Automatsko sažimanje teksta

Viher, Helena

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2017./ 2018.

Helena Viher

Automatsko sažimanje teksta

Završni rad

Mentor: dr.sc. Kristina Kocijan, doc.

Zagreb 2018.

Sadržaj

Sažetak

Automatsko sažimanje teksta podrazumijeva stvaranje sažetka uz pomoć računalnih algoritama koji iz dokumenata izlučuju samo najrelevantnije rečenice. Cilj ovog rada je prikazati razvoj automatskog sažimanja, pristupe koji se koriste pri stvaranju sažetka, tj. ekstrakciju i apstrakciju te načine evaluacije dobivenog rezultata kako bi se osigurala njegova kvaliteta te poboljšao rad samih algoritama. Naglašava se njegova važnost u raznim područjima zbog progresivno sve veće količine podataka koje nije moguće ručno obraditi, a također se prikazuju problemi koji se javljaju pri sažimanju te kako se s njima suočiti.

Ključne riječi: *automatsko sažimanje teksta, sažimanje, obrada prirodnog jezika, ekstrakcija, apstrakcija, evaluacija sažetaka*

Automatic Text Summarization

Abstract

Automatic text summarization is defined as a process of creating abstract using computer algorithms that extract only the most relevant sentences from a given document. Main purpose of this thesis is to review development of automatic summarization over the years, text summarization approaches, i.e. extraction and abstraction and evaluation methods of automatic summaries which are important part of quality insurance and help improve summarization. Thesis also focuses on importance of automatic text summarization in different fields caused by ever-growing amount of data that's impossible for humans to process without help of computers. Problems that appear during text summarization are also mentioned and solutions to these problems are suggested.

Key words: *automatic text summarization, summarization, natural language processing, extraction, abstraction, summary evaluation*

1. Uvod

Svakog dana u digitalnom okruženju objavljuje se na tisuće novih znanstvenih radova i drugih informacijskih izvora. Baš iz ovog razloga, individualno proučavanje novih radova u potrazi za relevantnim informacijama, čak i u uskom znanstvenom području, u većini slučajeva nije izvedivo jer zahtjeva previše vremena. Ručno izrađeni sažetci te ključne riječi koje se obično javljaju u sklopu samog rada vrlo često nisu dovoljni za određivanje njegove relevantnost jer autori koji ih izrađuju često nemaju znanja o tome kako standardni sažetak treba izgledati, a problemi se također mogu javiti u slučajevima kada neke od važnih informacija iz dokumenta ne pronalazimo u sažetku ili u ključnim riječima. Također nije rijetkost da sažetak dokumenta uopće nije dostupan, što je vrlo česta pojava kod tekstova koji nisu znanstvene prirode, npr. novinski članci, mrežne stranice itd.

Automatsko sažimanje riješilo bi ovaj problem te bi uvelike olakšalo pristup relevantnim informacijama. Osim što su ovakvi sažetci financijski pristupačniji od onih koji su ručno izrađeni, korisnik upitom sažetak može prilagoditi vlastitim potrebama. Noviji sustavi, osim sažimanja teksta, omogućuju i sažimanje zvučnih, slikovnih te multimedijskih sadržaja.

U ovom radu u drugom poglavlju bit će predstavljena važnost automatskog sažimanja te njegove prednosti i nedostaci. U trećem poglavlju bit će navedene vrste sažetaka, a detaljan pregled pristupa automatskom sažimanju slijedi u četvrtom poglavlju. Utjecaj konteksta i domene na automatsko sažimanje proučava se u petom, dok su metode pri istovremenom sažimanju više datoteka predstavljene u sedmom poglavlju. Konačno, u osmom poglavlju bit će navedeni sustavi i metode za evaluaciju sažetaka koji su od izuzetne važnosti za poboljšavanje rada sustava za automatsko sažimanje, a pregled rada te implikacije za budućnost bit će izneseni u zaključku.

2. Općenito o automatskom sažimanju

Prema Saggiono i Poibeau (2012), automatski sustavi za sažimanje se bave dvama glavnim pitanjima: kako odabrati važne informacije iz teksta te kako ih prikazati u sažetom obliku. Prvom pitanju moguće je pristupiti na mnoštvo načina, a neki od njih su statistički pristup, promatranje rečeničnih svojstava unutar teksta, pristupi temeljeni na strojnom učenju itd. Ovi pristupi detaljnije će biti prikazani u daljnjim poglavljima. Prikaz sažetaka javlja se u dva oblika, tj. u obliku ekstrakta te apstrakta, a ovi oblici razlikuju se po izgledu rečenica u nastalom sažetku (Allahyari et al., 2017). Sažetci nastali ekstrakcijom koriste rečenice koje su izravno preuzete iz originalnog dokumenta, a oni nastali apstrakcijom stvaraju potpuno nove rečenice te izgledom podsjećaju na ručno izrađene sažetke.

Kao što je već spomenuto u prethodnom poglavlju, automatsko sažimanje od izuzetne je važnosti za pronalaženje relevantnih podataka u obilju informacijskih izvora koji su dostupni u digitalnom okruženju. Torres-Moreno (2014) napominje činjenicu da sažetci uvelike skraćuju vrijeme čitanja što ujedno i istraživačima olakšava postupak odabira dokumenata koje će koristiti u istraživanju. Automatsko sažimanje također olakšava proces indeksiranja te povećava efikasnost indeksa, a sami sažetci mogu biti prilagođeni potrebama korisnika na temelju postavljenog upita. Također je važno napomenuti da je jedna od najvećih prednosti automatskog sažimanja veća objektivnost algoritma za sažimanje u usporedbi s ljudskim sažimateljem.

Nažalost, tehnologija obrade prirodnog jezika još uvijek nije dovoljno razvijena za stvaranje savršenih ekstrakata te izuzetno koherentnih apstrakata, no područje automatskog sažimanja uvelike je napredovalo od svojih početaka u 1950-im godinama (Ježek i Steinberger, 2008). Još jedan veliki nedostatak automatskog sažimanja je visoka cijena izrade sustava koji je jedino isplativ dugoročno.

3. Vrste sažetaka

Sažimanje dokumenata može se kategorizirati na mnoštvo načina, a razlog tome su dvije važne činjenice: velika količina različitih tipova dokumenata koji proizlaze iz različitih izvora te raznolike potrebe korisnika. Iz ovog razloga isti oblik sažimanja neće raditi u svim situacijama (Torres-Moreno, 2014).

Prema tome Torres-Moreno (2014) sažetke općenito, bez obzira na to jesu li nastali ručno ili automatski, kategorizira prema različitim kriterijima kao što su:

1. **funkcija:**
    - indikativni sažetak (*engl. indicative summary*) koji pruža informacije o temi o kojoj dokument govori, a izgledom podsjećaju na sadržaj,
    - informativne sažetke (*engl. informative summary*) čiji je cilj prikazati sadržaj te po potrebi objasniti argumente u tekstu, a zapravo se radi o skraćenom dokumentu;
2. **broj dokumenata koji se sažimaju:**
    - sažetak jednog dokumenta (*engl. single document summary*),
    - sažetak dva ili više dokumenata (*engl. multidocument*) koji najčešće govore o istoj temi;
3. **žanr:**
    - sažetci novinskih članaka (*engl. news summary*),
    - specijalizirani sažetci (*engl. specialized summary*) koji se odnosne na jednu specijaliziranu domenu, npr. znanost, tehnologija, zdravstvo itd.,
    - književni sažetci (*engl. literary summary*),
    - enciklopedijski sažetci (*engl. encyclopedic summary*),
    - sažetci društvenih mreža (*engl. social network summary*) koji se odnose na sažetke blogova i drugih izuzetno kratkih dokumenata poput objava na Twitteru i sl.;

4. tip:

- ekstrakt (*engl. extract*) koji podrazumijeva sažetak koji je nastao slaganjem fragmenata rečenica iz izvornog teksta,
- apstrakt (*engl. abstract*) koji se definira kao sažetak koji je nastao preradom i parafraziranjem izvornog teksta,
- kompresija rečenica (*engl. sentence compression*) koji ima jednak broj rečenica kao i izvornik, no rečenice su skraćene i pojednostavljene uz pomoć algoritama;

5. sažimatelj:

- autorski sažetak (*engl. author summary*) čiji je autor ujedno i autor izvornog dokumenta te odražava njegovo osobno mišljenje,
- stručni sažetak (*engl. expert summary*) čiji je autor stručnjak unutar određene domene o kojoj izvorni dokument govori, ali najvjerojatnije nema specijalizaciju za pisanje sažetaka,
- profesionalni sažetak (*engl. professional summary*) koji je napisao profesionalni sažimatelj koji najčešće nije stručnjak u području o kojem dokument govori, ali je vrlo dobro upoznat s tehnikama pisanja, normama i standardima koji se koriste pri izradi sažetaka;

6. kontekst:

- općeniti sažetak (*engl. generic summary*) koji daje opći pregled dokumenta i najvažnijih informacija unutar njega, ali ignorira informacijsku potrebu korisnika,
- sažetak koji je nastao prema upitu (*engl. query-guided summary*) čiji je zadatak, ako što mu i ime govori, odgovoriti na informacijski upit korisnika;

7. ciljana publika:

- sažetci za publiku bez profila (*engl. without a profile*) koji su neovisni o potrebama određenih skupina korisnika, a njihov sadržaj temeljen je jedino na informacijama unutar izvornog dokumenta,

- sažetci temeljeni na profilu korisnika (*engl. based on a user profile*) koji su usmjereni na korisnike unutar jedne specijalizirane domene poput sporta, politike, ekonomije, zdravstva itd.

Torres-Moreno (2014) također spominje podjelu prema broju jezika na kojima je pisan izvorni dokument, a razlikuje jednojezične, višejezične sažetke te međujezične (*engl. cross-lingual*) sažetke.

Kumar et al. (2016) sažetke dijele na nešto drugačiji način te različite dimenzije sustava za sažimanje kategoriziraju prema vrsti unosa, zadatku te tipu sažetka koji je dobiven procesom automatskog sažimanja. Vrsta unosa podrazumijeva sažimanje jednog ili više dokumenata. Pri tome autori naglašavaju da se pri sažimanju više datoteka odjednom najčešće radi o datotekama koje govore o istoj temi, odnosno pripadaju istoj domeni. Prema zadatku sažetke dijele na općenite, specifične za domenu ili izrađene prema upitu. Općeniti sažetci sažimaju tekstove bez obzira na to iz koje domene potječu ili kojom temom se bave te sve dokumente promatraju kao homogene tekstove. Na ovaj način radi većina sustava za sažimanje (Nenkova i McKeown, 2011) s obzirom da je teže stvoriti sustave koji sažetke izrađuju prema upitu ili određenoj domeni te je za njihovu izradu potrebno mnoštvo jezičnih resursa, baza znanja (*engl. knowledge base*), baza podataka i sl. (Kumar et al., 2016). Posljednja kategorija koju Kumar et al. (2016) spominju je konačni proizvod koji je nastao procesom sažimanja, tj. ekstrakt i apstrakt.

4. Pristupi automatskom sažimanju

Dva glavna pristupa automatskom sažimanju teksta su ekstrakcija i apstrakcija (Allahyari et al., 2017). Iako je cilj oba pristupa stvoriti kratak i informativan sažetak, način na koji se do tog sažetka dolazi te njegov konačni izgled i sadržaj vrlo su različiti.

Kako bi ove razlike bile jasnije i razumljivije, potrebno je detaljnije objasniti značenje oba pojma. Chettri i Chakraborty (2017) ekstrakt definiraju kao tip sažetka koji je nastao odabirom nekoliko rečenica iz dokumenta uz pomoć jasno definiranih uvjeta na temelju kojih im je pridodana određena razina važnosti unutar izvornog teksta. Za razliku od ekstrakata, čije su rečenice direktno preuzete iz dokumenta, cilj apstrakta je prikazati važne informacije na novi način. Sažetak koji je nastao ovom tehnikom ne sadrži riječi i fraze iz izvornog teksta, već koncepte i ideje predstavlja u drugačijem obliku.

Kada govorimo o stvaranju sažetka izuzetno je važno spomenuti razinu obrade koja se provodi. Ježek i Steinberger (2008) prema razini obrade razlikuju pristup na površinskoj razini (*engl. surface level*) te na dubljoj razini. Prvim pristupom se iz dokumenta u obliku ekstrakta izlučuju značajke teksta poput statistički i pozicijski značajnih riječi, riječi iz domene te iz korisničkog upita, ključnih fraza, itd. Pri obradi na dubljoj razini potrebno je izvršiti semantičku analizu kako bi se utvrdili odnosi između riječi i fraza, a sažetak se izrađuje uz pomoć sinteze prirodnog jezika. Na ovaj način moguće je stvoriti i apstrakt i ekstrakt (Ježek i Steinberger, 2007).

Saggion et al. (2013) pristupe automatskom sažimanju dijele na sličan način te razlikuju površni pristup (*engl. superficial approach)* i pristup temeljen na znanju (*engl. knowledge-based approach*). Površni pristup obuhvaća najranije metode korištene za automatsko sažimanje tzv. klasične metode sažimanja. U njih ubrajamo statistički pristup, pristup temeljen na svojstvima te grafovima, a uključuju i Edmundsonov empirijski pristup koji podrazumijeva stvaranje sažetaka uz pomoć strojnog učenja. Pristupi temeljni na znanju podrazumijevaju korištenje složenih i sveobuhvatnih leksičkih resursa te primjenjuju teorije organizacije diskursa u općem ili specifičnom kontekstu.

4.1. Ekstrakcija

Kao što je već spomenuto, ključni koncept sažimanja ekstrakcijom je identificiranje i izlučivanje važnih rečenica iz dokumenta te sastavljanja istih u sažetak (Kumar et al., 2016). Sažetak nastao na ovaj način zapravo je zbirka izlučenih rečenica, a njegova dužina ovisi o stupnju kompresije (Haiduc, Aponte i Marcus, 2010). Ekstrakcija je trenutno najjednostavniji i najpristupačniji, pa i s time najčešće korišteni pristup automatskom sažimanju teksta (Aliguliyev, 2009; Ko i Seo, 2008).

4.1.1. Statistički pristup

U ranijim istraživačkim radovima koji su se bavili automatskim sažimanjem istraživači su vjerovali da će se važnije riječi češće pojavljivati unutar dokumenta u usporedbi s ostalim riječima (Luhn, 1958). Ovo mišljenje potaknulo je stvaranje niza sustava za automatsko sažimanje koji su važne rečenice izlučivali uz pomoć frekvencije riječi (Klassen, 2012).

Osim frekvencije riječi, u statističkom pristupu za procjenu važnosti rečenica koristi se i mjera frekvencije pojavnica-inverzne frekvencije dokumenata (*engl. Term Frequency-Inverse Document Frequency, tf-idf*). Ova mjera se tradicionalno koristila za prikupljanje informacija o čestoti riječi i fraza unutar korpusa određene domene (Jurafsky i Martin, 2009). Kod automatskog sažimanja zadatak tf-idf metode je odgovoriti na pitanje imaju li sve riječi i fraze, koje se jednako često javljaju u nekom tekstu, jednaku važnost. Npr. svi novinski članci koji izvještaju o razaranjima izazvanim potresima sadržavati će riječ „potres“, pa će prema tome težina te riječi biti veća (*engl. term weight*). Zahvaljujući tome, riječima koje se često pojavljuju unutar dokumenta moguće je smanjiti težinu uz pomoć usporedbe čestote riječi s proporcionalnom čestotom unutar zbirke dokumenata, odnosno korpusa (Kumar et al., 2016). Zbog ovog svojstva tf-idf pristup postao je jednim od najčešće korištenih tehnika za sažimanje ekstrakcijom (Hovy i Lin, 1998).

Velika prednost statističkog pristupa je njegova neovisnost o jeziku kojim je pisan dokument (Haiduc, Aponte i Marcus, 2010). No, ukoliko pretpostavka da se važnije riječi

javljaju češće ne vrijedi za određeni tekst, sustavi koji se koriste statističkim pristupom mogu naići na probleme osobito ako pri stvaranju sažetka nije korištena tf-idf mjera.

4.1.2. Pristup temeljen na svojstvima

Važnost rečenica također se može procijeniti identificiranjem svojstava koja odražavaju relevantnost tih rečenica (Kumar et al, 2016) Rečenice s ovim svojstvima imaju prednost biti odabrane za sažetak pri procesu sažimanja. Edmundson (1969) navodi tri takva svojstva:

- **smještaj rečenice unutar teksta:** prve rečenice u dokumentu sadrže najvažnije informacije i daju općeniti pregled teme o kojoj dokument govori
- **prisutnost riječi iz naslova**
- **prisutnost rečeničnih priloga tj. modifikatora i konektora (*engl. cue words*):** riječi i fraze poput „zaključno“, „rezultat“, „ukratko“, „sažetak“, „cilj“ itd.

Gupta i Lehal (2010) opisuju još nekoliko često korištenih svojstava uz pomoć kojih se može odrediti važnost određene rečenice unutar teksta, a to su:

- **dužina rečenice:** vrijedi pretpostavka da će kraće rečenice sadržavati manje važnih informacija, a izuzetno duge rečenice se također izbjegavaju
- **težina riječi:** ima sličnu ulogu kao i u statističkom pristupu te se odnosi na rečenice koje sadrže više riječi koje su česte unutar dokumenta
- **osobne imenice:** rečenice koje sadrže imena osoba i organizacija smatrat će se važnijima.

Ostala svojstva koja mogu utjecati na percepciju važnosti rečenice navodi Rajkhowa (2015), a to mogu biti:

- **font**: pri čemu je u obzir uzeta činjenica da riječi pisane podebljanim, podcrtanim, kosim i velikim slovima najčešće sadrže važne informacije
- **smještaj paragrafa unutar teksta**: slično smještaju rečenice unutar teksta, paragrafima na početku dokumenta daje se prednost pri odabiru za sažetak

- **međusobna sličnost rečenica**: rečenice se međusobno uspoređuju pri čemu se u obzir uzimaju njihove ključne riječi
- **pristrane riječi** (*engl. biased words*): prethodno definirane riječi koje su prije sažimanja prikupljene u liste, a odnose se na određenu domenu
- **zamjenice**: rečenice koje sadrže zamjenice neće se pojaviti u sažetku, već se moraju proširiti u odgovarajuću imenicu.

4.1.3. Pristup temeljen na grafovima

U pristupu temeljenom na grafovima rečenice i riječi prikazane su u obliku čvorova i vrhova koji povezuju elemente u tekstu koji su u semantičkom odnosu (Chettri i Chakraborty, 2017). Algoritmi iterativnih grafova, kao što su HITS i Googleov PageRank, izvorno su razvijeni i korišteni kao alati za istraživanje strukture poveznica mrežnih stranica te njihovo rangiranje (Ježek i Steinberger, 2008). Kasnije su uspješno korišteni i u drugim područjima poput društvenih mreža, u analizi citata i sl.

U obradi prirodnog jezika korišten je model algoritma pod nazivom TextRank (Mihalcea i Tarau, 2005). Ovaj algoritam radi na istom principu kao i ostali algoritmi iterativnih grafova, tj. svaka se rečenica označava tjemenom na grafu, a veze između njih stvorene su pomoću poveznica između rečenica. Nakon toga, sličnost para rečenica računa se tako da se dobiveni grafovi preklope, a njihovo preklapanje može se odrediti i prema broju preklapanja njihovih leksičkih reprezentacija. Iterativni dio algoritma se zatim primjenjuje na graf rečenica i nakon dodavanja ocjena svakoj od rečenica one s najvišim ocjenama koriste se u sažetku (Ježek i Steinberger, 2008).

4.1.4. Empirijske metode i strojno učenje

S pojavom strojnog učenja 90-ih godina prošlog stoljeća javila se ideja za korištenje ove metode pri automatskom sažimanju (Das i Martins, 2007). Najraniji sustavi sažetke su stvarali uz pomoć statistike, no bili se neovisni o svojstvima riječi i rečenica. Ovakvi sustavi primarno su koristili nekritičke Bayesove (*engl. naive-Bayes*)

metode. Nešto kasnije pri sažimanju u obzir su se počela uzimati neka od rečeničnih svojstava, a algoritmi za učenje više nisu donosili pretpostavke neovisne o tim svojstvima. Ostali pristupi koji spadaju u ovu kategoriju uključuju skriveni Markovljev model (*engl. hidden-Markov model*), model stabala odlučivanja, metode koje koriste neuronske mreže te svojstva dobavljena iz drugih izvora npr. česte riječi u upitima postavljenim mrežnim pretraživačima.

Već spomenute nekritičke Bayesove metode opisali su Kupiec et al. (1995) koristeći se pri tome Edmundsonovim (1969) idejama. Ova metoda koristi klasifikacijske funkcije koje svaku rečenicu kategoriziraju prema tome može li se ona koristiti u sažetku ili ne s obzirom na njezinu važnost unutar dokumenta. Pri tome se u obzir uzima duljina rečenice i prisutnost velikih slova.

Nekoliko godina kasnije (Aone et al., 1990 navedeno u Mani, 1999) sustav DimSum osim nekritičkog Bayesovog klasifikatora za procjenu važnosti rečenica u dokumentu uvodi i formule za računanje frekvencije riječi (*engl. term frequency*) te algoritam za računanje inverzne frekvencije po dokumentima (*engl. inverse document frequency*) s ciljem definiranja ključnih riječi unutar dokumenta. Pri računanju inverzne frekvencije korišteni su veliki korpusi iz domene kojoj pripada tekst koji se sažima.

Sredinom 90-ih godina Lin i Hovy (1998) proučavali su ideju vrednovanja rečenica s obzirom na samo jedno njezino svojstvo, tj. smještaj rečenice unutar teksta. Naime, za većinu dokumenata vrijedi pretpostavka da se najvažnije informacije u tekstu uvijek nalaze na istim mjestima, npr. u naslovu, na početku teksta, u apstraktu itd. Istraživači su također primijetili da položaj ovih informacija nerijetko varira s obzirom na domenu.

Lin (1999) se nadovezuje na ideju korištenja rečeničnih svojstava pri strojnom učenju te primjećuje da su različita rečenična svojstva međusobno zavisna, a taj problem pokušao je riješiti uz pomoć modela stabla odlučivanja (*engl. decision trees*). U svom istraživanju utjecaj svojstava na odluku izlučivanja rečenica promatrao je na velikom uzorku podataka koji je pružio TIPSTER-SUMMAC program za evaluaciju, a podaci su uključivali javno dostupne zbirke tekstova klasificirane prema temama (Das i Martins, 2007). Lin (1999) potom u svrhu istraživanja uvodi niz svojstava:

- **riječi iz upita** (*engl. query signature*): rečenica se ocjenjuje prema broju riječi iz upita prisutnih u njoj
- **najistaknutije riječi iz korpusa** (*engl. IR signature*): odnosi se na broj istaknutih riječi iz korpusa pronađenih u rečenici
- **brojčani podaci**: rečenice se ocjenjuju prema tome sadrže li brojčane podatke ili ne
- **osobne imenice:** podrazumijeva prisutnost imena osoba, gradova, mjesta, organizacija itd.
- **zamjenice i pridjevi**
- **nazivi dana u tjednu**
- **nazivi mjeseci**
- **citati**.

Pri evaluaciji automatskih sažetaka uz pomoć referentnih sažetaka, sažetci nastali uz pomoć modela stabla odlučivanja, u usporedbi s ostalim modelima, ostvarili su vidljivo bolje ocjene u gotovo svim tematskim područjima. Za područja u kojima ovaj model nije ocjenjen visokim ocjenama, Lin (1999) je smatrao da sadrže neovisna svojstva. Nadalje, njegova analiza svojstava potvrdila je važnost istaknutih riječi iz korpusa čime je ujedno i potvrđena Luhnova (1958) rana pretpostavka o njihovoj važnosti pri odabiru rečenica koje će se izlučiti za sažetak.

Nešto kasnije, Conroy i O'leary (2001) automatskom sažimanju pristupaju na nov način te definiraju skriveni Markovljev model. Za razliku od prijašnjih modela koji pri vrednovanju rečenicama pridodaju mnoštvo svojstava, skriveni Markovljev model ih zadržava samo tri: smještaj rečenice u dokumentu, broj riječi u rečenici i vjerojatnost riječi u rečenici s obzirom na riječi u cijelom dokumentu. Rad algoritma pritom postaje sekvencijalan, odnosno odvija se u tri faze: dvije koje su sažimateljske i jedna koja je nesažimateljska.

Ranih 2000-ih godina, na inicijativu Konferencije za razumijevanje dokumenata (*engl. Document Understanding Conference, DUC*), istraživači su započeli rad na sustavima za sažimanje individualnih novinskih članaka (Das i Martins, 2007). Na

temelju rezultata ovog projekta zaključeno je da sustavi koji ostvaruju visoke ocjene pri evaluaciji još uvijek nisu u mogućnosti izlučiti rečenice sa značajnom statističkom važnosti unutar teksta. Za razliku od njih, rečenice izlučene uz pomoć jednostavnijih statističkih modela bile su izuzetno statistički značajne unatoč tome što su sustavi lošije ocjenjivani pri evaluaciji. Ova pojava pripisana je činjenici da se u člancima, u pravilu, najvažniji dijelovi teksta nalaze u prvih nekoliko paragrafa, a statistički sustavi koristili su baš ovo rečenično svojstvo pri odabiru rečenica za sažetak. Iz ovog razloga, DUC je nakon 2002. godine odustao od jednodatotečnih sažimanja novinskih članaka.

Kao odgovor na problem statističke značajnosti, Svore, Vanderwende i Burges (2007) predstavljaju algoritam za sažimanje temeljen na neuronskim mrežama te podacima i svojstvima pribavljenim iz drugih izvora. Oni pri tome koriste zbirku dokumenata prikupljenih s CNN-ove mrežne stranice, a svaki je od dokumenata sadržavao naslov, vrijeme nastanka, tri do četiri ručno istaknute informacije iz teksta te tekst članka. Zadatak njihovog sustava bio je stvoriti tri istaknute informacije iz teksta koje su kasnije evaluirane uz pomoć dvije metrike tj. usporedbe ručno izrađenih istaknutih informacija i onih izrađenih automatski te analize individualnih rečenica. Sam algoritam je treniran uz pomoć oznaka i svojstava rečenica iz članaka za koje su smatrali da utječu na važnost rečenica.

4.2. Apstrakcija

Dok se danas većina istraživanja o automatskom sažimanju bavi pitanjima odabira relevantnih informacija iz teksta, manji broj njih bavi se stvaranjem novih, dosljednih i kohezivnih tekstova iz izvornog dokumenta (Saggion et al., 2013). Razlog tome je što je apstrakcija, za razliku od ekstrakcije, puno složeniji problem. Dodatno, kako bi ovakva reinterpretacija teksta uopće bila moguća, potrebno je razviti sustav koji ima mogućnost korištenja tehnika obrade prirodnog jezika koje su još uvijek u razvoju. Sustav bi također morao s lakoćom moći odrediti glavnu ideju dokumenta.

Metode koje se bave pitanjem razvitka ovakvog sustava nazivaju se apstraktne ili neekstraktivne, a obuhvaćaju područja poput generiranja naslova (*engl. headline*

generation), kompresije rečenica (*engl. sentence compression*) te izreži-i-umetni (*engl. cut-and-paste*) sažimanja.

4.2.1. Generiranje naslova

Cilj generiranja naslova je stvoriti kratki, ali informativni naslov za određeni dokument, npr. za novinski članak (Zajic et al., 2007). Ovo područje je tradicionalno bilo usko povezano uz ekstrakciju. No, veliki nedostatak ovog pristupa bila je nemogućnost iskorištavanja korpusa za učenje te s time i stvaranja sažetka u „malom omjeru“, zbog čega su istraživači počeli koristiti pristupe temeljene na strojnom učenju (Jin i Hauptmann, 2000).

Witbrock i Mittal (1999) ovaj pristup koriste za učenje korelacije riječi u dokumentu i naslovu, no pri tome ih oboje ograničavaju na isti površinski niz (*engl. surface string*). Na njih se nadovezuju Jin i Hauptmann (2000) olakšavajući pri tome navedeno ograničenje. Zajic et al. (2007) predlažu dva pristupa: pristup temeljen na sintaktičkim stablima te robusniji sakriveni Markovljev model.

4.2.2. Kompresija rečenica

U procesu automatskog sažimanja uobičajeno je da stručnjaci nakon ekstrakcije dodatno uređuju rečenice kako bi sažetak bio što jezgrovitiji i dosljedniji (Jing, 2000). Cilj algoritma za kompresiju rečenica je oponašati ovaj postupak te sažeti rečenice unutar teksta izbacivanjem nepotrebnih dijelova (Saggion et al., 2013). Ovakav sustav također bi trebao imati mogućnost sintaktičke transformacije, zamjene složenih fraza s po potrebi općenitijim ili specifičnijim frazama, kombiniranja dviju skraćenih rečenica, preslagivanja izlučenih rečenica itd. (Jing i McKeown, 2000).

Statistički pristup kompresiji rečenica predstavili su Knight i Marcu (2000). Oni su smatrali da je to prvi korak za rješavanje problema sažimanja jednog ili više dokumenata. Dva modela koja koriste su model kanala sa šumom (*engl. noisy-channel*

model) koji kompresiju rečenica uči uz pomoć referentnog poravnatog korpusa te modela stabla odlučivanja koji je kondicioniran kontrolnim izjavama.

Kako bi kompresija rečenica bila što kvalitetnijima, osobito u profesionalnom okruženju, ponekad nije dovoljno samo izbaciti određene dijelove rečenica, već je potrebno izvršiti složene lingvističke izmjene poput premještanja dijelova rečenice te stvaranja potpuno novih fraza. Jing (2000) koristi jedan od takvih sustava te za postizanje kompresije brisanjem i skraćivanjem fraza koristi višestruke izvore znanja: referentni poravnati korpus koji sadrži izvorne rečenice te rečenice skraćene od strane stručnjaka za sažimanje, leksikon velikog razmjera, WordNet i sintaktički parser. Njegov algoritam za kompresiju odvija se u pet koraka:

1. **sintaktički parsing:** tekst se analizira i označava
2. **gramatička provjera:** određuju se dijelovi koji se ne smiju brisati
3. **provjera konteksta:** sustav odlučuje koji dijelovi rečenice su najviše povezani s glavnom temom o kojoj se govori u dokumentu
4. **korpusno dokazivanje:** sustav uz pomoć referentnog korpusa određuje koje fraze bi stručnjak najvjerojatnije odstranio
5. **krajnja odluka:** na temelju rezultata prijašnjih koraka odlučuje se kako će izgledati konačna kompresija rečenica.

4.2.3. Izreži-i-umetni sažimanje

Osnovni cilj ovog sustava je umanjiti razlike između automatskih i ručno izrađenih sažetaka koje se javljaju iz dva razloga (Jing i McKeown, 2000):

1. većini sustava je još uvijek teško odrediti koji su dijelovi dokumenta važni
2. većina sustava za sažimanje uz to koristi prilično loše tehnike za generiranje jezika te se većinom još uvijek temelje na ekstrakciji.

Problemi se također javljaju kod izlučenih rečenica u ekstraktu koje mogu biti nepovezane, nedosljedne pa čak i varljive te mogu čitatelja dovesti do krivih zaključaka bez obzira na činjenicu da ih je sustav za sažimanje proglasio prikladnim za sažetak.

Jing i McKeown (2000) predlažu računalni model koji bi riješio ovaj problem. Njihov sustav za sažimanje temelji se na šest operacija koje stručnjaci koriste pri ručnom sažimanju, a dobivene su analizom ručno izrađenih sažetaka iz različitih domena. Te operacije su:

- **skraćivanje rečenica:** uključuje brisanje nepotrebnih riječi, fraza i umetnutih rečenica
- **kombiniranje rečenica:** pri čemu se više rečenica spaja u jednu, u nekim slučajevima se koristi u kombinaciji sa skraćivanjem rečenica, a ponekad uključuje i parafraziranje
- **sintaktička transformacija:** može se javiti u prve dvije operacije, npr. rečenični subjekt se seli s kraja na početak rečenice
- **leksičko parafraziranje:** događa se kada se fraze parafraziraju
- **generalizacija ili specifikacija:** javlja se kod zamjene fraza ili rečenica s općenitijim ili specifičnijim verzijama
- **izmjena redoslijeda:** podrazumijeva izmjenu redoslijeda rečenica u sažetku, npr. posljednja rečenica može postati prva, itd.

Sustav radi na sljedeći način: u prvoj fazi sustav za sažimanje uz pomoć ekstrakcije izlučuje ključne rečenice iz dokumenta nakon čega moduli za skraćivanje i kombiniranje rečenica implementiraju gore navedene operacije. Sustav, osim izlučenih ključnih rečenica, preuzima i izvorni dokument, a osim njih koristi se i drugim jezičnim resursima i alatima kao što su: korpus, program za automatsko rastavljanje tekstova, referentni sustav, WordNet i opširni leksikon dobiven kombinacijom više izvora.

5. Utjecaj konteksta i domene na automatsko sažimanje

U većini slučajeva, sustav za sažimanje ima pristup dopunskoj građi koja pomaže pri odabiru najvažnijih dijelova teksta unutar dokumenta (Nenkova i McKeown, 2011). Ova dopunska građa također pomaže kod određivanja konteksta i domene te pri računanju težine riječi. Dopunska građa može se javiti u raznim oblicima, a neki od njih su:

- razni korpusi i zbirke dokumenata vezani uz domenu u kojoj se nalazi dokument
- citati u znanstvenim radovima
- oznake kojima su označene poveznice i multimedijski sadržaji na mrežnim stranicama
- komentari na mrežnim stranicama i blogovima koji raspravljaju o najzanimljivijim dijelovima teksta
- korisnički upiti kod pretraživanja sadržaja.

5.1. Medicina i zdravstvo

Automatsko sažimanje izuzetno je korisno u domeni zdravstva te liječnicima omogućava lakši pristup relevantnim podacima iz liječničkog kartona pacijenta, a također olakšava pronalaženje relevantnih informacija o određenim bolestima ili lijekovima (Becher, Endres-Niggemeyer i Fichtner 2002). Sažetci također mogu biti od koristi za same pacijente koji informacije o svojim zdravstvenim problemima pretražuju na mrežnim stranicama (Kaicker et al., 2010). Nadalje, automatsko sažimanje od velike je pomoći kod pretraživanja velikih biomedicinskih baza podataka, kao što je MEDLINE, koji sadrži preko 20 milijuna članaka (Kumar et al., 2016).

Jedan od prvih sustava za automatsko sažimanje u domeni zdravstva bio je Centrifuser koji je sažetke slagao prema upitu korisnika. Pri tome su dokumenti prikazani uz pomoć strukture stabla (*engl. tree data structure*) iz čijih se grana izlučuju rečenice relevantne za korisnikov upit (Elhadad et al., 2005). Na prijedlog Fiszman et al. (2009) izgrađen je još jedan sustav za sažimanje koji generira apstraktne sažetke kako bi liječnicima pomogao pronaći najvažnije informacije o određenoj bolesti iz MEDLINE baze podataka.

Još jedan pristup automatskom sažimanju u medicinskoj domeni je uporaba ontologije. Ontologija omogućava međusobno povezivanje informacija na temelju njihovih zajedničkih svojstava (Khelif, Dieng Kuntz i Barbry, 2007). Jedan od sustava koji se koristi ovom tehnikom za sažimanje je ULMS (Verma, Chen i Lu, 2007) koji radi na principu uparivanja riječi iz sličnih rečenica.

5.2. Znanstveni članci

Postupak kojim se sažimaju znanstveni članci naziva se utjecajno sažimanje (*engl. impact summarization*). Pri odabiru rečenica iz rada koje će se koristiti u sažetku promatraju se radovi autora koji su citirali rad koji se želi sažeti te se na temelju najčešće citiranih dijelova rada identificiraju oni najvažniji (Nenkova i McKeown, 2011). Kako bi se u obzir uzela važnost svake rečenice u sažetku, ovaj pristup koristi izračun vjerojatnosti riječi iz izvornog dokumenta, a konačni proizvod, tj. sažetak znanstvenog članka, kombinacija je izračuna utjecaja te vjerojatnosti riječi.

5.3. Novinski članci

Najraniji sustav koji se bavio sažimanjem novinskih članaka bio je SUMMONS, a javio se 90-ih godina prošlog stoljeća (McKeown i Radev, 1995). Zadaća mu je bila sažimanje članaka o terorističkim napadima. Sagrađen je uz pomoć sustava za razumijevanje poruka upravljanog predlošcima (*engl. template-driven message*

understanding system) pod nazivom MUC-4 (Chinchor i Sundheim, 1992) koji bi prvo obradio odabrani tekst te potom popunio predložak pomoću kojega bi stvorio sažetak.

Na sličan način radio je i sustav RIPTIDES (White et al., 2001) koji se bavio sažimanjem članaka o prirodnim nepogodama. On je baš kao i SUMMONS sažetke stvarao uz pomoć predložaka popunjenih informacijama iz teksta, no pri tome sustav bi ih rasporedio u strukturu koja odgovara određenoj nepogodi. Nakon toga je svakoj od izlučenih rečenica unutar predloška pridodana ocjena koja predstavlja njezinu važnost, a rečenice s najvišim ocjenama iskorištene su za sažetak.

Za sažimanje novinskih izvješća na mrežnim stranicama razvijen je Newsblaster (McKeown, Barzilay i Blair-Goldensohn, 2002). Ovaj sustav pri sažimanju statističkim pristupom identificira rečenice koje se često pojavljuju u novinskim člancima. Rečenice se nakon toga spajaju u sažetak koristeći se pri tome strojnim učenjem (*engl. machine learning*).

Za generiranje sažetaka iz više članaka i izvješća i to u domeni prirodnih nepogoda, Li et al. (2010) predlažu korištenje ontologijski obogaćenog sažimanja. Sažetak, dobiven iz više članaka i izvješća, izrađen je prema upitu korisnika. Svaka od rečenica u tekstovima najprije je ontologijski obrađena u odgovarajućoj domeni, a rečenice koje najviše odgovaraju upitu koriste se u sažetku. Sličan koncept pod nazivom nejasna ontologija (*engl. fuzzy ontology*) predložili su Lee, Chen i Jian (2003) za stvaranje sažetaka vremenskih prognoza. Ovaj pristup pogodniji je za domene s nesigurnim ishodima.

Daniel, Radev i Allison (2003) istraživali su utjecaj sporednih događaja u novinskom članku na kvalitetu sažimanja. U njihovoj studiji ocjenjivači su imali zadatak iz članka odabrati rečenice koje pripadaju ranije određenom skupu sporednih događaja. Potom su uz pomoć tih rečenica algoritmom za sažimanje stvoreni sažetci. Na temelju rezultata autori su zaključili da uključivanje sporednih događaja u postupku sažimanja uvelike poboljšava kvalitetu dobivenog sažetka. Iz ovog razloga predloženo je koristiti sustave za automatsko grupiranje sporednih događaja.

Najveći doprinos za sažimanje novinskih članaka imali su Kumar et al. (2014) koji su u postupak odabira rečenica za sažetak uveli kontekstualna pitanja poput: „tko?“,

„što?“ i „kada?“. Zahvaljujući ovim pitanjima kvaliteta sažetaka novinskih članaka vidljivo se poboljšala.

5.4. Mrežne stranice i blogovi

Pri sažimanju mrežnih stranica svakako bi se u obzir trebale uzeti poveznice te dijelovi teksta koji sadrže informacije o poveznicama (Nenkova i McKeown, 2011). Poveznice najčešće sadrže oznake u kojima se nalazi opis sadržaja mrežne stranice ili nekih njezinih dijelova. Slične oznake moguće je pronaći i u multimedijskim sadržajima koji nerijetko čine veći dio stranice. Multimedijske stranice bez oznaka predstavljaju velik problem sustavima za sažimanje jer je tada izuzetno teško razlikovati dobar i kvalitetan sadržaj od lošijeg.

Kod sažimanja blogova prvenstveno se želi doći do autorovog mišljenja (Kumar et al., 2016). Pri sažimanju najčešće se koristi frekvencijski pristup, no zanimljivo je to što se u većini slučajeva koristi frekvencija riječi iz komentara, a ne autorove objave na blogu jer se smatra da će se u komentarima raspravljati najvažnije ideje u tekstu (Nenkova i McKeown, 2011). Rečenice koje su potaknule raspravu koristit će se u sažetku.

5.5. Elektronska pošta

Kako bi sažimanje elektronske pošte bilo uspješno, važno je da sustav za sažimanje prepoznaje karakteristike ovog načina komuniciranja te jedinstvenog lingvističkog žanra koji istovremeno ima svojstva i govornog i pisanog jezika (Nenkova i McKeown, 2011). Također je od izuzetne važnosti uzeti u obzir činjenicu da je elektronska pošta zapravo vrlo interaktivne prirode te sandučić obično sadrži više konverzacija između jedne ili više osoba kroz neko dulje vremensko razdoblje. Iz ovog razloga često nije moguće protumačiti značenje i smisao odgovora bez poruke na koju je odgovoreno.

U ranijim radovima koji su se bavili sažimanjem razgovora preko elektronske pošte isključivo su korišteni sustavi temeljeni na principu ekstrakcije (Nenkova i McKeown, 2011). Tako na primjer sustav koji opisuju McKeown, Shrestha i Rambow (2007) izlučuje pitanja i odgovore iz poruka. Newman i Blitzer (2003) opisuju sličan sustav koji najprije grupira poruke, a zatim rečenice iz tih poruka ocjenjuje na temelju njihovih svojstava nakon čega su stvoreni sažetci za svaku grupu.

5.6. Sažimanje prema upitu

Automatsko sažimanje prema upitu najlakše se može definirati kao sustav koji iz dokumenta sažima samo one informacije koje su relevantne za korisnikov upit (Chettri i Chakraborty, 2017). Pri odabiru rečenica, osim prisutnosti riječi iz upita, sustav u obzir uzima i njihovu važnost unutar konteksta u kojem se javljaju.

Za ovu vrstu sažimanja najčešće se koriste tehnike slične onima za sažimanje vijesti (Nenkova i McKeown, 2011) te pristupi temeljeni na grafovima (Otterbacher, Erkan i Radev, 2009). Ostali pristupi sažimanju prema upitu uglavnom su razvijeni za određenu vrstu upita. Tako npr. postoje sustavi specijalizirani za stvaranje biografskih sažetaka u kojima je upit ime osobe za koju se biografija želi izraditi. Stvaranje sažetaka prema upitu također je od izuzetne važnosti za mrežne pretraživače u svrhu stvaranja isječaka teksta kod prikaza rezultata pretraživanja (Nenkova i McKeown, 2011).

6. Istovremeno sažimanje više dokumenata

U nekim slučajevima javlja se potreba za istovremenim sažimanjem više datoteka. Dva najčešća pristupa za automatsko istovremeno sažimanje više dokumenata su metoda temeljena na klasterima i metoda temeljena na grafovima (Gupta i Lehal, 2010). Uz to Kumar et al. (2016) navode još jednu metodu koja se temelji na diskursu.

6.1. Metoda temeljena na klasterima

Ova metoda rečenice prema sličnosti grupira u klastere (*engl. cluster*) koji mogu predstavljati različite podteme unutar dokumenta (Kumar et al., 2016). Svaka od rečenica je potom ocjenjena prema važnosti, a iz svakog klastera u svrhu stvaranja sažetka izlučuje se po jedna rečenica s najvišom ocjenom.

Radev et al. (2004) koristili su ovu metodu u svom sustavu za sažimanje pod nazivom MEAD. Pri tome koriste takozvane centroide koji predstavljaju pojmove s visokom tf-idf mjerom koji su reprezentativni za klaster kojem pripadaju. Usporedbom rečenica iz klastera s pripadajućim centroidom uz pomoć kosinusne sličnosti najsličnije rečenice odabiru se za sažetak. Za razliku od prijašnjih sustava, MEAD ne koristi modul za generiranje jezika, a svi dokumenti kao model „vreće riječi“ (*engl. bag of words*) što sustav čini skalabilnim i neovisnim o domeni.

6.2. Metoda temeljena na grafovima

Kao što je već objašnjeno u poglavlju 4.1.3., sustavi koji koriste metodu temeljenu na grafovima pri analizi teksta izrađuju simbolički prikaz veza između riječi i rečenica u obliku grafa. Nakon toga se, u većini slučajeva, sličnost između rečenica računa uz pomoć kosinusne sličnosti (Erkan i Radev, 2004). Važnost rečenice računa se usporedbom te sličnosti.

Pristupi temeljeni na grafovima izuzetno su dobri u prepoznavanju važnih rečenica iz zbirke dokumenata, a dobiven graf također omogućuje razlikovanje tema unutar nepovezanih podgrafova (Kumar et al., 2016). No, ovakvi sustavi imaju veliki nedostatak, tj. u potpunosti ovise o sličnosti i ne razumiju značenje i odnos između rečenica koje se analiziraju.

6.3. Metoda temeljena na diskursu

Metoda temeljena na diskursu važnost rečenice određuje na temelju semantičkog odnosa između jezičnih jedinica (Kumar et al, 2016). Teorijski model koji na ovaj način pristupa sažimanju ustanovio je Radev (2000) pod nazivom teorija međudatotečne strukture (*engl. Cross-Document Structure Theory, CST*). U ovom modelu riječi, fraze i rečenice mogu biti povezane jedino ako između njih postoji semantička veza.

U starijim studijama smatralo se da je ova metoda izuzetno korisna za pronalaženje najvažnijih rečenica unutar zbirke dokumenata. U jednoj takvoj studiji (Blair-Goldensohn et al., 2002) najprije je stvoren sažetak uz pomoć sustava za sažimanje pod nazivom MEAD (Radev, Blair-Goldensohn i Zhang, 2001) nakon čega je od stručnjaka zatraženo da odrede CST odnose u rečenicama unutar zbirke. Konačno, rečenice s niskim ocjenama zamijenjene su s rečenicama koje imaju veću semantičku povezanost. Rezultat ovog pristupa bio je izuzetno koherentan sažetak.

Utjecaj ove metode na proces sažimanja također su istraživali Jorge i Pardo (2010). U svojoj studiji sažimali su novinske članke koje su prethodno anotirali stručnjaci, a CST odnosi korišteni su za pronalaženje ponavljanja, komplementarnosti i proturječnosti između rečenica. Međutim, velik nedostatak ove metode je činjenica da CST odnosi moraju biti ručno izrađeni.

Iz ovog razloga javio se niz pokušaja automatizacije stvaranja CST odnosa. Tako su ga, na primjer, Zahri i Fukumoto (2011) odredili uz pomoć klasifikatora strojne vektorske potpore (*engl. Support Vector Machine, SVM*). Pri tome je za utvrđivanje težine rečenica korišten PageRank algoritam, dok se usmjerenost (*engl. directionality*)

algoritma odredila uz pomoć CST odnosa. Najzad, na temelju ovih odnosa, povezane rečenice po potrebi su uređene kako bi se izbjeglo ponavljanje.

U srodnoj studiji, Kumar et al. (2012) predložili su korištenje Genetic-CBR klasifikatora za prepoznavanje semantičkih odnosa iz dokumenata koji nisu anotirani. Za rangiranje rečenica korištene su dvije tehnike: tehnika nejasnog rasuđivanja (*engl. fuzzy reasoning*) i tehnika modela glasanja (*engl. voting model*). Ove tehnike pritom koriste prethodno prepoznate CST odnose. Obje navedene studije pokazale su da sustavi koji koriste CST odnose pri sažimanju stvaraju kvalitetnije sažetke od onih koji se koriste metodama temeljenim na klasterima ili grafovima.

6.4. Sažimanje višejezičnih datoteka

Problem sažimanja dokumenata pisanih na više jezika spomenuli su Hovy i Lin (1998) a potom i McKeown, Klanvas i Evans (2005). Ovaj oblik sažimanja bio bi izuzetno koristan za stvaranje sažetka više novinskih izvora pisanih različitim jezicima. U svom radu McKeown, Klanvas i Evans (2005) razmatraju situaciju u kojoj se novinski članak želi sažeti na određenom jeziku, a za izradu sažetka dostupno je više datoteka pisanih tim jezikom, ali i onih koji su pisani drugim jezicima.

Za stvaranje ovakvog sažetka prvo je potrebno prevesti članke na ciljani jezik uz pomoć sustava za strojno prevođenje. Tekstovi se potom pretražuju kako bi se utvrdilo postoje li slične rečenice u tekstovima koji su prevedeni i onima na željenom jeziku, a ukoliko su takve rečenice pronađene one zamjenjuju automatski prijevod u sažetku kako bi te rečenice bile što više gramatički točne. Korištenje datoteka na stranom jeziku olakšava određivanje važnosti rečenica unutar teksta zbog činjenice da su na taj način sustavu dostupne dodatne informacije koje u nekim slučajevima nije moguće pronaći u tekstovima koji su pisani željenim jezikom.

7. Evaluacija automatskih sažetaka

Evaluacija sažetka najvažniji je zadatak sustava za automatsko sažimanje te je prijeko potrebna za određivanje točnosti i korisnosti sažetka, a s time i za poboljšavanje sustava za sažimanje. Evaluacija sažetaka kreće od ručnih do automatskih metoda, uključujući i sve između.

Metode evaluacije mogu se klasificirati u dvije skupine, a to su: ekstrinzična i intrinzična evaluacija (Rajkhowa, 2015). Ekstrinzična evaluacija ocjenjuje korisnost i iskoristivost u kontekstu zadatka za koji je sažetak izrađen (Haiduc, Aponte i Marcus, 2010). Ocjenjivanje je prema tome indirektno, a korisnost i iskoristivost sažetka uspoređuju se s izvornim dokumentom. Intrinzična pak evaluacija podrazumijeva procjenu kvalitete sažetka uz pomoć postavljenih normi. Ova metoda prvenstveno procjenjuje rad algoritma za rudarenje teksta (*engl. text mining*) kao izolirane jedinice unutar sustava za sažimanje. Intrinzična evaluacija može se podijeliti u dvije kategorije (Ježek i Steinberger, 2009): evaluacija sadržaja koja procjenjuje kvalitetu odabira ključnih podataka iz teksta te evaluacija kvalitete teksta koja prikuplja podatke o čitljivosti, gramatičkim pogreškama te koherentnosti sažetka.

Evaluacija sažetka nastalog ekstrakcijom puno je lakša od evaluacije onog koji je nastao apstrakcijom zbog činjenice da se rečenice iz ekstrakta jednostavno mogu usporediti s onima iz izvornog teksta jer rečenice pri ekstrakciji nisu izmijenjene već samo izlučene iz dokumenta. No, bez obzira na to, evaluacija sažetaka izuzetno je teška. Naime, osim činjenice da ne postoji točna definicija savršenog sažetka koja vrijedi za svaki dokument i svaku situaciju, osobito uzimajući u obzir činjenicu da je danas većina sustava za sažimanje usko specijalizirana za određeno područje ili domenu, evaluacija svojstava poput razumljivosti, dosljednosti i čitljivosti izuzetno je teška, čak i za ljudske ocjenjivače. Kod ručne evaluacije problem se javlja zbog subjektivnosti te svaki pojedinac ima vlastito mišljenje o tome kako dobar sažetak treba izgledati i što sve treba sadržavati. Provođenje ručne evaluacije također je izuzetno skupo i zahtjeva mnogo

vremena, pa je njeno provođenje u velikim razmjerima u većini slučajeva neizvedivo (Lin, 2004). Kod strojne evaluacije problemi se pak javljaju zbog nepostojanja standarda za evaluaciju automatskog sažimanja, što nije problem u ostalim područjima koja se bave obradom prirodnog jezika, npr. parsiranju (Das i Martins, 2007).

Baš iz ovog razloga pri evaluaciji koristi se mnoštvo različitih kriterija za evaluaciju sažetaka. Dva najpraktičnija kriterija koji se koriste pri evaluaciji su: preciznost (*engl. precision*) i odziv (*engl. recall*), a pomažu pri izračunu sličnosti između referentnog ručno izrađenog i automatski stvorenog sažetka (Gholamrezazadeh, Salehi i Gholamzadeh, 2010). Preciznost (P) se računa kao broj rečenica koje se javljaju u ručnom te automatskom sažetku podijeljen s ukupnim brojem rečenica u automatskom sažetku. Odziv (R) se pak računa kao ukupan broj rečenica u oba sažetka podijeljen s brojem rečenica u ručno izrađenom sažetku. Mjera pod nazivom *F-score* kombinira ova dva kriterija, a najjednostavnija formula za nju glasi:

F = (2*P*R*)/P+R

U nekim slučajevima se za evaluaciju sažetaka koriste još dva dodatna kriterija, a to su: omjer kompresije (CR) i omjer retencije (RR). Formule za njih glase:

CR = dužina sažetka / dužina izvornog teksta
RR = informacije u sažetku / informacije u izvornom tekstu

Dobar sažetak trebao bi imati niski omjer kompresije, dok bi omjer retencije trebao biti visok.

7.1. Ručna evaluacija

Ručna evaluacija danas je još uvijek najjednostavniji (Allahyari et al., 2017) i najčešći (Rajkhowa, 2015) način procjene kvalitete automatskih sažetaka. Uz pomoć ovog oblika evaluacije moguće je ocijeniti složenije karakteristike sažetka poput:

sintaktičke točnosti, semantičke koherentnosti, logičke organizacije i redundancije koje nije moguće ispitati uz pomoć sustava za automatsku evaluaciju. No, baš kao i kod evaluacije automatskih sažetaka, ne postoji općeprihvaćeni standard za ručnu evaluaciju te je na raznim konferencijama predloženo nekoliko različitih pristupa ručnoj evaluaciji koje su odigrale veliku ulogu u izradi evaluacijskih metoda. Sažetci su uvijek ocjenjivani ocjenama između 1 i 10.

Lin i Hovy (2002.) opisuju jedan od njih. Ovaj postupak evaluacije korišten je na Konferenciji za razumijevanje dokumenata (*engl. Document Understanding Conference, DUC*) 2001. godine gdje je korišteno sučelje za evaluaciju sažetaka (*engl. Summary Evaluation Envirnoment, SEE*) kao podrška ručnoj evaluaciji. Ocjenjivači iz američkog Nacionalnog instituta za standarde i tehnologiju (*engl. National Istitute of Standards and Technology, NIST*) imali su zadatak usporediti „idealne" ručno izrađene sažetke sa sažetcima koji su izrađeni uz pomoć sustava za automatsko sažimanje. Oba sažetka bila su rastavljena na jedinice od kojih je svaka predstavljala jednu rečenicu unutar sažetog teksta, a te jedinice bile su prikazane jedna pored druge unutar SEE sučelja. Ocjenjivači su uspoređivali međusobnu sličnost, gramatika, koherentnost i dosljednost jedinica, a kriterij odziva također je korišten u svrhu evaluacije. Na temelju rezultata dobivenih nakon završetka projekta izrađena je studija u kojoj su primijećena velika odstupanja između individualnih ocjenjivača (Das i Martins, 2007). Naime, ocjene dodijeljene usporedbom jedinica razlikovale su se u 18% slučajeva za sažetke nastale iz jednog dokumenta, dok su se ocjene za sažetke nastale iz više dokumenata razlikovale u 7.6% slučajeva. Također su zabilježena neslaganja između ocjenjivača oko stvaranja sažetaka. Oko 40% ocjenjivača nije se slagalo oko sažetaka nastalih iz jednog dokumenta, a u 29% slučajeva razilazila su se mišljena o sažetcima nastalim iz više dokumenata. Kako bi se izbjegla ovako velika neslaganja pri evaluaciji sažetaka Lin i Hovy (2002) predložili su korištenje automatske metrike inspirirane BLEU metrikom za evaluaciju strojnog prevođenja (Papineni et al., 2001) pod nazivom akumulirajući podudarni rezultati n-grama (*engl. accumulative n-gram matching score, NAMS*).

Jedna od novijih metoda predstavljena je na Konferenciji za analizu teksta (*engl. Text Analysis Conference, TAC*) 2008. godine na kojoj su evaluirani sustavi za sažimanje

prema upitu (Dang i Owczarzak, 2008). Nakon zadavanja upita sustavu, ocjenjivači su imali zadatak evaluirati kvalitetu nastalog sažetka. 2009. godine na konferenciji ručno izrađeni referentni sažetci ostvarili su prosječnu ocjenu 8.8/10 zbog čega se ona smatra gornjom granicom koju sažetci mogu postići (Saggion et al., 2013).

7.2. Strojna evaluacija

S početkom u ranim 2000-im predložen je niz metoda za automatiziranje evaluacije automatskih sažetaka (Radev et. al., 2003). Većina ovih metoda temeljena je na usporedbi referentnih i automatski izrađenih sažetaka. Danas se za strojnu evaluaciju najčešće koriste dva pristupa: ROGUE i PYRAMID (Saggion et al. 2013).

7.2.1. ROGUE

Skup mjera za strojnu evaluaciju automatskih sažetaka pod nazivom ROGUE (*engl. Recall Oriented Understudy for Gisting Evaluation)* prvi je predstavio Lin (2004). Kvaliteta sažetka se izračunava usporedbom ručno izrađenih referentnih sažetaka s onima koji su nastali uz pomoć sustava za automatsko sažimanje uz pomoć *n-grama* (niza elemenata *n*). Uobičajeno je da se koristi više od jednog referentnog sažetka što omogućava veću fleksibilnost i pravedniju evaluaciju. Ova metoda izuzetno je uspješna u evaluacija sažetaka nastalih ekstrakcijom, a javlja se u nekoliko varijanti:

- ROGUE-n: temelji se na mjeri odaziva pri čemu se promatra broj preklapanja *n-grama* u referentnim i automatskim sažetcima;
- ROGUE-L: mjeri vrijednost najduljeg pod-niza zajedničkog i referentnom i automatski stvorenom sažetku, a temelji se na pretpostavci da što je zajednički niz dulji, to je sličnost veća;
- ROGUE-SU: ova metoda u obzir uzima tokene koji su unigrami i bigrami te omogućava ubacivanje riječi između prve i druge riječi u bigramu što omogućava stvaranje tokena iz riječi koje nisu uzastopne.

7.2.2. PYRAMID

PYRAMID je kao skup mjera za evaluaciju sažetaka nastao zbog nemogućnosti ROGUE sustava da ocjeni sažetke nastale abstrakcijom (Saggion et al., 2013). Naime, ROGUE sažetke ocjenjuje tako da u automatskom sažetku traži rečenice i fraze iz referentnog sažetka. Ukoliko ih ne pronalazi, sažetak ocjenjuje niskom ocjenom. Kako sažetci nastali abstrakcijom ne sadrže rečenice preuzete izravno iz dokumenta, ROGUE sustav ih ne može pravilno ocijeniti.

PYRAMID sustav, za razliku od njega, prvo iz referentnih sažetaka izlučuje važne informacije, a to se može izvršiti ili automatski ili uz pomoć stručnjaka. Ove informacije nazivaju se jedinicama sažimateljskog sadržaja (*engl. Summarization Content Units, SCU*). Nakon izlučivanja informacijama pridaje se težina, a ocjenu dobivaju s obzirom na to u koliko se referentnih sažetaka pojavljuju. Nakon ovog koraka vidljiva je analogija s piramidom, naime, SCU s niskim ocjenama čini temelj piramide, a nekolicina njih ocjenjena visokim ocjenama čine njen vrh. Potom je svaki SCU povezan je s listom lingvističkih izraza. U posljednjem koraku sustav izračunava broj i težinu SCU koji se javljaju u sažetku a koji se ocjenjuje, te mu se na temelju njih dodjeljuje konačna ocjena.

8. Zaključak

Automatsko sažimanje teksta danas je od neosporive važnosti za znanstvenu zajednicu, ali i za sve pojedince koji su u potrazi za relevantnim informacijama. Upotreba sažetaka nastalih na ovaj način također je izuzetno raširena u svakodnevnom životu. Sustave za automatsko sažimanje koriste gotovo svi mrežni pretraživači te mnoge mrežne stranice na kojima se najčešće sažimaju novinski članci.

Ekstrakcija je daleko najraširenija metoda koja se koristi za stvaranje automatskih sažetaka, a razlog tome je relativno jednostavna implementacija te jednostavniji algoritmi za sažimanje u usporedbi s onima koji se koriste pri apstrakciji. Ovakvi sustavi jednostavno izlučuju rečenice iz dokumenta uz pomoć raznih metoda za određivanje važnosti rečenica. Sustavi za stvaranje apstraktnih sažetaka koji relevantne informacije prikazuju u drugačijem obliku od onoga koji se javlja u originalnom dokumentu još uvijek su u razvoju prvenstveno zbog njihove ovisnosti o tehnologijama obrade prirodnog jezika.

Napredak jezičnih tehnologija općenito ima veliki utjecaj na razvoj sustava za automatsko sažimanje jer se radi o tekstovima pisanim prirodnim jezikom, a na njihov razvoj utječe još niz znanstvenih područja i disciplina kao što su prikupljanje informacija, umjetna inteligencija, strojno prevođenje itd. Sustavi za evaluaciju nastalih sažetaka također su od izuzetne važnosti za poboljšanje rada sustava za sažimanje, no važno je napomenuti da je evaluacijske sustave potrebno prilagoditi vrsti sažetka te njenoj domeni i kontekstu kako bi konačna ocjena bila što točnija, a kako bi ocjenjivanje bilo ravnopravno, potrebno je razviti međunarodne standarde za evaluaciju.

Daljnji razvoj sustava za automatsko sažimanje teksta sigurno će imati još pozitivnih utjecaja na društvo te je iz ovog razloga izuzetno važno da zemlje i organizacije ulažu u napredak samih sustava, ali i srodnih disciplina.

9. Literatura

1. Aliguliyev, R. M. (2009) A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Systems with Applications, 36(4), str. 7764-7772.
2. Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E. D.; Gutierrez, J. B.; Kochut, K. (2017) Text Summarization Techniques: A Brief Survey. International Journal of Advanced Computer Science and Application, 8(10), str. 397-405.
3. Aone, C.; Okurowski, M.; Gorlinsky, J.; Larsen, B. (1990) A trainable summarizer with knowledge acquired from robust NLP techniques. U: Mani, I. (1999) Advances in Automatic Text Summarization. Cambridge: The MIT Press.
4. Becher, M.; Endres-Niggemeyer B.; Fichtner G. (2002) Scenario forms for web information seeking and summarizing in bone marrow transplantation. U: Proceedings of the Conference on Multilingual Summarization and Question Answering. Stroudsburg: Association for Computational Linguistics, str. 1-8. URL: https://dl.acm.org/citation.cfm?id=1118846 (23-8-2018)
5. Blair-Goldensohn, S.; Radev, D.R.; Zhang, Z. (2001) Experiments in single and multi-document summarization using MEAD. Ann Arbour, str. 48-109.
6. Blair-Goldensohn, S.; Radev, D.R.; Zhang, Z. (2002) Towards CST-enhanced summarization. U: Proceedings of the 18th National Conference on Artificial Intelligence. Edmonton: American Association for Artificial Intelligence Menlo Park, str. 439-445. URL: http://tangra.cs.yale.edu/~radev/si/papers/aaai02.pdf (20-8-2018)
7. Blitzer, J. C; Newman, P. S. (2003) Summarizing archived discussions: a beginning. U: Proceedings of the 8th international conference on Intelligent user interfaces. New York: ACM.
8. Chettri, R; Chakraborty, U. (2017) Automatic Text Summarization. International Journal of Computer Applications, 161(1). URL:

https://www.ijcaonline.org/archives/volume161/number1/chettri-2017-ijca-912326.pdf (15-8-2018)

9. Chinchor, N. i Sundheim, B. (1993) MUC-5 evaluation metrics. U: Proceedings of the 5th conference on Message understanding. Stroudsburg: Association for Computational Linguistics, str. 69-78. URL: https://dl.acm.org/citation.cfm?id=1072026 (24-8-2018)
10. Conroy, J. M.; O'leary, D. P. (2001) Text summarization via Hidden Markov Models. U: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, str. 406-407.
11. Dang, H. T.; Owczarzak, K. (2008): Overview of the tac 2008 opinion question answering and summarization tasks. U: Proceedings of the TAC 2008 Workshop. Gaithersburg: National Institute of Standards and Technology. URL: https://tac.nist.gov/publications/2008/additional.papers/update_summ_overview08.proceedings.pdf (12-8-2018)
12. Daniel, N., D. Radev and T. Allison, 2003. Sub-event based multi-document summarization. U: Proceedings of the HLT-NAACL on Text Summarization Workshop. Stroudsburg: Association for Computational Linguistics, str. 9-16. URL: https://dl.acm.org/citation.cfm?id=1119469 (24-8-2018)
13. Das, D.; Martins A. F. T. (2007) A Survey on Automatic Text Summarization. Language Technologies Institute. URL: http://www-cgi.cs.cmu.edu/afs/cs.cmu.edu/Web/People/afm/Home_files/Das_Martins_survey_summarization.pdf (18-8-2018)
14. Edmundson, H. P. (1969) New Methods in Automatic Extracting. Journal of the ACM, 16(2), str. 264-285.
15. Elhadad, N.; Kan, M.-Y.; Klavans, J. L.; McKeown, K. R. (2005) Customization in a unified framework for summarizing medical literature. Artificial Intelligence in Medicine, 33(2), str. 179-198.

16. Erkan, G.; Radev, D. R. (2004) LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research. 22 (1), str. 457-497.
17. Fiszman, M.; Mishira, R.; Bian, J.; Weir, C. R.; Jonnalagadda, S.; Mostafa, J.; Del Fiol, G. (2014) Text summarization in the biomedical domain: A systematic review of recent research. Journal of Biomedical Informatics, 52, str. 457-467.
18. Gholamrezazadeh, S.; Salehi, M. A.; Gholamzadeh, B. (2009) A Comprehensive Survey on Text Summarization Systems. Computer Science and its Applications. URL: https://www.researchgate.net/profile/Mohsen_Salehi2/publication/262561680_Taxonomy_of_Contention_Management_in_Interconnected_Distributed_Systems/links/567c407c08aebccc4e011aa7.pdf (11-8-2018)
19. Gupta, V.; Lehal G.S. (2010) A survey of text summarization extractive techniques. Journal of Emerging Technologies in Web Intelligence, 2(3), str. 258-268.
20. Haiduc, S.; Aponte, J.; Marcus A. (2010) On the Use of Automated Text Summarization Techniques for Summarizing Source Code. U: Proceedings of the 2010 17th Working Conference on Reverse Engineering. DC: IEEE Computer Society Washington, str. 35-44. URL: https://www.academia.edu/7329658/On_the_Use_of_Automated_Text_Summarization_Techniques_for_Summarizing_Source_Code (16-8-2018)
21. Hovy, E.; Lin, C.-Y. (1998) Automated text summarization and the SUMMARIST system. U: Proceedings of a workshop on held at Baltimore, Maryland. Stroudsburg: Association for Computational Linguistics, str. 197-214. URL: https://dl.acm.org/citation.cfm?id=1119121 (20-8-2018)
22. Ježek, K.; Steinberger, J. (2008) Automatic Text Summarization (The state of the art 2007 and new challenges). Znalosti, 30(2), str. 1-12.
23. Ježek, K.; Steinberger, J. (2009) Evaluation measures for Text Summarization. Computing and Informatics, 28, str. 1001-1026.

24. Jin, R.; Hauptmann, A. G. (2000) Title Generation for Spoken Broadcast News using a Training Corpus. Sixth International Conference on Spoken Language Processing. URL: https://www.isca-speech.org/archive/icslp_2000/i00_2680.html (20-8-2018)
25. Jing, H. (2000) Sentence reduction for automatic text summarization. U: Proceedings of the sixth conference on Applied natural language processing. Stroudsburg: Association for Computational Linguistics, str. 310-315. URL: https://dl.acm.org/citation.cfm?id=974190 (20-8-2018)
26. Jing, H.; McKeown, K. R. (2000) Cut and paste based text summarization. U: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. Stroudsburg: Association for Computational Linguistics, str. 178-185. URL: https://dl.acm.org/citation.cfm?id=974329 (20-8-2018)
27. Jorge, M. L. D. R. C.; Pardo, T.A.S. (2010) Experiments with CST-based multidocument summarization. U: Proceedings of the Workshop on Graph-Based Methods for Natural Language Processing. Stroudsburg: Association for Computational Linguistics, str.74-82. URL: https://dl.acm.org/citation.cfm?id=1870490.1870502 (23-8-2018)
28. Jurafsky, D.; Martin, J. H. (2009) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. 2. izd. New Jersey: Prentice Hall.
29. Kaicker, J.; Bebono, V. B.; Dang, W.; Buckley, N.; Thabane, L. (2010) Assessment of the quality andvariability of health information on chronic pain websites using the DISCERN instrument. BMC Medicine, 8(59), str. 1-8.
30. Khelif, K; Dieng-Kuntz, R.; Barbry, P. (2007) An Ontology-based Approach to Support Text Mining and Information Retrieval in the Biological Domain. An Ontology-based Approach to Support Text Mining and Information Retrieval in the Biological Domain, 13(12), str. 1881-1907.

31. Klassen, P. P. (2012) Calculating LLR topic signatures with dependency relations for automatic text summarization. Magistarski rad. Seattle: University of Washington.
32. Knight, K.; Marcu, D. (2000) Statistics-based summarization – step one. U: Proceeding of the 17th National Conference of the American Association for Articial Intelligence. Palo Alto: AAAI Press, str. 703-710. URL: https://www.aaai.org/Papers/AAAI/2000/AAAI00-108.pdf (18-8-2018)
33. Ko, Y.; Seo, J. (2008) An effective sentence extraction technique using contextual information and statistical approaches for text summarization. Patter Recognition Letters, 29(9), str. 1366-1371.

    Kumar, Y. J.; Salim, N.; Abuobieda, A. (2012) A Genetic-CBR Approach for Cross-Document Relationship Identification. U: Advanced Machine Learning Technologies and Applications. Cairo: Springer, str. 182-192.
34. Kumar, Y. J.; Salim, N.; Abuobieda, A.: Albaham, A. T. (2014) Multi document summarization based on news components using fuzzy cross-document relations. Applied Soft Computing, 21, str. 265-279.
35. Kumar, Y. J.; Goh, O. S.; Basiron, H.; Choon, N. H.; Suppiah, P. C. (2016) A Review on Automatic Text Summarization Approaches. Journal of Computer Science, 12(4), str. 178-190.
36. Kupiec, J.; Pedersen, J.; Chen, F. (1995) A trainable document summarizer. U: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, str. 68-73. URL: https://www.cis.upenn.edu/~nenkova/Courses/cis430/trainableSummarizer.pdf (12-8-2018)
37. Lee, C.; Chen, Y.; Jian, Z. (2003) Ontology-based fuzzy event extraction agent for Chinese e-news summarization. Expert Systems with Applications, 25(3), str. 431-447.
38. Li, L.; Wu, K.; Li, J.; Li, T. (2010) Ontology enriched multi-document summarization in disaster management. U: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval.

New York: ACM, str. 819-820. URL: https://dl.acm.org/citation.cfm?id=1835632 (24-8-2018)

39. Lin, C.-Y. (1999). Training a selection function for extraction. U: Proceedings of CIKM ’99, New York: ACM Press, str. 55–62Lin, C.-Y. (2004) Rouge: A package for automatic evaluation of summaries. U: Proceedings of the ACL-04 Workshop. Stroudsburg: Association for Computational Linguistics, str. 74-81. URL: https://www.aclweb.org/anthology/W04-1013 (10-8-2018)
40. Lin, C.-Y.; Hovy, E. (2002) Manual and automatic evaluation of summaries. U: Proceedings of the ACL-02 Workshop on Automatic Summarization. Stroudsburg: Association for Computational Linguistics, str. 45-51. URL: https://dl.acm.org/citation.cfm?doid=1118162.1118168 (11-8-2018)
41. Luhn, H. P. (1958) A Business Intelligence System. IBM Journal of Research and Development, 2(4), str. 314-319.
42. Mani, I. (1999) Advances in Automatic Text Summarization. Cambridge: The MIT Press.
43. McKeown, K.; Barzilay, R.; Blair-Goldensohn (2002) The Columbia Multi-Document Summarizer for DUC 2002. Workshop on Automatic Text Summarization. URL: https://duc.nist.gov/pubs/2002papers/columbia_hatzivass.pdf (24-8-2018)
44. McKeown, K.; Klavans, J.; Evans, D. K. (2005). Similarity-based multilingual multi-document summarization: Technical Report CUCS-014-05. New York: Columbia University.
45. McKeown, K.; Radev, D. R. (1997) Generating Natural Language Summaries from Multiple On-Line Sources. Computational Linguistics, 24(3), str. 470-500.
46. McKeown, K.; Shrestha, L., Rambow, O. (2007) Using Question-Answer Pairs in Extractive Summarization of Email Conversations. CICLing 2007: Computational Linguistics and Intelligent Text Processing. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.421.5859&rep=rep1&type=pdf (25-8-2018)

47. Mihalcea, R.; Tarau, P. (2005) An Algorithm for Language Independent Single and Multiple Document Summarization. U: Natural language processing, IJCNLP 2005 : second international joint conference, Jeju Island, Korea, October 11-13, 2005 : proceedings. Berlin: Springer. URL: https://www.aclweb.org/anthology/I05-2004 (21-8-2018)
48. Nenkova, A.; McKeown, K. (2011) A Survey of Text Summarization Techniques. Foundations and Trends in Information Retrieval 52(2-3), str. 103-233. of the ACL'02 Workshop on Automatic Summarization. URL: https://dl.acm.org/citation.cfm?id=1118163 (21-8-2018)
49. Osborne, M. (2002) Using maximum entropy for sentence extraction. U: Proceedings of the ACL-02 Workshop on Automatic Summarization. Stroudsburg: Association for Computational Linguistics, str 1-8. URL: https://dl.acm.org/citation.cfm?doid=1118162.1118163 (20-8-2018)
50. Otterbacher, J.; Erkan, G.; Radev, D. (2009) Biased lexrank: Passage retrieval using random walks with question-based priors. Information Processing and Management, 45, str. 42–54.
51. Papineni, K.; Roukos, S.; Ward, T. (2001) Bleu: a method for automatic evaluation of machine translation. U: Proceedings of ACL '02. Stroudsburg: Association for Computational Linguistics, str. 311-318. URL: https://www.aclweb.org/anthology/P02-1040.pdf (11-8-2018)
52. Radev, R. D.; Jing, H.; Styś, M.; Tam, D. (2004) Centroid-based summarization of multiple documents. Information Processing & Management, 40(6), str. 919-938.
53. Radev, D. R. (2000) A common theory of information fusion from multiple text sources step one: Crossdocument structure. U: Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue. Stroudsburg: Association for Computational Linguistics, str. 74-83. URL: https://dl.acm.org/citation.cfm?id=1117745 (11-8-2018)
54. Rajkhowa, A. (2015) Automatic Text Summarization. Advances in Computer Science and Information Technology, 2(10), str. 100-103.

55. Saggion, P.; Poibeau, T. (2013) Automatic text summarization: Past, present and future. In Multi-source, Multilingual Information Extraction and Summarization, Berlin: Springer.
56. Svore, K.; Vanderwende, L.; Burges, C. (2007) Enhancing single-document summarization by combining RankNet and third-party sources. U: Proceedings of the EMNLP-CoNLL. Prague: The Association of Computational Linguistics, str. 448–457.
57. Torres-Moreno, J. M. (2014) Automatic Text Summarization: Cognitive Science and Knowledge Management. 1.izd. London: Wiley-ISTE.
58. Verma, R.; Chen, P.; Lu, W. (2007) A Semantic Free-text Summarization System Using Ontology Knowledge, IEEE Transactions on Information Technology in Biomedicine, 5(4), str. 261-270.
59. White, M.; Korelsky, T.; Cardie, C; Ng, V.; Pierce, D.; Wagstaff, K. (2001) Multidocument summarization via information extraction. U: Proceedings of the first international conference on Human language technology research. Stroudsburg: Association for Computational Linguistics, str. 1-7. URL: https://dl.acm.org/citation.cfm?id=1072206 (24-8-2018)
60. Witbrock, M.; Mittal V (1999) Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries. U: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, str. 315-316. URL: http://www.di.ubi.pt/~jpaulo/competence/papers/(1999)Witbrock&Mittal.pdf (16-8-2018)
61. Zahri, N. A. H. B.; Fukumoto, F. (2011) Multi-document summarization using link analysis based on rhetorical relations between sentences. U: Computational Linguistics and Intelligent Text Processing. Berlin: Springer Science and Business Media.
62. Zajic, D.; Dorr, B. J.; Lin, J.; Schwartz, R. (2007) Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. Information Processing and Management, 43(6), str. 1549.-1570.