

Skaitmenizuotų tekstų santraukų rengimo sistemos: panašumai, skirtumai ir taikymas lietuvių kalbai

Jurgita Lasytė, Bronius Tamulynas

Anotacija. Informacinei atskirčiai mažinti bei žmogaus laikui taupyti yra sukurti automatiniai santraukų rengimo instrumentai, kurių esmę sudaro specialiosios kompiuterinės santraukų komponavimo programos. Automatinis teksto santraukos rengimas yra vienas iš svarbių kompiuterinės lingvistikos tyrimo objektų, siekiant gauti prasmingesnę ir kokybiškesnę rezultatą. Apie šias technologijas Lietuvoje kol kas turime mažai informacijos, nors užsienyje kompiuterinės lingvistikos specialistai jas kuria ir nuolat tobulina. Šiuo metu didelė dalis santraukų automatinio rengimo sistemų (SARS) neapsiriboja vien skaitmeniniu tekstu, bet siekia apimti įvairialypę informaciją, cirkuliuojančią pasauliniame interneto tinkle – tekstą, vaizdinę ir garsinę medžiagą, paveikslėlius ir pan. Tyrimui buvo atrinktos automatinės *teksto* santraukų rengimo sistemos ir technologijos, kurios yra daugiau artimos ir reikšmingos kompiuterinės lingvistikos sričiai. Nustatyta jų praktinė nauda ir galimybės jas taikyti lietuvių kalbos tekstams. Atrinkę ir susipažinę su keliomis populiariausiomis SARS technologijomis (angl. *summarizers*), atskleisime skaitmeninio teksto santraukų rengimo skirtumus, atsirandančius kuriant santraukas įvairių kalbų ar žanrų tekstams su skirtingo tipo sistemomis.

Reikšminiai žodžiai: kompiuterinė lingvistika, automatinis santraukų rengimas, teksto glaudinimas, santraukos įskaitomumas, santraukos prasmingumas.

Įvadas

Pastaraisiais metais ypač padaugėjo informacijos, pasiekiančios žmogų per internetą. Informacinei atskirčiai mažinti ir žinių suvokimo efektyvumui padidinti buvo sukurti automatiniai santraukų rengimo instrumentai. Jų esmę sudaro specialiosios kompiuterinės santraukų komponavimo programos. Pastaruoju metu SARS tampa ypač svarbiu kompiuterinės lingvistikos tyrimo objektu. Apie panašias sistemas Lietuvoje kol kas turime mažai informacijos, nors užsienyje jos intensyviai kuriamos ir tobulinamos. Kadangi mūsų šalyje kol kas nėra sukurta panašaus tipo sistemų, todėl yra svarbu susipažinti su jų taikymo praktine nauda. Dabartinės automatinės santraukų rengimo technologijos neapsiriboja vien skaitmeniniais teksta, bet siekia apimti įvairialypę informaciją, cirkuliuojančią pasauliniame interneto tinkle – tekstą, vaizdinę ir garsinę medžiagą, paveikslėlius ir pan. Šiame straipsnyje aprašomos ir tiriamos automatinės *teksto* santraukų rengimo technologijos ir sistemos, kurios labiausiai artimos ir reikšmingos kompiuterinės lingvistikos sričiai. Tyrimui buvo atrinktos 8 plačiau žinomos ir dažniau naudojamos SARS technologijos ir sistemos (angl. *summarizers*) bei pabandyta jas taikyti lietuvių kalbos tekstams. Atskleisti skaitmenizuoto teksto santraukų sistemų skirtumai, kurie atsiranda kuriant santraukas skirtingų kalbų ar žanrų tekstams.

Skaitmenizuoto teksto analizės ir apdorojimo metodai santraukoms rengti

Pirmieji įrašai apie kompiuterizuotas tekstų santraukos technologijas ir sistemas randami XX amžiaus šeštojo ir septintojo dešimtmečių moksliniuose darbuose bei publikacijose. Jas aprašė Luhn ir Edmundson (Mitkov, 2005). Po keleto dešimtmečių pertraukos, patobulėjus kalbos apdorojimo technologijoms, išsiplėtus kompiuterio atminčiai ir padidėjus kompiuterio spartai bei atsiradus poreikiui

skaitmenizuoti didelius tekstų rinkinius (tekstynus, internetinės informacijos srautus ir pan.), susidomėjimas SARS technologijomis (Mitkov, 2005) vėl atgijo. Vienas iš aktualiausių ir daugiausiai ginčų keliantis klausimas yra pačio žodžio *santrauka* samprata. Mokslininkai, dirbantys bei rašantys apie SARS, pačią santrauką dažnai apibūdina skirtingai. Paprastai naudojama bendroji santraukos samprata:

„santrauka – tekstas, kuris gaunamas iš vieno ar daugiau tekstų, glaustai pateikiantis svarbiausią informaciją ir yra ne ilgesnis kaip pusė pradinio originalaus teksto“ (Mitkov, 2005).

SARS kūrimo tikslai nuo XX amžiaus 6-ojo dešimtmečio mažai pasikeitė – tebesiekama pateikti ilgų tekstų trumapas, bet informatyvias santraukas. Gal būt išskirtina šiuolaikinių SARS technologijų savybė reikėtų laikyti jų gebėjimą apimti ne tik tekstą, bet ir garsinę bei vaizdinę medžiagą, internetinius straipsnius ar portalus (Mitkov, 2005).

Tekstui apdoroti dažniausiai taikomas statistika paremtas modelis – renkami statistiniai duomenys apie žodžių vartojimą, kurie naudojami teksto apdorojimo procese. Santraukoms generuoti ir jas formuoti įprasta išskirti šiuos metodus.

1. Pozicinis arba vietos metodas. Daugumos žanrų tekstai paklūsta būtent tam žanrui priskiriamoms taisyklėms (antraštės, pavadinimai, pastraipos ir pan.). Naudodama šį metodą sistema vadovaujasi principu – kas yra pavadinime, antraštėje ir pirmoje pastraipoje – svarbūs dalykai. Kitose pastraipose esanti informacija laikoma ne tiek svarbia, nes deklaruojama, kad esminė informacija pateikiama teksto pradžioje.

2. Raktinio (reikšminio) žodžio metodas. Įvairiuose žanruose tam tikri žodžiai ir frazės aiškiai parodo svarbumą

(pvz., „reikšmingas“, „šiam darbe“, „svarbiausias“ ir pan.), todėl sakiniai su reikšminiais žodžiais įtraukiami į santrauką.

3. *Žodžio ir frazės dažnio metodas* pagrįstas besikartojančių žodžių dažniu. Kuo dažniau žodis kartojasi, tuo svarbesni sakiniai su tais žodžiais.

4. *Reikšminių žodžių (užklauso) ir dalinio pavadinimo sutapimo metodas* pažymi sakinių pagal reikalingų žodžių kiekį sakinyje. Reikalingi žodžiai yra tokie, kurie yra teksto pavadinimuose, antraštėse, ar sutampa su vartotojo užklausoje įvestais reikšminiais žodžiais.

5. *Riškumo arba leksinio sujungimo metodas*. Žodžiai gali būti sujungiami įvairiais būdais, įskaitant kartojimą, sinonimiką, semantinius ryšius, kurie atsispindi žodynuose. Naudojant šį metodą pasitelkiamas papildomas informacijos šaltinis (žodynas), ir atrenkami atitinkami sakiniai (Mitkov, 2005, pp.585-587; Mani, 2001, pp.13-14, 18).

Praktinis santraukų automatinio rengimo sistemų taikymas

Straipsnyje pateikiami rezultatai, gauti atlikus tyrimus su lietuviškais ir angliškais mokslinio, publicistinio ir grožinio stiliaus tekstais. Kiekvienos santraukų sistemos reikalavimai yra skirtingi, todėl jos geriausiai veikia galingsniuose ir talpesniuose naujesnės kartos kompiuteriuose, kurie naudoja tiek senesnės kartos *Windows XP* ar naujausios kartos *Windows 7 OS* operacines sistemas su interneto ryšiu¹. Tyrimai buvo atliekami su sistemomis, kurios gali apdoroti anglų ir lietuvių kalbos tekstus, išskyrus naujienų portalus bei garso ir vaizdo medžiagą: *Automatic Text Summarizer*; *Copernic Summarizer*; *Intellexer Summarizer*; *Pertinence Summarizer*; *QuickJist Summarizer*; *Shvoong Summarizer*; *Subject Search Summarizer*; *Microsoft Office MS Word 2003 AutoSummarize*. Iš pasirinktųjų – keturios laisvai prieinamos ir nemokamos sistemos. Kitos trys veikia tik internete. Jas galima suskirstyti į dvi grupes: pagal kainą – nemokamos ir mokamos; pagal įdiegimą – įdiegiamosios ir internetinės.

Didesnė dalis kompiuterizuoto teksto santraukų sistemų skiria tam tikrus tekstų žanrus ar stilius, kuriems jos geba rengti santraukas. Žanrų bei stilių neskiria: *Automatic Text Summarizer*, *Copernic Summarizer*, *QuickJist Summarizer*, *Subject Search Summarizer*, *Shvoong Summarizer* ir *MS Office Word 2003 AutoSummarize*. Kitos dvi sistemos (*Intellexer Summarizer*, *Pertinence Summarizer*) skiria tam tikrus teksto žanrus ar stilius: *Intellexer Summarizer* skirtas bendrojo, mokslinio, ekonomikos, politikos ir licencijoms; *Pertinence Summarizer* (internetinis) – be žanro (nežinomam žanrui), chemijos, finansų, teisės, spaudos, telekomunikacijoms, medicinos bei kitoms sritims.² Galima teigti, kad SARS pagal teksto žanrą yra skirstomos į priklausomas arba nepriklausomas nuo teksto žanro.

Kiekvienai SARS yra būdingas atitinkamas suspaudimo (glaudinimo) lygis. Dauguma jų turi vienodus teksto glaustumo laipsnius. Kai kuriose iš jų šis laipsnis vertina-

mas procentais, kitose – žodžių, sakinių ar tiesiog ženklių skaičiumi. Dažniausias ir populiariausias teksto santraukos glaudinimo matas yra procentinis. Procentiniu glaudinimu vadinamas atvejis, kai originaliam pradiniam tekstui yra parengiama norimos procentinės apimties santrauka. Kai kurios santraukų sistemos, leidžia pasirinkti ir įvairesnius glaudinimo lygius, tokius kaip 6 %, 12 %, t. y. tokius, kurių pageidauja vartotojas. Dažniausiai tokie papildomi pasirinkimai naudojami mokamose santraukų sistemose. Procentiniu glaudinimu yra paremtos: *Copernic Summarizer*; *Intellexer Summarizer*; *Pertinence Summarizer*; *QuickJist Summarizer*; *Shvoong Summarizer*; *MS Office Word AutoSummarize*.

Antras pagal populiarumą santraukos glaudinimo būdas yra santraukos apimties nustatymas sakiniiais. Dalis SARS leidžia pasirinkti sakinių skaičių, iš kurių bus sudaryta santrauka. Glaudinimą pagal sakinių kiekį naudoja: *Automatic Text Summarizer*; *Intellexer Summarizer*; *Subject Search Summarizer*; *MS Office Word AutoSummarize*. Vartotojas pats gali pasirinkti, kiek sakinių nori matyti iš tam tikro teksto – tai gali būti nuo 3 sakinių iki 100 ir daugiau, priklausomai nuo teksto apimties. Toks santraukos rengimo būdas yra patogus siekiant gauti trumpesnę ir informatyvesnę santrauką.

Trečiasis būdas, kai vartotojas santraukos apimtį pats pasirenka pagal žodžių skaičių. Paprastai tokioms santraukoms parengti siūlomi 100, 250, 1000 bei kitokie žodžių skaičiaus variantai. Santraukos rengimas pagal žodžių kiekį yra panašus į santraukos parengimą paremtą sakinių kiekiu. Santraukų pagal žodžių kiekį glaudinimo būdą naudoja dvi santraukų rengimo sistemos – *Pertinence Summarizer* ir *MS Office Word AutoSummarize*. *Pertinence Summarizer* šį būdą naudoja, nes spaudžiant tekstą procentiniu teksto glaudinimo metodu automatiškai pateikiamas ir žodžių skaičius, kuris atitinka procentinį žodžių santykį. *MS Office Word AutoSummarize* tiesiog leidžia pasirinkti iš visų pateiktų sakinių mus tenkinantį skaičių. Tačiau, jei santraukos rengimas sakiniiais mūsų netenkina, galima rinktis pagal procentinę teksto apimtį arba žodžių kiekį.

Kiek mažiau žinomas būdas yra santraukos rengimas pagal ženklų kiekį. Jį naudoja tik kelios sistemos. Taip rengiamos santraukos paprastai naudoja nuo 1000 ženklų iki 90 000 ženklų. Galima pasirinkti neribotą ženklų skaičių, tačiau tada jau nebus santrauka, o tik atkartotas tekstas. Tam, kad būtų sukurta santrauka, atitinkanti tam tikrą procentinį santykį, patartina žinoti kiek originalo tekste yra spaudos ženklų. Santraukas glaudina pagal spaudos ženklų skaičių – *Subject Search Summarizer* ir *MS Office Word AutoSummarize*. Didžiausias skirtumas tarp šių sistemų yra tas, kad *MS Office Word AutoSummarize* leidžia atitinkamai pasirinkti 100 arba mažiau žodžių ir 500 arba mažiau žodžių.

Kaip parodė tyrimai, dalis SARS turi du glaudinimo būdus, iš kurių vartotojas gali pasirinkti jam tinkamesnį. Dažniausia jos naudoja procentinį glaudinimo būdą, bet šalia pateikiama glaudinimo sakiniiais ar žodžiais galimybė. Reikia paminėti ir tai, kad *Pertinence Summarizer*, pateikdama santrauką procentais, automatiškai nurodo ir žodžių kiekį parengtoje santraukoje. *MS Office Word AutoSummarizer* nurodo ne tik parengtos santraukos apimtį sakiniiais ir žo-

¹ Interneto ryšys reikalingas toms SARS, kurios pateikia santraukas naršyklės lange.

² Dėl kitų sričių tekstų santraukų rengimo reikia konsultuotis su *Pertinence Summarizer* kūrėjais.

džiais, bet ir pradinio dokumento apimtį sakiniiais ir žodžiais, bei suteikia daugiau santraukos glaudinimo būdų, nei kitos santraukų sistemos.

Kalbos atpažinimas. Santraukų rengimo sistemos taip pat atsižvelgia į originalo teksto kalbą. Beveik visos SARS atpažįsta daugiau nei vieną kalbą. Daugiausiai kalbų atpažįsta ir palaiko *Subject Search Summarizer* – 36 kalbas, tarp jų ir lietuvių kalbą. Pagrindinis jos, kaip ir kitų sistemų, reikalavimas, kad simboliai-hieroglifai (kinų, japonų kalbų rašmenys) būtų užrašomi lotynišku alfabetu, kitaip santrauka nebus daroma, nes ji nesupras parašyto teksto.

Mažiau kalbų atpažįsta *Shvoong Summarizer* – 21 kalbos tekstus. Ji taip pat atpažįsta ir lietuvių kalba rašytus tekstus. Trečia pagal atpažįstamų kalbų skaičių yra *Automatic Text Summarizer* atpažįsta ne tik 13 kalbų tekstus, bet taip pat ir įvairių kitų kalbų tekstus parašytus lotynišku alfabetu. *Pertinence Summarizer* atpažįsta 10 kalbų tekstus, tarp jų ir lietuvių kalbos tekstus. *Copernic Summarizer* atpažįsta tik 4 kalbų tekstus, o neatpažintus ženklus ši sistema paverčia neatpažįstamais ženklais. *QuickJist Summarizer* atpažįsta viena kalba daugiau nei *Copernic Summarizer* – 5 kalbas. Suradusi neatpažįstamų ženklų tekste, *Copernic Summarizer* juos paverčia neįskaitomais simboliais. Taip pat ši sistema gali sugeneruoti santraukas ir kitiems tekstams, parašytiems lotynišku alfabetu. Mažiausiai kalbų atpažįsta *Intellexer Summarizer*, kurios pagrindinė ir vienintelė kalba yra anglų kalba. Bet tai tik tuo atveju, jei naudojama *English* versija. Be angliškosios *Intellexer Summarizer* versijos taip pat yra prancūzų ir vokiečių, kurios santraukas atlieka taip pat tik viena kalba. Neatpažintus simbolius tekste *Intellexer Summarizer* gali interpretuoti kitais atsitiktiniais simboliais. *MS Office Word AutoSummarize* taip pat tekstą atpažįsta tik tada, kai tekstas yra angliškas arba teksto kalba pakeičiama į anglų kalbą, paliekant originalo rašmenis. Kitu atveju *MS Office Word AutoSummarize* visiškai neteikia teksto santraukos.

Teksto įkėlimas. Kiekviena minėta SARS tekstą įkelia skirtingais būdais. Įkėlimo patogumą gana sunku nustatyti, nes tai priklauso nuo vartotojo pasirinkimo ir poreikio. Galima išskirti kelis įkėlimo būdais: įkeliamas dokumentas, kurio santrauką norima gauti; kopijuojamas dokumentas, kurio santrauką norima gauti; pažymimas teksto turinys ir su komanda COPY arba greitųjų klavišų kombinacija Ctrl+C išskiriamas atitinkamas santraukai rengti tinkamas *MS Word* dokumentas.

Automatic Text Summarizer ir *Shvoong Summarizer* tekstą atpažįs ir panaudos tik tada, jei jis kopijuojamas į naršyklės langą. *Copernic Summarizer* ir *Intellexer Summarizer* tekstą gali nuskaityti iš *MS Word*'o ar kito tekstinio dokumento bei internetinio puslapio. *Pertinence Summarizer* santrauką pateikia keliais atvejais, jei *MS Word*'o ar tekstinis failas bus įkeltas į internetinį sistemos langą ar nukopijuotas. *Pertinence Summarizer* taip pat gali parengti ir internetinio puslapio santrauką, tereikia kaip ir į *Intellexer Summarizer* ar *Copernic Summarizer* langus įrašyti ar kopijuoti internetinio puslapio adresą. Tuo tarpu *Subject Search Summarizer* ir *QuickJist Summarizer* santraukas pateikia tik tada, kai pažymėtas visas tekstas ir pasirinkta COPY komanda arba greitųjų klavišų Ctrl+C kombinacija. *MS Office Word AutoSummarize* santrauką

pateikia meniu eilutėje pasirinkus *Tools – AutoSummarize* (liet. *Įrankiai – Automatiškai apibendrinti*).

Santraukų saugojimas. SARS, sukūrusi santrauką gali ją ir išsaugoti. Šią savybę turi visos anksčiau minėtos sistemos, tačiau tai atlieka skirtingai. Santraukų saugojimo būdai: *MS Word*'o dokumentai; HTML dokumentai; tekstiniai dokumentai; XML dokumentai; kopijos iš programos ar interneto lango į *MS Word* dokumentą; programos langų kopijos kaip ekrano vaizdas (angl. *printscreen*).

Kiekviena SARS suteikia skirtingas saugojimo galimybes. *Automatic Text Summarizer*, *QuickJist Summarizer*, *Pertinence Summarizer* leidžia tik pasižymėti parengtą santrauką ir ją nukopijuoti į *MS Word*'o ar tekstinį dokumentą. *Copernic Summarizer* leidžia santrauką išsaugoti keturiais formatais – *MS Word*'o, tekstinio dokumento, HTML dokumento, XML dokumento. *Intellexer Summarizer* leidžia saugoti tik dviem formatais – tekstinio dokumentu ir HTML dokumentu. *Subject Search Summarizer* ir *Shvoong Summarizer* santrauką saugo tik kopijuotą aktyvaus programos lango vaizdą. Taip pat galima kopijuoti ir *Pertinence Summarizer* bei kitomis santraukų sistemomis parengtą tekstą. *MS Office Word AutoSummarize* leidžia tekstą išsaugoti įvairiais formatais – ne tik kopijuojant, XML, HTML, *.txt ar *.doc bet ir *.rtf formatu, *MS Word 2007*, *Works 4.0–8.0* formatais, senesnės versijos *MS Word* formatais – *MS Word 97* ir t. t.

SARS praktinio taikymo rezultatų analizė

Praktinio taikymo tyrimui buvo panaudoti 6 tekstai – 3 angliški tekstai, ir 3 lietuviški trijų žanrų tekstai – moksliniai, publicistiniai ir grožiniai. Anglišių ir lietuviškų tekstų apimtis sakiniiais ir žodžiais pateikiama 1 lentelėje. Santraukos kokybės vertinimą atliko 5 nepriklausomi ekspertai, kurie perskaitę pateiktas tekstų santraukas, jas vertino pagal įvairius kokybės rodiklius. Ekspertinę analizę atliko 18–22 metų amžiaus studentai, studijuojantys pagal verslo vadybos, turizmo ir viešbučių vadybos, maisto technologijos ir inžinerijos, automobilių techninio eksploatavimo ir automatikos studijų programas.

Santraukos kokybei ištirti buvo naudojami šie kriterijai: teksto suspaudimo laipsnis, gautos santraukos apimtis ir prasmės perteikimas, teksto suprantamumas ir įskaitomumas. Taip pat buvo įvertinta bendra santraukos kokybė, atsižvelgiant į visus minėtus kriterijus.

1 lentelė. Tyrime naudotų anglišių ir lietuviškų tekstų suvestinė

Kalba	Mokslinis tekstas		Grožinis tekstas		Publicistinis tekstas	
Angliškas tekstas	1417 ž.	70 sak.	2699 ž.	81 sak.	580 ž.	31 sak.
Lietuviškas tekstas	1132 ž.	71 sak.	1489 ž.	130 sak.	220 ž.	13 sak.

Praktiniam SARS tyrimui buvo paimtos aštuonios, anksčiau aprašytos ir nagrinėtos sistemos: *Automatic Text Summarizer*, *Shvoong Summarizer*, *Copernic Summarizer*, *Intellexer Summarizer*, *Pertinence Summarizer*, *QuickJist Summarizer*, *Subject Search Summarizer* ir *MS Word AutoSummarize*. Kiekvienai SARS buvo naudojami tie patys tekstai bei panašus glaudumo laipsnis. Tiriamiems ir analizuojamiems tekstams buvo taikomas 25 %, 30 % ir

50 % glaustumas, kuris buvo parenkamas atsižvelgiant į sistemos galimybes ir leistinas parinktis. Tiriamiesiems tekstams buvo taikoma ir 5 % bei 10 % suspaudimas, siekiant surasti santraukų kokybės skirtumus tarp gautų 50 %, 30 %, 25 %, 10 % ir 5 % santraukų.

Kadangi *Pertinence Summarizer*, *Shvoong Summarizer*, *Subject Search Summarizer* ir *Automatic Text Summarizer* negali glaudinti pagal procentinę išraišką, todėl joms buvo nurodomas sakinių skaičius, kuris atitinka 25 % ar 50 % viso sakinių kiekio arba naudojama 30 % bei 50 % suspaudimas. Sakinių glaudinimo būdas buvo tiriamas *Automatic Text Summarizer* ir *Subject Search Summarizer* programoms, o 30 % glaustumas buvo taikomas *Pertinence Summarizer* ir *Shvoong Summarizer* sistemoms. Kitos sistemos gali parengti 10 %, 25 %, 30 %, 50 % ir kitokio laipsnio glaustumo santraukas. *MS Word AutoSummarizer* negali atlikti kitokių glaudinimo nei 10 %, 25 %, 50 %. Likusios SARS leido pasirinkti suspaudimo laipsnį 5 %.

Santraukų apimtys, kai glaudinama pagal sakinių kiekį, skiriasi angliškiems ir lietuviškiems tekstams. Naudojant 50 % suspaudimą, angliškojo teksto santraukos apimtis skyrėsi nuo 2 iki 12 sakinių. Ilgiausias angliško mokslinio teksto santraukas parengia *Copernic Summarizer* ir *Intellexer Summarizer*, o trumpiausią – *QuickJist Summarizer*. Ilgiausias santraukas *Copernic Summarizer* ir *Intellexer Summarizer* parengia lietuviškiems ir angliškiems tekstams. Iš tyrimo naudotų sistemų *Shvoong Summarizer* vienintelė santraukų sistema, kuri nepateikė mokslinio teksto santraukos, kai taikėme 50 % bei 30 % glaustumo kriterijų, nors kitokio žanro lietuviškų tekstų santraukos buvo pateikiamos be jokių kliūčių. Ši sistema taip pat nepateikė ir įvairaus glaudinimo atvejų lietuviškam moksliniam tekstui. Kadangi *Automatic Text Summarizer* ir *Subject Search Summarizer* negali atlikti procentinio glaudinimo, sakinių skaičius, kuris sudaro santrauką buvo gautas paėmus 50 % originalaus teksto sakinių, kurie kai kuriais atvejais skyrėsi nuo kitų SARS pateikiamų sakinių santraukoje skaičiaus. *Subject Search Summarizer* pateikia papildomus sakinius tinkamus santraukai, reikšminius žodžius ar kitą papildomą teksto informaciją, todėl santraukos ilgis, gali būti ilgesnis 2–3 sakiniais nei pasirinktoji pradinė santraukos apimtis. Iš gautų duomenų galima teigti, jog angliškų tekstų santraukų apimtis sakiniais santraukų sistemose yra vienoda, arba skiriasi nežymiai. Kitaip nei anglų kalbos santraukose, lietuvių kalbos santraukos ilgis sutampa tik publicistinio teksto santraukose, o mokslinio ir grožinio tekstų santraukos savo apimtimi skiriasi, nors naudota toks pats glaustumo kriterijus ir tas pats tekstas. Taip yra todėl, kad trumpesnio ir paprastesnio teksto santraukas parengti lengviau, nei ilgesnio ir sudėtingesnio. Iš tikrųjų santraukų apimtys skyrėsi nežymiai ir nebuvo sistemų, pateikiančių vienodo ilgio santraukas, kas pasitaikė atliekant angliškų tekstų glaudinimą.

Vidutinė angliško grožinio teksto santrauka tyrimo metu, naudojant 50 % glaustumo laipsnį, apytikriai gali būti prilyginta 38,75 sakiniais, angliško mokslinio teksto vidutinė santrauka – 29,88 sakiniai, angliško publicistinio teksto santrauka – 15,125 sakiniai. Vidutinis lietuviško grožinio teksto santraukos ilgis naudojant 50 % glaustumą yra

57,25 sakiniai, mokslinio – 30,71, o publicistinio – 6,125 sakiniai. Nors originalus angliškas grožinis tekstas buvo ilgesnis nei lietuviškas grožinis tekstas, tačiau gauta lietuviška teksto santrauka yra ilgesnė nei atitinkamo angliško teksto santrauka.

Pastebėta, jog ne visose santraukose su tuo pačiu glaudinimo laipsniu sakinių skaičius yra vienodas. Tarkime, veikimu ir santraukų rengimo principais panašios *Copernic Summarizer* ir *Intellexer Summarizer* ne visada pateikia vienodos apimties santraukas. Grožinių tekstų anglų ir lietuvių kalbomis santraukos, kurias pateikia minėtosios sistemos yra sudarytos iš vienodo skaičiaus sakinių, tačiau kitų žanrų santraukų apimtys skiriasi – *Copernic Summarizer* pateikia trumpesnes santraukas, nei *Intellexer Summarizer*.

Pagal procentinį 25 % glaustumą grožinio angliško teksto santraukos sakinių skaičiumi daugeliu atveju yra vienodos ar panašios. Pastebima, kad *Pertinence Summarizer* parengta santrauka yra tik vienu sakiniu ilgesnė, nors buvo naudojama 30 % glaudinimas. Tačiau *Shvoong Summarizer* parengta santrauka yra trumpesnė už 25 % atitinkamo teksto santraukos sakinių skaičių, nors naudojama 30 % glaudinimas. Trumpiausią teksto santrauką sakinių skaičiumi pateikia *QuickJist Summarizer*, nors jam šio bandymo metu buvo priskirta 25% glaustumas. *Subject Search Summarizer* santrauka yra ilgesnė dviem sakiniais, dėl to, kad pridedama pora papildomų sakinių, apibūdinančių kurią nors teksto dalį ar veiksmą, ar pateikiant papildomą informaciją apie tekstą ir reikšminius jo žodžius. Į *Copernic Summarizer* ir *Intellexer Summarizer* santraukos ilgį neįskaičiuojamos reikšminių žodžių eilutės, kurios nurodo pagrindinius teksto žodžius.

Pastebėta, kad *QuickJist Summarizer* pateikiamos 25 % santraukos lietuvių ir anglų kalbų tekstuose yra trumpiausios, o ilgiausiomis galima laikyti *Subject Search Summarizer* pateikiamas santraukas. Tačiau *Subject Search Summarizer* santraukos ilgis buvo pasirinktas pagal *MS Word AutoSummarize* pateiktų sakinių 25 % santraukos glaudinimą, kadangi pati sistema nenaudoja jokių procentinių glaudinimo būdų.

2 lentelė. Lietuviško teksto santraukų apimtis remiantis sakinių skaičiumi, naudojant 50 % kompresiją

Programa	Grožinis tekstas	Mokslinis tekstas	Publicistinis tekstas
Automatic Text Summarizer	65 sak.	35 sak.	6 sak.
Copernic Summarizer	67 sak.	40 sak.	9 sak.
Intellexer Summarizer	66 sak.	34 sak.	6 sak.
MS Word AutoSummarize	56 sak.	31 sak.	5 sak.
Pertinence Summarizer	39 sak.	26 sak.	6 sak.
QuickJist Summarizer	53 sak.	21 sak.	5 sak.
Shvoong Summarizer	44 sak.	-	7 sak.
Subject search Summarizer	68 sak.	28 sak.	5 sak.

Kiekvienos SARS pateikiamos santraukos nors glaudinamos ir tokiu pačiu – 50 % laipsniu, bet kiekvienoje santraukoje žodžių skaičius yra skirtingas. Kaip minėta anksčiau, *Shvoong Summarizer* nepateikia lietuviško mokslinio teksto santraukos, todėl šios sistemos santrauka nėra nagrinėjama. Iš turimų duomenų matyti, jog santraukos apimtis žodžiais daugiausia skiriasi 653 žodžiais, o mažiausiai – 6 žodžiais. Vidutiniškai visų tekstų santraukų ilgis yra 654

žodžiai santraukai, tačiau santraukos ilgis priklauso nuo teksto ilgio bei glaudinimo laipsnio, naudoto tekstams. Tam tikroje dalyje santraukų (ir lietuviškų, ir angliškų) žodžių skaičius atitinkamo stiliaus tekste yra panašus ir skiriasi vidutiniškai 30 žodžių. Tačiau santraukos, kurios yra ilgesnės, dažniausiai yra suprantamesnės ir prasmingesnės. Esant mažesniai 30 % glaudinimui sistemoms *Shvoong Summarizer* ir *Pertinence Summarizer* santraukų ilgiai skiriasi 46 žodžiais. *Pertinence Summarizer* pateikiamos santraukos ilgis žodžiais ir kitos sistemos, tokios kaip *Intellexer Summarizer* žodžių skirtumas yra palyginti labai mažas (51 žodis, o palyginus su *Automatic Text Summarizer* dar mažesnis – 41 žodis).

Skirtumai tarp to paties glaustumo teksto santraukoms svyruoja priklausomai nuo to, kokius santraukų rengimo būdus jos naudoja. Sistemos, kurios santraukas rengia pagal sakinių kiekį automatiškai pateikia didesnį žodžių kiekį, nes dažniausiai į santrauką įtraukiami ilgiausi teksto sakiniai. Nors *Automatic Text Summarizer* ir *Subject Search Summarizer* glaudinimui pasirinktas vienodas sakinių skaičius, atitinkantis 25 % viso teksto, tačiau šių sistemų pateiktos santraukos skiriasi 234 žodžiais. Toks skirtumas sistemoms naudojančioms tą patį glaudinimo būdą yra pakankamai didelis. Tačiau, esant ir kitokiam glaudinimo laipsniui santraukose kartais pastebimi gana dideli žodžių skaičiaus skirtumai. Eksperimentai parodė, kad trumpiausias santraukas žodžiais pateikia *QuickJist Summarizer*, kuris, kaip ir likusios sistemos santrauką rengia naudodamas 25 % glaudinimo laipsnį. Vidutiniškai nuo kitų santraukų *QuickJist Summarizer* parengta santrauka skiriasi ~260 žodžiais. Tai, palyginus su kitomis sistemomis ir kitokiais glaudinimo lygiais yra didelis skirtumas, nes vidutinis skirtumas yra ~27,33 žodžiai. Kaip ir žodžių kiekis santraukoje, taip ir vidutinė jos apimtis kiekvienu atveju yra skirtinga. Skirtumai, kaip jau minėta, dažniausiai priklauso nuo glaustumo laipsnio, santraukos ilgio ir sistemos galimybių. Skirtumai tarp vidutinio santraukos sakinių ilgio žodžiais nėra dideli – kai kuriais atvejais mažesni nei naudojant 25 % ar 30 % glaudinimo laipsnį. Naudojant 50 % glaudinimo laipsnį vidutinis sakinių ilgis skiriasi nuo 2 iki 10 dešimties žodžių sakiniui. Tai nėra daug, nors bendras sakinių skaičius santraukose skiriasi. Iš gautų duomenų galima teigti, jog SARS santraukose ilgiausius sakinius pateikia, priklausomai nuo teksto stiliaus, jo apimties bei kalbos. Tendencija, jog trumpiausias 50 % suspaudimas santraukas pateikia *QuickJist Summarizer*, išlieka beveik visose lietuvių ir anglų kalbų santraukose. *QuickJist Summarizer* pateikiamos santraukos yra trumpesnės nei kitos, tačiau dėl to santrauka nepraranda prasmingumo ir suprantamumo.

Iš pateiktų duomenų matome, jog vidutiniškai ilgiausius sakinius santraukoje pateikia *Shvoong Summarizer*. Kaip jau minėta anksčiau, taip yra todėl, kad šiuo atveju *Shvoong Summarizer* naudoja 30 % glaudinimą, ir ima didesnę teksto dalį, nei kitos santraukų sistemos. Nors *Pertinence Summarizer* taip pat naudoja 30 % glaudinimą, tačiau šios sistemos pateikiama santrauka nuo 25 % teksto santraukos skiriasi nežymiai. *Automatic Text Summarizer* ir *Subject Search Summarizer* vidutinis sakinių ilgis skiriasi 8,67 žodžio (plg. su *Pertinence Summarizer* ir *Shoong Summarizer*, kur skirtumas lygus 24,84 žodžiams).

Kiekvienas tekstas, kaip ir kiekviena santrauka turi prasmę, kuri išreiškiama žodžiais, sakiniais, pastraipomis. Tačiau dažnai kyla klausimas, ar pasinaudojus SARS iš gautosios santraukos bus įmanoma adekvačiai nustatyti teksto turinį, neperskaičius originalaus teksto. Iš gautų rezultatų, galima teigti, jog teksto prasmė santraukose priklauso nuo glaustumo laipsnio. 50 % glaustumo santraukos yra aiškesnės, prasmingesnės, nors ir atitinkamai ilgesnės. Naudojant 25 %, 10 %, 5 % glaustumą, gaunamos santraukos yra trumposnės, tačiau ne visada informatyvios, prasmingos ir suprantamos. Dažnai vyrauja padiriki sakiniai, bendras kontekstas sunkiau suprantamas. Ypač tai pastebima grožinio žanro kūriniuose, kur sakiniai ištraukiami iš bendro konteksto ir pateikiami kaip santraukos dalis. Mokslinio žanro santraukoje sakinių „išsiderinimas“ yra mažiau pastebimas, bet, atidžiau skaitant, jį galima išvengti. *Subject Search Summarizer* mokslinio teksto santraukoje pateikia keletą vienodų sakinių, kuriuos kitos SARS santraukoje pateikia tik vieną kartą.

Naudojant mažesnę nei 25 % glaustumą tekstas tampa sunkiau suprantamas, o naudojant žemiausią 5 % glaustumą gaunamas tekstas yra beprasmis ir mažai adekvatus originalui. Iš to galima teigti, jog geriausiai teksto prasmę atspindi nuo 25 % iki 50 % glaustintos santraukos. Tai patvirtina išvadą, jog kuo daugiau spaudžiamas tekstas, tuo mažiau prasmės lieka santraukoje.

Teksto santraukos suprantamumas (nesigilinant į prasmę, o tik į sakinių išdėstymą). Santraukos suprantamumas yra vienas iš svarbiausių kokybės kriterijų, kaip esminis reikalavimas gerai santraukai. Deja tobulą santrauką gauti nėra lengva. Tiek lietuviškuose, tiek angliškuose santraukų tekstuose pasitaiko įvairių klaidų. Dažnai SARS programose pateikiamų santraukų sakiniai būna ištraukti iš konteksto, prasideda mažosiomis raidėmis. Pasitaikė keli atvejai, kai buvo sujungiami žodžiai, romėniški skaičiai painiojami su raidėmis. Tokie dalykai pastebimi angliškame grožiniame tekste, kuriame romėnišku skaitmeniu žymimas skaičius „I“ supainiojamas su angliškuoju raidės „I“ simboliu. Pavyzdžiui, *QuickJist Summarizer* grožinio teksto santraukoje yra sakiny *I HAD just been looking long and sadly at Holbein's ploughman, and was walking through the fields, musing on rustic life and the destiny of the husbandman*. Sistemos, kurios nepainiojo skaičiaus vienas su žodžiu „I“ yra *Automatic Text Summarizer* ir *Shvoong Summarizer*. Kitos sistemos „I“ priskiria prie šalia esančio sakinio, kaip įvardį „aš“, dar kitos – „suklijuoja“ su prieš tai einančiu žodžiu, taip sukurdamos nesamą žodį. Žodžių suliejimų pasitaiko ir *Subject Search Summarizer*, *Automatic Text Summarizer*.

Dalis SARS sakinius tiesiog ima iš teksto, neskirdamos nei mažųjų, nei didžiųjų raidžių, taip sudarydamos įspūdį, jog sakiny ne tik, kad nebaigtas, bet ir trūksta jo dalies, kuri atskleistų kokią nors prasmę. Pavyzdžiui, *Subject Search Summarizer* lietuviško grožinio teksto santraukos sakiny: *jis jau matė daug pilių ir daug moterų (bet nė viena negali prilygti tai, kuri jo laukia po dviejų dienų)*. Pasitaiko atvejų, kai keli žodžiai sujungiami į vieną, pavyzdžiui, romėniškas skaičius „I“ sujungiamas su *HAD* ir santraukoje pateikiama kaip vienas žodis *IHAD*, lietuviško teksto santraukoje randamas toks žodžių sujungimo pavyzdys: *Pir-*

masis skyriusJo. Žinoma, tokie atvejai nesukelia daug nepatogumų skaitant, bet anglų kalboje painiojant romėniškai rašomą skaičių „I“ su anglišku „aš“ kartais iškreipiama prasmė. Panašių atvejų pasitaiko beveik visose SARS, todėl negalima tvirtinti, jog kuri nors iš jų pateikė santrauką be trūkumų.

Santraukos įskaitomumas. Dar vienas svarbus santraukos rodiklis yra santraukos įskaitomumas. Net jei santrauka būtų ir ideali, ypač svarbu kaip įgyvendinta galimybė perskaityti santrauką. Santraukų įskaitomumo vertinimas – ar yra neatpažintų ženklų, ar pakeičiamos raidės ir tai, kas dar gali turėti įtakos į santraukos įskaitomumą. Iš gautų rezultatų galima teigti, kad geriausiai įskaitomos tiek lietuvių, tiek anglų kalbomis yra santraukos parengtos *MS Word AutoSummarize*, *Automatic Text Summarizer*, *Pertinence Summarizer*, *Shvoong Summarizer*, *Subject Search Summarizer*. Gerai įskaitomas angliškas santraukas pateikia ir *Copernic Summarizer*, *Intellexer Summarizer* bei *QuickJist Summarizer*, tačiau šių sistemų lietuviškų tekstų santraukų įskaitomumas gerokai nusileidžia kitoms sistemoms. *QuickJist Summarizer* sunkiau įskaitomos santraukos yra tada, kai santrauka skaitoma tos sistemos lange. Ją nukopijavus į *MS Word* dokumentą, santrauka yra įskaitoma, joje nebūna neatpažintų ženklų, kitaip nei programos lange. *Copernic Summarizer* ir *Intellexer Summarizer* parengtos santraukos (net ir į *MS Word* dokumento kopijos), yra sunkiai įskaitomos. *Copernic Summarizer* lietuviškoje santraukoje panaikina lietuviškus rašmenis – š virsta s, ž – z; q – a; ū, ū – u; i – i; é, e – e; ši sistema lietuviškus rašmenis padaro lotyniškais. *Intellexer Summarizer* lietuviškose santraukose lietuviški rašmenys paverčiami neįskaitomais, tenka spėti, kokia tai raidė. Pvz., *Copernic Summarizer*:

iatobulejus – ištobulėjus; dideliu pakitimu – didelių pakitimų; iakoje – šakoje; tekstynu – tekstynų; pa~velgti – pažvelgti; atskleid~ia – atskleidžia; ~mones – žmones, ~vaig~des - žvaigždės ir t. t.

Intellexer Summarizer:

dÄ–l – dėl; sÄ«nÄ³ – sūnų; u,tvara – užtvarą; gra,u – gražu; triusiancios – triūsiančios; panaÄ«Ä«s – panašūs; io – šio.

3 lentelė. Lietuvių kalbos teksto santraukos įskaitomumo įvertinimas³

Programa	Grožinis tekstas	Mokslinis tekstas	Publicistinis tekstas
Automatic Text Summarizer	5	5	5
Copernic Summarizer	2	2	2
Intellexer Summarizer	2	2	2
MS Word AutoSummarize	5	5	5
Pertinence Summarizer	5	5	5
QuickJist Summarizer	3	3	3
Shvoong Summarizer	5	-	5
Subject Search Summarizer	5	5	5

Atlikti tyrimai parodė, kad visos sistemos gali rengti puikiai įskaitomas angliškas tekstų santraukas. Tuo tarpu geriausiai įskaitomas lietuviškas santraukas rengia tik *MS Word AutoSummarize*, *Automatic text summarizer*, *Subject*

Search Summarizer, *Pertinence summarizer*, *Shvoong summarizer* sistemos. Prasčiau įvertinamos *QuickJist summarizer* gautos lietuviškų tekstų santraukos, kadangi šios programos lange lietuviškos santraukos nėra lengvai įskaitomos, atsiranda neįskaitomų ženklų, tačiau pasinaudojus funkcija *Copy Clipboard* nukopijuotas tekstas yra taip pat gerai įskaitomas, kaip ir santrauka gauta *MS Word AutoSummarize* sistema. Blogiausiai įvertinamas *Intellexer summarizer* ir *Copernic Summarizer* sistemomis parengtų santraukų įskaitomumas, kadangi tekstuose tenka spėlioti raides ar žodžius.

Santraukų kokybė. Santraukų kokybė yra vienas svarbiausių geros santraukos rodiklių. Ją sudaro šie parametrai: santraukos apimtis, perteikiamo teksto prasmė, santraukos įskaitomumas, jos naudingumas, aiškumas, konkretumas, pagrindinių faktų pateikimas ir prasmingumas. Bendras integralinis vertinimas pagal santraukos įskaitomumą, prasmingumą, jos apimtį, gaunamas kai santrauką skaitęs asmuo įvertina ją penkiabalėje skalėje.

Pagal gautus tyrimo rezultatus nustatyta, kad blogiausiai vertinamos *Subject Search Summarizer* parengtos santraukos. Šios sistemos pateiktos santraukos jas vertinusių ekspertų požiūriu yra blogiausios. Geriausiomis santraukų sistemomis būtų galima vadinti *MS Word AutoSummarize* bei *Automatic Text Summarizer*. Jos pagal visus rodiklius veikia geriausiai ir jų pateikiamos santraukos yra pakankamai geros. Šių sistemų geresnį įvertinimą lėmė ir tai, kad lietuviškos santraukos yra įskaitomos ir pakankamai prasmingos, naudingos ir informatyvios. *QuickJist Summarizer* taip pat gali būti priskiriamas prie geriau santraukas rengiančių sistemų, nors lietuviškų tekstų santraukose pastebimi keli trūkumai. *Copernic Summarizer* ir *Intellexer Summarizer* pateikia pakankamai geras santraukas anglų kalbos tekstams, tačiau lietuvių kalbos santraukos šiose sistemose yra prastesnės kokybės nei angliškos santraukos, todėl joms buvo skirtas žemesnis balas.

Apibendrinant rezultatus galima teigti, jog geriausios santraukų sistemos, tinkančios lietuvių kalbai yra *Automatic Text Summarizer*, *MS Word AutoSummarize* ir *QuickJist Summarizer*, kurios pateikia geriausios kokybės lietuviškas santraukas. Blogiausias lietuviškas santraukas pateikia *Subject Search Summarizer*, *Pertinence Summarizer*, *Copernic Summarizer*, *Intellexer Summarizer*, nors anglų kalbos santraukos gautos šiomis sistemomis (išskyrus *Subject Search Summarizer*) yra pakankamai geros. Vidutiniškos kokybės lietuviškas santraukas pateikia *Shvoong Summarizer*.

Išvados

SARS tyrimui naudota angliški ir lietuviški mokslinio, publicistinio ir grožinio stiliaus tekstai su aštuoniomis santraukų rengimo sistemomis: *Automatic text summarizer*, *Copernic summarizer*, *Intellexer summarizer*, *MS Word AutoSummarize*, *Pertinence summarizer*, *Subject Search Summarizer*, *QuickJist summarizer*, *Shvoong summarizer*. Nustatyta, kad automatinės teksto santraukų sistemos pakankamai gerai veikia su bet kokio stiliaus ar žanro anglų kalbos tekstais. Analizuojant lietuvių kalbos tekstus išskyla problemos dėl pačio teksto nuskaitymo ir jo programinės interpretacijos, gautos santraukos įskaitomumo ir supran-

³ Įskaitomumas buvo vertinamas penkiabalėje sistemoje, kur 5 – labai gerai, 4 – gerai, 3 – vidutiniškai, 2 – blogai, 1 – labai blogai.

tamumo bei prasmės perteikimo galimybių. Naudojant didesnę teksto glaudinimą, žymi dalis SARS nesugeba santraukose adekvačiai perteikti pagrindinio teksto turinio, neatspindi jo prasmės. Todėl, norėdami gauti pakankamai gerą santrauką, turėtume naudoti ne didesnę negu 50 % glaudumo laipsnį, atskirais atvejais – 25–30 %, bet ne didesnius. Pastebėta, kad *Copernic summarizer* ir *Intellexer summarizer* sistemų parengtas santraukas lietuvių kalbos tekstui yra sunku įskaityti, žodžių reikšmė dažnai gali būti nuspėjama tik pagal bendrą sakinio kontekstą. Šios santraukų sistemos neatpažįsta lietuviškų raidžių, tačiau puikiai veikia su visais teksta parašytais lotynišku alfabetu. Lietuvių kalbai tinkančios santraukų sistemos yra *MS Word AutoSummarize*, *Automatic Text Summarizer* ir *QuickJist Summarizer*. Jos pateikia santraukas, kurios yra geresnės kokybės nei kitų santraukų sistemų, išlaiko informatyvumą ir prasmingumą.

Nustatyta, kad prastos kokybės lietuviškas santraukas pateikia *Subject Search Summarizer*, *Pertinence Summarizer*, *Copernic Summarizer*, *Intellexer Summarizer* sistemos, nors anglų kalbos santraukos, gautos šiomis sistemomis (išskyrus *Subject Search Summarizer*) yra vertinamos pakankamai gerai. Vidutiniškos kokybės lietuviškas santraukas pateikia *Shvoong Summarizer*. Reikia pripažinti, kad automatinės teksto santraukų sistemos yra nepaprastai naudingos, nes sumažina laiko sąnaudas sutrumpindamos tekstą pagal vartotojo pageidavimus. Šias sistemas pritaikius lietuvių kalbai, ar sukūrus lietuvišką automatinę skaitmeninio teksto santraukų rengimo sistemą, palengvėtų ir taptų spartesnis tekstų skaitomumas gimtąją kalbą, jos leistų efektyviau atrinkti informaciją bei taupyti didelės apimties tekstų peržvalgos trukmę.

Jurgita Lasytė, Bronius Tamulynas

Summarizers – Similarities, Differences and Application in the Lithuanian Language

Summary

Summarizer is a useful system, which allows compressing text and represents it in a shorter way. In Lithuania summarizers are new technological developments and information about them is scarce, although these systems are created and developed by the specialists of computational linguistics in other countries. In the research English and Lithuanian texts were used, which allowed seeing differences among output summaries. In the English texts, summaries using all summarizers were quite informative, useful and did not miss the main information from the original texts, except cases when a high compression rate was used. While summarizing Lithuanian texts, some summarizers confront with language recognition problem – the output summaries in Lithuanian are with unrecognizable symbols, or in individual cases some symbols are missing. Summaries in the Lithuanian language are not of the same quality as the summaries in English. It was also noticed, that Lithuanian summaries with higher compression rate are losing the meaning of the original text.

To conclude, it was noticed that commercial summarizers give better summaries in the English language, but in the Lithuanian language better summaries are given by noncommercial summarizers. Furthermore, summaries received using high compression rate are losing their information and meaning. Analyzed summarizers are best fitted for the English language, therefore using summarizer for other language than English stipulates various problems – from language recognition to rendering of text meaning. It is very important to adapt different languages to summarizers, as this would significantly improve summarizers. Findings of this paper can be used in creating a Lithuanian summarizer or improving other summarizers.

Straipsnis įteiktas 2010 01
Parengtas spaudai 2010 10

Apie autorius

Jurgita Lasytė, Kauno technologijos universiteto magistrantė.
El. paštas: jurgita.lasyte@gmail.com

Bronius Tamulynas, technikos mokslų daktaras, Kauno technologijos universiteto Informatikos fakulteto Kompiuterių tinklų katedros docentas.
Mokslinės veiklos sritys: kompiuterinio kalbų vertimo technologijos, kompiuterinė lingvistika, intelektualųjų sistemų modeliavimas, lingvistinių duomenų struktūros.
Adresas: Kauno technologijos universitetas, Informatikos fakultetas, Studentų g. 50-414, 51368 Kaunas.
El. paštas: bronius.tamulynas@ktu.lt

Literatūra

1. Mani, I., 2001. *Automatic Summarization*. Philadelphia: John Benjamins Publishing Company.
2. Mitkov, R., 2005. *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, pp.583-598.

Šaltiniai

1. *Automatic Text Summarizer*. Search and Information Extraction Research Lab. Language Technologies Research Center, International Institute of Information Technology. Interaktyvus. Prieiga per internetą: <<http://search.iit.ac.in/~jags/summarizer/index.cgi>> [Žiūrėta 2009-05-01].
2. *Copernic Summarizer*. Copernic Inc. Interaktyvus. Prieiga per internetą: <<http://www.copernic.com/en/products/summarizer/>> [Žiūrėta 2009-05-01].
3. *Intellexer Summarizer*. Intellexer. Interaktyvus. Prieiga per internetą: <<http://summarizer.intellexer.com/>> [Žiūrėta 2009-05-01].
4. *MS Word AutoSummarize – Word Help*. Microsoft Corporation. Interaktyvus. Prieiga per internetą: <<http://www.microsoft.com/education/autosummarize.msp>> [Žiūrėta 2009-05-01].
5. *Pertinence Summarizer*. Pertinence Mining. Interaktyvus. Prieiga per internetą: <<http://www.pertinence.net/ps/main.jsp?ui.lang=>>> [Žiūrėta 2009-05-01].
6. *QuickJist Summarizer*. Vitaly Demin. Interaktyvus. Prieiga per internetą: <<http://www.quickjist.com/>> [Žiūrėta 2009-05-01].
7. *Shvoong Summarizer*. Shvoong Ltd., USA. Interaktyvus. Prieiga per internetą: <<http://www.shvoong.com/summarizer/>> [Žiūrėta 2009-05-01].
8. *Subject Search Summarizer*. Kryloff technologies, Inc. Interaktyvus. Prieiga per internetą: <<http://www.kryltech.com/summarizer.htm>> [Žiūrėta 2009-05-01].

