Eestikeelsete tekstide sisukokkuvõtjast EstSum

Kaili Müürisep
Tartu Ülikool

1. Sissejuhatus

Automaatne sisukokkuvõtete tegemine tekstist on protsess, mille käigus luuakse tekstist olemasoleva põhjal uus lühendatud versioon, mis sisaldab ainult kasutajale vajalikku informatsiooni.

Lühikesed ja informatiivsed ülevaated dokumentide, ajalehe- ja teaduslike artiklite jms sisust osutuvad tänapäeva tohutu informatsioonitulva juures asendamatuteks. Neid on näiteks tarvis nii mobiiltelefonide või pihuarvutite puhul, mille ekraanidelt on ebamugav liiga pikka teksti lugeda; Interneti otsingumootorites, et väljastatavate päringute hulgas oleks lihtsam orienteeruda; kiire ülevaate saamiseks juhul, kui lugejal on vaja läbi töötada suurem hulk tekste; või hoopis teksti inimkõnena ettelugemisel arvuti poolt, kuna lühem tekst on kuulajale arusaadavam ja mugavam. Omaette uurimisvaldkond on suulise kõne lühendamine, mille käigus eemaldatakse pausid jutus, kordused ja eneseparandused.

Taoliste arvutiprogrammi abil koostavate kokkuvõtete tegemisel kerkivad esile järgnevad probleemid: mida üldse lugeda tähtsaks informatsiooniks? kui palju informatsiooni on vaja eraldada? millised on need omadused või kriteeriumid, mille põhjal kõige tähtsamaid lauseid välja valida? kuidas vältida seosetute lausete kaasamist? kuidas vältida tarbetuid kordusi?

Töö eesti keele sisukokkuvõtjaga on käinud juba mitu aastat, kuid enamasti on see piirdunud diplomi- ja bakalaureusetöö tasemel eksperimentidega (Lippur 2000; Mutso 2005, vt ka Müürisep, Mutso 2005). EstSum on esialgu orienteeritud ainult veebi-uudiste ja elektrooniliste ajaleheartiklite sisukokkuvõtmisele.

2. Mis on sisukokkuvõte

Radev jt annavad sisukokkuvõtte mitterange definitsiooni: “Sisukokkuvõte on tekst, mis on saadud ühe või rohkema teksti töötlemisel; mis annab edasi originaalteksti(de) olulist informatsiooni ja mis pole pikem kui pool originaaltekstist, enamasti oluliselt

lühem." (Radev jt 2002: lk 399) Teksti mõistet on siin kasutatud üsna vabalt: selle all mõeldakse nii tavalist teksti kui ka kõnet, multimeediadokumente, hüperteksti jne.
Sisukokkuvõtte peaeesmärk on esitada teksti peamised ideed väiksemas mahus.
Sisukokkuvõtteid on väga mitmeid tüüpe. Neid eristatakse selle järgi, kas saadud sisukokkuvõtte laused on originaaltekstist välja valitud (väljavõte, ingl k *extract*) või on nad genereeritud (ülevaade, ingl k *abstract*). Väljavõtte laused on täielikult kopeeritud originaaltekstist ning sisu antakse edasi autori täpses sõnastuses. Ülevaate puhul on sisukokkuvõttes lauseid, mida originaaltekstis ei esine. Tavaliselt on ülevaadetes kasutatud parafraseerimist (asesõnade asendamist, osalausete eemaldamist jms). Üldiselt võimaldavad ülevaated sisu tihedamini kokku pakkida kui väljavõtted, samas on neid automaatselt keerulisem genereerida.
Teine sisukokkuvõtete liigitus jagab sisukokkuvõtted indikatiivseteks ja informatiivseteks. Indikatiivne sisukokkuvõte peab andma arusaama, millest on dokumendis juttu ilma detailidesse laskumata. Informatiivsed sisukokkuvõtted peavad edastama lühidalt kogu olulise informatsiooni.
Sisukokkuvõtteid saab liigitada ka teema põhjal: eristatake üldisi ja teemale orienteeritud (*topic-oriented*, ka *user-focused*) sisukokkuvõtteid, kus viimased annavad informatsiooni lugeja huvidele vastavatel teemadel. See, millisel kujul lugeja huvi on ilmutatud, sõltub süsteemist: arvesse võidakse võtta kasutajaprofiili ja seniseid harjumusi, kasutaja koostatud päringut või ka loomulikus keeles esitatud küsimust.

3. Sisukokkuvõtmise meetodid

Klassikaline sisukokkuvõtmise protsess on kolmeetapiline (Mani 2001: lk 13):
1. Analüüs. Toimub sisendi analüüs ja selle sisekujule viimine,
2. Transformatsioon. Toimub teisendus sisendi sisekujult kokkuvõtte sisekujule.
3. Süntees. Sisukokkuvõte sisekuju muudetakse loomuliku keele tekstiks.
Tegelikes süsteemides erinevate etappide piirid nii selged pole ning sageli nimetatakse neid etappe teiste nimedega.
Enamik tänapäeva sisukokkuvõttesüsteeme kasutab endiselt väljavõtte tüüpi sisukokkuvõtete tegemise metoodikat, s.t originaaltekstist valitakse välja laused, mis kannavad süsteemi meelest olulist informatsiooni ja esitatakse need kasutajale.
Selliste süsteemide loomise ajalugu ulatub 50ndatesse aastatesse. Levinuim tehnika

seisneb selles, et iga lause jaoks arvutatakse selle lause skoor ehk kaal, mis põhineb lause asukohal tekstis, sõnade ja fraaside sagedustel, võtmefraaside esinemisel jne. Suurima skooriga laused kaasatakse sisukokkuvõttesse. Uuemad meetodid kasutavad tunnuste leidmiseks masinõppimise meetodeid või oluliste tekstipassaažide määramiseks süntaktilist analüüsi, samuti uuritakse pigem sõnadevahelisi seoseid kui sõnade hulki.

Lausetele kaalu arvutamise meetod põhineb klassikaks saanud Edmundsoni paradigmal (Edmundson 1969), kes kasutas lausete väljavalimisel nende hindamiseks valemit (1), kus W(s) on lause s kaal, C(s) on selle lause märgusõnade skoor (*cue words*), K(s) võtmesõnade skoor (*key words*), L(s) asukoha skoor (*location*) ja T(s) pealkirja sõnade (*title words*) skoor, α, β, γ ja δ on konstandid

(1) $W(s) = \alpha C(s) + \beta K(s) + \gamma L(s) + \delta T(s)$

Märgusõnade skoor. Märgusõnadeks või fraasideks loetakse sõnu, mis viitavad, et autor ise on selles lauses sisu kokku võtnud, nt "kokkuvõtteks", "järelikult", "selles artiklis" aga ka kesk- ja ülivõrdes omadussõnad "parim", "edukam", kõikvõimalikud hinnangut väljendavad sõnad ("õnnestus", "edukas") jpt. Kui sellised sõnad või fraasid esinevad lauses, siis saab see lause lisapunkte. Kui aga lauses leidub sõnu, mis võivad välistada lause sobimise sisukokkuvõttesse, nt "juhuslik", "vaevalt", siis lause kaalu vähendatakse. Mitmed hilisemad uuringud on näidanud, et märgusõnade ja -fraaside kasutamine on eriti õigustatud teaduslike tekstide sisukokkuvõtmisel. Ainult seda meetodit kasutades on võimalik leida 55% olulistest lausetest tekstis (Teufel, Moens 1997).

Võtmesõnade skoor. Olulised laused sisaldavad sõnu, mis esinevad tekstis mõnevõrra sagedamini. Muidugi tuleb seejuures arvestada sõnade üldist sagedust. Samas eksperimendid näitavad, et võtmesõnade skoori arvestamine ei pruugi sisukokkuvõtja tulemust parandada (Marcu 2003).

Positsiooniskoor. Paljud sisukokkuvõtjad eeldavad, et laused, mis paiknevad teksti algul, on olulisemad kui tagapool paiknevad. Eksperimendid näitavad, et selle eeldusega töötavad sisukokkuvõtjad on parimad ühe tunnuse põhjal lauseid valivatest kokkuvõtjatest (Marcu 2003). Kuigi see meetod sõltub paljuski teksti žanrist ja sisukokkuvõtte pikkusest, on see üks tõhusamaid meetodid 33% pakkimise korral.

Pealkirjas esinevate sõnade skoor. Sõnad, mis esinevad teksti pealkirjas on ilmselt temaatilised ning neid sõnu sisaldavad laused peaksid olema olulised.

Edmundson määras nende parameetrite väärtused käsitsi, samuti konstantide α, β, γ, ja δ väärtused. Tema eksperimendid näitasid, et märgusõnade, asukoha- ja pealkirja sõnade skooride kasutamine andis parima tulemuse, eraldi katsetatuna oli edukaim asukohapõhine meetod ja nõrgim võtmesõnade meetod. Tänapäeval on välja töötatud mitmed meetodid parameetrite väärtuste määramiseks masinõppimise meetodil

Uuemad meetodid võtavad arvesse lausete süntaktilist ja semantilist infot. **Kohesioonipõhised meetodid** eeldavad, et olulised on need laused, millel on kõige rohkem seoseid teiste lausetega. Üheks kohesioonipõhiliseks meetodiks on näiteks leksikaliste ahelate (*lexical chain*) meetod (Barzilay, Elhadad 1997). Leksikalisse ahelasse kuuluvad omavahel semantiliselt tugevalt seotud sõnad (nt kapsas, porgand, juurvili jmt). Ahelad genereeritakse automaatselt tesauruste või Wordneti abil. Tähtsateks lauseteks loetakse need, mille mõni sõna kuulub tugevasse ahelasse. Ahela tugevus arvutatakse tema pikkuse ja liikmete sageduse põhjal.

Teine näide kohesioonipõhisest meetodist on seotuse (*connectedness*) meetod (Mani 2001: 95-102): sõnadest moodustatakse graafi tipud. Tippude vahel luuakse kaared vastavalt sellele, kas antud sõnade vahel on grammatilisi, koreferentsi- või leksikalise sarnasuse seoseid. Laused, mis sisaldavad enim seotud sõnu, loetakse olulisteks.

Teksti diskursuse struktuuri põhised meetodid. Teksti hierarhilist diskursuse struktuuri saab kasutada oluliste lausete leidmiseks. Kui lõik *p* arendab edasi lauset *l*, siis on mõistlik eeldada, et sisukokkuvõte peaks sisaldama ainult lauset *l*. See meetod on osutunud eriti sobivaks väga lühikeste sisukokkuvõtete tegemisel (Marcu 2000).

4. EstSumi kirjeldus

Eesti keele sisukokkuvõtja EstSum kasutab lausete väljavalimismeetodit ehk siis genereerib väljavõtte. Hetkel on EstSum orienteeritud veebis avaldatud uudiste ja ajaleheartiklite indikatiivsetele sisukokkuvõtetele.

EstSumi kavandamisel oli eeskujuks rootsi keele sisukokkuvõtja SWESUM[1] (Dalianis jt 2003).

EstSum koosneb kolmest moodulist: HTML-konverter, lausestaja ja väljavõtete tegija.

HTML-konverter eemaldab sisukokkuvõtte jaoks ebaolulised HTML-märgendid, normaliseerib ristuvad märgendid, eemaldab tabelid ja konverteerib sisendi SGML-

[1] http://swesum.nada.kth.se/index-eng.html

formaati. SGML-formaadis märgendatakse pealkirjad ja alapealkirjad, autorid, pildiallkirjad. Samuti märgendatakse oluline šriftiinformatsioon, eristades rasvast, kald- ja tavalist kirja .

Lausestaja kasutab reeglipõhist meetodit sisendi töötlemisel, lause alguse ja lõpu märgendamiseks kasutatakse 30 Perli regulaaravaldist.

EstSum arvutab sisukokkuvõtte pikkust kahel viisil. Esimene moodus on tavaline lausepõhine meetod, mille korral sisukokkuvõte 30% tähendab, et teksis on 30% esialgsetest lausetest. Samas on pikemad laused informatsioonirikkamad ning tegelikult ei pruugi teksti pikkus nii palju lüheneda. Teine võimalus on arvutada sisukokkuvõtte pikkus sõnades. Sel viisil saadud sisukokkuvõtted on tõepoolest 30% esialgse teksti pikkusest.

EstSum kasutab oluliste lausete väljavalimiseks informatsiooni lausete asukoha, formaadi ja sõnavara kohta. Lausetele skoori arvutamiseks kasutatakse Edmundsoni valemile (1) sarnast valemit (2):

(1) (2) $W(s)=\alpha P(s)+\beta F(s)+\gamma K(s)$

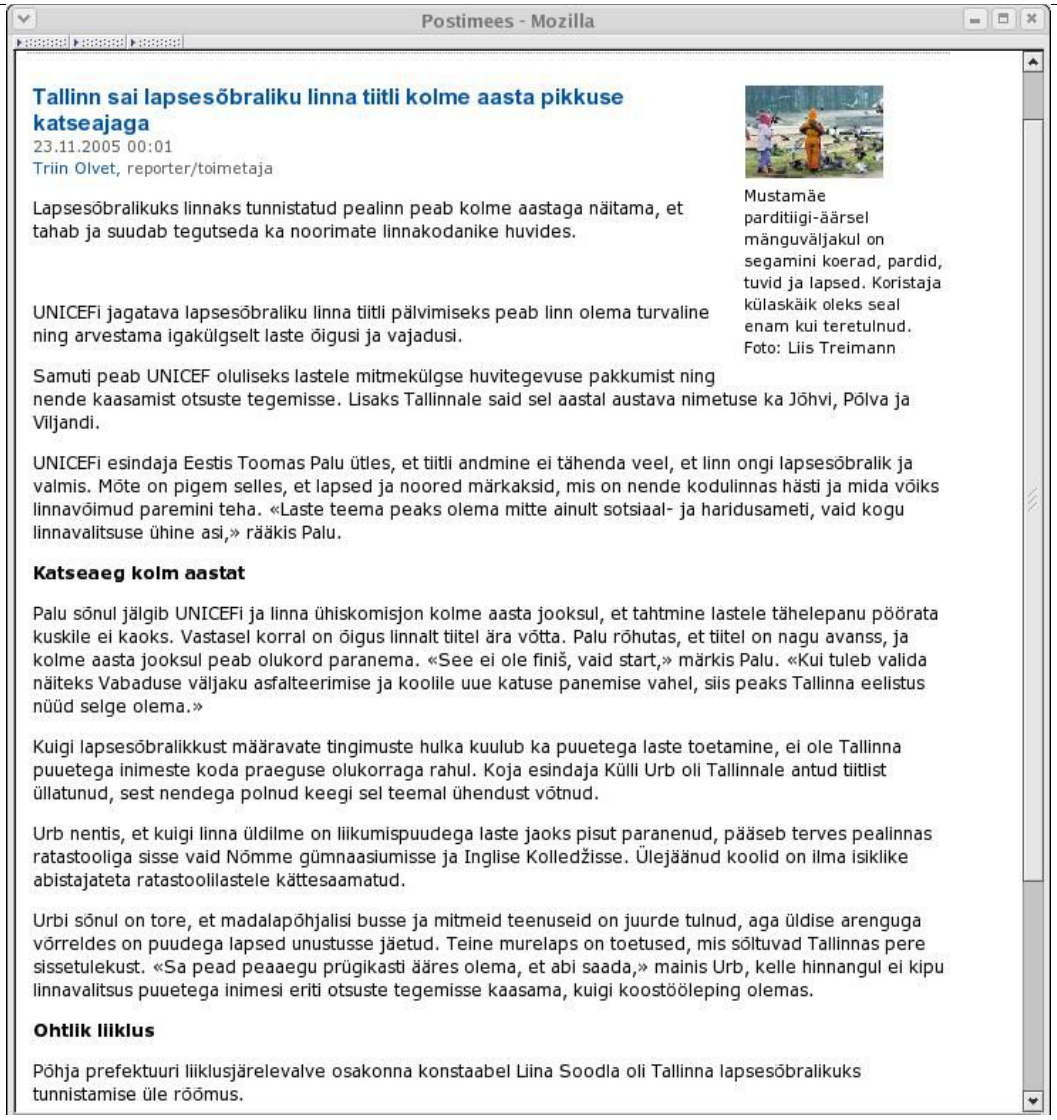
W(s) on lause s kaal, P(s) on positsioonipõhine skoorifunktsioon, F(s) formaadipõhine skoorifunktsioon and K(s) on sõnasageduste põhine skoorifunktsioon; α, β ja γ on konstandid. EstSumis puudub märgusõnade arvestamise võimalus, sest see nõuab eestikeelsete ajaleheartiklite sõnastuse põhjalikumat analüüsi, kuid kasutatakse erinevalt algsest valemist ära formaadiinformatsiooni.

Tunnuste kaalud ja konstandid α, β ja γ on määratud käsitsi kasutades selleks väikest treeningkorpust (20 teksti). Ehkki kasutatud korpus on suhteliselt väike, võimaldas see siiski määrata parameetrite esialgsed väärtused. Täpsem korpuse ja algoritmi kirjeldus on toodud (Müürisep, Mutso 2005).

Lausete asukohta uurides selgus, et olulised laused paiknevad pealkirja järel. Esimene lause oli sisukokkuvõttes 100% juhtudest, teine ja kolmas 65% juhtudest. Suurendati ka nende lausete skoore, mis järgnesid alapealkirjale, samuti lõigu esimese, teise ja kolmanda lause ning artikli viimase lause skoori.

EstSum loeb tähtsateks neid lauseid, milles kasutatakse rasvast või kaldkirja. Formaadipõhine skoorifunktsioon arvestab ka lause kirjavahemärkidega: hüüu- ja küsimärgid vähendavad lause kaalu, samuti jutumärgid. Täielikult välistatakse pildiallkirjade lisamise sisukokkuvõttesse.

Praegune EstSumi versioon ei kasuta lingvistilist moodulit, seepärast kasutatakse



Joonis 1. Katke EstSumi sisendtekstist
http://www.postimees.ee/231105/esileht/siseuudised/183960_print.php

võtmesõnade statistika tegemiseks sõnavorme, mitte sõnade põhivorme.

Võtmesõnade tuvastamiseks kasutatakse kahte meetodit: 1) leitakse sõnad, mis on artiklis väga sagedased, kuid mitte nii sagedased üldises sagedustabelis; 2) pealkirjas ja alapealkirjades leiduvad sõnad loetakse olulisteks.

Samas, treeningkorpuse lähemal uurimisel ilmnes, et ainult 48% lausetest, mis sisaldasid pealkirjas leiduvaid sõnu, esinesid ka sisukokkuvõttes. Tekstis sagedasti esinevate sõnadega laused olid sisukokkuvõttes ainult 25% juhtudest.

Et leida konstantide α, β ja γ väärtusi, testisime, millise tulemuse annaks iga skoorifunktsioon eraldi. Ilmnes, et olulisim on positsiooni arvestav skoorifunktsioon.

Tallinn sai lapsesõbraliku linna tiitli kolme aasta pikkuse katseajaga

Lapsesõbralikuks linnaks tunnistatud pealinn peab kolme aastaga näitama, et tahab ja suudab tegutseda ka noorimate linnakodanike huvides.

UNICEFi jagatava lapsesõbraliku linna tiitli pälvimiseks peab linn olema turvaline ning arvestama igakülgselt laste õigusi ja vajadusi.

UNICEFi esindaja Eestis Toomas Palu ütles, et tiitli andmine ei tähenda veel, et linn ongi lapsesõbralik ja valmis.

Palu sõnul jälgib UNICEFi ja linna ühiskomisjon kolme aasta jooksul, et tahtmine lastele tähelepanu pöörata kuskile ei kaoks.

Urb nentis, et kuigi linna üldilme on liikumispuudega laste jaoks pisut paranenud, pääseb terves pealinnas ratastooliga sisse vaid Nõmme gümnaasiumisse ja Inglise Kolledžisse.

Urbi sõnul on tore, et madalapõhjalisi busse ja mitmeid teenuseid on juurde tulnud, aga üldise arenguga võrreldes on puudega lapsed unustusse jäetud.

Põhja prefektuuri liiklusjärelevalve osakonna konstaabel Liina Soodla oli Tallinna lapsesõbralikuks tunnistamise üle rõõmus.

Joonis 2. EstSumi poolt genereeritud sisukokkuvõte

Formaati arvestav skoor erineb ainult üksikutel lausetel ning see iseseisvana lausete valikuks ei sobi. Võtmesõnasid arvestav skoorifunktsioon oli ebatäpsem, ning seetõttu määrati konstantide , ja väärtusteks vastavalt 0,4, 0,4 ja 0,2.

Selliste väärtuste korral 51% EstSumi poolt leitud sisukokkuvõtete lausetest kattus inimese poolt valitutega.

Joonistel 1 ja 2 on toodud näide lähtetekstist ja 30-protsendilisest sisukokkuvõttest. Selle artikli puhul on näha, et EstSum eelistab lõikude esimesi lauseid. Formaadipõhine skoorifunktsioon aitab vältida jutumärkides tsitaatide kaasamist sisukokkuvõttesse. Ainult positsiooni ja võtmesõnade baasil otsustades oleks sisukokkuvõtte viimaseks lauseks «Kõik need mänguväljakud ja noortekeskused on väga vahvad.»

EstSumi loodud sisukokkuvõte ei ole sidus. Selles artiklis ei kaasatud sisukokkuvõttesse asesõnadega lauseid, kuigi enamasti valitakse välja ka mõni lause, milles on ilma igasuguste selgitusteta asesõnad *ta* või *see*. Näiteväljavõttes kasutatakse isiku tähistamiseks perekonnanime Urb, lähtetekstis oli tema täisnimi ja tiitel korralikult antud, kuid see lause ei osutunud valituks.

Artikli viimane osa käsitles Tallinna liikluskultuuri, millega konstaabel Soodla ei olnud rahul, kuid sisukokkuvõttesse sattus tendentslikult ainult positiivne lause.

5. Tulemuste hindamine

Kuidas hinnata sisukokkuvõtte headust? Mis teeb ühe sisukokkuvõtte heaks ja teise halvaks? Tavaliselt toimub automaatselt genereeritud sisukokkuvõtte hindamine sel viisil, et seda võrreldakse inimes(t)e poolt koostatud sisukokkuvõttega (väljavõttega) ning leitakse kattuvate lausete osakaal. Hea, kui neid inimesi, kes sisukokkuvõtteid käsitsi koostavad, oleks mitu. Samas on sisukokkuvõtja töö hindamisel hea teada fakti, et kahe inimese poolt koostatud väljavõtetes kattuvad ainult 70% lausetest (Hassel 2003).

EstSumi hindamiseks loodud korpus koosnes 11 tekstist, milles oli keskmiselt 23 lauset. EstSumi poolt valitud laused kattusid 60% ulatuses inimese poolt valitud lausetega. Parimal juhul oli samu lauseid 85% ja halvimal juhul ei kattunud ükski (väga lühike artikkel).

6. Järeldused ja plaanid edaspidiseks

Eestikeelsete tekstide sisukokkuvõtja EstSum on tegelikult ikka veel eksperimentaalses arengujärgus ning selle arendamiseks on vaja teha palju tööd.

Olulisimad neist oleks lingvistilise mooduli (morfoloogiaanalüsaator, morfoloogiline ühestaja ja süntaksianalüsaator) ühendamine EstSumiga. Eelkõige võimaldaks see paremat võtmesõnade statistikat, mis praegu on EstSumi nõrgim koht. Semantikal põhineva heuristika kasutuselevõtt vajaks Wordneti (või sellel baseeruva sõnastiku) ühendamist sisukokkuvõtjaga. Süntaktilise informatsiooni olemasolu lubaks lauseid automaatselt lühendada, eemaldades näiteks osalauseid, samuti oleks see eelduseks anafooride lahendamisele. Süntaktiliselt analüüsitud sisend ja lahendatud anafoorid võimaldaksid katsetada ka keerukamaid kohesioonipõhiseid sisukokkuvõttemeetodeid.

Samas on ilmselge, et 10-20 teksti käsitsi koostatud sisukokkuvõtte põhjal tehtud üldistused ei ole piisavalt täpsed. Vaja on suuremat test- ja treeningkorpust, mis võimaldaks parameetrite väärtused leida statistiliste masinõppimise meetoditega.

Kirjandus

Barzilay,Regina, Elhadad, Michael 1997. Using Lexical Chains for Text

Summarization, in *Proceedings of the Intelligent Scalable Text Summarization Workshop*, ACL, Madrid, 1997

Dalianis, Hercules, Hassel, Martin, Wedekind Jürgen, Haltrup, Dorte, de Smedt, Koenraad, Lech, Till Christopher 2003. Automatic text summarization for the Scandinavian languages. In Holmboe, H. (ed.) Nordisk Sprogteknologi 2002: *Årbog for Nordisk Språkteknologisk Forskningsprogram 2000-2004*, pp. 153-163. Museum Tusculanums Forlag

Edmundson, H. P. 1969. New methods in automatic abstracting. In: *Journal of the Association for Computing Machinery 16 (2)*. 264-285. Reprinted in: Mani, I.; Maybury, M.T. (eds.) *Advances in Automatic Text Summarization*. Cambridge, Massachusetts: MIT Press. 21-42.

Hassel, Martin 2003. Exploitation of Named Entities in Automatic Text Summarization for Swedish. In the *Proceedings of NODALIDA '03* - 14th Nordic Conference on Computational Linguistics, May 30-31 2003, Reykjavik, Iceland.

Lippur, Andres 2000. Automaatne sisukokkuvõtete tegemine eestikeelsetele tekstidele. *Bakalaureusetöö*. Tartu Ülikool. Arvutiteaduse Instituut.

Mani, Inderjeet 2001. *Automatic summarization*. Amsterdam: John Benjamins Publishing Co.

Marcu, Daniel 2000. The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach. *Computational Linguistics*, 26 (3), pages 395-448

Daniel Marcu 2003. Automatic Abstracting, *Encyclopedia of Library and Information Science*, pp 245-256

Mutso, Pilleriin 2005. Automaatne sisukokkuvõtete tegemine. *Diplomitöö*. Tartu Ülikool, Arvutiteaduse Instituut.

Müürisep, Kaili, Mutso, Pilleriin 2005. ESTSUM - Estonian newspaper texts summarizer. *Proceedings of The Second Baltic Conference on Human Language Technologies*. Tallinn. Pp. 311-316

Radev, Dragomir R., Hovy, Eduard, McKeown, Kathleen 2002. Introduction to the special issue on summarization. *Computational Linguistics*. Vol 28(4). 399-408.

Simone Teufel, Marc Moens 1997. Sentence Extraction as a Classification Task. *Proceedings of Intelligent and Scalable Text Summarization Workshop*. Madrid, 58 -65.

Lisa. Automaatselt genereeritud 10% sisukokkuvõte artiklist

Eestikeelsete tekstide sisukokkuvõtjast EstSum

Kaili Müürisep

Automaatne sisukokkuvõtete tegemine tekstist on protsess, mille käigus luuakse tekstist olemasoleva põhjal uus lühendatud versioon, mis sisaldab ainult kasutajale vajalikku informatsiooni.

Radev jt annavad sisukokkuvõtte mitterange definitsiooni: «Sisukokkuvõte on tekst, mis on saadud ühe või rohkema teksti töötlemisel; mis annab edasi originaalteksti(de) olulist informatsiooni ja mis pole pikem kui pool originaaltekstist, enamasti oluliselt lühem.» (Radev jt 2002: lk 399).

Klassikaline sisukokkuvõtmise protsess on kolmeetapiline (Mani 2001: lk 13):

Märgusõnadeks või fraasideks loetakse sõnu, mis viitavad, et autor ise on selles lauses sisu kokku võtnud, nt «kokkuvõtteks», «järelikult», «selles artiklis» aga ka kesk- ja ülivõrdes omadussõnad «parim», «edukam», kõikvõimalikud hinnangut väljendavad sõnad («õnnestus», «edukas») jpt.

Olulised laused sisaldavad sõnu, mis esinevad tekstis mõnevõrra sagedamini.

Paljud sisukokkuvõtjad eeldavad, et laused, mis paiknevad teksti algul, on olulisemad kui tagapool paiknevad.

Sõnad, mis esinevad teksti pealkirjas on ilmselt temaatilised ning neid sõnu sisaldavad laused peaksid olema olulised.

Eesti keele sisukokkuvõtja EstSum kasutab lausete väljavalimismeetodit ehk siis genereerib väljavõtte.

EstSum kasutab oluliste lausete väljavalimiseks informatsiooni lausete asukoha, formaadi ja sõnavara kohta.

W(s) on lause s kaal, P(s) on positsioonipõhine skoorifunktsioon, F(s) formaadipõhine skoorifunktsioon and K(s) on sõnasageduste põhine skoorifunktsioon; α, β ja γ on konstandid.

Tunnuste kaalud ja konstandid α, β ja γ on määratud käsitsi kasutades selleks väikest treeningkorpust (20 teksti).

Joonistel 1 ja 2 on toodud näide lähtetekstist ja 30-protsendilisest sisukokkuvõttest.