Mark Fišel, PhD ja Lehar Oha, 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.

Kristin Petersel

17.05.2022