TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Taaniel Saarnik

Ajaseoste automaatne tuvastamine tekstis

Bakalaureusetöö (9 EAP)

Juhendaja Siim Orasmaa

Tartu 2021

Ajaseoste automaatne tuvastamine tekstis

Lühikokkuvõte:

Selles lõputöös kirjeldatakse ajalehetekstidest sündmusi ja sündmustevahelisi ajaseoseid tuvastavate mudelite loomist. Ajalehetekstides esineb palju erinevaid sündmusi ja ajaväljendeid. Selleks, et arvuti suudaks neid erinevates keeletöötlus rakendustes kasutada, tuleb need tekstis mingil viisil märgendada. Manuaalselt suurte tekstide märgendamine on aga väga tülikas ja aeganõudev protsess ja seepärast oleks suur abi süsteemist, mis oskaks seda automaatselt teha. Käesoleva lõputöö käigus uuritakse selle probleemi varasemaid lahendusi ning luuakse ka mudeleid, mis automatiseerivad selle protsessi.

Võtmesõnad: NLP, keeletehnoloogia, ajaseosed, ajaväljendid, sündmused, ajasemantika, märgendamine, TimeML, EstTimeMLCorpus, EstBERT, EstNLTK

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

Automatic Detection of Temporal Relations in Text

Abstract:

The aim of this thesis is to create machine learning models that identify events described in newspaper texts and identify temporal relations between events and temporal expressions. Newspaper texts contain many different events that can be interpreted as being connected via temporal relations. To be able to use that information in different natural language processing tasks, it is crucial to tag these words somehow. Manual tagging is very troublesome and time-consuming task and that is why automatic system would be very beneficial. This thesis describes earlier solutions to this problem and a process of creating machine learning models to solve this.

Keywords: NLP, language technology, temporal relations, temporal expressions, events, temporal semantic, tagging, TimeML, EstTimeMLCorpus, EstBERT, EstNLTK

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

Sisukord

Sissejuhatus 5

1. Märgendusstandardid 6

1.1 TimeML 6

1.2 Eesti keelele kohandatud TimeML 7

2. Varasemad tööd tekstide TimeML märgenduses 9

2.1 Ingliskeelsetel tekstidel 9

2.2 Eesti keelel 11

2.3 Huvipakkuvad eestikeelsete tekstide jaoks loodud mudelid 13

2.3.1 EstNLTK NerTagger 13

2.3.2 EstBERT 13

3. Töökäik 14

3.1 Andmete töötlemine 14

3.1.1 Sündmuste kihid 14

3.1.2 Ajaväljendite kihid 17

3.2 Mudelite loomine 19

3.2.1 Sündmusi tuvastavad mudelid 19

3.2.2 Ajaseoseid tuvastavad mudelid 20

4. Analüüs 23

4.1 Sündmuste tuvastamine 23

4.2 Ajaseoste tuvastamine 24

4.2.1 Sündmuste ja dokumendi loomise aja vahelised ajaseosed 24

4.2.2 Ajaväljendite ja sündmuste vahelised ajaseosed 25

4.2.3 Peamiste sündmuste vahelised ajaseosed 25

4.2.4 Erineval süntaksipuu tasemel olevate sündmuste vahelised suhtetüübid 26

5. Kokkuvõte 27

Viidatud kirjandus 29

Lisad 31

I. Kirjutatud kood ja kasutatud failid 31

II. Litsents 32

Sissejuhatus

Paljudes tekstides kirjeldatakse mingeid sündmuseid. Need sündmused võivad, kuid sageli ei pruugi, esineda kindlas ajalises järjekorras, varasemad sündmused eespool ja hilisemad järgnemas nendele. Vahepeal võidakse viidata varem toimunud või tulevastele sündmustele. Mõnel juhul võib ajaväljend selgelt kirjas olla, kuid ajalisi seoseid saab luua ka teadaolevate sündmustega. Lugedes saab lugeja aru, millal mingi sündmus aset leidis või võtab ning suudab seostada sündmusi teiste teadaolevate sündmustega. Need teadmised aitavad luua loetud tekstide kohta kokkuvõtteid ja vastata tekstiga seotud küsimustele. Selleks, et arvuti suudaks tekstis esinevaid sündmusi ja ajaväljendeid kasutada mitmete loomuliku keele töötlusülesannete jaoks, tuleb tekstile lisada märgendusi. Märgendusstandardite järgi tekstide manuaalne märgendamine on aeganõudev tegevus. Selleks, et palju vajalikku aega mitte raisata, võib mõelda selle töö automatiseerimisele.

Selle lõputöö eesmärgiks on treenida mudeleid, mis suudaks etteantud eestikeelsetest ajalehetekstidest tuvastada sündmusviiteid, ajaseoseid sündmusviidete vahel ning ajaseoseid sündmusviidete ja ajaväljendite vahel. Ajalehetekstid saadakse Eesti TimeML korpusest (vt ptk 1), kus on vajalikud märgendused juba tehtud. Korpuses olevaid kirjeid kasutatakse juhendatud masinõppes.

Käesolev töö koosneb neljast osast. Esimeses osas tutvustatakse Eesti TimeML korpuses kasutusel olevaid märgendusstandardeid. Teises osas käsitletakse varasemaid inglise ja eesti keelel loodud mudeleid. Kolmandas osas selgitatakse selles lõputöös kasutatud meetodeid andmete töötlemiseks ja mudeli treenimiseks. Neljandas osas antakse ülevaade loodud mudelite tulemustest.

Lisa I sisaldab linki koodirepositooriumile, kus on kättesaadav bakalaureusetöö ülesannete täitmiseks kirjutatud kood, katsetused ja kasutatud failid.

1. Märgendusstandardid

Käesolevas töös kasutatakse masinõppes Siim Orasmaa poolt loodud Eesti TimeML korpust[1]. Eesti TimeML korpuses on sündmuste ja nendega seotud ajaseoste märgendamiseks kasutatud TimeML märgendusstandardit [1], mida Orasmaa kohandas eesti keele jaoks. See peatükk keskendub TimeML märgendusstandardile ja korpuses kasutatava märgenduse erinevustele.

1.1 TimeML

TimeML märgendusstandard võimaldab tekstis esinevaid sündmusi, ajaväljendeid ja nendevahelisi seoseid kindlate märgenditega tähistada [1]. Sündmuste jaoks kasutatakse EVENT märgendit, ajaväljendite jaoks TIMEX3 märgendit, ajalist järjekorda kirjeldavad sõnad tähistatakse SIGNAL märgendiga ning TLINK märgendiga tähistatakse sündmuste ja ajaväljendite omavahelisi suhteid. Järgnevates lõikudes olevad märgendite selgitused on peamiselt refereeritud J. Pustejovsky jt [1] TimeML märgendussüsteemi tutvustavast konverentsiartiklist.

EVENT märgendiga on ära märgitud tekstis esinevad sündmused. Peamiselt tähistatakse sellega sündmusi kirjeldavaid tegusõnu. Lisaks selgesti eristatavate tegusõnadega edasi antud sündmustele märgistatakse sellega ka sõnu, mis kirjeldavad olukordi, asjaolusid või on tegemist sündmuse tegusõnast tuletatud käändsõnavormiga. EVENT märgenditel on ka kirjas sündmuse klass. Sündmused jaotatakse kaheksasse klassi: *occurrence* ehk juhtum (nt ehitama, ostma, laskma), *state* ehk seisund (nt kuuluma, asuma, armuma), *reporting* ehk teavitamine (nt ütlema, hõiskama, teavitama), *i-action* ehk kavatsetud tegevus (nt põhjendama, proovima, andma), *i-state* ehk kavatsetud seisund (nt oskama, tunduma, lootma), *aspectual* ehk mingi sündmuse faasi muutus (nt hakkama, lõpetama, alustama) ja *perception* ehk taju (nt jälgima, nägema, tundma).

TIMEX3 märgendiga tähistatakse ajaväljendeid, olgu need numbrilised või sõnalised. Tähistatakse nii täpseid ajaväljendeid (kellaaeg, kuupäev, aasta), täpsustamata ajaväljendeid (*järgmine kuu/päev, eile*, *homme*) ning kestvusi (*3 nädalat*, *terve aasta*, *3 tundi).* Täpsustamata ajaväljendite täpsema väärtuse võib saada dokumendi loomise aega (DCT – ing *document creation time*) kasutades (nt lauses „**Homme** kogunetakse Tallinna lauluväljakule.“), kuid vahest tuleb vaadata ka teisi tekstis esinevaid ajaväljendeid (nt lauses „**2000. aastal** oli

[1] https://github.com/soras/EstTimeMLCorpus

sündimus 25% suurem kui **aasta varem**.*“)*. TIMEX3 märgendis on kirjas ajaväljendi tüüp ehk kas on tegemist kuupäevaga (ing *date*), kellaajaga (ing *time*) või perioodiga (ing *duration*). Lisaks ajaväljendid normaliseeritakse ehk märgendusse lisatakse ISO formaadil põhinev väärtus. Näiteks ajaväljendi „18. detsember 2004“ puhul on tegemist kuupäevaga ning ISO formaadi järgi [2] saab seda esitada kui „2004-12-18“. Kellaaja tüübiga on ajaväljend „täna kell 20:30“ ning kui varasemast on teada kuupäev, siis seda saab esitada kui „2004-12-18T20:30:00“, kus kuupäev ja kellaaeg on eraldatud tähega T. Kui kuupäeva pole teada, siis märgitakse lihtsalt kellaaeg. Perioodi tähistavate ajaväljendite väärtustel on alguses täht P ja sellele järgneb kuupäeva või aja väärtus. Lubatud on ära jätta märgendi komponente (aasta, kuu, päev, tund, minut, sekund), kui seda ajaväljendiga pole mainitud ehk ajaväljendile „5 päeva“ saab panna väärtuseks lihtsalt „P5D“.

SIGNAL märgendit kasutatakse ajaväljendite ajaliste seoste ja prepositsioonide märkimiseks. Sellised sõnad tekstis on näiteks *enne, pärast, samal ajal* ning inglise keelsed väljendid *for, during, on, at*. Vajadusel lisatakse märgendisse ka eitust või kordust kirjeldav sõna.

LINK märgendid seovad tekstis esinevad sündmused ja ajalised väljendid omavahel nii, et nendega saab ka tuua esile sündmuste toimumise järjekorra. LINK märgenditega on võimalik ajalise suhte tähistamiseks siduda omavahel sündmusi või sündmust ja aega (TLINK) (inglise keeles suhe sündmusel *saw* ja ajaväljendil *at noon* lauses „*She saw him at noon.*“), tähistada seoseid EVENT ja LINK märgendite vahel, kus võib muutuda sündmuse toimumise kindlusaste (SLINK) (inglise keeles suhe sõnadel *wants, to* ja *teach* fraasis *„Tom wants to teach.*“ ja lauses „*She forgot to take it.*“ sõnad *forgot, to, take*) ning seos aspektiivse sündmuse ja selle argument sündmuse vahel (ALINK) (inglise keeles *stopped* ja *reading* lauses „*Tom stopped reading.*“).

1.2 Eesti keelele kohandatud TimeML

Eesti TimeML korpuse koostamisel on eeskuju võetud TimeML märgendusest ning järgnevates lõikudes on märgendite selgitused refereeritud S. Orasmaa artiklist [3]. Korpuses on kasutusel EVENT, TIMEX3 ja TLINK märgendid, mida on kohandatud eesti keele jaoks.

EVENT märgendusel tuli lahendada nii-öelda lagunemisprobleem ehk kuidas mitmesõnalisi sündmusi märkida. See probleem tekib sündmustel, kus põhisõnaks olev tegusõna on puhtalt grammatiline (nt sõna *olema*) või semantiliselt nõrk (nt sõna *tegema*). Otsustati, et kui sõna *olema* on koos olekut tähistava nimi-, omadus- või määrsõnaga, siis see lisatakse mitmesõnalise sündmuse märgendusse (sõna *olema* ja olekut tähistav sõna). Aga kui esineb

infinitiivse tegusõnaga (da-, vat- ja ma-tegevusnimi), siis sõna *olema* jäetakse märgendusest välja.

Erinevalt inglise keelest, lisati sündmuse märgendusse ka põhisõnaga koos käivad modaalverbid ja otsustati märgendada ka taustsündmusi, kui need on seotud märgendatud ajalise väljendiga või on seotud märgendatud sündmusega ning esinevad tekstis mitu korda. Kui vaadata lauset „Esimeses rallisõidus saavutas Toomas suure vahemaa ja hoidis seda nädala lõpuni.“, siis põhisündmusteks oleksid *saavutas* ja *hoidis* ning taustsündmuseks oleks sõna *rallisõidus*.

Eesti keeles ei ole kasutusel ajaväljenditega seotud prepositsioone, vaid see info antakse edasi käändelõppudega. Näiteks ingliskeelses ajaväljendis „*on Friday*“ esineb prepositsioon *on*, eesti keeles kasutatakse sobivat käändelõppu ning saadakse ajaväljend *„reedel*“. Sõnade sees märgendamist ei soovitud teha ja seega SIGNAL märgendust ei kasutatud.

LINK märgendustest kasutati ainult TLINK märgendusi (vt ptk 1.1). SemEval-2010 ülesande TempEval-2[2] (vt ptk 2.1) eeskujul jaotati relatsioonid neljaks: seosed sündmuste ja ajaväljendite vahel, sündmuste ja dokumendi loomise aja vahel, kahe järjestikuse lause peamiste sündmuste vahel ja samas lauses olevate sündmuste vahel, kus üks on süntaksipuus kõrgemal. TLINK märgendustel on kirjas ka relatsioonitüüp. Kasutusel on järgmised relatsioonitüübid:

1. BEFORE – ajaliselt eelnev
2. AFTER – ajaliselt järgnev
3. SIMULTANEOUS – täpne ajaline kattuvus
4. IDENTITY – täpne ajaline kattuvus ja on tegemist sama sündmusega
5. IS_INCLUDED – ajaline sisaldumine (A on Bs)
6. INCLUDES – ajaline kaasaarvamine (B sisaldab A)
7. BEFORE-OR-OVERLAP – kahtlus eelneva ja kattuva vahel
8. OVERLAP-OR-AFTER – kahtlus järgneva ja kattuva vahel
9. VAGUE – selgusetu ajaline suhe

Relatsioonitüüpide valikul lähtuti eesti keele iseärasustest.

[2] https://www.aclweb.org/anthology/S10-1010

2. Varasemad tööd tekstide TimeML märgenduses

Varasemad katsed tekstile automaatselt genereerida TimeML märgendusi on peamiselt toimunud ingliskeelsetel korpustel. Sündmuste, ajaväljendite ja ka ajaseoste tuvastamiseks on kasutatud nii masinõppe kui ka reeglipõhist lähenemist.

2.1 Ingliskeelsetel tekstidel

B. Arnulphy jt [4] said inglise keelsetest tekstidest sündmuste tuvastamisel kõige parema tulemuse CRF mudeliga, millele olid nad lisanud ka K-Nearest Neighbors (KNN) algoritmi. Nende sõnul peaks KNN aitama lugeda sündmuste toimumiste jada lausetes samaks, kus ühte lausesse on lisatud mingi sõna vahele. Lisaks aitab see numbriliste väärtustega ja sünonüümidega arvestamist. Mudelile anti ette sõnad, nende lemmad ja sõnaliigid ning pandi IOB märgistust ennustama (B – sündmuse algus, I – sündmuse fraasi osa, O – pole sündmus). CRF-KNN mudeliga õnnestus neil saada nii täpsuseks (ing *precision*), saagiseks (ing *recall*) kui ka F1-skooriks (ing *F1-score*) 0,86.

Inglise keelsetest tekstidest sündmuste ja ajaväljendite vaheliste suhete tuvastamisega on tegeletud SemEval-2007 TempEval-1 võistlusel, kus oli kuus osalejat [5]. Kasutati kolme kindlat suhtetüüpi: *enne* (ing *before*), *peale* (ing *after*), *kattuv* (ing *overlap*). Lisaks neile oli veel kolm suhtetüüpi suhete jaoks, mis olid mitmetähenduslikud (*before-or-overlap*, *overlap-or-after*) ja olukordadeks, kus suhet pole võimalik määrata (*vague*). Lisatud suhtetüüpide tõttu kasutati ranget ja vabamat hindamissüsteemi. Rangema puhul peeti õigeks ainult täpseid ennustusi, kuid vabama süsteemi puhul anti ennustusele väärtus arvude 0 ja 1 vahel (näiteks suhtetüübile *before-or-overlap* ennustades *before* saadi väärtuseks 0,5). Suhete tuvastamine oli jagatud kolmeks osaks. Esimeses ülesandes tuli tuvastada sündmuste ja ajaväljendi vahelisi suhteid ainult väljenditel, mis esinesid samas lauses. Teiseks ülesandeks oli dokumendi loomisaja ja dokumendis esinevate sündmuste suhete tuvastamine. Kolmandas ülesandes tuli leida kõrvuti olevate lausete peamiste sündmuste vahelisi suhteid. Võistlusel osalenud mudelite tulemused on näha tabelis 1.

Tabel 1. TempEval-1 osalenud mudelite tulemused kolmes ülesandes. Kaldkriips eraldab range ja vaba hinnangusüsteemi tulemust (M. Verhagen jt, 2009 [5], kohandatud)

Mudel	**Ülesanne 1** **Täpsus, saagis, F-skoor**	**Ülesanne 2** **Täpsus, saagis, F-skoor**	**Ülesanne 3** **Täpsus, saagis, F-skoor**
CU-TMP	61/63	75/76	54/58
LCC-TE	59/61, 57/60, 58/60	75/76, 71/72, 73/74	55/58
NAIST	61/63	75/76	49/53
USFD	59/60	73/74	54/57
WVALI	62/64	80/81	54/64
XRCE-T	53/63, 25/30, 34/41	78/84, 57/62, 66/71	42/58

Tabelis 1 on kirjas mudelite range ja vabama hindamissüsteemi tulemused. Täpsust ja saagist pole märgitud, kui need kattuvad F-skooriga. Tulemuste poolest paistab silma WVALI, mis toetus varem valmistatud süsteemile, mis kasutab teadmiste põhiseid ja statistilisi meetodeid [5]. USFD mudel, mille tulemused ei jäänud ülesannete parimatest tulemustest palju alla, kasutas kõige sirgjoonelisemat lähenemist klassifitseerimisülesandele. Mudelile etteantud andmed pärinesid sündmuste ja ajaväljendite märgistustest või dokumendist kergelt saadud informatsioonist, põhjalikku analüüsi lingvistilise automaatanalüüsi tööriistaga tekstile ei tehtud. Iga ülesannet katsetati erinevate masinõppe algoritmidega, mille seast valiti iga ülesande jaoks parim.

SemEval-2010 TempEval-2 oli varem mainitud võistluse edasiarendus, ülesandeid oli võimalik lahendada mitmel keelel ning ülesanded olid jagatud rohkemateks osadeks [6]. Relatsioonide tuvastamise ülesanne olid jagatud neljaks: suhted samas lauses olevate ajaväljendite ja sündmuste vahel (A), suhted dokumendi sündmuste ja dokumendi loomisaja vahel (B), suhted kõrvuti olevate lausete peamiste sündmuste vahel (C), suhted samas lauses olevate sündmuste vahel, kus üks sündmus on lause süntaksipuus kõrgemal kui teine (D). Mudelid said sisendiks sündmuste või sündmuse ja ajaväljendi paare ning pidid neile sobiva suhtetüübi määrama. Hindamiseks jagati õigete vastuste arvu vastuste koguarvuga. Tabelis 2 on kuvatud relatsioonide tuvastamise ülesannete tulemused.

Tabel 2. TempEval-2 relatsioonide tuvastamises osalenud mudelite tulemused (M. Verhagen jt, 2010 [6], kohandatud)

Mudel	**A**	**B**	**C**	**D**
JU_CSE	0,63	0,80	0,56	0,56
NCSU-indi	0,63	0,68	0,48	0,66
NCSU-joint	0,62	0,21	0,51	0,25
TIPSem	0,55	0,82	0,55	0,59
TIPSem-B	0,54	0,81	0,55	0,60
TRIOS	0,65	0,79	0,56	0,60
TRIPS	0,63	0,76	0,58	0,59
USFD2	0,63	-	0,45	-

TempEval-2007 toimunud võistlusel osales mudel USFD ning USFD2 mudeliga on tulemuse paranemist märgata esimeses ülesandes, kuid kolmandas on see langenud. Heade tulemuste poolest paistab silma mudel TRIOS, mis sai kõigis neljas ülesandes teistega võrreldes kõrged tulemused. Halvemast küljest jäi aga silma NSCU-joint, mis sai B ja D ülesandes palju väiksemad tulemused. TempEval-2 analüüsis on ka võrreldud tulemusi TempEval-1 tulemustega (vt tabel 3).

Tabel 3. TempEval-1 ja TempEval-2 tulemuste võrdlus (M. Verhagen jt, 2010 [6], kohandatud)

		A	**B**	**C**
TempEval-1	**Keskmine**	0,59	0,76	0,51
	Standardhälve	0,03	0,03	0,05
TempEval-2	**Keskmine**	0,61	0,70	0,53
	Standardhälve	0,04	0,22	0,05

Tabelis 3 paistab silma TempEval-2 ülesandes B saadud kehvem tulemuste keskmine ja suurem standardhälve, kuid selle põhjuseks on NSCU-joint mudeli kehv tulemus. Kui NSCU-joint tulemust mitte arvestada, siis keskmiseks saab 0,78 ning standardhälveks 0,05. Ülesannete tulemused seega paranesid, kuid mitte märgatavalt.

2.2 Eesti keelel

Eesti keelsetel tekstidel on S. Orasmaa [7] loonud ajaväljendite tuvastaja. Tegemist on reeglipõhilise süsteemiga, kus kasutatav märgenduskeel toetub TimeML märgendusstandardile. Reeglite abil tuvastati tekstides väiksemad ajaväljendeid, millest

võimalusel loodi omakorda pikemaid fraase (nt. „tuleval reedel“ + „kell kaks“). Tuvastatud ajaväljendid normaliseeriti vastavalt märgenduskeelele. Süsteemi arendati ja testiti ajakirjandustekstide peal. Ajaväljendite eraldamisel saadi saagiseks 82,0 ning täpsuseks 98,4 ja ajaväljendite liigitamisel saavutati täpsuseks 97,2 ning semantika normaliseerimisel täpsuseks 87,4.

2.3 Huvipakkuvad eestikeelsete tekstide jaoks loodud mudelid

Eestikeelsetel tekstidel on loodud mitmeid mudeleid. Käesoleva töö raames pakuvad eriti huvi mudelid, millega on võimalik teha nimeüksuste tuvastamist. Algselt võimaldavad need tekstides tuvastada isikuid, organisatsioone ja asukohti. Mudelit ümber treenides oleks võimalik tuvastada sündmusi.

2.3.1 EstNLTK NerTagger

Eestikeelsete tekstide töötlemiseks loodud Pythoni teegis EstNLTK 1.6 on olemas nimeüksuste tuvastaja. Tegemist on A. Tkachenko jt [8] poolt loodud nimeüksuste tuvastajaga, mida on kohandatud EstNLTK tarbeks. Tkachenko jt poolt loodud mudel treeniti 572-l ajalehe artiklil ning treenimisel kasutati kümne alamhulga ristvalideerimist. Eesmärgiks oli mudelile õpetada tekstides tuvastama isikuid, organisatsioone ja asukohti. Masinõppe tunnustena kasutati morfoloogilise analüüsi tulemusi, tähemärkide järjestust ning olemasolu ja varem tuvastatud sõnade märgendeid. Mudeli saagiseks saadi 87,5, täpsuseks 86,6 ning F1-skooriks 87,0. EstNLTK teegis oleva mudeli ümbertreenimiseks loodud meetodeid[3] saab kasutada, et treenida mudelit tuvastama sündmusi. Samuti on võimalik valida, milliseid tunnuseid kasutatakse.

2.3.2 EstBERT

BERT-i puhul on tegemist suurel märgendamata korpusel treenitud mudeliga, mis on treenimisel tuvastanud sisendkeele üldisi seaduspärasusi. Mõned mitmekeelsed BERT mudelid toetavad ka eesti keelt, kuid H. Tanvir jt [9] lõid eestikeelsetel tekstidel keelespetsiifilise EstBERT mudeli. Sel hetkel kõige suurimast saadaolevast eestikeelsete tekstide korpustest Estonian National Corpus 2017[4] saadi peale andmete puhastamist treenimiseks 3,3 miljonit tekstidokumenti. Nende tekstide peale kokku oli 75,7 miljonit lauset ja 1154 miljonit sõna. Mudel peenhäälestati sõnaliikide, nimeüksuste, teksti rubriigi ja meeleolu tuvastamise ülesannetele ja hinnati nende ülesannete täitmist. Nimeüksuste tuvastamisele peenhäälestamiseks ja hindamiseks kasutati Tkachenko jt [8] poolt loodud korpust. Nimeüksuste tuvastamisel saadi saagiseks 90,38, täpsuseks 88,42 ja F1-skooriks 89,39. Käesolevas töös kasutatakse nimeüksuste tuvastamise lähenemist EstBERT-iga sündmuste tuvastamiseks ning EstBERT-ist saadavaid sõnade vektoreid relatsioonide masinõppes.

[3] https://github.com/estnltk/estnltk/blob/devel_1.6/estnltk/taggers/estner/ner_training.ipynb

[4] https://doi.org/10.15155/3-00-0000-0000-0000-071E7L

3. Töökäik

Enne mudelite loomist tuli tegeleda andmete töötlemisega. Selleks kasutati Pythoni teeki EstNLTK 1.6[5]. Tegemist on S. Laur jt [10] poolt eesti keelele mõeldud loomuliku keele töötlemise (NLP – ing *natural language processing*) teegi uuendatud versiooniga. Teek võimaldab tekste segmenteerida (nt sõnedeks, sõnadeks, lauseteks), teha morfoloogilist analüüsi (nt lemmad, sõnavorm, sõnaliik), teha süntaktilist analüüsi ja kasutada mitmeid valmisolevaid tööriistu NLP ülesannete lahendamiseks. Käesolevas töös otsustati korpuses olevad artiklid ja märgendused viia EstNLTK 1.6 *Text* objektide kujule. Lisaks vaikimisi loodud *Text* objekti kihtidele (segmenteerimine) lisati kihid sündmuste ja ajaväljendite märgenduste jaoks. EstNLTK *Text* objektid salvestati JSON-failidesse, mis võimaldas korpuses olevat infot kergemal kujul käsitleda. Sündmustega ja ajaväljenditega relatsioonid otsustati salvestada eraldi JSON-failidesse. Peale andmetöötlust sai luua mudelid tekstist sündmuste ja relatsioonide tuvastamiseks.

3.1 Andmete töötlemine

Tööd alustati Eesti TimeML korpuse tekstide, märgenduste ja relatsioonide konverteerimisega EstNLTK *Text* objektideks. Korpuses on 80 artiklit, mille peale kokku on märgitud 4366 sündmust ja 995 ajaväljendit. Artiklite ja nende märgistuste sisselugemiseks kasutati korpuses olemasolevaid meetodeid[6]. Esimese asjana loodi artiklitest *Text* objektid ning siis lisati neile märgendused kihtide (ing *layers*) kaudu. Korpusest saadud käsitsi lisatud märgendusi nimetame kuldstandardiks.

3.1.1 Sündmuste kihid

Algselt loodi kiht *gold_events*, mis sisaldab tervet infot käsitsi märgendatud sündmuste kohta. Kihis on kirjas sündmuse sõne, lause ID tekstis, sõna ID lauses, sündmuse ID, väljend, märgendus ja sündmuse klass. Tabelis 4 on esitatud osa *Text.gold_events* väljundist artikli puhul, milles on lause „Kontserdil on koos Eesti nooremate lauljate paremik.“.

[5] https://github.com/estnltk/estnltk/tree/devel_1.6 kasutati EstNLTK 1.6.7 + osaliselt arendusharu koodi

[6] https://github.com/soras/EstTimeMLCorpus/blob/master/exported_corpus_reader.py

Tabel 4. Lõik artikli *Text* objekti *gold_events* kihist

text	sentence_ID	word_ID_in_sentence	event_ID	expression	event_annotation	event_class
Kontserdil	12	0	e53	"Kontserdil"	EVENT OCCURRENCE	OCCURRENCE
on	12	1	e54	"on koos"	EVENT STATE multiword="true"	STATE
koos	12	2	e54	"on koos"	EVENT multiword="true"	STATE

See kiht võimaldab hoida artikli sündmuste märgenduste infot kergesti loetavas ja ligipääsetavas kohas. Kihti kasutati relatsioonide õppimisel ja teiste kihtide loomisel.

Loodi kolm kihti, mis järgisid IOB[7] (ing *inside, outside, beginning*) märgendussüsteemi. Neid kihte kasutati, et treenida nimeüksuste tuvastajaid (*NERtagger*[8]), mis suudaksid tekstist tuvastada sündmusi kirjeldavaid väljendeid. Kihtides on artikli kõik sõned ja neile vastavad IOB märgendused.

Algselt olid kihid tehtud korpuses kasutatud sõnestuse järgi. Hiljem EstNLTK NERtaggerit treenides selgus, et EstNLTK enda ja korpuse tekstide sõnestustes esines erinevusi, mis tekitasid treenimisel probleeme. Sõnestuse erinevused, mis silma paistsid, olid seotud aja/arvude vahemikega, kus oli kasutatud sidekriipsu, või võõrsõnad, mille käändelõpp oli lisatud ülakomaga. Näiteks „200-300“ ja „party’lgi“ on EstNLTK poolt sõnestatud „200“, „-“, „300“ ja „party“, „’“, „lgi“, kuid korpuses „200-300“ ja „party’lgi“. Selle probleemi parandamiseks loodi IOB kihid kasutama *Text.words* kihis olevaid sõnesid.

Sündmuste tuvastamisel katsetati kolme IOB märgendusskeemi, mis erinesid üksteisest märgendiga kaasaantud info ja mitmesõnaliste väljendite märgendamise reeglite poolest. Tabelitel 5, 6 ja 7 on esitatud lause „Kontserdil on koos Eesti nooremate lauljate paremik.“ märgendused nende märgendusskeemide järgi.

[7] https://www.geeksforgeeks.org/nlp-iob-tags/

[8] https://github.com/estnltk/estnltk/blob/version_1.6/tutorials/taggers/ner_tagger.ipynb

Tabel 5. Lause märgistusviis *gold_word_events_synced* kihis

text	nertag	sentence_ID
Kontserdil	B-EVENT	12
on	B-EVENT	12
koos	I-EVENT	12
Eesti	O	12
nooremate	O	12
lauljate	O	12
paremik	O	12
.	O	12

Kihis *gold_word_events_synced* on sündmust kirjeldavate fraaside algus tähistatud B-EVENT märgiga, fraasi ülejäänud osad I-EVENT märgiga ning muud sõned O märgiga.

Tabel 6. Lause märgistusviis *gold_word_events_w_classes_synced* kihis

text	nertag	sentence_ID
Kontserdil	B-EVENT_OCCURRENCE	12
on	B-EVENT_STATE	12
koos	I-EVENT_STATE	12
Eesti	O	12
nooremate	O	12
lauljate	O	12
paremik	O	12
.	O	12

Kihis *gold_word_events_w_classes_synced* on märgenditele lisainfona juurde lisatud ka sündmuse klass.

Tabel 7. Lause märgistusviis *gold_word_events_only_main_synced* kihis

text	nertag	sentence_ID
Kontserdil	B-EVENT_OCCURRENCE	12
on	B-EVENT_STATE	12
koos	O	12
Eesti	O	12
nooremate	O	12
lauljate	O	12
paremik	O	12
.	O	12

Kihis *gold_word_events_only_main_synced* on sündmuse märgendus, mis sisaldab ka klassi, jäetud mitmesõnaliste väljendite puhul ainult peasõna juurde. Peasõnaks peetakse fraasi sõna, mille tipp on lause süntaksipuus kõrgemal.

3.1.2 Ajaväljendite kihid

Sarnaselt sündmuste kihile *gold_events*, loodi ajaväljendite jaoks kiht *gold_timexes*, mis sisaldab tervet infot käsitsi märgendatud ajaväljendite kohta. Kihis on kirjas ajaväljendi sõne, lause ID tekstis, sõna ID lauses, ajaväljendi ID, väljend, märgistus, ajaväljendi tüüp ja selle väärtus. Tabelis 8 on esitatud osa *Text.gold_words* väljundist artikli puhul, kus on lause „Aastast 1893 kuni tänavuse suveni oli tornis rist.“.

Tabel 8. Lõik artikli Text objekti *gold_timexes* kihist

text	sentence_ID	word_ID_in_sentence	timex_ID	expression	timex_annotation	type	value
Aastast	4	0	t6	"Aastast 1893 kuni"	TIMEX DATE 1893 multiword="true"	DATE	1893
1893	4	1	t6	"Aastast 1893 kuni"	TIMEX multiword="true"	DATE	1893
kuni	4	2	t6	"Aastast 1893 kuni"	TIMEX multiword="true"	DATE	1893
tänavuse	4	3	t7	"tänavuse suveni"	TIMEX DATE 2002-SU multiword="true"	DATE	2002-SU
suveni	4	4	t7	"tänavuse suveni"	TIMEX multiword="true"	DATE	2002-SU

Kiht tegi kergeks ajaväljendite info hoidmise ja kasutamise. Nendest andmetest loodi ajaväljendite kiht, kus mitmesõnaliste väljendite kirjed olid tehtud üheks.

Algsest *gold_timexes* kihist loodi *gold_timexes_phrases* kiht. Fraaside kirjete koondamine üheks kirjeks võimaldas võrrelda EstNLTK automaatse ajaväljendite tuvastaja tulemusi kuldstandardiga. Tabelis 9 on näha, kuidas tabelis 8 kuvatud fraaside „Aastast 1893 kuni“ ja „tänavuse suveni“ kirjed võeti kokku.

Tabel 9. Lõik artikli *Text* objekti *gold_timexes_phrases* kihist

text	sentence_ID	word_ID_in_sentence	timex_ID	expression	timex_annotation	type	value
['Aastast', '1893', 'kuni']	4	0	t6	"Aastast 1893 kuni"	TIMEX DATE 1893 multiword="true"	DATE	1893
['tänavuse', 'suveni']	4	3	t7	"tänavuse suveni"	TIMEX DATE 2002-SU multiword="true"	DATE	2002-SU

Fraaside kaupa ajaväljendite kirjed tegid kergeks kattuvatele EstNLTK automaatse ajaväljendite tuvastaja kirjetele lisada *timex_ID* väärtused.

Lisaks loodi EstNLTK automaatse ajaväljendite tuvastajaga kiht *timexes*. Automaatse kihi kirjetele lisati juurde *timex_ID* atribuut ning kuldstandardiga kattuvatele ajaväljenditele lisati sellele sobib väärtus. Ajaväljendite *timex_ID* väärtuste ülekandmisel kontrolliti *timexes* ja kuldstandardi ajaväljendite algus- ja lõpuindeksi väärtusi tekstis. Kõige parem tulemus saadi jälgides kuldstandardi kirjeid, mille puhul võrreldavate ajaväljendite algusindeksite ja ka lõpuindeksite vahe oli väiksemvõrdne 15ga või algus- või lõpuindeksid kattusid. Nende kirjete seast valiti kõige väiksema indeksite erinevustega kirje *timex_ID* väärtus. Tabelis 10 on näha *timexes* kihi lõpptulemust artikli puhul, milles on lause „Aastast 1893 kuni tänavuse suveni oli tornis rist.“.

Tabel 10. Lõik artikli *Text* objekti automaatsest ajaväljendite kihtist *timexes* (6 veergu eemaldatud)

text	tid	type	value	temporal_function	timex_ID
['Aastast', '1893']	t4	DATE	1893	False	t6
['tänavuse', 'suveni']	t5	DATE	2002-SU	True	t7

Ajaväljendite *timex_ID* väärtuste ülekandmisel ei õnnestunud igale *timexes* kihi kirjele anda kuldstandardi *timex_ID* väärtust. Selle põhjuseks oli see, et kõiki *timexes* kihis olevaid ajaväljendeid polnud kuldstandardi kihis märgitud, selliseid ajaväljendeid esines 18s artiklis ning neid kokku oli 34. Automaatne ajaväljendite tuvastaja oli ajaväljenditeks pidanud arve, millel puudus ajaline tähendus või oli tegemist vanusega või sünniaastaga, mida kuldstandardi kihis polnud, sest seda ei peetud käsitsi märgendamisel oluliseks. Lisaks leidis see ka teistsuguses käändes olevaid ajaväljendeid. Näiteks peale kuldstandardi kihis märgitud ajaväljenditele „praegu“, „Praegu“ ja „praegu“ leidis see ka ajaväljendi „praeguse“. Kuigi

ajaväljendi „praeguse“ väärtus on teistes käändes olevatega sama, jäeti sellele *timex_ID* väärtuseks *None*, sest tegemist on siiski erineva sõnaga tekstis. Ajaväljendeid, mida kuldstandardi kihis esines, kuid automaatses kihis mitte, oli 86 ning neid esines 46s artiklis.

3.2 Mudelite loomine

Käesolevas töös loodi mudeleid kahe keeletöötluse ülesande lahendamiseks: sündmuste tuvastamine ja ajaseoste määramine. Järgnevates lõikudes räägitakse mudelite loomisest ning nende sisend- ja väljundandmetest.

3.2.1 Sündmusi tuvastavad mudelid

Sündmuste tuvastamisel prooviti kahte lähenemist. Kuna sündmuste õppimine toetus IOB märgendustele, siis oli võimalik kasutada EstNLTK NERtagger'it, mida sai ümber treenida sündmuste tuvastamisele. Teine lähenemine toetus EstBERT[9] mudelile, mis oli juba suurel eesti keelsel korpusel treenitud. Selleks, et EstBERT mudelit kasutada sündmuste tuvastamiseks, tuli seda peenhäälestada. Mõlema mudeli treenimisel kasutati ristvalideerimist täpselt samasuguste treening- ja testandmete jaotustega. Ristvalideerimist tehti 10 alamhulgaga ehk treenimiseks anti ette 72 artiklit ning mudeleid testiti ülejäänud 8 artikli peal. Mudelite hindamiseks arvutati testimishulgal õigsust (ing *accuracy*), täpsust (ing *precision*), saagist (ing *recall*) ja F1-skoori (ing *F1-score*).

EstNLTK NERtagger'i treenimisel sai ette anda *Text* objekti kihi nime, kus kirjeteks oli artikli sõned ja nende IOB märgistused. Selliseid sobivaid kihte sai tehtud kolm nagu peatükis 5.1.1 kirjeldatud sai. Treenimist jooksutati *gold_word_events_synced* (tavaline IOB), *gold_word_events_w_classes_synced* (IOB koos sündmuse klassiga) ja *gold_word_events_only_main_synced* (IOB ainult peasõnal koos klassiga) kihtidel.

Eeltreenitud EstBERT'il põhineva mudeli loomisel võeti eeskuju Tobias Sterbaki artiklist[10], kus ta õpetab BERT mudelit peenhäälestama nimeüksuste tuvastamiseks. Artiklite tekstidele rakendati sõnapiiride tuvastamist ehk üksustamist (tokeniseerimist). Korpuses olev tekstide sõnestus polnud EstBERT mudeli jaoks sobiv ning seega tuli tekstid ümbersõnastada EstBERT mudeli enda sõnestajaga. Ümbersõnastamine tagas selle, et sisendandmed oleksid samasugusel kujul nagu need olid mudeli eeltreenimisel. Sõned võisid muutuda mitmeks osaks ja seetõttu

[9] https://huggingface.co/tartuNLP/EstBERT

[10] https://www.depends-on-the-definition.com/named-entity-recognition-with-bert/

tuli uuendada ka märgendite järjendit. Lausete ja nende märgendite järjendite elemendid muudeti numbrilisteks ning järjendite pikkused ühtlustati 75 peale. Pikkuste ühtlustamiseks lisati lausete järjendite lõppu 0.0 ja märgendite puhul märgend PAD. Lisaks loodi maskeering, et ignoreerida pikkuste ühtlustamiseks lisatud märgendeid. Optimiseerijana (ing *optimizer*) kasutati AdamW õppimise määraga (ing *learning rate*) 3e-5 ning epohhide (ing *epoch*) arvuks määrati 40. Igal epohhil arvutati valideerimishulgal F1-skoor, mida kasutati mudeli treenimise varaseks peatamiseks (ing *early-stopping*) juhul kui F1-skoor ei paranenud viiel järjestikusel epohhil.

3.2.2 Ajaseoseid tuvastavad mudelid

Ajaseoste tuvastamine on jagatud neljaks osaks: ajaseosed sündmuste ja dokumendi loomise aja vahel, ajaseosed samas lauses olevate ajaväljendite ja sündmuste vahel, ajaseosed kõrvuti olevate lausete peamiste sündmuste vahel ja ajaseosed samas lauses olevate sündmuste vahel, kus üks sündmus on lause süntaksipuus kõrgemal kui teine. Iga ülesande jaoks loodi eraldi mudel.

Mudelite treenimisel kasutavateks tunnusteks valiti EstBERT'iga loodud sündmuste ja ajaväljendite vektoreid. Vektorid sõnade jaoks saadi BERT-i peidetud kihtidest. Katsetamiseks võeti nelja viimase kihi vektorid kokku liidetuna, nelja viimase kihi vektorite summa, viimase kihi vektori, eelviimase kihi vektori ja kõikide kihtide vektorite summa. Kuna on sündmusi ja ajaväljendeid, mis koosnevad mitmest sõnast, katsetati neile vektorite valimisel erinevaid lähenemisi. Sündmuste puhul prooviti:

- ainult fraasi peasõna vektorit.
- fraasi sõnade vektorite keskmist väärtust
- fraasi sõnade vektorite kaalutatud keskmist väärtust, kus kõrgem kaal on peasõnal
- fraasi sõnade vektorite kaalutatud keskmist väärtust, kus kaalud on sõnaliikide järgi (suurimast kaalust alates: üheselt tegusõnaks liigitatud sõnad, nud- ja tud-vormid, käändsõnad, ülejäänud sõnaliigid)

Mitmesõnaliste ajaväljendite puhul otsustati kasutada ainult väljendi sõnede vektorite keskmist väärtust. Mudelite väljundiks olid vektoritele vastavad ennustatud relatsioonitüübid.

Iga ülesande jaoks kasutati kõiki olemasolevaid artikleid, kuid nendest saadud sisendandmete hulk varieerus. Sündmuste ja dokumentide loomise aegade vahel olevaid relatsioone kirjeldavaid vektoreid saadi 4010 (sündmuste vektorid), ajaväljendite ja sündmuste

relatsioonide jaoks saadi 559 vektorit (vastavate sündmuste ja ajaväljendite vektorid kokku panduna), kõrvuti olevate lausete peamiste sündmuste relatsioonide jaoks saadi 2552 vektorit (vastavate sündmuste vektorid kokku panduna) ning samas lauses olevate sündmuste relatsioonide jaoks saadi 3127 vektorit (vastavate sündmuste vektorid kokku panduna). Tabelis 11 on näha relatsioonitüüpide jaotuseid erinevate ülesannete andmetes.

Tabel 11. Relatsioonitüüpide jaotused

	Sündmus-dokumendi loomise aeg	**Ajaväljend-sündmus**	**Peamised sündmused**	**Sama lause sündmused**	**Kokku**
SIMULTANEOUS	1	148	228	318	695
BEFORE-OR-OVERLAP	384	63	195	417	1059
AFTER	463	14	370	389	1236
INCLUDES	903	8	326	363	1600
VAGUE	521	2	393	545	1461
OVERLAP-OR-AFTER	190	26	154	141	511
BEFORE	1516	5	554	646	2721
IS_INCLUDED	32	293	321	304	950
IDENTITY	-	-	11	4	15
Kokku	4010	559	2552	3127	10148

Iga ülesande andmetes paistab silma, et esineb relatsioonitüüpe, mida on teistega võrreldes palju vähem. Mudelite treenimisel katsetati relatsioonitüüpide jaotust korrigeerida alavalimise (ing *undersampling*) ja ülevalimisega (ing *oversampling*).

Masinõppe mudelitena kasutati Pythoni teegi Scikit-learn[11] klassifitseerijaid. Katsetati selles teegis olevaid tugivektormasinaid LinearSVC[12] ja SVC[13], lähima naabri klassifikaatorit

[11] https://scikit-learn.org/stable/

[12] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

[13] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

KNeighborsClassifier[14], otsustusmetsa RandomForestClassifier[15] ja närvivõrku MLPClassifier[16]. Klassifitseerijate peal prooviti EstBERT-i peidetud kihtide valimiste ning sündmuste ja ajaväljendite vektorite valimise lähenemiste kombinatsioone.

[14] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[15] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[16] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

4. Analüüs

Selles peatükis on välja toodud loodud mudelite tulemused. Mudelite tulemusi on analüüsitud ning võimaluse korral üksteisega või varasemate tööde tulemustega võrreldud. Varasemate töödega võrdlustes tasub meeles pidada, et võrreldakse mudeleid, mis on loodud erinevate keelte jaoks. Seetõttu esines treeningandmete koguses ja märgendusviisides erinevusi.

4.1 Sündmuste tuvastamine

EstNLTK ümbertreenitud NERtagger'il saadud tulemused on näha tabelis 12. Tabelis on kuvatud ristvalideerimisel (10 alamhulka) saadud tulemuste keskmised.

Tabel 12. EstNLTK NERtagger'i tulemused sündmuste tuvastamisel

	Tavaline IOB	**IOB koos sündmuse klassiga**	**IOB ainult peasõnal koos klassiga**
Õigsus	0,94	0,91	0,93
Täpsus	0,91	0,74	0,74
Saagis	0,79	0,63	0,68
F1-skoor	0,85	0,68	0,71

Kuna tegemist on klassifitseerimisülesandega ning klasside elementide kogused on tasakaalust väljas, on mudelite hindamisel parem vaadata täpsust, saagist või nendest kahest saadud F1-skoori. Õigsuse järgi oleksid need mudelid üksteisega päris võrdsed, kuid teisi mõõdikuid vaadates selgub nende vahe. Kõige parema tulemuse saadi tavalise IOB märgendusega kihti kasutades. Selle mudeli saagiseks saadi 0,79 ja täpsuseks 0,91, mille tulemusel saadi F1-skooriks 0,85. Mudelid IOB koos sündmuse klassiga ja IOB ainult peasõnaga said teineteisele lähedased F-skoorid (vastavalt 0,68 ja 0,71) ning mõlema mudeli puhul on saagis ja täpsus märgatavalt langenud.

EstBERT'il põhineva mudeli tulemusi on näha tabelis 13. Tabelis kuvatud tulemused on ristvalideerimisel (10 alamhulka) saadud tulemuste keskmised.

Tabel 13. EstBERT mudeli tulemused sündmuste tuvastamisel

	Tavaline IOB	IOB koos sündmuse klassiga	IOB ainult peasõnal koos klassiga
Õigsus	0,95	0,94	0,95
Täpsus	0,88	0,78	0,79
Saagis	0,87	0,77	0,79
F1-skoor	0,87	0,77	0,79

EstBERT mudeli täpsus ja saagis on iga märgistussüsteemi korral peaaegu võrdsed. Võrreldes EstNLTK NERtaggeriga, paistab silma tavalise IOB kihi puhul suurem F-skoor ja saagis, kuid väiksem täpsus. Ülejäänud kihtidel on aga nii saagis kui ka täpsus ning seega ka F-skoor märgatavalt paranenud. Tavalise IOB märgistuse F-skoor, milleks on 0,87, on ka suurem peatükis 2.1 mainitud B. Arnulphy jt poolt loodud mudeli saavutatud F-skoorist 0,86.

4.2 Ajaseoste tuvastamine

Ajaseoste tuvastamise ülesannete jaoks kasutati erinevaid mudeleid. See tähendab, et ei loodud mudelit, mis oleks saanud kõigi nelja ajaseoste tuvastamise ülesandega korraga hakkama. Järgnevatel tabelitel kuvatud tulemused on saadud mudelite ristvalideerimise (10 alamhulka) tulemusena.

4.2.1 Sündmuste ja dokumendi loomise aja vahelised ajaseosed

Artiklis esinevate sündmuste ja selle artikli loomisaja vaheliste suhete tuvastamisel tuli kõige paremini toime MLPClassifier ja selle tulemused on näha tabelis 14. Parim tulemus saadi sündmusi kirjeldavate vektoritega, mis olid saadud EstBERT mudeli eelviimasest peidetud kihist valides sündmust kirjeldavaks vektoriks sündmuse peasõna vektori. Mudeli muudetud parameetriteks oli *alpha*=2.8, *solver*="lbfgs", *max_iter*=5000 ja *random_state*=0.

Tabel 14. Sündmuste ja dokumentide loomise aegade vahelisi ajaseoseid tuvastava mudeli tulemused

	Saagis	Täpsus	F1-skoor
Mikrokeskmine	0,66	0,66	0,66
Makrokeskmine	0,49	0,43	0,44
Kaalutatud keskmine	0,66	0,67	0,66
TempEval-2 hinnang (õiged / kõik)	0,66		

Tunduvalt väiksemad makrokeskmise väärtused on põhjustatud tasakaalust väljas olevate andmete tõttu. Kaalutatud keskmiste väärtused arvestavad relatsioonitüüpide arvu ja on seega suuremad. TempEval-2 võitlusel osalevate mudelitega võrreldes saavutas see mudel kehvema tulemuse. Võistlusel saavutatud tulemused olid vahemikus 0,68 – 0,82 (ühe erandiga) ning käesolev mudel saavutas hindeks 0,66.

4.2.2 Ajaväljendite ja sündmuste vahelised ajaseosed

Samas lauses olevate ajaväljendite ja sündmuste ajaseoste tuvastamist sooritas kõige paremini MLPClassifier ja selle tulemused on tabelis 15. Vektorid saadi EstBERT-i nelja viimase peidetud kihi summeerimisel ning sündmuste ja ajaväljendite puhul valiti peasõna vektor. Mudelile määrati parameetriteks *alpha*=3, *solver*="sgd", *max_iter*=10000, *random_state*=0.

Tabel 15. Ajaväljendite ja sündmuste vahelisi ajaseoseid tuvastava mudeli tulemused

	Saagis	**Täpsus**	**F1-skoor**
Mikrokeskmine	0,60	0,60	0,60
Makrokeskmine	0,38	0,28	0,30
Kaalutatud keskmine	0,60	0,65	0,62
TempEval-2 hinnang (õiged / kõik)	0,60		

Mudel kannatab samuti relatsioonitüüpide tõttu, mille kirjeid on vähe. TempEval-2 võistlusel olid mudelite hinnangud 0,54 – 0,65 vahel. Käesolev mudel saavutas hinnanguks 0,60.

4.2.3 Peamiste sündmuste vahelised ajaseosed

Kõrvuti olevate lausete peamiste sündmuste vaheliste ajaseoste tuvastamisega sai kõige paremini toime MPLClassifier ning selle saavutatud tulemused on tabelis 16. Sündmuste vektorid saadi EstBERT-i nelja viimase peidetud kihi summeerimisel ning sündmuste põhisõnade vektoreid valides. Mudelil seadistatud parameetriteks oli *alpha*=2.8, *solver*="lbfgs", max_*iter*=10000 ja *random_state*=0.

Tabel 16. Peamiste sündmuste ajaseoseid tuvastava mudeli tulemused

	Saagis	Täpsus	F1-skoor
Mikrokeskmine	0,51	0,51	0,51
Makrokeskmine	0,52	0,42	0,43
Kaalutatud keskmine	0,51	0,53	0,52
TempEval-2 hinnang (õiged / kõik)	0,51		

Kuigi relatsioonitüübid pole tasakaalus, saagise makrokeskmine pole mikro- ja kaalutatud keskmisest väiksem. TempEval-2 võistlusel osalenud mudelite hinnangud olid 0,45 – 0,58 vahel. Loodud mudel sai hinnanguks 0,51.

4.2.4 Erineval süntaksipuu tasemel olevate sündmuste vahelised suhtetüübid

Samas lauses olevate sündmuste vahel, kus üks sündmus on lause süntaksipuus kõrgemal kui teine, sai ajaseoste tuvastamisega kõige paremini hakkama samuti MLPClassifier, mille tulemused on näha tabelis 17. Sündmuste vektorid saadi EstBERT-i eelviimasest peidetud kihi vektoritest, kus leiti sündmuste vektorite keskmise. Mudeli parameetriteks valiti *alpha*=1, *solver*="lbfgs", *max_iter*=10000 ja *random_state*=0.

Tabel 17. Erineval süntaksipuu tasemel olevate sündmuste ajaseoseid tuvastava mudeli tulemused

	Saagis	Täpsus	F1-skoor
Mikrokeskmine	0,46	0,46	0,46
Makrokeskmine	0,39	0,38	0,38
Kaalutatud keskmine	0,46	0,47	0,46
TempEval-2 hinnang (õiged / kõik)	0,46		

Mudeliga saavutatud TempEval-2 hinnang 0,46 jääb alla võistlusel osalenutele (v.a üks erand). Võistlusel parema tulemuse saanud osalenute hinnangud jäid 0,56 – 0,66 vahele.

5. Kokkuvõte

Käesoleva töö eesmärgiks oli luua juhendatud masinõppel põhinevad mudelid eestikeelsetest ajalehetekstidest sündmuste tuvastamiseks ja nendes tekstides olevate sündmuste vaheliste ajaseoste ning sündmuste ja ajaväljendite vaheliste ajaseoste tuvastamiseks. Selleks vajalikud ajaleheartiklid saadi Eesti TimeML korpusest, kus oli sündmused ja ajaväljendid ning nende suhted käsitsi märgendatud. Töös sai antud ülevaade märgendusstandardist TimeML ja sellest inspireeritud eesti keelele mõeldud TimeML märgendusstandardist. Lisaks käsitleti varasemaid katsetusi nii sündmuste kui ka sündmuste ja ajaväljendite ajaseoste tuvastamisel. Kirjeldatud on ka töö eesmärgi saavutamiseks tehtud tööd (andmete töötlemine ja mudelite loomine) ja loodud mudelite analüüs.

Bakalaureusetöö raames viidi EstTimeML korpus EstNLTK 1.6 Text objektide kujule ning saadi valmis kaks sündmusi tuvastavat mudelit ja neli mudelit ajaseoste tuvastamiseks. Sündmusi tuvastavaid mudeleid treeniti BIO märgendusformaati kasutades. Ühe mudeli puhul on tegemist ümberõpetatud EstNLTK 1.6 nimeüksuste tuvastajaga ja teine on nimeüksuste tuvastamisele peenhäälestatud EstBERT mudel. Lisaks lihtsalt sündmuste tuvastamisele katsetati ka sündmuse klassi tuvastamist. Sündmuste tuvastamisel sai EstNLTK mudel F1-skooriks 0,85 ning EstBERT mudel 0,87. Koos klassiga tuvastamisel said mudelid F1-skooriks 0,68 ja 0,77. Tuvastades ainult sündmuse peasõna koos selle klassiga saadi F1-skoorideks 0,71 ja 0,79. Ajaseoseid tuvastavad mudelid kasutasid sisendandmetena sündmuste ja ajaväljendite vektoreid, mis olid saadud EstBERT mudeli peidetud kihtidest. Sündmuste ja dokumentide loomise aegade vahel ajaseoste tuvastamisel saavutas mudel F1-skooriks 0,66. Samas lauses olevate ajaväljendite ja sündmuste ajaseoste tuvastamisel saadi F1-skooriks 0,60. Kõrvuti olevate lausete peamiste sündmuste ajaseoste tuvastamisel saadi F1-skooriks 0,51 ning sündmuste vahel, kus üks sündmus on lause süntaksipuus kõrgemale kui teine saadi F-skooriks 0,46.

Käesoleva bakalaureusetöö edasiarendusena oleks võimalik luua eraldi mudel tuvastatud sündmuste klasside määramiseks. Sel moel oleks sündmuste tuvastamine ja neile klassi määramine eraldatud ehk mudelite seadistamine oleks kergendatud. Sellise lähenemisega võib olla võimalik saada sündmused koos nende klassiga ilma, et sündmuste leidmine või korrektse klassi määramine kannataks. Ajaseoste tuvastamisel saaks katsetada sisendandmetele sündmuste ja ajaväljendite ümber olevate sõnade vektorite lisamist, muu lisainfo kaasamist või relatsioonitüüpide hulga lihtsustamist. Samuti võib proovida mudeli loomist, mis suudaks kõiki

nelja suhtetüübi tuvastamise ülesannet täita. Võimalik ka muid masinõppemeetodeid katsetada. Suurim edasiarendus oleks süsteemi loomine, mis suudaks ajaseoseid ära tunda automaattuvastatud sündmuste ja ajaväljendite vahel. See võimaldaks luua rakenduse, mis sisendiks antud tekstis leiab sündmused ja ajaväljendid ning ka nendega soetud ajaseosed. Selline rakendus lihtsustaks käsitsi märgendust, et luua veelgi rohkem treeningandmeid ja seega treenida veelgi paremaid mudeleid.

Viidatud kirjandus

[1] Pustejovsky, J., Castaño, M. J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., Katz, G., Radev, R. D. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New Directions in Question Answering. Papers from 2003 AAAI Spring Symposium, Stanford University.* Stanford, CA, USA: 2003, 28-34. https://www.researchgate.net/publication/221441154_TimeML_Robust_Specification_of_Event_and_Temporal_Expressions_in_Text (06.12.2020)

[2] SAS Help Center: Working with Dates and Times by Using the ISO 8601 Basic and Extended Notations. https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/leforinforref/p1a0qt18rxydrkn1b0rtdfh2t8zs.htm (23.04.2021)

[3] Orasmaa, S. Towards an Integration of Syntactic and Temporal Annotations in Estonian. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Iceland: ELRA, 2014, 1259-1266. https://www.aclweb.org/anthology/L14-1439/ (03.04.2021)

[4] Arnulphy, B., Claveau, V., Tannier, X., Vilnat, A. Supervised Machine Learning Techniques to Detect TimeML Events in French and English. *Natural Language Processing and Information Systems. NLDB 2015. Lecture Notes in Computer Science, vol 9103.* Springer, Cham. 2015. https://doi.org/10.1007/978-3-319-19581-0_2 (06.12.2020)

[5] Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., Pustejovsky, J. The TempEval challenge: identifying temporal relations in text. *Language Resources and Evaluation 43*. 2009, 161-179 https://doi-org.ezproxy.utlib.ut.ee/10.1007/s10579-009-9086-z (04.04.2021)

[6] Verhagen, M., Sauri, R., Caselli, T., Pustejovsky, J. SemEval-2010 Task 13: TempEval-2. *Proceedings of the 5th International Workshop on Semantic Evaluation*, 57-62. Uppsala, Sweden: Association for Computational Linguistics, 2010, 57-62. https://www.aclweb.org/anthology/S10-1010/ (14.04.2021)

[7] Orasmaa, S. Automaatne ajaväljendite tuvastamine eestikeelsetes tekstides. *Eesti rakenduslingvistika ühingu aastaraamat 8*, 153-169. http://dx.doi.org/10.5128/ERYa8.10 (04.04.2021)

[8] Tkachenko, A., Petmanson, T., Laur, S. Named Entity Recognition in Estonian. *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*. Sofia, Bulgaria: Association for Computational Linguistics, 2013, 78-83. https://www.aclweb.org/anthology/W13-2412/ (24.04.2021)

[9] Tanvir, H., Kittask, C., Eiche, S., Sirts, K. EstBERT: A Pretrained Language-Specific BERT for Estonian. 2020. https://arxiv.org/abs/2011.04784 (13.04.2021)

[10] Laur, S., Orasmaa, S., Särg, D., Tammo, P. EstNLTK 1.6: Remastered Estonian NLP Pipeline. *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020, 7152-7160. https://www.aclweb.org/anthology/2020.lrec-1.884/ (24.04.2021)

Lisad

I. Kirjutatud kood ja kasutatud failid

Töö käigus kirjutatud kood, katsetused ja kasutatud failid on kättesaadavad lehel https://github.com/TaanielS/Ajaseoste-automaatne-tuvastamine-tekstis.

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Taaniel Saarnik**,

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

   Ajaseoste automaatne tuvastamine tekstis,

   mille juhendaja on **Siim Orasmaa**,

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Taaniel Saarnik

07.05.2021