EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 4, 5–16

STATISTILISE KEELEMUDELI ADAPTEERIMINE EESTI KEELE KÕNETUVASTUSES

Tanel Alumäe

Ülevaade. Artiklis käsitletakse eesti keele suure sõnavaraga kõnetuvastuse statistilise keelemudeli adapteerimist. Adapteerimise lähteandmeteks on väike teemaspetsiifiliste lausete korpus. Adapteerimise käigus leitakse varjatud semantika analüüsi (LSA) abil suurest dokumendikorpusest antud teemale lähedaseimad tekstid. Saadud tekstide põhjal konstrueeritakse uus teemaspetsiifiline unigramm-mudel ning see kombineeritakse üldise *N*-gramm-mudeliga, mille tulemusena saadakse teemale adapteeritud *N*-gramm-mudel. Artiklis võrreldakse morfeeme, sõnu ja lemmasid adapteerimismudeli põhiüühikutena.

Meetodit testitakse raadiouudistesalvestuste tuvastamisel. Tuvastuse esimeses faasis leitakse üldise keelemudeli abil igale uudisnupule tuvastushüpoteesid, mida kasutatakse keelemudeli adapteerimiseks. Tuvastuse teises faasis kasutatakse adapteeritud keelemudelit uute tuvastushüpoteeside saamiseks. Tulemused näitavad, et adapteerimisega saavutatakse oluline tuvastuskvaliteedi paranemine. Selgub, et morfeemidepõhisel adapteerimisel saavutatud 10-protsendiline vigade vähenemine on statistiliselt oluliselt suurem kui sõna- või lemmapõhisel adapteerimisel saadud muutused. Artiklis analüüsitakse ka saadud tulemuste võimalikke põhjuseid.*

Võtmesõnad: kõnetuvastus, keelemudeli adapteerimine, LSA, lemmatiseerimine, morfeemid

1. Sissejuhatus

Suure sõnavaraga kõnetuvastuses kasutatakse statistilist keelemudelit sõnade aprioorse kontekstuaalse tõenäosuse hindamiseks. Tüüpiliselt kasutatakse keelemudelina *N*-gramm-mudelit, kus eeldatakse, et sõna kontekstuaalne tõenäosus sõltub ainult temale eelnevast *N*-1 sõnast. Tavaliselt piirdutakse trigramm-mudelitega

* Artikkel on valminud riikliku programmi "Eesti keele keeletehnoloogiline tugi (2006–2010)" rahalisel toel.

(*N*=3). Statistiline keelemudel saadakse suure tekstikorpuse (kümned kuni sajad miljonid sõnad) analüüsi põhjal.

Statistilise keelemudeli adapteerimine on ülesanne, kus väikese olemasoleva teemaspetsiifilise korpuse põhjal kohandatakse üldist keelemudelit nõnda, et ta sobib paremini antud teemale. Kohandamise tulemusena peaks kõikide teemaga lähedalt seotud olevate sõnade ja sõnakombinatsioonide tõenäosused suurenema ning antud teemast semantiliselt kaugel olevate sõnade tõenäosused vähenema.

Viimastel aastatel on mitme keele kõnetuvastuse keelemudeli adapteerimiseks edukalt kasutatud nn varjatud semantika analüüsi (ingl *latent semantic analysis*, LSA) (vt nt Bellegarda 1998). Meetod kohandab järk-järgult keelemudelit vastavalt hiljuti tuvastatud sõnadele, kasutades sõnade dokumentides koosesinemise statistikat keelemudelis olevate sõnade unigramm-tõenäosuse ümberarvutamiseks.

See lähenemine aga ei pruugi olla otstarbekas flekteeruvate ja aglutinatiivsete keelte puhul. Sellistes keeltes on erinevate sõnavormide arv väga suur, mistõttu statistilises keelemudelis kasutatakse sõnade asemel morfeeme (Alumäe 2006) või tekstikorpuse statistika põhjal leitud morfeemilaadseid ühikuid (Siivola jt 2003). Et rakendada standardset LSA-põhist adapteerimismeetodit, peaks ka LSA-mudelis dokumentide esitamiseks kasutama morfeeme. Kuna LSA kasutab dokumentide esitamiseks järjestamata sõnade esinemissageduse ehk nn *bag-of-words* meetodit, tekkis autoril kahtlus, et dokumendi esitamine morfeemide esinemissagedusena annab dokumendi sisust vähem aimu, kui dokumendi esitamine sõnade või lemmade esinemissagedusena.

Selles artiklis tutvustatakse LSA-põhist statistilise keelemudeli adapteerimismeetodit, kus LSA-mudelis kasutatavad ühikud ei pruugi kattuda keelemudeli ühikutega. See annab võimaluse kasutada semantiliste seoste modelleerimiseks sõnu, lemmasid, morfeeme või muid ühikuid. Treeningu käigus viiakse dokumendikorpuses olevad tekstid soovitud kujule, st leitakse neis olevate valitud ühikute sisaldus. Selle põhjal arvutatakse LSA sarnasusmudel. Kõnetuvastus koosneb siis kahest faasist: esimeses faasis leitakse igale tuvastatavale lausele *N* parimat lausekandidaati. Parimad lausekandidaadid viiakse LSA-mudeliga ühilduvale kujule (st näiteks lemmatiseeritakse) ning selle põhjal leitakse aktiivse teema kujutis LSA-ruumis. Seejärel leitakse antud teemale kõige lähedasemad treeningdokumendid ning nende põhjal kohandatakse esimeses faasis kasutatud üldist keelemudelit. Saadud adapteeritud keelemudeli abil arvutatakse esimeses faasis saadud lausekandidaatidele uued keelemudelipõhised skoorid ning selle põhjal leitakse uued parimad lausehüpoteesid.

2. LSA

Varjatud semantika analüüs (LSA) (Landauer 1998) on korpusepõhine matemaatiline meetod sõnade ja dokumentide semantilise sarnasuse arvutamiseks ja esitamiseks. LSA ülesandeks on *M* sõnast koosneva sõnahulga *V* ja *N* dokumendist koosnevad dokumendikorpuse *T* kujutamine vektorruumis, nii et iga sõna hulgas *V* ja iga dokument hulgas *T* oleks esitatavad selles ruumis. Selleks konstrueeritakse kõigepealt $M \times N$ maatriks W, mille iga elemendi W_{ij} väärtus on sõna w_i esinemisarvu kaal dokumendis d_j. Kasutatav kaalufunktsioon peaks arvesse võtma nii sõna

tähtsust antud dokumendis kui ka sõna semantilist informatiivsust. Üheks selliseks kaaluks (Bellegarda 1998), mida ka antud töös kasutatakse, on

$$(1)\quad W_{ij} = (1 - \varepsilon_i) \log_2 \left(1 + \frac{c_{ij}}{n_j} \right) ,$$

kus c_{ij} on sõna w_i esinemisarv dokumendis d_j, n_j on dokumendis d_j olevate sõnade arv ning ε_i on sõna w_i normaliseeritud entroopia korpuses T. Normaliseeritud entroopia vastab sõna w_i negatiivsele semantilisele informatiivsusele ning see arvutatakse järgmiselt:

$$(2)\quad \varepsilon_i = -\frac{1}{\log(N)} \sum \frac{c_{ij}}{t_j} \log \frac{c_{ij}}{t_j} .$$

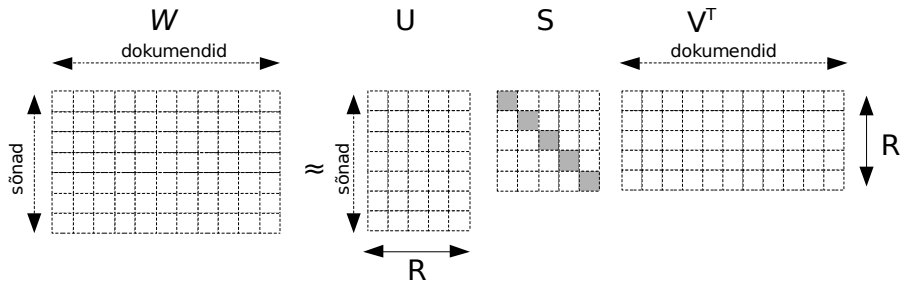
Siin $t_i = \sum c_j$ on sõna w_i esinemisarv korpuses T. Ühtlaselt ja paljudes dokumentides esinevate sõnade (näiteks side- ja asesõnad) entroopiaväärtus on lähedane ühele, ning suhteliselt vähestes dokumentides esinevate sõnade entroopiaväärtus lähedane nullile (nt allpoolkirjeldatud eksperimentides on sõna *ja* normaliseeritud entroopia 0,96 ning sõna *kvantarvuti* vastav väärtus 0,17). Sellise kaalufunktsiooni kasutamise tulemusena rõhutatakse maatriksis W rohkem selliseid sõnu, mille semantiline informatiivsus on suur (ning entroopia seega väike): on ju selge, et kui kahes dokumendis kasutatakse sagedasti sõna *ja*, siis ei pruugi nad olla sisult sarnased; kui aga mõlemas kasutatakse sagedasti sõna *kvantarvuti*, siis on dokumentide sisuline lähedus palju tõenäolisem.

Saadud maatriksi W veerud kirjeldavad korpuses olevaid dokumente skaleeritud sõnade esinemissagedustega. Igale dokumendile vastab M-mõõtmeline veeruvektor, kus M on tüüpiliselt 20 000–60 000. Kahe dokumendi võrdlemiseks tuleks siin võrrelda kahe dokumendivektori vastavaid elemente. Dokumendivektori suurest dimensionaalsusest tingituna on see üsna arvutusmahukas, samuti on olemas oht, et sisult sarnased dokumendid kasutavad pisut erinevat sõnavara (nt sünonüüme), mille tulemusena dokumentide kaugus vektorruumis tuleks ikkagi küllalt suur. Seetõttu on leitud (Deerwester jt 1990), et dimensionaalsuse vähendamiseks ning kompaktsete tunnuste esiletoomiseks on kasulik rakendada maatriksile W nn lõigatud singulaarlahutust (ingl *truncated singular value decomposition*, SVD).

Järguga R lõigatud singulaarlahutuse tulemusena lahutatakse esialgne maatriks W kolmeks faktoriks

$$(3)\quad W \approx \hat{W} = USV^T,$$

kus U on vasakpoolsete singulaarvektorite u_i maatriks ($M \times R$), S on R singulaarväärtusest koosnev diagonaalmaatriks ning V on parempoolsete singulaarvektorite v_j maatriks ($N \times R$). Maatriks $\hat{W}$ on R järku parim lähendus algsele maatriksile W. Vektorid u_i esitavad sõnu w_i ja vektorid v_j dokumente d_j saadud LSA-ruumis (vt joonis 1).



Joonis 1. Singulaarlahutus sõna–dokument maatriksist

Kui enne singulaarlahutuse rakendamist esitati dokumente *M*-mõõtmeliste maatriksi veeruvektoritega, siis nüüd vastab igale dokumendile *R*-mõõtmeline vektor $v_i S$. Kuna *R* valitakse tüüpiliselt vahemikust 100...300 (autori eksperimentides *R*=200), siis saavutatakse sellise teisendusega oluline dimensionaalsuse vähenemine. Samuti on leitud, et lõigatud SVD leiab algse maatriksi *W* kõige olulisemad komponendid ning ignoreerib kõrgemat järku struktuure, mida võib pidada müraks (Yang 1995). Teisisõnu, vektor $v_i S$ esitab kompaktselt dokumendi d_j sõnavara.

3. Keelemudeli adapteerimine

Keelemudeli adapteerimiseks kasutatakse üldise keelemudeli marginaalide adapteerimist vastavalt teemaspetsiifilistele lausetele semantiliselt lähedasemate dokumentide analüüsist saadud unigramm-tõenäosustele. Semantiliselt lähedasemad dokumendid leitakse eelnevalt kompileeritud LSA-mudeli põhjal.

3.1. Lähedaste dokumentide leidmine

Semantiliselt lähedaste dokumentide leidmiseks tuleb esmalt olemasolevatest teemaspetsiifilistest lausetest e adapteerimislausetest koosnev pseudodokument viia LSA-ruumiga ühilduvale kujule. Selleks tuleb kõigepealt leida lausetes esinevate erinevate ühikute (morfeemide, sõnade või lemmade) arv ning siis arvutada uue pseudodokumendi vektor $\tilde{d}_p$ vastavalt valemile (1). Saadud pseudodokumendi kujutis LSA-ruumis on siis

$$\tilde{v}_p = \tilde{d}_p^T U S^{-1}. \quad (4)$$

Seejärel võib arvutada kauguse pseudodokumendi vektori $\tilde{v}_p$ ja kõikidele korpuses olevatele dokumentidele vastavate vektorite v_i vahel, leides nurga vektorite $\tilde{v}_p S$ ja $v_i S$ vahel:

$$K(\tilde{v}_p, v_i) = \angle(\tilde{v}_p S, v_i S) = \arccos \frac{\tilde{v}_p S^2 v_i^T}{\|\tilde{v}_p S\| \cdot \|v_j S\|} \quad (5)$$

Kauguse põhjal järjestatakse kõik korpuses olevad dokumendid ning keelemudeli adapteerimiseks valitakse *L* antud teemale kõige lähedasemat dokumenti.

3.2. Kaalutud unigramm-mudeli arvutamine

Keelemudeli kohandamisel saadud dokumendikauguste teadmisel oleks kasulik, kui semantiliselt lähedasemad dokumendid annaksid adapteerimisel suuremat kaalu kui kaugemad dokumendid. Selleks arvutatakse *L* lähedaseimas dokumendis esineva *i*-nda keeleühiku kaalutud arv järgmiselt:

$$(6)\quad c_{adapt}(i) = \sum_{j \in C_L} \left(1 - \frac{K(\tilde{v}_p, v_j)}{\pi}\right) c_{ij} \ ,$$

kus C_L on *L* lähedaseimale dokumendile vastav indeksite hulk. Saadud murdosalistest keeleühiku esinemisarvudest koostatakse unigramm-keelemudel, kasutades nn loomulikku diskonteerimist (Ristad 1995). See diskonteerimismeetod valiti seetõttu, et teda on hõlbus kasutada murdarvuliste esinemisarvude puhul ning ta andis arendusfaasis häid tulemusi.

3.3. Kiire marginaalide adapteerimine

Kiire marginaalide adapteerimine (ingl *fast marginal adaptation*, FMA) (Kneser jt 1997) on meetod, millega saab üldise *N*-gramm keelemudeli kiiresti adapteerida teemaspetsiifilisele tekstile. Meetod kasutab algse keelemudelina segakorpusel treenitud keelemudelit. Adapteerimise käigus muudetakse algset keelemudelit nii, et selle marginaaljaotus oleks võrdne teemaspetsiifilistel tekstidel treenitud unigramm-mudeliga. Selgub, et see on samaväärne üldises *N*-gramm-mudelis olevate tõenäosuste skaleerimisega

$$(7)\quad P_{adap}(w \mid h) = \frac{\alpha(w) P_{BG}(w \mid h)}{Z(h)} \ ,$$

kus $P_{adap}(w|h)$ on adapteeritud sõna *w* kontekstuaalne tõenäosus, $P_{BG}(w|h)$ on sõna *w* esialgne kontekstuaalne tõenäosus ning *Z(h)* on normaliseerimisfaktor, mis tagab tõenäosuste summeerumise üheks. Skaleerimisfaktor $\alpha(w)$ on ligikaudselt

$$(8)\quad \alpha(w) \approx \left(\frac{P_{adap}(w)}{P_{BG}(w)}\right)^{\beta} \ ,$$

kus $P_{adap}(w)$ on sõna *w* adapteerimisdokumentide põhjal leitud unigramm-tõenäosus, $P_{BG}(w)$ sõna *w* unigramm-tõenäosus üldise keelemudeli järgi, ning β valitud faktor vahemikus 0...1, mis eksperimentides seati võrdseks 0,5-ga. Skaleerimisfaktori ülesandeks on keeleühikute tõenäosuse suurendamine või vähendamine, vastavalt ühiku suhtelisele tõenäosusele adapteerimiskorpuses võrreldes üldise segakorpusega. Normaliseerimisfaktor on esitatav kujul

$$(9)\quad Z(h) = \sum_{w} \alpha(w) P_{BG}(w \mid h) \cdot$$

4. Eksperimendid

4.1. Eksperimentaalse ülesande kirjeldus

Eksperimentaalseks ülesandeks võeti Eesti Raadio (ER) uudiste transkribeerimine. Kõnematerjal koosneb Vikerraadio täistunni lühiuudistest, mis on osa valmivast ER uudistekorpusest. Eksperimentideks valiti juhuslikult mõned olemasolevad uudistesaadete salvestused. Saadetes olev kõne transkribeeriti käsitsi ning segmenteeriti uudisnuppudeks ja lauseteks. Segmendid, mis ei sisaldanud kõnet vaid muud helimaterjali (nt ava- ja lõppsignatuurid), eemaldati. Meetodi testimiseks kasutati 21 minutit kõnematerjali, mis koosneb 44 uudisnupust ja 193 lausest. Erinevatele parameetritele parimate väärtuste leidmiseks kasutati 21 minutit kõnematerjali, mis koosneb 20 uudisnupust ja 101 lausest. Keskmine lausete arv ühes uudisnupus oli seega 4,6.

4.2. Akustilised mudelid

Kuna terve ER uudistekorpus ei ole veel transkribeeritud, ei saanud akustiliste mudelite treenimiseks kasutada uudistes olevat kõnet. Selle asemel kasutati mudelite treenimiseks eesti keele SpeechDat-tüüpi kõneandmebaasi (Meister jt 2003). Treenimiseks kasutati ainult andmebaasis olevaid aktsepteeritava kvaliteediga salvestusi, mille koguarv on 2969, kestusega 241 tundi. Erinevate kõnelejate arv on 1332. Kõne on digitaliseeritud 8 kHz sagedusel kasutades 8-bitist A-law-kodeeringut.

Akustiliste mudelite treenimiseks rakendati vaba lähtekoodiga SphinxTrain[1] tarkvarapaketti. Mudelite hulgas on 25 foneemmudelit, 5 erinevat müramudelit ning vaikuse mudel. Akustiliste tunnustena kasutati mel-sagedusskaalaga kepstrumkoefitsiente (ingl *mel frequency cepstrum coefficients*, MFCC), mis arvutati 130 Hz–3400 Hz laiusest sagedusribast. Sämplimisakna laius oli 25,6 ms ning sämplimisperiood 10 ms. Igast aknast arvutati 512-punktilise kiire Fourier' teisenduse (ingl *fast Fourier transform*, FFT) abil 31-ribaline filterpank, mida omakorda kasutati 13 kepstrumkoefitsiendi arvutamiseks. Akustiliste mudelitena kasutatakse kolme emiteeriva olekuga nn paremalt-vasakule topoloogiaga Markovi peitmudeleid.[2] Väljundvektorid on 39-kohalised ning koosnevad 13 kepstrumkoefitsiendist ning neile vastavatest delta- ning delta-delta-koefitsientidest.[3] Treenimise käigus saadakse seotud olekutega trifoonmudelid, millel on kokku 8000 olekut. Igat olekut modelleeritakse kaheksa Gaussi jaotuse summaga.

Enne tuvastamist adapteeriti SpeechDat-kõneandmebaasi põhjal treenitud mudeleid u 15 minuti ER uudistekorpuses oleva käsitsi transkribeeritud kõne põhjal.

Kõnetuvastussüsteemi häälduss õnastik koostati automaatselt sõnade ortograafia põhjal, kasutades väikest arvu kontekstitundlikke reegleid ning käsitsi leitud reegleid tähtsamate võõrnimede ortograafia eestipärasele kujule viimiseks.

[1] Vt http://cmusphinx.org (01.01.2007).

[2] Paremalt-vasakule topoloogiaga Markovi peitmudeliteks nimetatakse sellise topoloogiaga mudeleid, kus puuduvad nn *skip*-üleminekud, st mudeli suvalisest olekust saab liikuda ainult järgmisesse olekusse, või samasse olekusse tagasi. Sellise mudeli käivitamisel külastatakse alati kindlas järjekorras kõiki mudeli olekuid.

[3] Delta- ja delta-delta-koefitsiendid võrduvad vastavate koefitsientide esimest ja teist järku tuletistega.

4.3. Keelemudel

Üldise statistilise keelemudeli treenimiseks kasutati järgnevaid osasid Tartu Ülikooli (TÜ) eesti keele segakorpusest (Kaalep, Muischnek 2005): ajalehed Postimees (33 miljonit sõna), Eesti Ekspress (7,5 miljonit sõna), Maaleht (4,3 miljonit sõna), ilukirjandus (4,2 miljonit sõna), ajakirjad Akadeemia (7 miljonit sõna) ning Kroonika (0,6 miljonit sõna), riigikogu stenogrammid (13 miljonit sõna). Lisaks sellele koguti Internetist Eesti Päevalehe artiklite korpuse (93 miljonit sõna) ja etv24.ee uudistekorpuse (4,8 miljonit sõna).

Keelemudeli komplekteerimiseks kasutati tarkvarapaketti SRILM (Stolcke 2002). Kõik tekstid töödeldi eelnevalt eesti keele morfoloogiaanalüsaatori ja -ühestaja (Kaalep, Vaino 2001) abil, mille põhjal segmenteeriti sõnad morfeemideks. Keelemudeli sõnavaraks on 60 000 morfeemi, mis valiti maksimaalse tõepära meetodil kõigi alamkorpuste segust TÜ tasakaalustatud korpuse põhjal. Kasutades saadud sõnavara, kompileeriti kõikidele alamkorpustele oma trigramm-mudel. Trigramm-mudelite arvutamisel kasutati SRILM-is realiseeritud modifitseeritud Kneser-Ney diskonteerimist. Kõik trigramm-mudelid kompileeriti lõpuks üheks üldiseks mudeliks, kasutades TÜ tasakaalustatud korpust optimaalsete kombineerimiskaalude leidmiseks.

Saadud keelemudeli entroopia-mõõde (ingl *perplexity*) kõnetuvastuseksperimentides kasutatavate uudislõikude transkriptsioonide suhtes on 128. Keelemudeli sõnavara katvus on 99,5%.

4.4. LSA-mudeli konstrueerimine

Semantiliste seoste modelleerimiseks konstrueeriti kolm LSA-mudelit: sõnade-, lemmade- ja morfeemidepõhine. Lähteandmetena kasutati ülaltoodud ajalehekorpuste ja etv24.ee artikliteks segmenteeritud tekste. Kokku on nendes korpustes u 500 000 artiklit. Lemmapõhise mudeli konstrueerimiseks asendati kõik artiklites olevad sõnad vastavate lemmadega, kasutades morfoloogiaanalüsaatorit. Morfeemidel põhineva mudeli konstrueerimiseks segmenteeriti sõnad morfeemideks. Sõna- ja lemmapõhise mudeli sõnavaraks võeti 60 000 kõige sagedasemat vastavat ühikut. Morfeemipõhise mudeli sõnavaraks võeti samad 60 000 morfeemi, mis olid eelnevalt valitud statistilise keelemudeli sõnavaraks. Sõnapõhise mudeli esialgses sõna–dokument maatriksis on u 84 miljonit nullist erinevat elementi. Lemmapõhise mudeli vastav väärtus on 79 miljonit ning morfeemipõhisel mudelil 117 miljonit.

Algsete maatriksite konstrueerimiseks ning dokumentide läheduse arvutamiseks kasutati autori loodud tarkvara. Maatriksite singulaarlahutuse arvutamiseks rakendati tarkvarapaketti PROPACK.[4]

[4] http://sun.stanford.edu/~rmunk/PROPACK/ (01.01.2007).

4.5. Kõnetuvastuse metoodika

Kõnetuvastuseksperimentideks kasutati tarkvarapaketti CMU Sphinx 3.6.3.[5] Tuvastuse esimeses faasis genereeriti igale lausungile 1000 parimat lausehüpoteesi. Iga uudisnupu parimatest lausehüpoteesidest konstrueeriti pseudodokumendid, mis projitseeriti LSA-ruumi. Igale uudisnupule leiti selle põhjal 1200 sarnaseimat dokumenti. Nende dokumentide põhjal konstrueeriti igale uudisnupule teemaspetsiifiline unigramm-mudel, mille abil adapteeriti üldises trigramm-mudelis olevaid kontekstuaalseid tõenäosusi. Seejärel arvutati igale esimeses faasis saadud lausehüpoteesile uued keelemudelipõhised skoorid. Need kombineeriti olemasolevate akustilise mudeli põhiste skooridega ning selle põhjal leiti igale lausungile uus parim lausehüpotees.

4.6. Tulemused

Kõnetuvastuse kvaliteedi hindamiseks kasutati tähevigade osakaalu (ingl *letter error rate*, LER), mis on defineeritud kui

$$(10)\ LER = \frac{S + D + I}{N} \cdot 100\ ,$$

kus S on asendusviga, D kustutusviga, I vahelekirjutusviga ning N lauses olev tegelik tähtede arv. Sõnadevahelist tühikut käsitleti eraldi tähena, et võtta arvesse võimalikke kokku-lahku-kirjutamise vigu. Vigade osakaalu mõõdeti enne ja pärast adapteeritud keelemudeli rakendamist. Tulemused on toodud tabelis 1. Tähevigade järel on toodud adapteerimisega saavutatud suhteline vigade vähenemise protsent, võrrelduna ilma adapteerimiseta saadud tulemustega.

Tabel 1. Tähevigade osakaal enne ja pärast adapteerimist, koos suhtelise vigade arvu muutusega võrrelduna enne adapteerimist saadud tulemustega

Adapteerimine	Tähevigade osakaal	Suhteline vigade arvu muutus
–	7,1%	
sõnapõhine	6,7%	–6%
lemmapõhine	6,6%	–7%
morfeemipõhine	6,4%	–10%

Tulemuste statistiliseks kontrolliks rakendati Wilcoxoni märgistatud astaktesti,[6] mille abil analüüsiti tuvastuskvaliteeti erinevate uudislugude kaupa enne ja pärast adapteerimist. Osutus, et kõik adapteeritud mudelid parandasid statistiliselt oluliselt adapteerimata süsteemi tulemusi. Morfeemipõhine adapteerimine osutus oluliselt paremaks kui teised adapteerimisviisid, lemma- ning sõnapõhisel adapteerimisel olulist vahet ei olnud.

Tabelis 2 on toodud mõne uudisnupu esimene lause ning morfeemipõhise LSA-mudeli abil leitud sarnaseimate artiklite pealkirjad. Iga uudisnupu kohta on toodud neli talle lähedaseimat artiklit.

[5] http://cmusphinx.org (01.01.2007).

[6] Wilcoxoni märgistatud astaktesti kasutatakse kahe valimi võrdlemiseks juhul, kui valimite jaotus erineb oluliselt normaaljaotusest. Test näitab, kas valimite vaheline erinevus on oluline või võib seda seletada juhusega.

Tabel 2. Mõned uudisnupud (toodud uudise esimene lause) ja neile lähedaste artiklite pealkirjad

Uudis	Lähedased artiklid
Seoses Tony Blairi visiidiga tõhustatakse tänasest alates Eesti piirikontrolli.	1. Leedu peaminister sõidab visiidile Riiga. 2. Kallas sõidab Briti peaministriga kohtuma. 3. Balti siseministrid kohtuvad Saaremaal. 4. Balti riikide sisepiiride valvamiseks pole raha.
Laskesuusatamise maailmakarikasarja avaetapil Östersundis ei pääsenud Eesti teatenelik starti.	1. Kuus hemoglobiinitasemega hädas olnud sportlast said loa võistelda. 2. Venemaa laskesuusatajad protestivad. 3. Teatesõite valitsesid Venemaa ja Norra. 4. Venelased kaebavad Lazutina ja Danilova karistused edasi.
Suurbritannia suurendab sõdurite arvu Afganistani missioonil.	1. Poola saadab Afganistani 1000 lisasõdurit. 2. Suurpealetung Afganistanis. 3. Suurbritannia viib Afganistani veel 3300 sõdurit. 4. Afganistani hakkab stabiliseerima 3000 välismaist sõjaväelast.
Kümmekond hariduse ja tervisega seotud organisatsiooni nõuab seksuaalkasvatuse muutmist kohustuslikuks õppeaineks.	1. Tervisekaitseorganisatsioonid soovivad eraldi õppeainet seksuaalkasvatuse jaoks. 2. Arstid nõuavad koolidesse kohustuslikku seksuaalkasvatust. 3. Õpilased õpetavad eakaaslasi tervist hoidma. 4. Hariduse andmine on õpetaja isiklik mure.

5. Arutelu

Katsetulemused näitasid, et morfeemipõhine adapteerimine annab kõnetuvastuses paremaid tulemusi, kui sõna- või lemmapõhine adapteerimine. See on mõneti vastuolus autori esialgse hüpoteesiga, mis kahtles morfeemide võimes kanda semantilist sisu. Siiski, morfeemide hea tulemus on seletatav mitme asjaoluga.
Esiteks, morfeemipõhise adapteerimismudeli sõnavara saab seada identseks kõnetuvastuses kasutatava keelemudeli sõnavaraga. See tähendab, et morfeemipõhise adapteerimismudeli efektiivne katvus tuvastuse esimesest faasist saadud lausehüpoteeside suhtes on 100%.

Teiseks, viimasel ajal on mitmed korpuslingvistid tõdenud käändevormide olulisust võrreldes lemmadega või lisaks lemmadele. Näiteks František Čermáki (2007) arvates "leiutatakse lemmatiseerimise käigus uusi kujutletavaid maailmu, mida tegelikult tihti ei eksisteeri". Lemmatiseerimise suurim oht on potensiaalne liigne üldistamine ning sellest tulenev suur informatsioonikadu, mis ka antud lemmapõhise lähenemise käigus võib toimuda. Selle argumendiga on seletatav ka sõnapõhise mudeli suhteliselt väike allajäämine lemmapõhisele mudelile, kuigi sõnapõhise mudeli katvus on tunduvalt väiksem.

Kolmandaks, dokumendi esitus morfeemide kogumina võib anda dokumendi sisust rohkem informatsiooni, kui esialgu tundub. Näiteks raudteejaamast rääki-

vas artiklis on ilmselt kõrge morfeemide *raud*, *tee*, *jaam* ja tõenäoliselt ka *jaama* sagedus. Kuna LSA kasutab nn *bag-of-words* lähenemist, ei ole dokumendivektorit vaadates muidugi selge, mis järjestuses neid morfeeme kasutatakse. Sellest hoolimata – dokumendid, kus esinevad süstemaatiliselt koos nimetatud morfeemid, räägivad ilmselt siiski raudteejaamaga seotud teemal, kuigi teoreetiliselt saaks nendest morfeemidest ka teisi sõnu moodustada.

6. Kokkuvõte

Artikkel kirjeldas eksperimente eesti keele üldise statistilise keelemudeli automaatse adapteerimisega. Välja pakutud meetod kasutab keelemudelist iseseisvat LSA-põhist dokumentide vektoriruumi olemasolevale teemaspetsiifilisele tekstile semantiliselt lähedaste tekstide leidmiseks suurest dokumendikorpusest. Saadud dokumentide põhjal koostatakse uus teemaspetsiifiline unigramm-mudel ning see kombineeritakse üldise *N*-gramm-mudeliga, kasutades kiiret marginaalide adapteerimist. Eksperimentide käigus loodi kolm erinevat – sõnadel, lemmadel ja morfeemidel põhinev – LSA-mudelit ning võrreldi nende efektiivsust eesti keele suure sõnavaraga kõnetuvastuses.

Esialgsed eksperimendid suhteliselt väikese hulga uudistesaadete salvestuste tuvastamisel näitasid, et kõigi kolme mudeliga saavutatakse oluline kõnetuvastuse kvaliteedi paranemine. Morfeemipõhise mudeliga saadud 10-protsendiline suhteline tähevigade vähenemine oli statistiliselt oluliselt suurem kui lemma- ja sõnapõhise mudeliga saadud muutused.

Tulevase töö ühe suunana on kavas suurendada eksperimentaalsete andmete mahtu. Samuti on kavas praegu kasutatav manuaalne uudiste segmenteerimine lauseteks ja uudisnuppudeks asendada automaatse segmenteerimissüsteemiga, mis looks realistlikuma rakendus-stsenaariumi.

Kuna uuringute käigus selgus, et morfeemid on sobivad ühikud keelemudeli adapteerimiseks, siis saab edaspidi eesti keele jaoks kasutada ka teisi levinud adapteerimismeetodeid, kus tüüpiliselt adapteerimismudeli sõnavara kattub keelemudeli sõnavaraga. Inglise keele puhul on häid tulemusi andnud PLSA (Federico 2002) ja PLDA (Tam, Schultz 2006). Eesti keele jaoks tuleks siis adapteerimisühikutena kasutada sõnade asemel morfeeme, nagu ka *N*-gramm-mudelis.

Kirjandus

Alumäe, Tanel 2006. Methods for Estonian Large Vocabulary Speech Recognition. Theses of Tallinn University of Technology. Thesis on informatics and system engineering 31. Tallinn: Tallinn University of Technology Press.

Bellegarda, Jerome R. 1998. A multispan language modeling framework for large vocabulary speech recognition. – IEEE Transactions on Speech and Audio Processing 6 (5), 456–467.

Deerwester, Scott; Dumais, Susan T.; Furnas George W.; Landauer, Thomas K.; Harshman, Richard 1990. Indexing by latent semantic analysis. – Journal of the American Society of Information Science 41 (6), 391–407.

Federico, Marcello 2002. Language model adaptation through topic decomposition and MDI estimation. – Proceedings of ICASSP, Vol. 1. Orlando, 773–776.

Kaalep, Heiki-Jaan; Muischnek, Kadri 2005. The corpora of Estonian at the University of Tartu: the current situation. – Proceedings of the Second Baltic Conference on Human Language Technologies. Tallinn, April 4-5, 2005. Tallinn: Tallinn University of Technology, 267–272.

Kaalep, Heiki-Jaan; Vaino, Tarmo 2001. Complete morphological analysis in the linguist's toolbox. – Tõnu Seilenthal, Anu Nurk, Triinu Palo (eds). Congressus Nonus Internationalis Fenno-Ugristarum, Pars V, Dissertationes sectionum: Linguistica. II. Tartu, 9–16.

Kneser, R.; Peters, J.; Klakow, D 1997. Language model adaptation using dynamic marginals. – Proceedings of Eurospeech, Vol. 4. Rhodos, 1971–1974.

Landauer, Thomas K.; Foltz, Peter W.; Laham, Darrell 1998. Introduction to latent semantic analysis. – Discourse Processes 25, 259–284.

Meister, Einar; Lasn, Jürgen; Meister, Lya 2003. Development of the Estonian SpeechDat-like database. – Proceedings of Eurospeech, Vol. 2. Genf, 1601–1604.

Ristad, Eric Sven 1995. A natural law of succession. – Tech. Rep. TR-495-95, Computer Science Department, Princeton University.

Siivola, Vesa; Hirsimäki, Teemu; Creutz, Mathias; Kurimo, Mikko 2003. Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. – Proceedings of Eurospeech. Genf, 2293–2296.

Stolcke, Andreas 2002. SRILM – an extensible language modeling toolkit. – Proceedings of ICSLP, Vol. 2. Denver, 901–904.

Čermák, František 2007. Some of the current problems of corpus and computational linguistics or fifteen commandments and general truths. – Plenary presentation. The 3rd Baltic Conference on Human Language Technologies. Kaunas, Lithuania, October 4-5, 2007.

Tam, Yik-Cheung; Schultz, Tanja 2006. Unsupervised language model adaptation using latent semantic marginals. – Proceedings of Interspeech 2006. ICSLP. Pittsburgh, 2206–2209.

Yang, Yiming 1995. Noise reduction in a statistical approach to text categorization. – Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval. Seattle, 256–263.

Tanel Alumäe (Küberneetika Instituut) on lõpetanud Tallinna Tehnikaülikooli info- ja kommunikatsioonitehnoloogia erialal. Kaitses samas ülikoolis 2002. a magistrikraadi ja 2006. a doktorikraadi. Alates 2003. a töötab TTÜ Küberneetika Instituudi foneetika- ja kõnetehnoloogia laboris. Uurimisvaldkondadeks on kõnetuvastus, keeletehnoloogia, statistilised meetodid, masinõpe.
tanel.alumae@phon.ioc.ee

STATISTICAL LANGUAGE MODEL ADAPTATION FOR ESTONIAN SPEECH RECOGNITION

Tanel Alumäe
Institute of Cybernetics at Tallinn University of Technology

This paper presents a statistical language model adaptation framework for Estonian large vocabulary speech recognition. Estonian is a highly inflected, agglutinative and compounding language. To reduce lexical variety, morphemes are used as basic units in a statistical language model. For language model adaptation, we use a small set of topic-specific sentences as an adaptation seed. Then, latent semantic analysis (LSA) is applied for finding semantically close texts from a large document corpus. The resulting adaptation corpus is used for compiling a topic-specific unigram language model for each story. The unigrams are combined with a background *N*-gram model using fast marginal adaptation, resulting in an adapted *N*-gram model. We compare words, lemmas and morphemes as basic units in the LSA model.

The method is tested on an Estonian broadcast news transcription task. In the first pass of the recognition, a general background language model is used for finding recognition hypotheses for all utterances. The hypotheses are then used as an adaptation seed to compile an adapted language model for each news story. In the second recognition pass, the adapted models are applied to find new recognition hypotheses. We observe a significant improvement in speech recognition quality after applying the adapted models. The 10% drop in letter error rate when using morpheme-based adaptation is significantly better than when using either word or lemma-based adaptation. The article also discusses some possible reasons behind this observation.

Keywords: speech recognition, language model adaptation, latent semantic analysis, fast marginal adaptation, morphemes, lemmatization