TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Arvutiteaduse instituut
Infotehnoloogia eriala

Keili Sellik

Automaatse sisukokkuvõtja töö hindamine

Bakalaureusetöö (4 AP)

Juhendaja: PhD Kaili Müürisep

Autor:.....“.....“ mai 2008
Juhendaja:.....“.....“ mai 2008

Lubada kaitsmisele
Professor ....“.....“ mai 2008

TARTU 2008

Sisukord

1. Sissejuhatus

Automaatne sisukokkuvõtete tegemine tekstist on protsess, milles luuakse olemasolevast tekstist uus, lühendatud versioon, mis sisaldab vaid tähtsamat või kasutajale vajalikku informatsiooni. Tänapäeva suure informatsioonihulga juures on lühendatud, kuid informatiivsed ülevaated kohati asendamatud. Neid vajatakse näiteks interneti otsingumootorites, pihuarvutites ja mobiiltelefonides. Eestis on automaatne sisukokkuvõtete tegemine alles arengujärgus. Käesolevas bakalaureusetöös kasutatakse eestikeelsete tekstide sisukokkuvõtjat EstSum, mis hetkel töötab ainult veebiuudiste ning ajaleheartiklite sisukokkuvõtmisega.

Kui sisukokkuvõtet on tarvis, siis on hea, kui seda on võimalus teha automaatselt vaid ühe hiireklikiga. Kindlasti tekib aga küsimus, kas sellise sisukokkuvõtte kvaliteet on sama hea, kui sellise puhul, kus inimene mõtleb terve teksti läbi ja valib siis laused, mis on tähtsad.

Uuringud on näidanud, et inimesed teevad suures osas omavahel kattuvaid sisukokkuvõtteid. Tekib küsimus, kui hea ja sobiv on automaatselt valminud sisukokkuvõte. Välja on töötatud erinevaid meetodeid hindamaks sisukokkuvõtjate tööd. Eestis on see suhteliselt uurimata valdkond ning pole informatsiooni selle kohta, milliseid meetodeid või programme tasuks ning milliseid on võimalik kasutada.

Käesolev bakalaureusetöö jaguneb kahte suuremasse ossa:

Esimeses osas antakse ülevaade erinevatest automaatsete sisukokkuvõtete hindamismeetoditest. Samuti selgitatakse hindamismeetodite jagunemist kahte laiemasse gruppi.

Teises osas kasutatakse esimeses osas toodud meetodite seast kahte, et hinnata automaatse sisukokkuvõtja EstSum tööd. Selleks on autor koostanud 50 artiklist sisukokkuvõtted ning samadest artiklitest on kokkuvõtted genereeritud ka sisukokkuvõtjaga EstSum.

2. Ülevaade meetoditest

2.1 Hindamismeetodite jagunemine

Esimene ning kõige laiem automaatsete sisukokkuvõtete hindamismeetodite jagunemine on Mani jt (Mani, Klein, House, Hirschman 2002) järgi sisemisteks (*intrinsic*) ning välimisteks (*extrinsic*). Välimised hindamismeetodid mõõdavad tõhusust ning aktsepteeritavust mingi konkreetse ülesande juures või praktilises süsteemis, nagu näiteks asjakohasuse määramine või lugemise mõistmine. Mitmed välised hindamismeetodid sisaldavad küsimise-vastamise ning mõistmise ülesandeid. Välimiste hindamismeetoditega mõõdetakse ka dokumendi sobivust mingisugusesse kindlasse teemasse. Sisemised hindamismeetodid aga vastupidiselt välistele mõõdavad süsteemi sisu järgi, tihti referentssisukokkuvõtete abil. Põhimõte on määramises, kui palju olulist informatsiooni on erineva lühendusprotsendi puhul sisukokkuvõtetes säilinud. Sisemised hindamismeetodid võivad mõõta ka kokkuvõtte sidusust selle subjektiivse loetavuse hindamise kaudu. Ka on võimalik hinnata grammatika õigsust, kontekstist välja rebitud anafooride olemasolu või korrastatud struktuuride (nt. nimekirjad, tabelid) lõhkumist.

2.2 Sisemised hindamismeetodid

Leidub kahte tüüpi sisemisi hindamismeetodeid – kvaliteedi ning informatiivsuse hindamine. Kvaliteedi hindamise puhul on oluline märkida, et see kaasab protsessi inimesi. Kõigis uurimustes ja uuringutes, milles on kaasatud inimesed, tekib küsimus, kas subjektid tõlgendavad hindamiskriteeriume õigesti ning järjekindlalt. Hindajate vahelist ühildumist on võimalik mõõta, loendades nende kordade osakaalu, kui hindajate otsus mingi lause või fragmendi suhtes on sama, või läheb konkreetses olukorras lahku. Selleks kasutatakse *Kappa* mõõtu (Mani 2001). *Kappa* arvutamiseks mõõdetakse, mitu hindajat märkisid lause kokkuvõttesse sobivaks ja mitu mitte. Valemis 2.1 on toodud *Kappa* mõõdu arvutusvalem:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.1}$$

Valemis tähistab P(A) nõustumiste arvu ning P(E) oodatavat nõustumiste arvu. K=1, kui kattumine hindajate otsuste vahel on täielik. Kvaliteedi hindamise puhul on üks kokkuvõtte

aspekte see, kui loetav see on. Seda saab mõõta vaid viisil, kus hindajad kokkuvõtteid loetavuse aspektist lähtuvalt mingite kriteeriumide alusel hindavad. On olemas ka selline kvaliteedi hindamise variant, kus hindajaid vaja ei ole, see on nn *off-line* hindamine – näiteks hinnatakse sel puhul kokkuvõtet stiili- või grammatikakontrollijaga. Kokkuvõtte tekstilise kvaliteedi hindamine ei pruugi alati näidata kokkuvõtte headust. Näiteks sellisel juhul, kui kokkuvõte on väga ladusalt kirjutatud, kuid sisaldab ebakorrektset infot või on kasutu. Samuti võivad erinevate rakenduste puhul nõuded kvaliteedile erineda. Eelnevast lähtudes eelistatakse tavaliselt informatiivsuse hindamist.

Informatiivsuse hindamisel vaadatakse näiteks, kui palju sisukokkuvõte algallikast pärit informatsiooni erinevate kokkusurumistasemete puhul sisaldab. Informatiivsuse hindamise alla kuulub ka sisukokkuvõtte hindamine referentssisukokkuvõtete abil – mõõdetakse, kui palju referentssisukokkuvõtte lausetest automaatne sisukokkuvõte sisaldab.

2.2.1 Referentssisukokkuvõtetest

Sisukokkuvõtjate töö hindamisel peab olema olemas mingi hulk sisukokkuvõtteid, mida loetakse õigeteks ehk ideaalseteks. See nn kuldne standard on tavaliselt käsitsi või poolautomaatselt koostatud. Käsitsi koostatud sisukokkuvõte võib olla vabas vormis kokkuvõte töö sisust, mis on tüüpiline näiteks akadeemilist uurimistööd käsitlevaile tekstidele. Samas võivad need sisukokkuvõtted olla loodud just automaatse sisukokkuvõtja töö hindamiseks ning koosneda lähtedokumendist ekstraheeritud lausetest. Probleem referentssisukokkuvõtete puhul on see, et referentssisukokkuvõtete kogus võib jääda ebapiisavaks, sest alati on võimalus, et süsteem loob sisukokkuvõtte, mis on kõikidest referentssisukokkuvõtetest täiesti erinev, kuid sellegi poolest hea sisukokkuvõte.

2.2.2 Meetod- Ekstraheeritud sisukokkuvõtete võrdlemine

Kui samast tekstist on nii käsitsi kui ka automaatselt lausete ekstraheerimise teel tehtud sisukokkuvõtted, siis on võimalik kahe kokkuvõtte läbivaatamise teel võrrelda, mitu protsenti lausetest langevad kokku. Selline käsitsi läbivaatamise meetod on aeganõudev, kuid kuna tavaliselt peetakse referentssisukokkuvõteteks käsitsi või poolautomaatselt koostatud kokkuvõtteid, saab sel moel ettekujutuse, kui palju langeb automaatne sisukokkuvõte samast etteantud tekstist käsitsi tehtud sisukokkuvõttega kokku. Saadud protsendi põhjal võib teha järelduse, kas automaatne sisukokkuvõte on hea või halb.

Hassel ja Danianis (Hassel, Danianis 2005) on läbi viinud uurimuse saamaks teada, kuidas on inimeste tehtud sisukokkuvõtted seoses ning kui suures osas need kattuvad. Nad palusid kahel grupil Rootsis ning ühel grupil Taanis teha sisukokkuvõtted, mille põhjal saaks teha järeldusi. Rootsi grupid tegid kokku 301 Rootsi uudiste kokkuvõtet ning Taani grupp tegi 135 Taani uudiste kokkuvõtet. Kokkuvõtete pikkus oli keskmiselt 32,5% originaaltekstist. Tulemuseks saadi, et umbes 30% tekstis esinevatest lausetest ühilduvad kõigil sisukokkuvõtteid koostanutel ning veidi alla 70% tekstis esinevatest lausetest ühilduvad enamusel sisukokkuvõtteid koostanutel.

2.2.3 Sisukokkuvõtete võrdlemine fragmentide abil

Kui nii referentssisukokkuvõte kui ka automaatselt genereeritu on koostatud mõlemad lausete ekstraheerimise teel, siis on sisukokkuvõtja töö hindamine suhteliselt lihtne: saab leida, mitu protsenti lausetest langesid mõlemas kokku. Samas ei pruugi selline hindamisviis olla kõige õiglasem. Võib juhtuda, et tekstis kannavad sama sisu edasi kaks erinevat lauset ning sisukokkuvõtetesse on sattunud just erinevad, kuid sama sisuga laused. Kui referentssisukokkuvõte või automaatne sisukokkuvõte ei kasuta originaallauseid või muudavad neid, siis selline automaatne lausete võrdlus pole võimalik. Selle asemel võrreldakse, kui palju langevad kokku väiksemad fragmendid mõlemas tekstis. Mida aga valida selliseks väiksemaks fragmendiks ning kui pikk või lühike see peaks olema?

Fragmentidena vaadatakse erinevates süsteemides erinevaid lause- või sõnaosi. (Hovy, Lin, Zhou, Fukumoto 2006). ROUGE (Lin ja Hovy 2003), mis on kõige sagedamini kasutatud automaatne sisukokkuvõtete hindamise pakett, kasutab võrdlemisel fragmentidena 17 erineva pikkusega n-grammi. On teada, et ROUGE ühildub hästi inim-

hinnangutega, mille üks põhjuseid võib olla see, et ROUGE vaatab erineva pikkusega fragmente võrdsetena. Lin ja Demner-Fushmani järgi saavad fragmendid hindamise süsteem POURPREs (Hovy, Lin, Zhou, Fukumoto 2006) skoori informatiivsuse alusel, mis on arvutatud iga sõna kohta eraldi. Selleks, et defineerida, mis moodustab sisufragmendi informatiivsuse põhjal, kasutatakse selleks treenitud inimesi. Defineeritakse näiteks sellised tuntud üksikud sidusad ühikud nagu „Ameerika Ühendriigid“ või „lennuk maandus“. Nenkova nimetab selliseid fragmente kokkuvõtte sisu ühikuteks (*Summary Content Units*). Semantilisi fragmente on ka mujal kasutatud. Riezer kasutab neid sellisel juhul, kui sisukokkuvõte tehakse lausete parsimise teel LFG struktuurideks ning seejärel jäetakse valitud kogus neist ära, kasutades *Maximim Entropy* mudelit (Hovy, Lin, Zhou, Fukumoto 2006).

2.2.4 Meetod- ROUGE

ROUGE on Perli programm. ROUGE (Lin 2004) nimi tuleb ingliskeelsest väljendist Recall-Oriented Understudy for Gisting Evaluation. ROUGE pakett sisaldab meetodeid, mille abil saab automaatsel viisil kindlaks teha sisukokkuvõtte kvaliteeti, võrreldes seda inimese tehtud kokkuvõtetega (referentssisukokkuvõtted). Hindamis-meetodites kasutab ROUGE ülekattuvaid ühikuid n-gramme, sõnajärjendeid ning ühilduvaid sõnapaare hinnatava sisukokkuvõtte ning referentssisukokkuvõtete vahel. ROUGE hindamispakett sisaldab nelja moodulit: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S:

ROUGE-N: n-grammide kokkusattumine.

N-gramm on järjend n sõnast või kirjavahemärgist. ROUGE-N võrdleb n-grammide kattumist hinnatava kokkuvõtte ja mingi arvu referentssisukokkuvõtete vahel. Seega on ROUGE-N mooduli poolt enam soositud see hinnatav kokkuvõte, mis sisaldab rohkem ühesuguseid sõnu erinevates referentssisukokkuvõttetes.

ROUGE-L: pikim ühine alamjärjend.

Järjend $Z=[z_1,z_2,...,z_n]$ on järjendi $X=[x_1,x_2,...,x_n]$ alamjärjend, kui leidub rangelt kasvav järjend $[i_1,i_2,...,i_k]$ X indeksite jaoks, nii et iga $j=1,2,...,k$ puhul saame $x_{ij}=z_j$. Kui on antud kaks järjendit X ja Y, siis nende pikim ühine alamjärjend (*longest common subsequence-LCS*) on alamjärjend, mis ühistest alamjärjenditest kõige pikem. Et pikimat ühist alamjärjendit kasutada kokkuvõtete hindamises, vaadatakse sisukokkuvõtte lauseid kui

sõnade järjendeid. Mida pikem on pikim ühine alamjärjend kahe kokkuvõtte lausetes, seda sarnasemad on need kaks kokkuvõtet.

ROUGE-W: kaalutud pikim ühine alamjärjend.

ROUGE-W on ROUGE-L täiendatud variant. Pikima ühise alamjärjendi meetodit on parandatud sellega, et lisatakse ka sellised pikimad ühised alamjärjendid, mis eelmise meetodi puhul on jäetud mälusse, kuid pole peetud vajalikuks kasutada.

ROUGE-S: vahele-jätmise-bigrammi koosesinemine.

Vahele-jätmise-bigramm on ükskõik missugune sõnapaar lausejärjestuses, mis lubab juhuslikke lünki. Vahele-jätmise-bigrammi koosesinemise meetod mõõdab bigrammide ülekattumist hinnatava sisukokkuvõtte ja referentssisukokkuvõtete vahel.

Ingliskeelne näide (Lin 2004):

Lause 1. police killed the gunman

Lause 2. police kill the gunman

Lause 3. the gunman kill police

Lause 4. the gunman police killed

Igas lauses on sõnade arvu järgi 6 vahele-jätmise-bigrammi. Näiteks lauses 1 on järgmised vahele-jätmise-bigrammid: “police killed”, “police the”, “police gunman”, “killed the”, “killed gunman”, “the gunman”. Lauses 2 on kolm vahele-jätmise-bigrammi sobivust lausega 1: “police the”, “police gunman”, “the gunman”. Lauses 3 on üks vahele-jätmise-bigrammi sobivus lausega 1: “the gunman”. Lauses 4 on kaks vahele-jätmise-bigrammi sobivust lausega 1: “police killed”, “the gunman”.

2.2.5 Meetod- baaselementide leidmine ja tuvastamine

Baaselementide abil etteantud kokkuvõtte hindamise algoritm (Hovy, Lin, Zhou, Fukumoto 2006) jaguneb kolme alaüksusesesse, mis on vastavuses kolme mooduliga, mida meetodi paketis kasutatakse. Sisendina võetakse sisukokkuvõte ja sellele antakse skoor. Pakett rakendab mooduleid kaks korda, kahes faasis: ettevalmistus ja skoorimine.

Ettevalmistusfaasis lahutab esimene moodul referentssisukokkuvõtted ideaalseteks baaselementideks. Teine moodul töötleb neid baaselemente ja ühildab semantiliselt identsed. Kolmas moodul määrab igale ideaalsele baaselemendile skoori.

Teises faasis lahutab esimene moodul hinnatava kokkuvõtte eraldi baaselementide nimekirjaks. Teine moodul võrdleb iga baaselementi ideaalsete baaselementide nime-

kirjaga. Kolmas moodul määrab skoori igale baaselemendile, mida on vaja hinnata ja arvutab üldskoori kõigi hinnatavate baaselementide kohta, mis sisaldusid kokkuvõttes.

Baaselementide abil sobitamise ja võrdlemise strateegiad saab jagada mitmesse klassi (kõige lihtsamast kõige keerulisemani):

1. leksikaalne omapära - sõnavormid peavad omavahel täpselt sobima
2. lemma omapära - sõna algvormid peab omavahel sobima
3. sünonüümi omapära - sõnad või nende ükskõik millised sünonüümid peavad omavahel sobima
4. (umbkaudsed) fraasi parafraasid peavad omavahel sobima
5. semantiline üldistus - sõnad, mis moodustavad baaselemendi, on asendatud nende semantiliste üldistustega ja siis sobitatud üldistustaseme valikuga.

Keerulised võrdlemise strateegiad peaksid ära tundma terve või osalise semantilise samaväärsuse. Näiteks “umbes 20 miljonit krooni” ja “ 19,8 miljonit krooni”.

Hovy jt (Hovy, Lin, Zhou, Fukumoto; 2006) käsitlevad baaselementidena fraasi põhja või kolmikut põhi-laiend-relatsioon. Põhja-laiendi relatsiooni leidmiseks kasutatakse erinevaid inglise keele parsereid:

BE-L: Charniak parser (valimispuu) + CYL kärpimisreeglid,

BE-F: Minipar (sõltuvuspuu relatsioonidega) + JF kärpimisreeglid,

Chunker: süntaktilise üksuse kuhjaja, mis sisaldab kärpimisreegleid,

Microsofti parser + kärpimisreeglid.

2.3 Välimised hindamismeetodid

Mani (Mani 2002) järgi on välimiste hindamismeetodite idee selles, et teha kindlaks kokkuvõtte tegemise kasulikkus mingis teises ülesandes või praktilises süsteemis, näiteks:

- Kui kokkuvõtte sisaldab mingisuguseid juhiseid, on võimalik mõõta instruktsioonide tõhusust nende täitmisel. Näiteks kui detailsest tehnilisest manuaalist, mis õpetab RAM mälu installeerima, on tehtud lühike, joonistega kokkuvõte, siis saab hinnata kokkuvõtte headust selle alusel, kas järgides kokkuvõtet on võimalik RAM mälu installeerida sama edukalt kui kasutades originaaljuhist.
- Võib uurida kokkuvõtte kasulikkust, lähtudes mingi informatsiooni vajadusest või eesmärgist, näiteks suurest kollektsioonist kellelegi oluliste dokumentide leidmisel või suunamisel.

- On võimalik hinnata kokkuvõtte tegija mõju süsteemile, milles see on tehtud, näiteks kui palju aitab või mõjutab sisu kokkuvõtmine küsimise-vastamise süsteemi.
- Veel üks võimalus on mõõta jõupingutuse suurust, mida on vaja, et hiljem kokkuvõtet muuta selliseks, et seda mingile vastuvõetavale ülesandest sõltuvale tasemele tuua.

2.3.1 Meetod- LDC Agreement

LDC puhul tehakse „kuldsel standardil“ põhinevad hinnangud väljaõpetatud märgendajate abil. Sellise meetodi on välja arendanud Pennsylvania ülikooli teadlased (Hobson jt 2007). LDC (Linguistic Data Cosortium) hinnanguid kasutatakse ka mõne teise hindamismeetodi juures baashinnangutena (nt järgmine meetod – relevantsuse ennustamine). LDC arvutamisel keskendutakse täpsuse mõõtmisele. Kokkuvõtte tegemise puhul on osad laused või fragmendid tekstist sellised, mida eeldatakse, et need jäävad erinevate kokkusurumisastmete puhul sisse. Täpsuse mõõtmisel vaadataksegi seda, kui suur on nende lausete, mida eeldati ja mis tõesti sisse jäid (nn „tõene positiivne“), ning nende lausete, mida ei eeldatud, ning mida ei jäetud sisse ka kokkuvõttesse (nn „tõene negatiivne“), summa üle kõigi hinnangute arvu. Täpsuse arvutamine valitakse põhjusel, et sellisel juhul on dokumendi jagunemine „tõesteks positiivseteks“ ning „tõesteks negatiivseteks“ umbes 50%. Üldisemate meetodite puhul on see arv väiksem. LDC puhul on probleemiks madal ühildumise tase märgendajate seas. See on tingitud „kuldsete standardite“ kasutamisest hinnangute tegemisel.

2.3.2 Meetod- Relevantsuse ennustamine

LDC alternatiiviks on relevantsuse ennustamine (*Relevance Prediction*), kus iga kasutaja teeb ise enda „kuldse standardi“ antud baasteksti alusel. Ühildumist mõõdetakse võrdlemise teel. Võrreldakse seda, mida kasutajad on pidanud tähtsaks täispikas baastekstis ning seda, mis kasutajad on valinud enda nn „kuldse standardi“ kokkuvõttesse. Kui kasutaja teeb lause või muu tekstifragmendi puhul otsuse info kasuks, mis on kooskõlas baastekstis tähtsaks peetud infoga, tähendab see, et kokkuvõttes on piisavalt informatsiooni ning kokkuvõte peaks saama kõrge hinnangu. Kui info ei ole baastekstis tähtsaks arvatud infoga kooskõlas, saab kokkuvõte madala skoori. Et relevantsuse ennustamise skoori arvutada, võetakse aluseks, et kui hindaja valitud info kokkuvõttes ühildub baastekstis

tähtsaks peetud infoga, on sellele kokkuvõttele hinnanguks antud 1, vastupidisel juhul 0. Väärtused arvutatakse kokku üle kõigi baasteksti hinnangute.

Hea on, kui hindaja, mõnel konkreetsel juhul ka näiteks süsteemi kasutaja, annab kõigepealt hinnangu baasteksti kohta ning alles siis individuaalse kokkuvõtte kohta. Selline järjekord kindlustab selle, et baastekstis ei peeta tähtsaks automaatselt seda, mis sisaldus kokkuvõttes. Sellist meetodit on katsetes kasutatud ning tulemused on näidanud, et see lähenemine on usaldusväärsem kui LDC, kuna ei tugine „kuldse standardi“ otsustel, mis on tehtud inimese poolt. Relevantsuse ennustamine sobib kokkuvõtete kasulikkuse illustreerimiseks reaalsetes olukordades nagu brauseri keskkond.

3. Katsed eestikeelsete tekstide sisukokkuvõtetega

Töö üheks oluliseks osaks oli uurida, milliste meetoditega on automaatset sisukokkuvõtet võimalik hinnata ning kui kättesaadavad on erinevad tarkvaralised lahendused. Selleks tegi autor käsitsi 50 Postimees online´is ilmunud artiklist 40% mahuga sisukokkuvõtted. Samuti on samadest artiklitest genereeritud sama mahuga automaatsed sisukokkuvõtted eestikeelsete tekstide sisukokkuvõtjaga EstSum.

3.1 Sisukokkuvõtjast EstSum

EstSum (Müürisep 2006) on eestikeelsete tekstide sisukokkuvõtja, mis kasutab lausete väljavalimismeetodit. Hetkel on EstSum orienteeritud veebis avaldatud uudiste ja ajaleheartiklite sisukokkuvõtetele.

EstSum koosneb kolmest moodulist: HTML-konverter, lausestaja ja väljavõtete tegija. HTML-konverter eemaldab sisukokkuvõtte jaoks ebaolulised HTML-märgendid, normaliseerib ristuvad märgendid, eemaldab tabelid ning konverteerib sisendi SGML-formaati. Lausestaja kasutab reeglipõhist meetodit sisendi töötlemisel, lause alguse ja lõpu märgendamiseks kasutatakse 30 regulaaravaldist.

EstSum arvutab sisukokkuvõtte pikkust kahel viisil - esimene moodus on tavaline lausepõhine meetod, mille korral 30% kokkuvõte tähendab, et tekstis 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. EstSum loeb tähtsaks neid lauseid, milles kasutatakse rasvast või kaldkirja. Formaadipõhine skoorifunktsioon arvestab ka lause kirjavahemärkidega: hüüu- ja küsimärgid vähendavad lause kaalu, samuti jutumärgid. Täielikult välistatakse pildiallkirjade lisamine sisukokkuvõttesse. 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.

Eestikeelsete tekstide sisukokkuvõtja EstSum on tegelikult hetkel veel eksperimentaalses arengujärgus.

3.2 Katsete sisu

Katsete sisuks oli 50st Postimees online'is[1] ilmunud artiklitest sisukokkuvõtete tegemine. Mahult moodustavad tehtud sisukokkuvõtted 40% vastavast originaalartiklist. Valitud artiklid on ilmunud ajavahemikus 1.- 3. aprill 2008. Artiklite valikul oli tähtsaks faktoriks see, et EstSum töötab hetkel just Postimees online tekstidega. Artiklid on rubriigist „Eesti uudised“. Sisukokkuvõtted on tehtud igast artiklist nii käsitsi kui ka eestikeelsete tekstide sisukokkuvõtjaga EstSum, et hiljem saadud tulemusi omavahel võrrelda. Valitud artiklite keskmiseks sõnade arvuks oli 189,04 sõna ning lausete arvuks keskmiselt 10,16 lauset. Täpne tabel kasutatud artiklite mahuga ning vastavate sisukokkuvõtete mahtudega on toodud lisas 2. Lisas 1 on toodud kasutatud originaalartiklid ning nii EstSumiga kui ka käsitsi tehtud sisukokkuvõtted.

3.3 Autori muljed katsete tegemisel

On suhteliselt iseenesestmõistetav, et oluliselt mugavam on, kui arvuti teeb sinu eest mingisuguse aeganõudva tegevuse ära. Sisukokkuvõtteid uudisteartiklitest tehes läks algus kiiresti, sest online uudised ei ole tavaliselt väga pikad ning olulise leidmine tundus loogiline. Samas selgus kiiresti tõsiasi, et kuna lauseid antud juhul poolitada või muuta ei tohtinud, sest EstSum seda ei tee, siis võib ette tulla päris mitmeid olukordi, kus laused tunduvad võrdväärsed või on parem lause liiga pikk, kuid otsus tuleb siiski vastu võtta. Kiiresti sai selgeks ka see, et uudiste puhul kordab esimene lõik pealkirja mõtet pikemalt ja see lõik tasub tavaliselt alles jätta. See on ka see lõik, mis tavaliselt tekstis on märgitud kas rasvaselt või kaldkirjas. Raske on kindlasti otsuse langetamine selliste lausete puhul, kus kasutatakse asesõnu ja sisu tundub oluline, eelnev isikut tutvustav lause on aga kas

1 www.postimees.ee

ebavajalik või liiga pikk. Samuti tekib dilemma juhuste juures, kus pool uudises esinevast tsitaadist tundub sisult oluline, pool aga mitte.

3.4 Meetodite valik

Et tehtud katseid analüüsida, tuli teha valik peatükis 2 kirjeldatud meetodite hulgast, mis oleksid reaalselt võimalikud ja esialgsel hinnangul kõige efektiivsemad. Peatükis 2 toodud meetoditest analüüsimiseks kõiki kasutada ei olnud võimalik. Väliseid hindamismeetodeid LDC Agreement´i ja relevantsuse hindamist ei ole võimalik veel kasutada, kuna Eestis on automaatne sisukokkuvõtete tegemine arenguga alles algusjärgus ning ei ole selliseid ülesandeid, kus sisukokkuvõtte headust väliste hindamismeetodite abil hinnata. Eestis ei kasutata automaatseid sisukokkuvõtteid veel mingiks kindlaks otstarbeks. Seega osutusid ainsateks reaalselt Eesti kontekstis kasutatavateks hindamismeetoditeks kirjeldatud sisemised hindamismeetodid.

Üheks kasutatavaks meetodiks on valitud käsitsi ekstraheeritud sisukokkuvõtete võrdlemine läbivaatamise teel, võttes aluseks, et referentssisukokkuvõtteks on autori käsitsi tehtud sisukokkuvõte, kuna ka automaatsete hindamismeetodite puhul kasutatakse referentssisukokkuvõtetena just inimeste tehtud sisukokkuvõtteid. Teise meetodina osutus valituks ROUGE, kuna see oli programmina Internetist kättesaadav. (http://berouge.com/default.aspx, eeldab programmi loojale soovkirja saatmist ning oma eesmärkide selgitamist, misjärel saadetakse link, millelt on võimalik programm endale alla laadida.)

3.5 Käsitsi automaatse sisukokkuvõtja töö hindamine

3.5.1 Tulemused

Kõik 50 käsitsi ja 50 EstSumiga tehtud kokkuvõtet algavad kõigil juhtudel sama, artikli esimese lausega. Esimene lause artiklis kordab tavaliselt pealkirjas väljaöeldud mõtet ning täiendab seda. Niisiis on see tõepoolest sisult samuti tähtis lause. Sõnade kattumise puhul oli aluseks võetud see, et autori tehtud sisukokkuvõte on referents-

sisukokkuvõte, ning kui EstSumi poolt automaatselt tehtud sisukokkuvõte sisaldas kõiki autori sisukokkuvõttes sisalduvaid sõnu, mis kattusid täielikult, sai EstSum skooriks 100%, kuigi võis tekkida olukord, kus lisaks sellele 100% kattumisele sisaldas EstSumi sisukokkuvõte veel lauseid.

Katsete tulemuseks on saadud, et EstSumi tehtud sisukokkuvõtted sisaldavad keskmiselt 65,29% referentssisukokkuvõttest, mis on üks tähtsamaid tulemusi. Kattumine on päris suur ning kui hinnata EstSumiga tehtud kokkuvõtteid headeks või halbadeks, on saadud tulemuse alusel võimalik teha järeldus, et EstSumi tehtud sisukokkuvõtted on head. Tabelis 1 on toodud eraldi kõigi artiklite sisukokkuvõtete protsentuaalne kattumine:

Tabel 1. Sõnade kattumine sisukokkuvõtetes sõna arvu järgi ning protsentuaalselt

Sõnade kattumine sisukokkuvõtetes		
Artikkel	***Kattunud sõnade arv***	***Protsent %***
1	51	61,1
2	63	100,0
3	76	100,0
4	18	50,0
5	54	100,0
6	45	60,0
7	32	49,2
8	55	53,9
9	17	31,5
10	66	91,7
11	127	74,7
12	46	100,0
13	65	65,7
14	21	50,0
15	35	100,0
16	18	38,3
17	111	68,5
18	41	40,6
19	47	63,5
20	13	18,6

21	53	100,0
Artikkel	***Kattunud sõnade arv***	***Protsent %***
22	49	100,0
23	80	63,0
24	32	51,6
25	39	70,9
26	68	76,4
27	67	78,8
28	21	33,3
29	53	100,0
30	15	22,4
31	12	27,9
32	45	78,9
33	29	53,4
34	20	51,3
35	71	70,3
36	50	56,8
37	71	79,8
38	20	50,0
39	68	57,1
40	25	59,5
41	24	46,2
42	30	50,8
43	36	52,2
44	44	100,0
45	33	64,7
46	52	68,4
47	41	100,0
48	37	100,0
49	25	50,0
50	24	33,3
Keskmine:		**65,29**

3.5.2 Näited

Kõige väiksem kattumine kahe sisukokkuvõtte vahel oli 18,6% 20. artiklis, kus kattus vaid esimene lause, kuigi vastavad sisukokkuvõtted ei olnud lühikesed. Autori tehtud sisukokkuvõte oli 101 ja EstSumi sisukokkuvõte 133 sõna pikk. Toon siia näiteks selle artikli mõlemad sisukokkuvõtted:

EstSum:
Tallinn hakkab ravimeid varustama venekeelse teabega
Linnavalitsuse korraldusel hakkab Tallinna sotsiaal- ja tervishoiuamet tagama ravimite toimeainepõhise venekeelse teabe kättesaadavust.
Ravimite venekeelne teave peab olema kättesaadav käsimüügiravimite osas 1. jaanuariks 2009 ja retseptiravimite osas 1. jaanuariks 2010.
Linnavalitsuse korraldusega antakse Tallinna sotsiaal- ja tervishoiuametile ülesanne koostada enamkasutatavate käsimüügi- ja retseptiravimite toimeainepõhine loetelu, korraldada vastava teabe tõlkimine vene keelde ja korraldada ravimite toimeainepõhise teabe venekeelsete infovoldikute koostamine, trükkimine ning levitamine.
Ravimite toimeainepõhine teave peab olema vene keeles kättesaadav käsimüügiravimite osas 1. jaanuariks 2009 ja retseptiravimite osas 1. jaanuariks 2010.

Autor:
Tallinn hakkab ravimeid varustama venekeelse teabega
Linnavalitsuse korraldusel hakkab Tallinna sotsiaal- ja tervishoiuamet tagama ravimite toimeainepõhise venekeelse teabe kättesaadavust.
Eesti Avatud Ühiskonna Instituut korraldas 2007. aasta juulis 15-74-aastaste Tallinna elanike uuringu, millest selgus, et 45 protsenti venekeelsest elanikkonnast oskab eesti keelt halvasti või ei valda üldse.
"Kuna Eesti apteekides müüdavatel ravimitel puudub venekeelne teave ravimi koostise, toimeainete sisalduse ning kasutamise ja säilitamise kohta, on paljudel Tallinna venekeelsetel elanikel raskusi ravimite õige kasutamisega," ütles abilinnapea Merike Martinson.

Kõige suurem kokkulangevus kahe samast artiklist tehtud sisukokkuvõtte vahel oli 100%. Selliseid 100% kattuvusega kokkuvõtteid oli mitu, enamus neist keskmisest sisukokkuvõtte pikkusest lühema pikkusega. Üks selline artikkel oli 2. artikkel. Toon siia näite sellest sisukokkuvõttest:

EstSum ja autor:
Valitsus raskendas välisametnikele altkäemaksu andmist
Valitsus kiitis täna heaks karistusseadustiku ja kriminaalmenetluse seadustiku muutmise seaduse eelnõu, millega võideldakse rahvusvahelise korruptsiooniga, raskendades välisriigi ametiisikutele altkäemaksu andmist.
Täna valitsuse poolt heaks kiidetud eelnõu ühe autori, justiitsministeeriumi nõuniku Tanel Kalmeti sõnul loevad maailma arenenud riigid äritegevuses välisriigi ametiisikule altkäemaksu andmist kuriteoks, mis kahjustab oluliselt vaba konkurentsi ja rikub häid äritavasid.
Konventsiooni täielik ja tulemuslik rakendamine Eestis on ka OECD-ga ühinemise tingimus.

3.5.3 Sisukokkuvõtete sisu ja eripärasuste analüüs

Kõik sisukokkuvõtted sisaldavad originaalartiklist esimest lauset. Postimees online´is on see lause, vahest ka lõik tavaliselt paksus kirjas ning seletab veidi pikemalt lahti pealkirja lause. Nagu ka lisas toodud tabelist näha, eelistab EstSum jätta sisukokkuvõtte pigem pikemaks, kui etteantud protsent võiks ette näha. Autor seevastu on eelistanud enamikel juhtudel, kus ühe lause sissejätmine tähendaks 40% oluliselt pikemaks jäämist, jätta lause pigem välja. Nii võiski tekkida olukord, kus EstSumi sisukokkuvõte sisaldas 100% autori poolt tehtud sisukokkuvõtte sisu, kuid EstSumil oli lisaks näiteks veel üks lause.

Ühe kokkuvõtte puhul oli huvitav see, et EstSum oli sisse jätnud vahepealkirja, mille lõpus oli hüüumärk, sisule see midagi juurde ei andnud. Samuti läheb samasse kategooriasse lause „Täiendatud kell 11.35- Eesti läheb Hiinasse turiste meelitama“, mille EstSum oli sisse jätnud, kuid mis on sisu koha pealt ebaoluline.

Oluline erinevus autori ja EstSumi kokkuvõtete vahel oli see, et autor jaoks oli keeruline jälgida ja otsustada, kas asesõnadega laused sisse jätta. Asesõnadega lausete

puhul oleks eelnev lause kindlasti isikut tutvustav, kuid kui sellisel juhul oleks läinud kokkuvõte liiga pikaks, või oli isikut tutvustav lause iseenesest tegelikult ebavajalik, siis autor jättis pigem asesõnaga lause samuti välja. EstSumi puhul on kokkuvõtetes oluliselt rohkem asesõnadega lauseid. Üldiselt on sisukokkuvõtted aga ka sisu poolest väga sarnased nagu seda näitab ka lausete kattumise protsent sõnade arvu järgi- 65, 29%. Autori märkmed sisukokkuvõtete võrdlemisel on toodud lisas 3.

3.7 ROUGE programmiga automaatse sisukokkuvõtja töö hindamine

ROUGE on Perli programm, millele saab ette anda hinnatava ja referents-sisukokkuvõtte, ning programm annab väljundiks skoori 0-1, kus 0 tähistab olukorda, kus midagi ei kattu ning 1 olukorda, kus kaks sisukokkuvõtet kattuvad 100 protsendiliselt. Kui ROUGE-le eriparameetreid ning käske ette mitte anda, kasutab programm vaikimisi ROUGE-L ehk pikima ühise alajärjendi järgi skoori arvutamist. Seda ei olnud ka autoril esialgu mõtet muuta, kuna EstSumi genereeritud sisukokkuvõtete ning Eesti keelega sobibki ROUGE alameetoditest just ROUGE-L. Seda põhjusel, et EstSum ei muuda lauseid ja sisukokkuvõttesse jäetakse sisse terve lause. ROUGE-L võrdleb pikimaid ühiseid alamjärjendeid. Alamjärjendid siinkohal ongi laused. Selle tõttu olid oodatavad tulemused sarnased neile, mis saadud käsitsi tehtud hindamise puhul.

Hiljem on tehtud ka 25 artikli sisukokkuvõtete analüüs teiste ROUGE moodulitega – ROUGE-N, ROUGE-W ja ROUGE-S, mis võrdlemisel kasutavad väiksemaid fragmente tekstist kui terve lause.

ROUGE genereerib kolm hinnangut – saagis (*recall),* täpsus *(precision)* ja f-mõõt *(F-measure)*. Täpsus ja f-mõõt on kasulikud, kui sisukokkuvõtte pikkus ei ole sunnitud kindla pikkusega. Nii oli antud olukorras kasulik võrrelda saagiste tulemusi nii ROUGE-L jaoks kui ka hiljem ROUGE teiste moodulite juures. Järgnevalt on toodud näide ROUGE-L väljundist ühe sisukokkuvõtte jaoks:

ROUGE-L Average_R: 0.52632 (95%-conf.int. 0.52632 - 0.52632)

ROUGE-L Average_P: 0.43478 (95%-conf.int. 0.43478 - 0.43478)

ROUGE-L Average_F: 0.47619 (95%-conf.int. 0.47619 – 0.47619),

kus Average_R on meile vajalik saagis, Average_P täpsus ning Average_ F f-mõõt skoor. ROUGE väljundid kõigi artiklite jaoks on toodud lisas 4.

3.7.1 ROUGE-L tulemused

ROUGE-L skoor tuli sarnane käsitsi hinnatud tulemusele. Kui ROUGE-L skoor teisendada protsentidesse ning ümardada ühe kümnendkohani nagu käsitsi hindamise puhul olid väärtused arvutatud, tuli tulemuseks 68,96% kattuvust. Erinevus käsitsi läbivaatamisega võib tuleneda sellest, kuidas automaadil on õnnestunud lauseteks jagamist teha, alapealkirjad on näiteks ilma punktita. Tabelis 2 on toodud ROUGE skoorid iga artikli sisukokkuvõtete jaoks eraldi:

Tabel 2. ROUGE-L skoorid iga artikli sisukokkuvõtete jaoks

Artikkel	***ROUGE-L skoor***	***Ümardus protsentidesse***
1	0.67708	67,7
2	1.00000	100,0
3	1.00000	100,0
4	0.52632	52,6
5	1.00000	100,0
6	0.65476	65,4
7	0.52055	52,0
8	0.61475	61,4
9	0.47143	47,1
10	0.90588	90,5
11	0.77228	77,2
12	1.00000	100,0
13	0.64957	65,0
14	0.50000	50,0
15	1.00000	100,0
16	0.40816	40,8
17	0.73158	73,2
18	0.63248	63,2

Artikkel	*ROUGE-L skoor*	*Ümardus protsentidesse*
19	0.70408	70,4
20	0.25974	26,0
21	1.00000	100,0
22	1.00000	100,0
23	0.67586	67,6
24	0.62857	62,9
25	0.73846	73,8
26	0.75510	75,5
27	0.79612	79,6
28	0.43662	43,7
29	1.00000	100,0
30	0.30380	30,4
31	0.36957	37,0
32	0.79104	79,1
33	0.56322	56,3
34	0.55814	55,8
35	0.78992	79,0
36	0.63208	63,2
37	0.82353	82,4
38	0.54717	54,7
39	0.60390	60,4
40	0.66667	66,7
41	0.48333	48,3
42	0.5479	54,8
43	0.71717	71,7
44	1.00000	100,0
45	0.75000	75,0
46	0.71910	71,9
47	1.00000	100,0
48	1.00000	100,0
49	0.59649	59,6
50	0.51163	51,2
Keskmine:	**0,70068**	**68,96**

3.7.2 ROUGE-L järeldused

Kuna kõikumine kahe erineva hindamismeetodiga ROUGE-L ja käsitsi saadud tulemuse vahel oli vaid 3,67%, võib teha järelduse, et mugavam on kindlasti kasutada automaatset hindamismeetodit, kuna selle kasutamine on oluliselt kiirem võrreldes käsitsi läbivaatamisega. Samas nõuab ka ROUGE kasutamine programmi süvenemist ning selle õppimist, samuti eeldab programm, et sisestatavad kokkuvõtted on teatud formaadis. Näiteks laialt levinud .doc laiendiga failid sisendiks ei sobi. Kõige paremini tundus sobivat txt fail, kuna seal ei teki dokumendi algusesse või lõppu mingisuguseid ühiseid osi, nagu näiteks html dokumendil:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0
Transitional//EN">
<HTML>
<HEAD>
     <META HTTP-EQUIV="CONTENT-TYPE" CONTENT="text/html;
charset=windows-1252">
     <TITLE>Kokkuv&otilde;te</TITLE>
     <META NAME="GENERATOR" CONTENT="OpenOffice.org 2.0
(Win32)">
     <META NAME="AUTHOR" CONTENT="Keili">
     <META NAME="CREATED" CONTENT="20080404;12480000">
     <META NAME="CHANGED" CONTENT="20080522;13240332">
     <STYLE>
     <!--
          @page { size: 21cm 29.7cm; margin: 2.5cm }
          P { margin-bottom: 0.21cm; direction: ltr; color:
#000000; widows: 2; orphans: 2 }
          P.western { font-family: "Times New Roman", serif;
font-size: 12pt; so-language: et-EE }
          P.cjk { font-family: "Times New Roman", serif;
font-size: 12pt }
          P.ctl { font-family: "Times New Roman", serif;
font-size: 12pt; so-language: ar-SA }
     -->
     </STYLE>
</HEAD>
<BODY LANG="en-US" TEXT="#000000" DIR="LTR">
<P LANG="et-EE" CLASS="western" STYLE="margin-bottom:
0cm">Uuring:
rahvas k&auml;rbib kulusid toidu arvelt</P>
```

```
<P LANG="et-EE" CLASS="western" STYLE="margin-bottom:
0cm"><BR>
</P>
<P LANG="et-EE" CLASS="western" STYLE="margin-bottom:
0cm">Faktum &amp;amp;
Ariko uuringu tulemustest selgub, et ligi kaks kolmandikku
Eesti
inimestest on kuue kuu taguse ajaga v&otilde;rreldes oma
tarbimist
piiranud ning esimese asjana on inimesed luubi alla
v&otilde;tnud
s&ouml;&ouml;gi - tervelt 59 protsenti k&uuml;sitlusele
vastanutest
on toidukulutusi v&auml;hendanud ning tarbimisharjumuste
muutusi on
juba n&auml;ha ka toidukauplustes.</P>
</BODY>
</HTML>
```

Sellisel juhul võtab ROUGE programm näiteks selle dokumendi algust ja lõppu kui teksti osa ning sisukokkuvõte saab suurema ROUGE skoori, kuna kattumist on rohkem. Samas sisukokkuvõtte sisu moodustab sellest dokumendist tegelikult alla poole. Selline tulemus on aga kindlasti petlik. Kui on aga soovi ja vajadust ROUGE programmiga tutvuda, siis on tegemist oma valdkonnas kasuliku programmiga.

3.7.3 ROUGE-N, ROUGE-W ja ROUGE-S tulemused

ROUGE-N jaoks on ette antud 4 erinevat võimalust, mille puhul ROUGE-N vaatab vastavalt võrdluses kas 1-, 2-, 3- või 4-sõnalisi järjendeid ehk n-gramme. Kõige lähemal ROUGE-L tulemusele on ROUGE-N skoor 2-sõnalise järjendi puhul, kattumine on sel juhul keskmiselt 67,8%. ROUGE-N 1-sõnalise järjendiga võrdlemisel annab tulemuseks 72,96% kattumise. ROUGE-N 3-sõnalise järjendi puhul on skoor 65,91% ning 4-sõnalise järjendi puhul 64,63%. Üleüldiselt kõige väiksema kattumise kahe sisukokkuvõtte vahel annab ROUGE-W moodul, kus keskmiseks skooriks tuli vaid 31,71%, mis erineb väga palju ROUGE-L tulemusest, mistõttu võib oletada, et Eesti keelele ROUGE-W ei sobi. ROUGE-S andis tulemuseks 56,84%. Need tulemused näitavad, et Eesti keele sisukokkuvõtete puhul, mis on tehtud lausete ekstraheerimise teel, sobib nendest kolmest alameetodist ROUGE-L kõrvale ka ROUGE-N, mis mõnel juhul, valides võrreldavaks n-grammiks väikse liikmete arvuga sõnajärjendi, võib anda isegi täpsema tulemuse.

ROUGE-N puhul on näha, et mida pikem sõnajärjend, seda väiksemaks kattumine kahe sisukokkuvõtte vahel muutub. Tabelis 3 on toodud protsentuaalne tulemus 1. kuni 25. artiklini iga meetodi järgi:

Tabel 3. ROUGE-N, ROUGE-W ja ROUGE-S skoorid esimese 25 artikli jaoks

Artikkel	*ROUGE-N (1) %*	*ROUGE-N (2) %*	*ROUGE-N (3) %*	*ROUGE-N (4) %*	*ROUGE-W %*	*ROUGE-S %*
1	67,7	63,2	61,7	60,2	30,7	45,0
2	100,0	100,0	100,0	100,0	43,4	100,0
3	100,0	100,0	100,0	100,0	42,8	100,0
4	52,6	48,6	47,2	45,7	28,4	27,0
5	100,0	98,6	97,1	95,5	43,7	100,0
6	65,5	61,4	59,8	59,3	28,8	42,6
7	53,4	47,2	43,7	41,4	22,7	28,3
8	62,3	50,4	46,7	43,7	24,0	37,9
9	52,9	37,7	32,4	31,3	20,9	26,9
10	90,6	88,1	85,5	84,1	40,1	80,8
11	74,8	68,7	67,0	65,8	28,6	55,5
12	100,0	100,0	100,0	100,0	45,6	100,0
13	67,5	62,9	61,7	59,6	25,5	44,5
14	52,0	44,9	43,8	42,6	25,1	26,3
15	100,0	100,0	100,0	100,0	48,4	100,0
16	42,9	37,5	36,2	34,8	20,8	17,9
17	73,8	67,7	64,9	63,1	27,0	53,8
18	65,0	52,6	50,4	48,2	24,9	41,7
19	71,4	63,9	60,4	57,9	29,7	46,6
20	27,3	19,7	17,3	16,2	11,8	7,4
21	100,0	100,0	100,0	100,0	45,2	100,0
22	100,0	98,4	96,7	94,9	45,4	100,0
23	67,6	61,1	60,1	59,2	27,0	45,0
24	62,9	53,6	50,0	49,3	28,8	39,2
25	73,9	68,8	65,1	62,9	33,5	54,7
Keskmine:	**72,96**	**67,8**	**65,91**	**64,63**	**31,71**	**56,84**

4. Kokkuvõte

Käesolevas bakalaureusetöös on uuritud sisukokkuvõtjate töö hindamismeetodeid ja tehtud reaalsed katsed nendest kahega. Enamus levinud hindamismeetodeid on sisemised hindamismeetodid, mis kasutavad oma töös inimese tehtud sisukokkuvõtteid ehk referentssisukokkuvõtteid. Neid kasutatakse võrdlemisel automaatsete sisukokkuvõtetega, võttes aluseks, et referentssisukokkuvõtted on nö ideaalsed. Töös selgus ka, et välimised hindamismeetodid Eesti konteksti ei sobi, kuna automaatne sisukokkuvõtete tegemine on alles arengujärgus ning seetõttu ei ole reaalseid rakendusi, kus automaatseid sisukokkuvõtteid kasutataks.

Töös leiti vastus küsimusele, kui suures osas kattuvad automaatsed sisukokkuvõtted käsitsi inimese poolt tehtutega. Kasutati EstSumiga genereeritud ning käsitsi tehtud sisukokkuvõtete võrdlemist kahe meetodiga. Katsete tulemusena selgus, et kahe meetodiga hindamisel saadud keskmine kattuvus sisukokkuvõtetes on vastavalt käsitsi läbivaatamisel 65,29% ning ROUGE-L puhul, mis on oma olemuselt sarnane käsitsi lausete võrdlemisega oli tulemus 68,96%. Samuti katsetati ka ROUGE programmi teisi mooduleid.

Abstract

Evaluation of Automatic Summarization

Bachelor Thesis

Keili Sellik

This paper studies different evaluation methods for automatic summarization. There are two types of methods - intrinsic and extrinsic. Intrinsic methods usually use „ideal“ summaries for evaluation. These are handmade summaries, which are considered to be ideal and they are compared to automatic summaries, taking into consideration that automatic summary is good, when it consists as much information from „ideal“ summary as possible. Extrinsic methods for evaluation are not usable in Estonian context at the moment, because automatic summaries are not far-spread in Estonia and there are no real situations where automatic summaries could be evaluated.

The paper also gives an answer to the question, how much do automatic summaries match with those that author had made manually. 50 Summaries made by both ways were evaluated with two methods and the result was that about 65,29% of summaries content matches in the case of manually compared and 68,96% match when using ROUGE-L. Other ROUGE methods were also tested.

5. Kasutatud allikad ja kirjandus

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

Müürisep, Kaili. *Eestikeelsete tekstide sisukokkuvõtjast EstSum*. Keel ja arvuti. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6.(Toim. M. Koit, R. Pajusalu, H. Õim) lk 115-125 Tartu 2006

Hovy, Eduard; Lin, Chin-Yew; Zhou, Liang; Fukumoto, Junichi. *Automated Summarization Evaluation With Basic Elements* In Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006), Genoa, Itaalia 2006

Mani, Inderjeet; Klein, Gary; House, David; Hirschman, Lynette. *SUMMAC: a text summarization evaluation*. Natural Language Engineering, 8(1): lk 43-68 2002

Lin, Chin-Yew. *ROUGE: a Package for Automatic Evaluation of Summaries.* In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004) Barcelona, Hispaania 2004

President Hobson, Stacy; Dorr, Bonnie J.; Monz, Cristof; Schwartz, Richard. *Task-based evaluation of text summarization using Relevance Prediction*. Information Processing and Management 43(6) 2007

Hassel, Martin; Dalianis, Hercules. *Generation of Reference Summaries*. In the proceedings of the 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Poznan, Poola 2005

ROUGE pakett kättesaadav: http://berouge.com/default.aspx, viimati külastatud 22.05.2008

Lisad

Lisa 1. Originaalartiklid, autori tehtud sisukokkuvõtted ja eestikeelsete tekstide sisukokkuvõtjaga EstSum genereeritud sisukokkuvõtted (CD)

Lisa 2. Originaalartiklite ja sisukokkuvõtete lausete ja sõnade arv

	Laused			***Sõnad***		
Artikkel	Originaal-artikkel	EstSum	Keili Sellik	Originaal-artikkel	EstSum	Keili Sellik
1	12	4	4	190	80	76
2	8	3	3	187	63	63
3	10	5	4	206	101	76
4	7	2	2	109	40	36
5	5	3	2	153	68	54
6	11	5	3	220	110	75
7	8	3	5	154	67	65
8	11	5	4	250	113	102
9	7	3	3	163	70	54
10	12	6	5	188	85	72
11	26	9	8	475	198	170
12	5	1	1	131	46	46
13	14	5	7	240	83	99
14	7	3	2	133	63	42
15	8	3	2	136	54	35
16	5	3	2	127	56	47
17	22	10	10	417	193	162
18	14	7	6	296	133	101
19	14	5	6	177	77	74
20	8	4	3	176	81	70
21	7	3	2	146	59	53
22	10	5	3	160	79	49
23	18	8	6	308	130	127
24	10	4	4	151	64	62
25	9	5	4	165	75	55
26	11	5	4	225	101	89
27	10	4	4	211	93	85
28	10	4	4	168	68	63
29	6	2	2	118	53	53
30	9	3	4	174	70	67

	Laused			***Sõnad***		
31	6	3	3	98	50	43
32	7	3	3	172	49	57
33	12	5	3	194	78	73
34	6	3	2	102	47	39
35	16	7	6	280	124	101
36	11	4	4	237	106	88
37	11	6	4	243	122	89
38	7	2	2	125	40	40
39	19	10	9	305	150	119
40	6	2	2	129	41	42
41	8	3	3	150	66	52
42	11	5	4	155	69	59
43	11	4	5	182	77	69
44	6	2	2	130	44	44
45	10	4	3	159	66	51
46	10	5	4	203	97	76
47	11	5	2	165	78	41
48	7	2	2	109	37	37
49	7	3	2	132	53	50
50	12	5	4	228	94	72
Keskmine	**10,16**	**4,3**	**3,76**	**189,04**	**81,24**	**69,28**

Lisa 3. Autori märkmed artiklite läbivaatamisel

Artikkel 1- 51 sõna kattuvad, esimene lause sama
Artikkel 2- kattumine 100%
Artikkel 3- 76 sõna algusest kattuvad, EstSum on juurde võtnud lõppu ühe lause.
Artikkel 4- 18 sõna kattuvad, lühikesest artiklist kattus niisiis vaid esimene lõik/lause pärast pealkirja. Teise lausena valisin mina ja EstSum erinevad.
Artikkel 5- 54 sõna kattuvad, sealhulgas esimene lause, estsum on lisaks ühe lause võtnud, muidu kattub 100 protsendiliselt
Artikkel 6- 45 sõna algusest kattuvad, EstSum mitu lauset pikem
Artikkel 7- 32 sõna kattuvad, taas esimene ja viimane lause.
Artikkel 8- 55 sõna kattuvad, sealhulgas esimene lausejärjestuses, mina ei kasutanud hüüumärgiga vahelauset (ei ole nagu lause, pigem paksus kirjas vahepealkiri)
Artikkel 9- 17 sõna kattuvad, ainult esimene lause

Artikkel 10- 66 sõna kattuvad, sealhulgas esimene lause, EstSum on jätnud sisse sellise lause nagu „Täiendatud kell 11.35 - Eesti läheb Hiinasse turiste meelitama! „, mis sisu kohapealt on täiesti ebaoluline.
Artikkel 11- 127 sõna kattuvad, sealhulgas esimene lause
Artikkel 12- 100% kattumine, kokkuvõte koosneb ühest, esimesest lausest.
Artikkel 13- 65 sõna kattuvad, sealhulgas esimene lause. Siin on huvitav see, et on kaks pikemat jutumärkides juttu, kus keegi midagi kommenteerib. EstSum on mõlemal juhul jutumärkides olevat osa poolitanud, mina olen aga kogu jutu alles jätnud.
Artikkel 14- 21 sõna (esimene lause) kattuvad, olgugi, et EstSumi kokkuvõte on lause võrra pikem.
Artikkel 15- 35 sõna kattuvad (minu poolelt on see 100%), kuid EstSum on jälle ühe lause lisaks pannud.
Artikkel 16- 18 sõna kattuvad, esimene lause
Artikkel 17- 111 sõna kattuvad, sealhulgas esimene lause. Kattumine oleks muidu nii pika artikli puhul üllatavalt suur, kuid EstSum on pidanud oluliseks siin artikli lõppu lisatud väljavõtteid seadustest.
Artikkel 18- 41 sõna kattuvad, sealhulgas esimene lause. Mina olen jätnud mitmed laused pigem sisse lisamata, sest palju oli kasutatud „tema“ vormi ja siis oleks pidanud ka eelneva isikut tutvustava lause sisse panema, kuid oluline oli ka pikkuse limiidist kinni pidada.
Artikkel 19- 47 sõna kattuvad, sealhulgas esimene lause.
Artikkel 20-13 sõna (esimene lause) kattuvad, kuigi tegelt päris pikk artikkel.
Arikkel 21- 53 sõna kattuvad, sealhulgas esimene lause, kusjuures tulemuseks minu kokkuvõtte suhtes tuleb 100%, kuid tegelikult on EstSumi variandis üks lause juures.
Artikkel 22- 52 sõna kattuvad, sealhulgas esimene lause, kusjuures sama asi, mis eelmisel juhul, isegi 2 lauset on juures
Artikkel 23- 80 sõna kattuvad, sealhulgas esimene lause.
Artikkel 24- 32 sõna kattuvad, sealhulgas esimene lause.
Artikkel 25- 39 sõna kattuvad, sealhulgas esimene lause.
Artikkel 26- 68 sõna kattuvad, sealhulgas esimene lause, ühe lause algus tundub kadunud olevat.
Artikkel 27- 67 sõna kattuvad, sealhulgas esimene lause.
Artikkel 28- 21 sõna kattuvad, ainult esimene lause. Taas tundub, et ühe lause algus on kadunud.
Artikkel 29- kattumine 100%.
Artikkel 30- 15 sõna kattuvad, ainult esimene lause.
Artikkel 31- 12 sõna kattuvad, ainult esimene lause.
Artikkel 32- 45 sõna kattuvad, sealhulgas esimene lause. Jälle on ühe lause algus justkui kadunud. Vaatasin originaalartiklit, seal on selle lause kaks osa eri ridadel, äkki on asi selles?
Artikkel 33- 39 sõna kattuvad, sealhulgas esimene lause.
Artikkel 34- 20 sõna kattuvad, ainult esimene lause. Lühikesed kokkuvõtted, aga kattub nii vähe.
Artikkel 35- 71 sõna kattuvad, sealhulgas esimene lause. Korrapära nagu ei ole.
Artikkel 36- 50 sõna kattuvad, esimene ja viimane lause.
Artikkel 37- 71 sõna kattuvad, sealhulgas esimene lause. EstSumi oma on oluliselt pikem.
Artikkel 38- 20 sõna kattuvad, ainult esimene lause- see on 1 lause kahest kummaski kokkuvõttes
Artikkel 39- 68 sõna kattuvad, sealhulgas esimene lause, EstSum on pikem ja seetõttu tundub kattumine suur

Artikkel 40- 25 sõna kattuvad, ainult esimene lause- 1 lause kahest lausest kummaski kokkuvõttes
Artikkel 41- 24 sõna kattuvad, ainult esimene lause. Mõlemad kokkuvõtted jätavad mulje, et on sisult erinevad, kui mitte arvata sisse esimest lauset.
Artikkel 42- 30 sõna kattuvad, sealhulgas esimene lause. Me EstSumiga peame tähtsaks erinevaid suundi tekstis.
Artikkel 43- 36 sõna kattuvad, sealhulgas esimene lause.
Artikkel 44- kattumine 100%, koosnevad kahest lausest kumbki.
Artikkel 45- 33 sõna kattuvad, sealhulgas esimene lause/lõik, ülejäänus lähevad arvamused lahku
Artikkel 46- 52 sõna kattuvad, sealhulgas esimene ja viimane lause
Artikkel 47- 41 sõna kattuvad, kattumine tuleb 100% kuigi tegelikult on EstSum pikem, sealhulgas esimene lause
Artikkel 48- 100% kattumine
Artikkel 49- 25 sõna kattuvad, ainult esimene lause.
Artikkel 50- 24 sõna kattuvad, esimene lõik, sealhulgas esimene lause. Huvitav on see, et vb paar aastat tagasi oleks tulemused olnud teised, sest artikkel räägib eestlaste ja venelaste erinevustest linna turistiatraktsioonide osas vms. Mina lähtusin pronksõduri jne päevakajalisusest. EstSum nimetab seda ainult esimeses lauses, sest seal on see sees.

Lisa 4. ROUGE-L väljundid

Artikkel 1:

ROUGE-L Average_R: 0.67708 (95%-conf.int. 0.67708 - 0.67708)
ROUGE-L Average_P: 0.65000 (95%-conf.int. 0.65000 - 0.65000)
ROUGE-L Average_F: 0.66326 (95%-conf.int. 0.66326 - 0.66326)

Artikkel 2:

ROUGE-L Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_P: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_F: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

Artikkel 3:

ROUGE-L Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_P: 0.76522 (95%-conf.int. 0.76522 - 0.76522)
ROUGE-L Average_F: 0.86700 (95%-conf.int. 0.86700 - 0.86700)

Artikkel 4:

ROUGE-L Average_R: 0.52632 (95%-conf.int. 0.52632 - 0.52632)
ROUGE-L Average_P: 0.43478 (95%-conf.int. 0.43478 - 0.43478)
ROUGE-L Average_F: 0.47619 (95%-conf.int. 0.47619 - 0.47619)

Artikkel 5:

ROUGE-L Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_P: 0.81395 (95%-conf.int. 0.81395 - 0.81395)
ROUGE-L Average_F: 0.89743 (95%-conf.int. 0.89743 - 0.89743)

Artikkel 6:

ROUGE-L Average_R: 0.65476 (95%-conf.int. 0.65476 - 0.65476)
ROUGE-L Average_P: 0.44355 (95%-conf.int. 0.44355 - 0.44355)
ROUGE-L Average_F: 0.52885 (95%-conf.int. 0.52885 - 0.52885)

Artikkel 7:

ROUGE-L Average_R: 0.52055 (95%-conf.int. 0.52055 - 0.52055)
ROUGE-L Average_P: 0.51351 (95%-conf.int. 0.51351 - 0.51351)
ROUGE-L Average_F: 0.51701 (95%-conf.int. 0.51701 - 0.51701)

Artikkel 8:

ROUGE-L Average_R: 0.61475 (95%-conf.int. 0.61475 - 0.61475)
ROUGE-L Average_P: 0.57252 (95%-conf.int. 0.57252 - 0.57252)
ROUGE-L Average_F: 0.59288 (95%-conf.int. 0.59288 - 0.59288)

Artikkel 9:

ROUGE-L Average_R: 0.47143 (95%-conf.int. 0.47143 - 0.47143)
ROUGE-L Average_P: 0.38372 (95%-conf.int. 0.38372 - 0.38372)
ROUGE-L Average_F: 0.42308 (95%-conf.int. 0.42308 - 0.42308)

Artikkel 10:

ROUGE-L Average_R: 0.90588 (95%-conf.int. 0.90588 - 0.90588)
ROUGE-L Average_P: 0.70000 (95%-conf.int. 0.70000 - 0.70000)
ROUGE-L Average_F: 0.78974 (95%-conf.int. 0.78974 - 0.78974)

Artikkel 11:

ROUGE-L Average_R: 0.77228 (95%-conf.int. 0.77228 - 0.77228)
ROUGE-L Average_P: 0.65000 (95%-conf.int. 0.65000 - 0.65000)
ROUGE-L Average_F: 0.70588 (95%-conf.int. 0.70588 - 0.70588)

Artikkel 12:

ROUGE-L Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_P: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_F: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

Artikkel 13:

ROUGE-L Average_R: 0.64957 (95%-conf.int. 0.64957 - 0.64957)
ROUGE-L Average_P: 0.80851 (95%-conf.int. 0.80851 - 0.80851)
ROUGE-L Average_F: 0.72038 (95%-conf.int. 0.72038 - 0.72038)

Artikkel 14:

ROUGE-L Average_R: 0.50000 (95%-conf.int. 0.50000 - 0.50000)
ROUGE-L Average_P: 0.37313 (95%-conf.int. 0.37313 - 0.37313)
ROUGE-L Average_F: 0.42735 (95%-conf.int. 0.42735 - 0.42735)

Artikkel 15:

ROUGE-L Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_P: 0.66667 (95%-conf.int. 0.66667 - 0.66667)
ROUGE-L Average_F: 0.80000 (95%-conf.int. 0.80000 - 0.80000)

Artikkel 16:

ROUGE-L Average_R: 0.40816 (95%-conf.int. 0.40816 - 0.40816)
ROUGE-L Average_P: 0.30303 (95%-conf.int. 0.30303 - 0.30303)
ROUGE-L Average_F: 0.34782 (95%-conf.int. 0.34782 - 0.34782)

Artikkel 17:

ROUGE-L Average_R: 0.73158 (95%-conf.int. 0.73158 - 0.73158)
ROUGE-L Average_P: 0.62332 (95%-conf.int. 0.62332 - 0.62332)
ROUGE-L Average_F: 0.67312 (95%-conf.int. 0.67312 - 0.67312)

Artikkel 18:

ROUGE-L Average_R: 0.63248 (95%-conf.int. 0.63248 - 0.63248)
ROUGE-L Average_P: 0.49007 (95%-conf.int. 0.49007 - 0.49007)
ROUGE-L Average_F: 0.55224 (95%-conf.int. 0.55224 - 0.55224)

Artikkel 19:

ROUGE-L Average_R: 0.70408 (95%-conf.int. 0.70408 - 0.70408)
ROUGE-L Average_P: 0.66990 (95%-conf.int. 0.66990 - 0.66990)
ROUGE-L Average_F: 0.68656 (95%-conf.int. 0.68656 - 0.68656)

Artikkel 20:

ROUGE-L Average_R: 0.25974 (95%-conf.int. 0.25974 - 0.25974)
ROUGE-L Average_P: 0.20833 (95%-conf.int. 0.20833 - 0.20833)
ROUGE-L Average_F: 0.23121 (95%-conf.int. 0.23121 - 0.23121)

Artikkel 21:

ROUGE-L Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_P: 0.88060 (95%-conf.int. 0.88060 - 0.88060)
ROUGE-L Average_F: 0.93651 (95%-conf.int. 0.93651 - 0.93651)

Artikkel 22:

ROUGE-L Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_P: 0.60784 (95%-conf.int. 0.60784 - 0.60784)
ROUGE-L Average_F: 0.75610 (95%-conf.int. 0.75610 - 0.75610)

Artikkel 23:

ROUGE-L Average_R: 0.67586 (95%-conf.int. 0.67586 - 0.67586)
ROUGE-L Average_P: 0.65772 (95%-conf.int. 0.65772 - 0.65772)
ROUGE-L Average_F: 0.66667 (95%-conf.int. 0.66667 - 0.66667)

Artikkel 24:

ROUGE-L Average_R: 0.62857 (95%-conf.int. 0.62857 - 0.62857)
ROUGE-L Average_P: 0.57143 (95%-conf.int. 0.57143 - 0.57143)
ROUGE-L Average_F: 0.59864 (95%-conf.int. 0.59864 - 0.59864)

Artikkel 25:

ROUGE-L Average_R: 0.73846 (95%-conf.int. 0.73846 - 0.73846)
ROUGE-L Average_P: 0.55814 (95%-conf.int. 0.55814 - 0.55814)
ROUGE-L Average_F: 0.63576 (95%-conf.int. 0.63576 - 0.63576)

Artikkel 26:

ROUGE-L Average_R: 0.75510 (95%-conf.int. 0.75510 - 0.75510)
ROUGE-L Average_P: 0.69159 (95%-conf.int. 0.69159 - 0.69159)
ROUGE-L Average_F: 0.72195 (95%-conf.int. 0.72195 - 0.72195)

Artikkel 27:

ROUGE-L Average_R: 0.79612 (95%-conf.int. 0.79612 - 0.79612)
ROUGE-L Average_P: 0.70690 (95%-conf.int. 0.70690 - 0.70690)
ROUGE-L Average_F: 0.74886 (95%-conf.int. 0.74886 - 0.74886)

Artikkel 28:

ROUGE-L Average_R: 0.43662 (95%-conf.int. 0.43662 - 0.43662)
ROUGE-L Average_P: 0.40789 (95%-conf.int. 0.40789 - 0.40789)
ROUGE-L Average_F: 0.42177 (95%-conf.int. 0.42177 - 0.42177)

Artikkel 29:

ROUGE-L Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_P: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_F: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

Artikkel 30:

ROUGE-L Average_R: 0.30380 (95%-conf.int. 0.30380 - 0.30380)
ROUGE-L Average_P: 0.30769 (95%-conf.int. 0.30769 - 0.30769)
ROUGE-L Average_F: 0.30573 (95%-conf.int. 0.30573 - 0.30573)

Artikkel 31:

ROUGE-L Average_R: 0.36957 (95%-conf.int. 0.36957 - 0.36957)
ROUGE-L Average_P: 0.31481 (95%-conf.int. 0.31481 - 0.31481)
ROUGE-L Average_F: 0.34000 (95%-conf.int. 0.34000 - 0.34000)

Artikkel 32:

ROUGE-L Average_R: 0.79104 (95%-conf.int. 0.79104 - 0.79104)
ROUGE-L Average_P: 0.91379 (95%-conf.int. 0.91379 - 0.91379)
ROUGE-L Average_F: 0.84800 (95%-conf.int. 0.84800 - 0.84800)

Artikkel 33:

ROUGE-L Average_R: 0.56322 (95%-conf.int. 0.56322 - 0.56322)
ROUGE-L Average_P: 0.55056 (95%-conf.int. 0.55056 - 0.55056)
ROUGE-L Average_F: 0.55682 (95%-conf.int. 0.55682 - 0.55682)

Artikkel 34:

ROUGE-L Average_R: 0.55814 (95%-conf.int. 0.55814 - 0.55814)
ROUGE-L Average_P: 0.42857 (95%-conf.int. 0.42857 - 0.42857)
ROUGE-L Average_F: 0.48485 (95%-conf.int. 0.48485 - 0.48485)

Artikkel 35:

ROUGE-L Average_R: 0.78992 (95%-conf.int. 0.78992 - 0.78992)
ROUGE-L Average_P: 0.65278 (95%-conf.int. 0.65278 - 0.65278)
ROUGE-L Average_F: 0.71483 (95%-conf.int. 0.71483 - 0.71483)

Artikkel 36:

ROUGE-L Average_R: 0.63208 (95%-conf.int. 0.63208 - 0.63208)
ROUGE-L Average_P: 0.54472 (95%-conf.int. 0.54472 - 0.54472)
ROUGE-L Average_F: 0.58516 (95%-conf.int. 0.58516 - 0.58516)

Artikkel 37:

ROUGE-L Average_R: 0.82353 (95%-conf.int. 0.82353 - 0.82353)
ROUGE-L Average_P: 0.59574 (95%-conf.int. 0.59574 - 0.59574)
ROUGE-L Average_F: 0.69136 (95%-conf.int. 0.69136 - 0.69136)

Artikkel 38:

ROUGE-L Average_R: 0.54717 (95%-conf.int. 0.54717 - 0.54717)
ROUGE-L Average_P: 0.56863 (95%-conf.int. 0.56863 - 0.56863)
ROUGE-L Average_F: 0.55769 (95%-conf.int. 0.55769 - 0.55769)

Artikkel 39:

ROUGE-L Average_R: 0.60390 (95%-conf.int. 0.60390 - 0.60390)
ROUGE-L Average_P: 0.51667 (95%-conf.int. 0.51667 - 0.51667)
ROUGE-L Average_F: 0.55689 (95%-conf.int. 0.55689 - 0.55689)

Artikkel 40:

ROUGE-L Average_R: 0.66667 (95%-conf.int. 0.66667 - 0.66667)
ROUGE-L Average_P: 0.69231 (95%-conf.int. 0.69231 - 0.69231)
ROUGE-L Average_F: 0.67925 (95%-conf.int. 0.67925 - 0.67925)

Artikkel 41:

ROUGE-L Average_R: 0.48333 (95%-conf.int. 0.48333 - 0.48333)
ROUGE-L Average_P: 0.38667 (95%-conf.int. 0.38667 - 0.38667)
ROUGE-L Average_F: 0.42963 (95%-conf.int. 0.42963 - 0.42963)

Artikkel 42:

ROUGE-L Average_R: 0.54795 (95%-conf.int. 0.54795 - 0.54795)
ROUGE-L Average_P: 0.48780 (95%-conf.int. 0.48780 - 0.48780)
ROUGE-L Average_F: 0.51613 (95%-conf.int. 0.51613 - 0.51613)

Artikkel 43:

ROUGE-L Average_R: 0.71717 (95%-conf.int. 0.71717 - 0.71717)
ROUGE-L Average_P: 0.68269 (95%-conf.int. 0.68269 - 0.68269)
ROUGE-L Average_F: 0.69951 (95%-conf.int. 0.69951 - 0.69951)

Artikkel 44:

ROUGE-L Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_P: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_F: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

Artikkel 45:

ROUGE-L Average_R: 0.75000 (95%-conf.int. 0.75000 - 0.75000)
ROUGE-L Average_P: 0.60811 (95%-conf.int. 0.60811 - 0.60811)
ROUGE-L Average_F: 0.67164 (95%-conf.int. 0.67164 - 0.67164)

Artikkel 46:

ROUGE-L Average_R: 0.71910 (95%-conf.int. 0.71910 - 0.71910)
ROUGE-L Average_P: 0.59813 (95%-conf.int. 0.59813 - 0.59813)
ROUGE-L Average_F: 0.65306 (95%-conf.int. 0.65306 - 0.65306)

Artikkel 47:

ROUGE-L Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_P: 0.52000 (95%-conf.int. 0.52000 - 0.52000)
ROUGE-L Average_F: 0.68421 (95%-conf.int. 0.68421 - 0.68421)

Artikkel 48:

ROUGE-L Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_P: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
ROUGE-L Average_F: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

Artikkel 49:

ROUGE-L Average_R: 0.59649 (95%-conf.int. 0.59649 - 0.59649)
ROUGE-L Average_P: 0.55738 (95%-conf.int. 0.55738 - 0.55738)
ROUGE-L Average_F: 0.57627 (95%-conf.int. 0.57627 - 0.57627)

Artikkel 50:

ROUGE-L Average_R: 0.51163 (95%-conf.int. 0.51163 - 0.51163)
ROUGE-L Average_P: 0.39640 (95%-conf.int. 0.39640 - 0.39640)
ROUGE-L Average_F: 0.44670 (95%-conf.int. 0.44670 – 0.44670)

Lisa 5. ROUGE-N, ROUGE-W ja ROUGE-S väljundid saagise jaoks

Artikkel 1:

1 ROUGE-1 Average_R: 0.67708 (95%-conf.int. 0.67708 - 0.67708)
1 ROUGE-2 Average_R: 0.63158 (95%-conf.int. 0.63158 - 0.63158)
1 ROUGE-3 Average_R: 0.61702 (95%-conf.int. 0.61702 - 0.61702)
1 ROUGE-4 Average_R: 0.60215 (95%-conf.int. 0.60215 - 0.60215)
1 ROUGE-W-1.2 Average_R: 0.30667 (95%-conf.int. 0.30667 - 0.30667)
1 ROUGE-S* Average_R: 0.45000 (95%-conf.int. 0.45000 – 0.45000)

Artikkel 2:

1 ROUGE-1 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-2 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-3 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-4 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-W-1.2 Average_R: 0.43361 (95%-conf.int. 0.43361 - 0.43361)
1 ROUGE-S* Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

Artikkel 3:

1 ROUGE-1 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-2 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-3 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-4 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-W-1.2 Average_R: 0.42751 (95%-conf.int. 0.42751 - 0.42751)
1 ROUGE-S* Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

Artikkel 4:

1 ROUGE-1 Average_R: 0.52632 (95%-conf.int. 0.52632 - 0.52632)
1 ROUGE-2 Average_R: 0.48649 (95%-conf.int. 0.48649 - 0.48649)
1 ROUGE-3 Average_R: 0.47222 (95%-conf.int. 0.47222 - 0.47222)
1 ROUGE-4 Average_R: 0.45714 (95%-conf.int. 0.45714 - 0.45714)
1 ROUGE-W-1.2 Average_R: 0.28421 (95%-conf.int. 0.28421 - 0.28421)
1 ROUGE-S* Average_R: 0.27027 (95%-conf.int. 0.27027 - 0.27027)

Artikkel 5:

1 ROUGE-1 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-2 Average_R: 0.98551 (95%-conf.int. 0.98551 - 0.98551)
1 ROUGE-3 Average_R: 0.97059 (95%-conf.int. 0.97059 - 0.97059)
1 ROUGE-4 Average_R: 0.95522 (95%-conf.int. 0.95522 - 0.95522)
1 ROUGE-W-1.2 Average_R: 0.43696 (95%-conf.int. 0.43696 - 0.43696)
1 ROUGE-S* Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

Artikkel 6:

1 ROUGE-1 Average_R: 0.65476 (95%-conf.int. 0.65476 - 0.65476)
1 ROUGE-2 Average_R: 0.61446 (95%-conf.int. 0.61446 - 0.61446)
1 ROUGE-3 Average_R: 0.59756 (95%-conf.int. 0.59756 - 0.59756)
1 ROUGE-4 Average_R: 0.59259 (95%-conf.int. 0.59259 - 0.59259)
1 ROUGE-W-1.2 Average_R: 0.28753 (95%-conf.int. 0.28753 - 0.28753)
1 ROUGE-S* Average_R: 0.42628 (95%-conf.int. 0.42628 - 0.42628)

Artikkel 7:

1 ROUGE-1 Average_R: 0.53425 (95%-conf.int. 0.53425 - 0.53425)
1 ROUGE-2 Average_R: 0.47222 (95%-conf.int. 0.47222 - 0.47222)
1 ROUGE-3 Average_R: 0.43662 (95%-conf.int. 0.43662 - 0.43662)

1 ROUGE-4 Average_R: 0.41429 (95%-conf.int. 0.41429 - 0.41429)
1 ROUGE-W-1.2 Average_R: 0.22663 (95%-conf.int. 0.22663 - 0.22663)
1 ROUGE-S* Average_R: 0.28311 (95%-conf.int. 0.28311 - 0.28311)

Artikkel 8:

1 ROUGE-1 Average_R: 0.62295 (95%-conf.int. 0.62295 - 0.62295)
1 ROUGE-2 Average_R: 0.50413 (95%-conf.int. 0.50413 - 0.50413)
1 ROUGE-3 Average_R: 0.46667 (95%-conf.int. 0.46667 - 0.46667)
1 ROUGE-4 Average_R: 0.43697 (95%-conf.int. 0.43697 - 0.43697)
1 ROUGE-W-1.2 Average_R: 0.24032 (95%-conf.int. 0.24032 - 0.24032)
1 ROUGE-S* Average_R: 0.37949 (95%-conf.int. 0.37949 - 0.37949)

Artikkel 9:

1 ROUGE-1 Average_R: 0.52857 (95%-conf.int. 0.52857 - 0.52857)
1 ROUGE-2 Average_R: 0.37681 (95%-conf.int. 0.37681 - 0.37681)
1 ROUGE-3 Average_R: 0.32353 (95%-conf.int. 0.32353 - 0.32353)
1 ROUGE-4 Average_R: 0.31343 (95%-conf.int. 0.31343 - 0.31343)
1 ROUGE-W-1.2 Average_R: 0.20912 (95%-conf.int. 0.20912 - 0.20912)
1 ROUGE-S* Average_R: 0.26874 (95%-conf.int. 0.26874 - 0.26874)

Artikkel 10:

1 ROUGE-1 Average_R: 0.90588 (95%-conf.int. 0.90588 - 0.90588)
1 ROUGE-2 Average_R: 0.88095 (95%-conf.int. 0.88095 - 0.88095)
1 ROUGE-3 Average_R: 0.85542 (95%-conf.int. 0.85542 - 0.85542)
1 ROUGE-4 Average_R: 0.84146 (95%-conf.int. 0.84146 - 0.84146)
1 ROUGE-W-1.2 Average_R: 0.40054 (95%-conf.int. 0.40054 - 0.40054)
1 ROUGE-S* Average_R: 0.80756 (95%-conf.int. 0.80756 - 0.80756)

Atikkel 11:

1 ROUGE-1 Average_R: 0.74752 (95%-conf.int. 0.74752 - 0.74752)
1 ROUGE-2 Average_R: 0.68657 (95%-conf.int. 0.68657 - 0.68657)
1 ROUGE-3 Average_R: 0.67000 (95%-conf.int. 0.67000 - 0.67000)
1 ROUGE-4 Average_R: 0.65829 (95%-conf.int. 0.65829 - 0.65829)
1 ROUGE-W-1.2 Average_R: 0.28631 (95%-conf.int. 0.28631 - 0.28631)
1 ROUGE-S* Average_R: 0.55470 (95%-conf.int. 0.55470 - 0.55470)

Artikkel 12:

1 ROUGE-1 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-2 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-3 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-4 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-W-1.2 Average_R: 0.45550 (95%-conf.int. 0.45550 - 0.45550)
1 ROUGE-S* Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

Artikkel 13:

1 ROUGE-1 Average_R: 0.67521 (95%-conf.int. 0.67521 - 0.67521)
1 ROUGE-2 Average_R: 0.62931 (95%-conf.int. 0.62931 - 0.62931)
1 ROUGE-3 Average_R: 0.61739 (95%-conf.int. 0.61739 - 0.61739)
1 ROUGE-4 Average_R: 0.59649 (95%-conf.int. 0.59649 - 0.59649)
1 ROUGE-W-1.2 Average_R: 0.25483 (95%-conf.int. 0.25483 - 0.25483)
1 ROUGE-S* Average_R: 0.44489 (95%-conf.int. 0.44489 – 0.44489)

Artikkel 14:

1 ROUGE-1 Average_R: 0.52000 (95%-conf.int. 0.52000 - 0.52000)
1 ROUGE-2 Average_R: 0.44898 (95%-conf.int. 0.44898 - 0.44898)

1 ROUGE-3 Average_R: 0.43750 (95%-conf.int. 0.43750 - 0.43750)
1 ROUGE-4 Average_R: 0.42553 (95%-conf.int. 0.42553 - 0.42553)
1 ROUGE-W-1.2 Average_R: 0.25077 (95%-conf.int. 0.25077 - 0.25077)
1 ROUGE-S* Average_R: 0.26286 (95%-conf.int. 0.26286 - 0.26286)

Artikkel 15:

1 ROUGE-1 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-2 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-3 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-4 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-W-1.2 Average_R: 0.48425 (95%-conf.int. 0.48425 - 0.48425)
1 ROUGE-S* Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

Artikkel 16:

1 ROUGE-1 Average_R: 0.42857 (95%-conf.int. 0.42857 - 0.42857)
1 ROUGE-2 Average_R: 0.37500 (95%-conf.int. 0.37500 - 0.37500)
1 ROUGE-3 Average_R: 0.36170 (95%-conf.int. 0.36170 - 0.36170)
1 ROUGE-4 Average_R: 0.34783 (95%-conf.int. 0.34783 - 0.34783)
1 ROUGE-W-1.2 Average_R: 0.20821 (95%-conf.int. 0.20821 - 0.20821)
1 ROUGE-S* Average_R: 0.17857 (95%-conf.int. 0.17857 - 0.17857)

Artikkel 17:

1 ROUGE-1 Average_R: 0.73684 (95%-conf.int. 0.73684 - 0.73684)
1 ROUGE-2 Average_R: 0.67725 (95%-conf.int. 0.67725 - 0.67725)
1 ROUGE-3 Average_R: 0.64894 (95%-conf.int. 0.64894 - 0.64894)
1 ROUGE-4 Average_R: 0.63102 (95%-conf.int. 0.63102 - 0.63102)
1 ROUGE-W-1.2 Average_R: 0.27020 (95%-conf.int. 0.27020 - 0.27020)
1 ROUGE-S* Average_R: 0.53790 (95%-conf.int. 0.53790 - 0.53790)

Artikkel 18:

1 ROUGE-1 Average_R: 0.64957 (95%-conf.int. 0.64957 - 0.64957)
1 ROUGE-2 Average_R: 0.52586 (95%-conf.int. 0.52586 - 0.52586)
1 ROUGE-3 Average_R: 0.50435 (95%-conf.int. 0.50435 - 0.50435)
1 ROUGE-4 Average_R: 0.48246 (95%-conf.int. 0.48246 - 0.48246)
1 ROUGE-W-1.2 Average_R: 0.24903 (95%-conf.int. 0.24903 - 0.24903)
1 ROUGE-S* Average_R: 0.41689 (95%-conf.int. 0.41689 – 0.41689)

Artikkel 19:

1 ROUGE-1 Average_R: 0.71429 (95%-conf.int. 0.71429 - 0.71429)
1 ROUGE-2 Average_R: 0.63918 (95%-conf.int. 0.63918 - 0.63918)
1 ROUGE-3 Average_R: 0.60417 (95%-conf.int. 0.60417 - 0.60417)
1 ROUGE-4 Average_R: 0.57895 (95%-conf.int. 0.57895 - 0.57895)
1 ROUGE-W-1.2 Average_R: 0.29675 (95%-conf.int. 0.29675 - 0.29675)
1 ROUGE-S* Average_R: 0.46560 (95%-conf.int. 0.46560 - 0.46560)

Artikkel 20:

1 ROUGE-1 Average_R: 0.27273 (95%-conf.int. 0.27273 - 0.27273)
1 ROUGE-2 Average_R: 0.19737 (95%-conf.int. 0.19737 - 0.19737)
1 ROUGE-3 Average_R: 0.17333 (95%-conf.int. 0.17333 - 0.17333)
1 ROUGE-4 Average_R: 0.16216 (95%-conf.int. 0.16216 - 0.16216)
1 ROUGE-W-1.2 Average_R: 0.11763 (95%-conf.int. 0.11763 - 0.11763)
1 ROUGE-S* Average_R: 0.07382 (95%-conf.int. 0.07382 - 0.07382)

Artikkel 21:

1 ROUGE-1 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

1 ROUGE-2 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-3 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-4 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-W-1.2 Average_R: 0.45200 (95%-conf.int. 0.45200 - 0.45200)
1 ROUGE-S* Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

Artikkel 22:

1 ROUGE-1 Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)
1 ROUGE-2 Average_R: 0.98361 (95%-conf.int. 0.98361 - 0.98361)
1 ROUGE-3 Average_R: 0.96667 (95%-conf.int. 0.96667 - 0.96667)
1 ROUGE-4 Average_R: 0.94915 (95%-conf.int. 0.94915 - 0.94915)
1 ROUGE-W-1.2 Average_R: 0.45413 (95%-conf.int. 0.45413 - 0.45413)
1 ROUGE-S* Average_R: 1.00000 (95%-conf.int. 1.00000 - 1.00000)

Artikkel 23:

1 ROUGE-1 Average_R: 0.67586 (95%-conf.int. 0.67586 - 0.67586)
1 ROUGE-2 Average_R: 0.61111 (95%-conf.int. 0.61111 - 0.61111)
1 ROUGE-3 Average_R: 0.60140 (95%-conf.int. 0.60140 - 0.60140)
1 ROUGE-4 Average_R: 0.59155 (95%-conf.int. 0.59155 - 0.59155)
1 ROUGE-W-1.2 Average_R: 0.27028 (95%-conf.int. 0.27028 - 0.27028)
1 ROUGE-S* Average_R: 0.45038 (95%-conf.int. 0.45038 - 0.45038)

Artikkel 24:

1 ROUGE-1 Average_R: 0.62857 (95%-conf.int. 0.62857 - 0.62857)
1 ROUGE-2 Average_R: 0.53623 (95%-conf.int. 0.53623 - 0.53623)
1 ROUGE-3 Average_R: 0.50000 (95%-conf.int. 0.50000 - 0.50000)
1 ROUGE-4 Average_R: 0.49254 (95%-conf.int. 0.49254 - 0.49254)
1 ROUGE-W-1.2 Average_R: 0.28792 (95%-conf.int. 0.28792 - 0.28792)
1 ROUGE-S* Average_R: 0.39172 (95%-conf.int. 0.39172 - 0.39172)

Artikkel 25:

1 ROUGE-1 Average_R: 0.73846 (95%-conf.int. 0.73846 - 0.73846)
1 ROUGE-2 Average_R: 0.68750 (95%-conf.int. 0.68750 - 0.68750)
1 ROUGE-3 Average_R: 0.65079 (95%-conf.int. 0.65079 - 0.65079)
1 ROUGE-4 Average_R: 0.62903 (95%-conf.int. 0.62903 - 0.62903)
1 ROUGE-W-1.2 Average_R: 0.33540 (95%-conf.int. 0.33540 - 0.33540)
1 ROUGE-S* Average_R: 0.54712 (95%-conf.int. 0.54712 - 0.54712)