Sažimanje teksta kao koristan alat u organizaciji, predstavljanju i pretraživanju informacija

Šerić, Marta

Sveučilište u Zadru

Odjel za informacijske znanosti
Preddiplomski sveučilišni studij Informacijske znanosti

Marta Šerić

Sažimanje teksta kao koristan alat u organizaciji, predstavljanju i pretraživanju informacija

Završni rad

Zadar, 2023.

Sveučilište u Zadru

Odjel za informacijske znanosti
Preddiplomski sveučilišni studij Informacijske znanosti

Sažimanje teksta kao koristan alat u organizaciji, predstavljanju i pretraživanju informacija

Završni rad

Studentica:
Marta Šerić

Mentor:
Doc. dr. sc. Ante Panjkota

Zadar, 2023.

Izjava o akademskoj čestitosti

Ja, **Marta Šerić,** ovime izjavljujem da je moj **završni** rad pod naslovom **Sažimanje teksta kao koristan alat u organizaciji, predstavljanju i pretraživanju informacija** rezultat mojega vlastitog rada, da se temelji na mojim istraživanjima te da se oslanja na izvore i radove navedene u bilješkama i popisu literature. Ni jedan dio mojega rada nije napisan na nedopušten način, odnosno nije prepisan iz necitiranih radova i ne krši bilo čija autorska prava.

Izjavljujem da ni jedan dio ovoga rada nije iskorišten u kojem drugom radu pri bilo kojoj drugoj visokoškolskoj, znanstvenoj, obrazovnoj ili inoj ustanovi.

Sadržaj mojega rada u potpunosti odgovara sadržaju obranjenoga i nakon obrane uređenoga rada.

Zadar, 10. svibnja 2023.

Sadržaj

1 UVOD 1

2 AUTOMATSKO SAŽIMANJE TEKSTA 2

2.1 METODE AUTOMATSKOG SAŽIMANJA TEKSTA 3

2.1.1 Ekstrakcijska metoda 4

2.1.2 Apstrakcijska metoda 4

2.1.3 Hibridna metoda 5

2.2 KLASIFIKACIJA NA TEMELJU VELIČINE UNOSA 7

2.3 OSTALE PODJELE AUTOMATSKOG SAŽIMANJA TEKSTA 7

3 PRIMJENA SAŽIMANJA TEKSTA U PODRUČJU INFORMACIJSKIH ZNANOSTI 10

3.1 ORGANIZACIJA INFORMACIJA I SAŽIMANJE TEKSTA 11

3.2 ULOGA AUTOMATSKOG SAŽIMANJA TEKSTA U PREDSTAVLJANJU INFORMACIJA I PREZENTACIJI INFORMACIJSKIH SADRŽAJA 13

3.3 AUTOMATSKO SAŽIMANJE TEKSTA U PODRUČJU PRETRAŽIVANJA INFORMACIJA 14

4 ALATI ZA SAŽIMANJE TEKSTA 17

4.1 ONLINE ALATI ZA AUTOMATSKO SAŽIMANJE TEKSTA 17

4.2 EVALUACIJA KREIRANIH SAŽETAKA 19

5 PROBLEMI KOD AUTOMATSKOG SAŽIMANJA TEKSTA 21

5.1 REDUNDANCIJA 21

5.2 FINANCIJSKA ULAGANJA 22

5.3 KOHEZIJA I KOHERENTNOST 22

5.4 DVOSMISLENE RIJEČI I SINONIMI 23

6 KRATKI OSVRT NA TRENDOVE U AUTOMATSKOM SAŽIMANJU TEKSTA 25

7 ZAKLJUČAK 27

8 POPIS LITERATURE 28

9 PRILOZI 36

9.1 POPIS SLIKA 36

Sažetak

Predmet rada je automatsko sažimanje teksta u kontekstu organizacije, predstavljanja i pretraživanja informacija. Automatsko sažimanje teksta primjenjivo je u brojnim područjima, a u radu se razmatra uloga sažimanja teksta u području informacijskih znanosti. Predstavljeni su različiti pogledi na sažimanje teksta, te su opisane osnovne metode sažimanja teksta. Osim dviju glavnih metoda, ekstrakcijske i apstrakcijske, spomenute su hibridna metoda kao i najaktualnije metrike evaluacije sažetaka. U radu su istaknuti nedostaci metoda sažimanja teksta i samog procesa sažimanja, a to su redundancija, financijska ulaganja, dvosmislenost riječi i sinonimi, te problem kohezije i koherencije kreiranih sažetaka. Automatsko sažimanje teksta korisno je u obradi tekstualnih datoteka, što je pogotovo važno uz eksponencijalni rast količine takvih podataka danas. Krajnji cilj područja sažimanja teksta je kreiranje kvalitetnih sažetaka sličnim onima kreiranim od strane čovjeka ali bez potrebe nadziranja procesa sažimanja. Naglašena je važnost evaluacije sažimanja teksta i mogućnost napretka u budućnosti po pitanju kvalitete sažetaka kreiranih alatima za sažimanje teksta uslijed razvoja NLP-a i umjetne inteligencije općenito.

Ključne riječi: Sažimanje teksta, ekstrakcijska metoda sažimanja, apstrakcijska metoda sažimanja, pretraživanje i predstavljanje informacija, organizacija podataka i informacija

1 Uvod

Nepobitno je kako se tekstualni sadržaj na internetu, kao što su časopisi, znanstveni članci, mailovi, pravni dokumenti i ostalo, eksponencijalno povećava svakog dana (El-Kassas, Salama, Rafea i Mohamed, 2021). Međutim, količina tekstualnog sadržaja može otežavati korisnikovu potragu za relevantnim i traženim informacijama. Da bi došao do traženih rezultata pretrage, korisnik se često susreće s mnogo neupotrebljivog i irelevantnog tekstualnog sadržaja (Nazari i Mahdavi, 2019). Težnja za što lakšim i bržim korištenjem i pretraživanjem informacija ukazala je na potrebu za sortiranjem i sažimanjem tekstova. Ručno sortiranje i organizacija cijelog internetskog sadržaja nije moguća zbog količine vremena, truda i troškova koji bi bili potrebni u njenoj realizaciji (El-Kassas, Salama, Rafea i Mohamed, 2021). Stoga se počinju primjenjivati računala, koja su se zbog svoje brzine pokazala kao odličan način automatizacije u organizaciji i pohrani informacija. Automatsko sažimanje teksta proces je kojim se krati originalan tekst ali sve glavne teme i ključne informacije ostaju iste. Osim kraćenja vremena potrebnog za čitanje i kraćom potragom za informacijama, sve informacije na određenu temu postaju uredno okupljene (Abualigah, Bashabsheh, Alabool i Shehab, 2020). Automatsko sažimanje teksta jednostavan je i snažan alat u obradi tekstualnih podataka, a primjenjiv je u medicini, ekonomiji, financijama, sportu, meteorologiji, pretraživanju i pronalaženje informacija. Iako je koristan i važan alat u potrazi za novim sadržajima, proces automatskom sažimanja teksta susreće brojne probleme i nedostatke.

Svrha rada jest razmotriti ulogu automatskog sažimanja teksta kao alata u području organizacije, predstavljanja i pretraživanja informacija.

Cilj rada jest prikazati tijek razvoja automatskog sažimanja teksta, objasniti metode automatskog sažimanja teksta, njegovu primjenu u organizaciji, predstavljanja i pretraživanju informacija, te navesti postojeće probleme kao i trendove.

Rad je podijeljen u sedam cjelina. Nakon uvoda slijedi cjelina automatsko sažimanje teksta u kojoj se predstavlja proces skraćivanja tekstualnih dokumenata pomoću raznih metoda. Slijedi cjelina primjene sažimanja teksta u području informacijskih znanosti i glavna poglavlja koja spajaju automatsko sažimanje teksta s organizacijom, predstavljanjem i pretraživanjem informacija. Nakon toga su nabrojani neki od alata za sažimanje teksta, te primjeri pomoću online alata za sažimanje teksta. Na kraju se nalaze cjeline o problemima s kojima se možemo susresti prilikom korištenja alata za sažimanje teksta i trendovi kretanja tehnologije automatskog sažimanja teksta u kojoj su spomenuta buduća očekivanja ovog područja, dok zaključak donosi sumarno najvažnije nalaze ovog rada.

2 Automatsko sažimanje teksta

Autori Gambhir i Grupta (2017) automatsko sažimanje teksta objašnjavaju kao kreiranje sažetka koji je dosljedan originalnom cjelovitom tekstu, sastoji se od manje riječi i sadrži sve ključne riječi i fraze kao i izvorni tekst. Smatraju da taj alat, koji se razvijao u posljednjih pedesetak godina, ima mnogo potencijala u brojnim segmentima, a posebice u organizaciji informacija na internetu.

Radev, Hovy i McKeown (2002) ističu da sažeti tekst može biti dvostruko kraći od izvornog teksta. Sažimanje teksta doprinosi smanjivanju broja nebitnih podataka nekog teksta, to omogućava brže čitanje i jednostavnije razumijevanje važnih koncepata, a to je i više nego potrebno u svijetu prezasićenom informacijama. Radev, Hovy i McKeown, (2002) također spominju da je automatsko sažimanje teksta nastalo primjenom stečenog znanja iz područja obrade prirodnog jezika (eng. Natural Language Processing - NLP), te smatraju da je za razumijevanje sažimanja teksta potrebno predznanje iz područja NLP-a.

Prijašnji napreci u automatskom sažimanju teksta bili su ograničeni zbog nedostatka dovoljno snažnih računala te problema s obradom prirodnih jezika. Sažeti tekstovi su također lakši za prevođenje jer zahtijevaju manje vremena od cjelovitih tekstova.

Sažimanje tekstova povećava našu efikasnost u potrazi za najbitnijim informacijama (Petrović i Bušelić, 2020). Dok pretražujemo podatke, više ćemo izvora provjeriti pregledavajući sažete radove jer ćemo jasno i sažeto pročitati o čemu se radi, koje su ključne riječi, teme i problematika rada. Time ćemo uštedjeti vrijeme i sama potraga za informacijama će nam biti olakšana.

Sažimanje teksta je proces automatskog stvaranje jednog vida komprimirane verzije zadanog teksta koji pruža korisne informacije (Fattah, 2014). Sažimanjem teksta određuju se najvažniji dijelovi koji nude ključne informacije o nekoj temi.

Sažimanje teksta, pored očite primjene u prezentiranju informacija i znanja, te pretraživanju ima bitnu ulogu u organizaciji informacija. Ona podrazumijeva obradu i uređenje podataka i informacija na način da budu jednostavnija za upotrebu i interakciju s korisnicima. Dobrom organizacijom informacija korisnici će sigurno biti zadovoljniji. Organizacija informacija podrazumijeva filtriranje dvostrukih fraza i irelevantnih informacija, te značajno smanjuje suvišne sažetke u rezultatima pretrage. Sažimanje teksta u organizaciji informacija korisno je u stvaranju sažetaka knjiga u digitalnim repozitorijima ili web stranica knjižnica kao i u bilo kojim drugim većim bazama podataka, kao osnovnim dijelovima informacijskih sustava. Autori dokumenata često izrade loš sažetak koji korisniku ne pomaže dokučiti o čemu se piše u dokumentu i koje su ključne riječi. Nije neobično ni da se sažetak dokumenta ne izradi

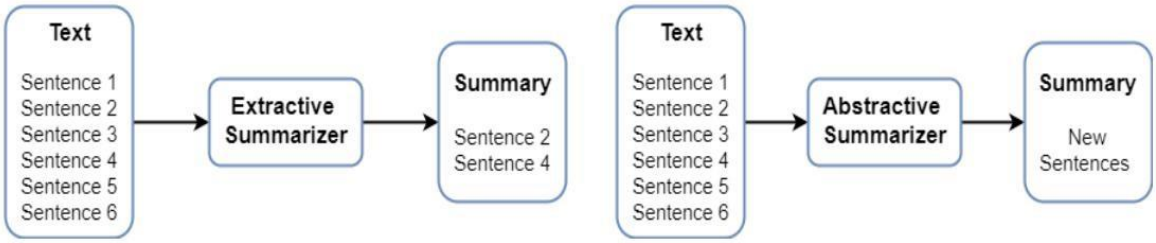
pa taj nedostatak automatsko sažimanje teksta direktno nadomješta. Sažimanje teksta je vrlo aktivno područje istraživanja i zato raste njegova primjena u sprječavanju širenja dezinformacija, smanjivanje redundancije i pomoć pri pronalaženju korisnih izvora informacija i podataka (Neto, Freitas i Kaestner, 2002). Korištenjem ove metode omogućena je bolja organizacija informacija tekstualnih dokumenata, njihovo jasnije i jednostavnije predstavljanje, te učinkovitiji pristup istima.

2.1 Metode automatskog sažimanja teksta

Tijekom proteklih nekoliko desetljeća razvijeno je mnogo metoda, a sve su imale isti cilj, a to je pokušaj kreiranja sažetaka što sličnijim onim izrađenim od strane čovjeka. Unatoč dosadašnjem trudu, istraživači još uvijek nisu postigli taj cilj (El-Kassas, Salama, Rafea i Mohamed (2021).

Isti autori opisali su trenutnu situaciju sa sažimanjem teksta izjavom da istraživači još uvijek sanjaju o sustavu automatskog sažimanja teksta za izradu sažetka koji pokriva sve glavne teme u ulaznom tekstu, ne uključuje suvišne ili ponovljene podatke, čitljiv je i kohezivan za korisnike.

Petrović i Bušelić (2020) spominju dvije osnovne metode automatskog sažimanja teksta koje se koriste u praksi: ekstrakcijska i apstrakcijska. Ekstrakcijska metoda koristi riječi i fraze iz teksta kako bi izradila sažetak dok apstrakcijska ima pristup stvaranja sažetka na način sličniji čovjeku, odnosno stvara nove rečenice kao što je prikazano na slici 1.



Slika 1 Konceptualni prikaz ekstrakcijske i apstrakcijske metode sažimanja teksta (Izvor: Yadav, D i Desai Yadav, A., 2022, str 3)

S druge strane, El-Kassas, Salama, Rafea i Mohamed (2021) ističu da pristup sažimanja teksta može biti apstrakcijski, ekstrakcijski ili hibridni koji je neki vid kombinacije prethodne dvije.

2.1.1 Ekstrakcijska metoda

Primjenom ekstrakcijske metode u kreiranju sažetaka koriste se iste riječi i fraze koje se pojavljuju u izvornom tekstu (El-Kassas, Salama, Rafea i Mohamed, 2021), što u konačnici rezultira razumljivijim sažetcima za korisnike (Tandel, Modi, Gupta, Wagle i Khedkar, 2016). Prednosti ekstrakcijske metoda u odnosu na apstrakcijsku su brzina i jednostavnost. Popularnost ove metode vidljiva je iz činjenice da većina preglednih radova fokusirana na ekstrakcijsku metodu (El-Kassas, Salama, Rafea i Mohamed, 2021). Njena primjenjivost nije proizašla isključivo iz njene jednostavnosti i brzine već iz činjenice da prethodno znanje iz područja prirodne obrade jezika nije njen preduvjet.

Iako je ekstrakcijska metoda vrlo popularna, dobiveni sažetci nisu ni približno zadovoljavajući zbog količine nedostataka kao što su redundancija u rečenicama, dužina rečenica veća od prosjeka, manjak semantike i kohezije, sukob različitih izvora u slučaju sažimanja više dokumenata, raspršenost informacija u rečenici (El-Kassas, Salama, Rafea i Mohamed, 2021). Međutim postoje istraživanja koja pokušavaju razviti ekstrakcijsku metodu i dokučiti rješenja za njegove nedostatke (Widyassari, Rustad, Shidik, Noersasongko, Syukur i Affandy, 2020).

Iako se pažnja preusmjerila na apstrakcijsku metodu, istraživanja provedena u zadnjih nekoliko godina pokazala su da je još uvijek velik interes za ekstrakcijsku metodu, što ukazuje da ima mjesta za napredak ekstrakcijske metode (Widyassari, Rustad, Shidik, Noersasongko, Syukur i Affandy, 2020).

2.1.2 Apstrakcijska metoda

Apstrakcijska metoda sažimanja teksta kreira sažetak u kojem su uključene riječi i fraze različite od onih koje se pojavljuju u izvornom dokumentu (Petrović i Bušelić, 2020). Sustav koji koristi ovu metodu prepoznaje važne rečenice iz izvornog teksta i parafrazira ih, što znači da sažetak sadrži sve bitne informacije ali u drugačijem obliku. Razlog tomu je opsežnija obrada prirodnog jezika i mnogo složeniji izrada algoritama od ekstrakcijskog. Velika prednost apstrakcijske metode jest ta da korištenjem drugačijih riječi od onih u tekstu, stvara originalni sažetak, svodeći time redundanciju na minimum (Sakhare, Kumar i Janmeda, 2018). Apstrakcijska metoda također stvara sažetke na način sličniji čovjeku, što je vrlo teško postići ekstrakcijskom metodom.

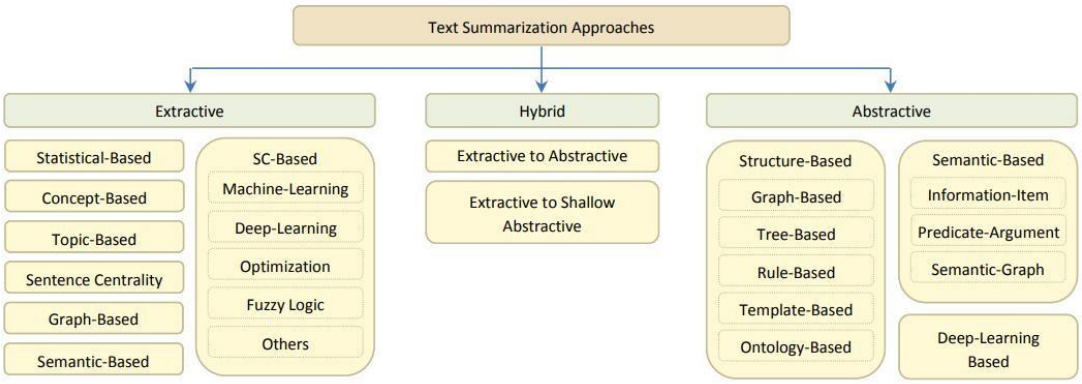
Unatoč prednostima apstrakcijske metode, puno se više koristi ekstrakcijska metoda zbog svoje jednostavnosti i čestog manjka jakih računalnih resursa koja su potrebna za primjenu

apstrakcijskih metoda (Allahyari, Pouriyeh, Assefi, Safaei, Trippe, Gutierrez i Kochut, 2017). Još jedan od nedostataka apstrakcijske metode jest taj da zahtijeva znanja iz područja prirodne obrade jezika, koja je sama po sebi još uvijek u razvoju (Wang i sur, 2017). Ako se ne ispune svi preduvjeti koji su potrebni za apstrakcijsku metodu, rezultat će biti sažetci niske kvalitete.

Apstrakcijska metoda u teoriju je bolja jer nudi bolje rezultate. Međutim preduvjeti koje ona zahtjeva nije lako zadovoljiti zbog trenutnog nedostatka znanja, resursa i vremena.

2.1.3 Hibridna metoda

Mnogi autori definiraju hibridnu metodu kao spoj ekstrakcijske i apstrakcijske metode. Slika 2 prikazuje podjelu pristupa sažimanja teksta prema autorima El-Kassas, Salama, Rafea i Mohamed (2021) koji dijele pristupe na ekstrakcijski, hibridni i apstrakcijski pristup.



Slika 2 Prikaz podjele pristupa sažimanja teksta i njihovih metoda (Izvor: El-Kassas, Salama, Rafea i Mohamed 2021., str. 8)

Autor Fattah (2014) spominje kako bi se rezultati sažimanja teksta mogli poboljšati upotrebom hibridne metode koja koristi različite statističke značajke. Ovim bi se alatom izvršavala sažimanja više dokumenata. Rezultati dobiveni istraživanjem ovog pristupa sažimanju dokazala su se donekle boljim od prethodnih metoda. Međutim, dobiveni rezultati nisu značajno bolji od rezultata ostalih suvremenih pristupa sažimanja više dokumenata (Fattah, 2014).

Hibridna je metoda korisna u generiranju sažetaka dokumenata, pogotovo onih koji su napisani na više jezika. Hibridna je metoda neovisna o jeziku dokumenta (Neto, Freitas i Kaestner, 2002). Zbog te prednosti autor Gupta (2013) predlaže korištenje hibridne metode na tekstovima napisanim na više jezika, npr. hindski i pandžapski.

Predložene su mnoge nove metode sažimanja koje se bave jezičnim značajkama i poboljšanju kvalitete sažetaka. Takvi sustavi zahtijevaju jače procesore i više memorije jer

trebaju pogoniti više lingvističkog znanja i kompleksne lingvističke tehnike (Gambhir i Gupta, 2017).

Hibridni sustav koristi različite tehnike kao što su ključne riječi ili fraze, tehnika podudaranja i tehnika temeljena na slučaju (El-Kassas, Salama, Rafea i Mohamed, 2021).

Hibridnim tehnikama važne informacije mogu se selektirati, povezati, komprimirati ili se neke informacije mogu izbrisati kako bi se dobile nove sažete informacije. Hibridni pristup može se razviti za izradu kvalitetnog sažetka kombiniranjem ekstrakcijskih i apstrakcijskih tehnika (Gambhir i Gupta, 2017).

Obično se sastoji od sljedećih koraka (Wang i sur, 2017, Lloret i sur, 2013):

1) prethodno obrađivanje,
2) ekstrakcijska rečenica pomoću ekstrakcijske metode,
3) generiranje sadržaja korištenjem i apstrakcijske i ekstrakcijske metode i
4) naknadno obrađivanje kako bi se utvrdilo da je generirani sadržaj vjerodostojan i da su sva pravila ispoštovana.

Neka od općenitih pravila su da rečenica mora imati minimalno tri riječi, rečenica mora sadržavati glagol, rečenica ne smije završavati članom (npr. eng "a", i "the"), prijedlogom, veznikom, nit i upitnom riječi (El-Kassas, Salama, Rafea i Mohamed, 2021). Autori Wang i sur. (2017) predlažu korištenje hibridne metode prilikom sažimanja dugačkih tekstova.

Hibridna metoda postiže bolje rezultate od korištenja jedne metode kada su u pitanju *precision*[1], *recall*[2] i F-mjera[3]. To je zato što je proces sažimanja teksta više-dimenzionalni problem koji spaja segmentaciju i tokenizaciju, vrijednosti riječi i rečenica, evaluaciju teksta, kosinusnu matricu sličnosti i identifikaciju optimalne kombinacije rečenica (Abualigah, Bashabsheh, Alabool i Shehab, 2020).

Autori Nazari i Mahdavi (2019) došli su do zaključka da bi korištenje različitih metoda na hibridan način bilo efikasnije. Kombiniranjem dviju metoda otklanjaju se njihovi nedostatci i poboljšava se kvalitete sažetka hibridnom metodom.

Autori Mahajani i sur. (2019) zaključili su da hibridna metoda ima obećavajuću budućnost, te savjetuju istraživačima da koriste i unaprjeđuju hibridnu metodu kako bi maksimalno iskoristili prednosti ekstrakcijske i apstrakcijske metode.

[1] *Preciznost (eng. precision) je u ovom kontekstu omjer broja odabranih točnih rečenica od ukupnog broja rečenica nekog teksta.*

[2] *Opoziv (eng. recall) je u ovom kontekstu omjer broja odabranih točnih rečenica od ukupnog broja točnih rečenica.*

[3] *F-mjera je harmonijska sredina preciznosti i odziva*

2.2 Klasifikacija na temelju veličine unosa

Veličina unosa predstavlja broj izvora koji unosimo u alat za sažimanje teksta, a mogu biti jedan dokument ili više dokumenta odjednom (El-Kassas, Salama, Rafea i Mohamed, 2021). Cilj sažimanje jednog dokumenta je kreiranje sažetka dokumenta zadržavajući sve važne informacije, te tako skratiti vrijeme potrebno za njegovo čitanje (Joshi, Wang i McClean, 2018). Isti autori navode da sažimanje više dokumenata odjednom ima isti cilj kao i sažimanje jednog dokumenta ali njegov sažetak se generira na temelju više ulaznih dokumenata.

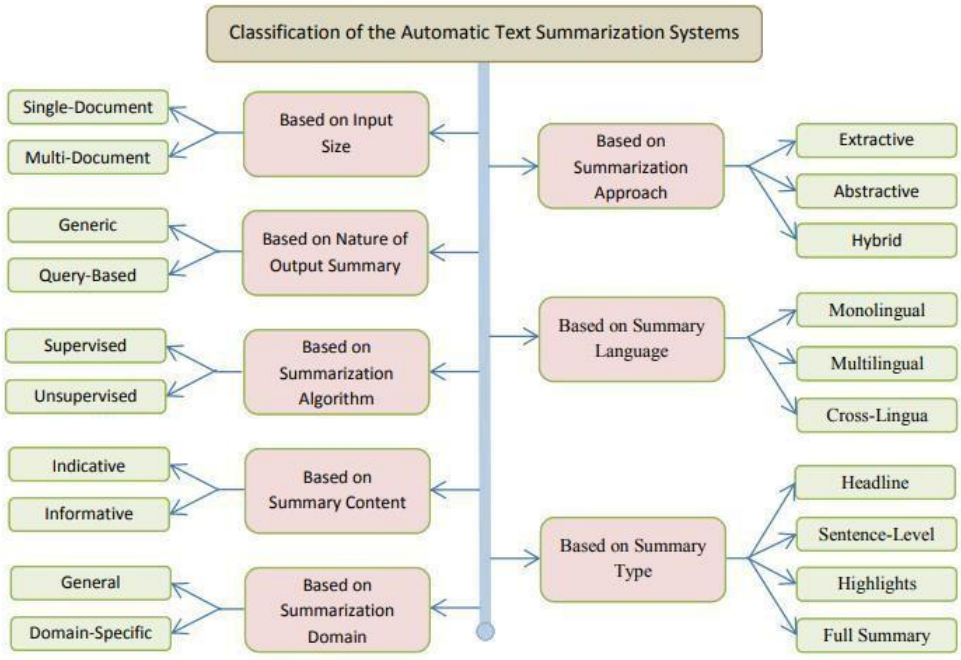
Uz razvoj alata za sažimanje teksta počelo se težiti sažimanju više dokumenata odjednom (Radev, Hovy i McKeown, 2002). Razlog tome jest pretrpanost informacija, redundancije, te neorganiziranost i potreba za bržim sažimanjem i sažimanjem više dokumenata odjednom. Sažimanjem više dokumenata nastaje jedan sažetak koji sadrži ključne teme i pojmove iz više različitih ali temom povezanih dokumenata. Stručnjaci gledaju leksičku povezanost i ostatak teksta kako bi utvrdili povezanost. Povezanost se mjeri brojem zajedničkih riječi ili fraza u tekstovima. Sustav koji to može omogućio bi čitatelju da sazna glavne teme teksta ili tekstova i da sazna sadržava li taj dokument/dokumenti ono što on treba. Sažetak jednog ili više dokumenata zasad još uvijek ne može zamijeniti originalni tekst već samo skratiti potragu za informacijama. Unatoč rastućom potrebom za sustavom koji bi mogao sažeti više različitih dokumenata odjednom, postoje dobri razlozi zašto još nisu postignuti očekivani rezultati. Glavni problemi takvih sustava su prepoznavanje i nošenja s redundancijom, uočavanje razlika između dokumenata i osiguravanje koherentnosti sažetka čak i onda kada dokumenti dolaze iz različitih izvora (Radev, Hovy i McKeown, 2002).

2.3 Ostale podjele automatskog sažimanja teksta

S obzirom na njegovu prirodu sažetak može biti generički sažetak ili sažetak relevantan za upit (Gong i Liu, 2001). Generički sažetak daje sveukupan uvid u smisao sadržaja dokumenta. Dobar generički sažetak treba sadržavati glavne teme dokumenta i svest redundanciju na minimum. Generički sažetak nema nikakav pregled tema i tretira dokument kao jedinstven tekst, pridodajući time svim informacijama istu razinu važnosti (Nenkova i McKeown, 2011). S druge strane, sažetak relevantan za upit prikazuje dokumente čiji je sadržaj usko vezan s pretraživanim upitom. Stvaranje sažetka relevantnog za upit je u biti proces dohvaćanja relevantnog upita rečenice i odlomka iz dokumenta koji je srodan s procesom dohvaćanja teksta (Gong i Liu, 2001). Sažetak usmjeren na upite također se naziva i sažetak usmjeren na temu ili korisnika. Takav sažetak uključuje sadržaj povezan s upitom, dok se opći smisao informacija prisutnih u dokumentu daje u generičkom sažetku (Gambhir i Gupta, 2017).

Važna podjela klasifikacije s obzirom na sadržaj sažetaka je indikativan ili informativan sažetak (El-Kassas, Salama, Rafea i Mohamed, 2021). Indikativni sažetak sadrži samo opće ideje ili informacije o izvornom tekstu, a njegov glavni cilj je informirati korisnike o opsegu unesenog teksta kako bi mogli donijeti odluku žele li pročitati izvorni tekst ili ne (Takeuchi, 2002). Informativni sažetak pak sadrži sve važne informacije i ideje koje se nalaze u izvornom tekstu, te pokriva sve teme izvornog teksta (Bhat i sur., 2018). Cilj informativnog sažetka je pokriti sve glavne teme izvornog teksta izostavljajući detalja (Takeuchi, 2002).

Zadnja važnija podjela je na temelju tipa sažetka, a ona se dijeli na naslov (Headline), razinu rečenice (Sentence level), istaknuto (Highlight) i cijeli sažetak (Full Summary). Dužina i sadržanost informacija sažetka ovise o namjeni sustava za automatsko sažimanje teksta. Generiranje naslova podrazumijeva kreiranje naslova koji je inače kraći od jedne rečenice (Dernoncourt, Ghassemi i Chang, 2018). Sažetak na razini rečenice generira jednu rečenicu iz ulaznog teksta, koja je obično apstrakcijska rečenica (Dernoncourt i sur., 2018). Sažimanje istaknutih dijelova teksta daje telegrafski stil i ekstremno sažet sažetak koji je obično u obliku nabrajanja, daje čitatelju kratak pregled glavnih informacija u ulaznom dokumentu (Woodsend i Lapata, 2010.). Generiranje cijelog sažetka obično se vodi duljinom sažetka ili omjerom kompresije (El-Kassas, Salama, Rafea i Mohamed, 2021). Većinu spomenutih podjela koje obuhvaća ovaj rad nalaze se na shemi ispod.



Slika 3 Prikaz klasifikacije sustava automatskog sažimanja teksta (Izvor: El-Kassas, Salama, Rafea i Mohamed 2021., str. 5)

Uz sve nabrojane podjele, postoje još mnogo podjela koje neće biti navedene jer prelaze okvire ovog rada. Ostale podjele mogu se pročitati u sljedećim radovima: (Nenkova, McKeown, 2012), (Suleiman, Awajan, 2020), (El-Kassas, Salama, Rafea, Mohamed, 2021), te (Gambhir, Gupta, 2017).

3 Primjena sažimanja teksta u području informacijskih znanosti

Postoje brojni oblici teksta kao npr. knjige, forumi, romani, novinski članci, mailovi, blogovi, recenzije, povijest bolesti pacijenata i sl. Alatima za sažimanje teksta moguće je sažeti sve oblike teksta, kako bi korisniku bilo lakše prepoznati sadrži li taj dokument ono što on traži. Dodatkom za prepoznavanje glasa moguće je sažeti i glasovne datoteke (El-Kassas, Salama, Rafea i Mohamed, 2021). One bi omogućile pregled informacija u audio dokumentima kao što su radijske emisije, glasovne poruke i sl.

Prednosti korištenja automatskog sažimanja teksta su brojni. U produžetku su nabrojane samo neke od pozitivnih strana korištenja automatskog sažimanja teksta:

Sažimanje teksta rješava probleme selektiranja najbitnijih dijelova rečenice i generira sažetak dokumenta (Fattah, 2014). Primjer takvog sažetka je Google i Google scholar koji ispod linka stranice s rezultatima imaju sažetak od dvije tri rečenice, takozvani „google snippet“. On služi čitatelju kao pokazatelj relevantnosti informacija koje se nalaze na stranici. S obzirom da je rezultata jedne pretrage nerijetko oko stotine tisuća, kvaliteta tih sažetaka treba biti dostojna zamjena punom originalnom tekstu kako bi čitatelju bio od pomoći. Još jedan od primjera su zdravstveni kartoni koji sadrže sve podatke nekog pacijenta. Medicinska dokumentacija ima više svrha ali primarno služi pravilnom liječenju pacijenata (Čizmić, 2009). Podaci koji se nalaze u kartonu ne služe samo zdravstvenim djelatnicima već i onima koji snose troškove liječenja i znanstvenicima. Sažimanjem zdravstvenih kartona podaci se mogu prikazati čitko i sažeto, a to omogućuje brži pronalazak primjerenog liječenja.

Brojni rezultati pretraživanja koje tražilice izbacuju korisniku ne znače ništa ako on ne može naći dokument koji će umanjiti njegovu informacijsku entropiju. Čitanjem sažetaka korisnik može zaključiti da ga tekst ne zanima ili da ne sadrži informacije koje traži i neće gubiti vrijeme na čitanje cijelog dokumenta. Stoga možemo zaključiti da kreiranje sažetaka smanjuje količinu vremena potrebnu za čitanje dokumenta (Neto, Freitas i Kaestner, 2002).

Izuzevši to što nam može uštedjeti vrijeme, automatskim sažimanjem teksta također je moguće kreirati sažeti pregled najbitnijih informacija iz elektroničkih mailova (Gambhir i Gupta, 2017). To znači da će korisnik moći pročitati sve ključne točke opširnih mailova pomoću jednog jezgrovitog sadržaja. Osim maila, moguće je sažeti i opširne poruke i razgovore zahvaljujući istraživanjima koja nastoje razviti i proširiti primjenu sažimanja na druge oblike teksta. Takozvano CQA sažimanje (eng. Community-based Question Answering) pruža pomoć korisnicima u pronalaženju informacija za kojima tragaju (Hsu, Suhara i Wang, 2022). CQA je

popularan alat kojim se koriste poznate tražilice kako bi korisnicima prikazali „najbolje odgovore“ na njihove upite. (Liu, Li, Cao, Lin, Han i Yu, 2008).

Još jedna od bitnih prednosti jest ta da zahvaljujući automatskom sažimanju teksta, rastući broj tekstualnog sadržaja više nije problem jer se on više ne sažimaju ručno (El-Kassas, Salama, Rafea i Mohamed, 2021). To znači da nekontrolirani rast informacija na internetu (ili u bazama informacijskih ustanova) ne predstavlja prepreku jer se one odvajaju, sortiraju i organiziraju strojno, što je brže i jeftinije.

Zadnja važna prednost koju valja istaknuti jest da se kreiranjem sažetka velikih tekstualnih datoteka smanjuje redundancija informacija (Nazari i Mahdavi, 2019). Valja napomenuti da nije svako ponavljanje informacija redundancija koja se mora izbjeći. Redundancija se manifestira ponavljanjem podataka, ali nije svako ponavljanje podataka redundantno ponavljanje (Xue, Lu, 2020). Dovoljno razvijenim alatima za automatsko sažimanje teksta moguće je smanjiti redundanciju informacija u teksta bez izostavljanja bitnih informacija.

3.1 Organizacija informacija i sažimanje teksta

Organizacija informacija je područje koje se bavi opisivanjem i indeksiranjem dokumenata i klasifikacijom koja osigurava sustave za predstavljanje i uspostavljanje reda u korist znanja i informacija (Taylor i Joudrey, 2004). Prije procesa indeksiranja potrebno je odrediti ključne riječi teksta. Ključne riječi predstavljaju glavni sadržaj dokumenta S primjenom sažimanja teksta proces indeksiranja može biti automatski jer postoje metode sažimanja koje rade po principu odabira ključnih riječi u tekstu i stvaranja sažetaka s njima kao glavnim točkama dokumenta (Shannaq i Adebiaye, 2015).

Autori Takeda i Takasu (2009) definiraju organizaciju informacija kao pomoć u korisnikovoj potrazi za informacijama. Primarna svrha organizacije informacija danas je pronalaženje i otkrivanje izvora primjerenih relevantnih informacija. Ukoliko web tražilica prikaže puno izvora koji su organizirani po relevantnosti tražitelju s nekoliko ključnih rečenica izvučenih iz dokumenta ispod, tražitelj će u kratkom roku pronaći ono što traži. Organizacija informacija i sažimanje neodvojivi su elementi u procesu traženja informacija.

Organizacija informacija općenito podrazumijeva: identifikaciju i skupljanje informacija koje su u suštini iste i razlikovanje onoga što nije sasvim isto (Svenonius i Willer, 2005). Primjenjujemo ju u mnogim djelatnostima i različitim ustanovama kako bi olakšali pronalaženje, razmjenu i korištenje informacija. Može se reći da je sažimanje teksta na neki način organizacija teksta.

Sažimanje teksta ima bitnu ulogu u organizaciji informacija i informacijskih sadržaja općenito. Ona podrazumijeva obradu i uređenje podataka i informacija tako da budu jednostavnija za upotrebu i interakciju s korisnicima. Dobrom organizacijom informacija korisnici će biti zadovoljniji jer vrijedi izreka „Dobra organizacija je pola obavljenog posla“. Organizacija informacija i sažimanje teksta korisno je u stvaranju sažetaka knjiga u digitalnim repozitorijima ili web stranica knjižnica kao i u bilo kojim drugim većim bazama podataka, kao osnovnim dijelovima informacijskih sustava.

U radu „Information Organization System for Duplicated Information Sources“ prikazani su problemi s kojim se susreću organizatori weba. Organizacija i održavanje informacija je proces koji korisnicima pruža bolja iskustva u pretraživanju i korištenju informacijskih izvora. Neki od zadataka su filtriranje i organizacija dvostrukih informacija koje se ponavljaju na internetu i dinamičnih informacijskih izvora. Primjer sustava koji zahtjeva organizaciju informacije na webu jesu blogovi. Spamovi na blogu, takozvani splogovi (eng. splogs), dvostruke su informacije koje je potrebno prijaviti i ukloniti s bloga. Splogovi čine 22% svih blogova, te je njihovo uklanjanje s bloga vrlo bitno za korisnike. Postoje algoritmi koji mogu naučiti raspoznati i ukloniti splogove (Takeda i Takasu, 2009).

Sažimanje teksta također se može koristiti u stvaranju sažetaka dokumenata koji se nalaze u velikim bazama podataka knjižnica, arhiva i ostalih informacijskih ustanova. Arhivi, Muzeji, Knjižnice (AKM) ustanove su poznate po tome što pohranjuju, organiziraju, zaštićuju i čine dostupnim brojne važne dokumente (Radić, 2017). Njihova je primarna svrha učiniti znanje koje one posjeduju vidljivim i dostupnim svim korisnicima, pogotovo istraživačima i znanstvenicima. Stoga je od izuzetne važnosti da se sva dostupna znanja organiziraju i stave na raspolaganje. Početkom 20. stoljeća algoritmi nisu bili dovoljno razvijeni za iscrpnu obradu informacija, stoga su ljudi bili primorani ručno organizirati velike količine informacija. Uz razvoj tehnologije, sada se to može učiniti računalima umjesto čovjeka brže, kvalitetnije i jeftinije (El-Kassas, Salama, Rafea i Mohamed, 2021).

Jedan od najvećih problema u informacijskim ustanovama jest subjektivnost. Korištenjem algoritma za automatsko sažimanje teksta dovodi do povećanja objektivnosti u usporedbi sa sažetcima koje je izradio čovjek (Takeda i Takasu, 2009). Problem subjektivnosti oduvijek je bio prisutan u javnim ustanovama kao što su knjižnice. Informacijski stručnjaci dužni su organizirati znanje na objektivan i logičan način kako bi korisnicima olakšali potragu za informacijama.

3.2 Uloga automatskog sažimanja teksta u predstavljanju informacija i prezentaciji informacijskih sadržaja

Prema časopisu Library and Information Science Research sažimanje i predstavljanje informacija predviđaju se kao jedna od 10 glavnih žarišta informacijskih znanosti 22. stoljeća (Chowdhury, 1999). Može se reći kako se to predviđanje polako ostvaruje i prije kraja ovog stoljeća.

Predstavljanje informacija podrazumijeva način organizacije web stranice na kojoj će se korisnik htjeti zadržati. Web stranice ne smiju: gomilati nepotrebne ili duplicirane informacije, biti dosadne, biti dizajnirane na način da je korisniku teško ili nemoguće pronaći informacije koje traži i sl. Sadržaji web stranica moraju biti sažeti, atraktivni i informativni korisniku. Predstavljanje informacija također podrazumijeva način na koji se podaci prikazuju i opisuju na internetu što podrazumijeva opisne jezike kao npr. HTML i XML, RDF o kojima neće biti riječi u ovom radu.

Predstavljanje ili prezentacija informacija podrazumijeva prikaz znanja na način da bude razumljiva široj publici, npr. provedena istraživanja dokazuju da prikaz veće količine informacija isključivo u obliku teksta smanjuje postotak razumijevanja korisnika (Van Berkel, Goncalves, Russo, Hosio i Skov, 2021). Taj se problem može lako riješiti upotrebljavanjem automatskog sažimanja teksta koje će smanjiti broj informacija i time olakšati razumijevanje korisnicima. U zadnjih nekoliko godina, razvio se KIGN (eng. Key Information Guide Network) koji ima za cilj usmjeriti proces generiranja sažetka. Uz kombinaciju s ekstrakcijskom metodom prikupljanja ključnih riječi, KIGN kodira ključne informacije za predstavljanje i implementira ih u apstrakcijsku metodu (Li, Xu, Li i Gao, 2018). Ovim postupkom postižu se odlični rezultati po pitanju sadržanosti informacijama generiranih sažetaka.

S druge strane prikaz podataka isključivo u grafovima i tablicama bit će jasna samo onim korisnicima koji ih znaju interpretirati. Takvi načini predstavljanja informacija ne bi trebali biti i jedini način. Kroz godine brojna istraživanja su posvećena proučavanju interpretacije informacija i efekta koji utječu na njihovo razumijevanje. Autori Van Berkel, Goncalves, Russo, Hosio i Skov (2021) došli su do zaključka da je najbolji način prezentacije informacija kombinacija teksta i vizualnog sadržaja, npr. tablice, grafovi, ilustracije i sl.

Sustavi za pretraživanje vizualnih informacija novitet su današnjih generacija koji omogućuje pretraživanje karakteristika slike (boja, oblik, tekstura i sl.) i daju korisniku na uvid rezultate sa sličnim karakteristikama (Hladilo, 2016). To znači da je danas korisnicima omogućeno pretraživane pomoću slika (Google images) i zvuka (aplikacija Shazam). Shazam nudi opciju pregleda svih pretraživanih rezultata pjesama. Svaki rezultat upita (ukoliko je aplikacija pronašla odgovarajući rezultat) pohranjuje se u korisnikovu knjižnicu gdje se može

preslušati isječak pjesma, najčešće refren. To se može smatrati načinom sažimanja glazbe jer štedi vrijeme korisnika.

Bukovac (2016, diplomski) u svom radu je pokazala kako je korištenje multimedijalnih sadržaja najbolji način predstavljaju informacija korisnicima. Ista autorica predlaže „skraćivanje“ kompliciranih tekstualnih informacija pretvaranjem tih informacija u info grafike. Korisnici bolje reagiraju na vizualni prikaz nego na tekst. Zamjenom teksta sa slikom uštedjet ćete korisniku vrijeme jer ste tekst saželi u sliku. Tu funkciju imaju Google obrasci koji nakon obrade prikupljenih anketiranih odgovara korisniku daju opciju da tisuće podataka prikaže u obliku grafova i tablica.

Prezentacija informacija odnosi se na način prikaz prilagođen korisniku (čovjeku). Korisnik svojim preferencijama i željama oblikuje elemente sučelja koje mu prezentira rezultate na određen način. Primjer personalizirane prezentacije informacija su sustavi preporuke na web stranicama, npr. 24sata nudi vijesti prilagođene korisnikovoj lokaciji (ukoliko korisnik dopusti tu informaciju), npr. članci o potresu u blizini te lokacije. Google nudi opciju uključivanja personaliziranih oglas kojim korisnik dopušta Googlovim alatima da sakupljaju podatke o njegovom spolu, dobi, interesima, bračnom statusu i sl. Rezultat su oglasi koje Googleov algoritam odabere za pojedinog korisnika. Sustav personalizacije također je prisutan i na aplikacijama. Brojne aplikacije, npr. Pinterest i Planta omogućuju korisniku da prilikom prvog korištenja aplikacije unese podatke koji će filtrirati i prikazati određene podatke vezane uz interese i preferencije korisnika. Pinterest tako nudi brojne teme za odabir (npr. makeup, fashion, quotes, food i sl.), dok Planta nudi odabir razine znanja koju korisnik posjeduju o vrtlarstvu. Korisnik u svakom trenu može promijeniti teme koje ga zanimaju.

Pretraživanjem određene teme rezultati će prikazati linkove ispod kojih se nalaze rečenice koje sadržavaju podatke vezane uz našu pretragu. Pretraživači traže vezu između korisnikovih pretraga i tekstualnog sadržaja weba. Stručnjaci nastoje unaprijediti algoritam za sažimanje kako bi razlikovao reprezentativne i važne rečenice od manje važnih (Allahyari, Pouriyeh, Assefi, Safaei, Trippe, Gutierrez i Kochut, 2017). Ograničenja pretraživača su slike, videozapisi i ostali ne-tekstualni sadržaji, stoga je takvo proširenje sažimanja popularno područje istraživanja.

3.3 Automatsko sažimanje teksta u području pretraživanja informacija

Internet kao univerzalni medij za prijenos podataka, informacija i znanja, ima zadatak prikazati ih na način na koji će se one najlakše pronaći (Putica, 2018). Tokom njegovog rasta i razvoja, koji se događao tijekom tri generacije, količina informacija na internetu dovela je do promjene u načinu dohvaćanja, sortiranja i klasificiranja informacija. To znači da se isti sadržaj

pretraživao i predstavljao drugačije prije 20 godina i sad. Korisnici su prije imali kataloge na raspolaganju za lakše pretraživanje informacija dok danas imamo mogućnost pretraživanja po naslovu, autoru, godini, ključnim riječima itd.

Pretraživanje informacija može se definirati kao način pronalaska informacija koje će smanjiti našu neizvjesnost o nekoj temi (Bagarić i Jokić-Begić, 2020).

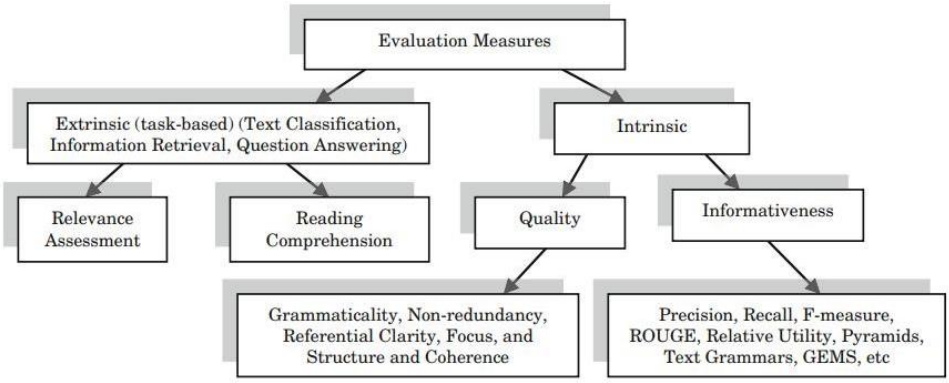
Informacije postoje na internetu radi njihova korištenja, stoga je izuzetno bitno da korisnici znaju pronaći znanja, te da su ona uvijek dostupna. Prilikom pretraživanja web izvora korisnik se može poslužiti različitim alatima čija je svrha olakšati i ubrzati proces pretraživanja informacija. Neki od tih alata su pretraživači, odnosno tražilice (eng. Search Engines) dok web preglednici (eng. Web Browsers) omogućuju njihovu jednostavnu primjenu.

Pretraživači ili tražilice su alati za otkrivanje informacija na Webu i jedan su od ključnih faktora procesa pretraživanja informacija. Najpopularniji današnji pretraživači (Google Search, Bing, Yahoo Search) nastali su kao odgovor na drastično povećanje informacija na Webu (Sudeepthi, Anuradha, Babu, 2012). Web preglednik je aplikacija potrebna za pristup uslugama koje se pružaju putem weba i pomoću njega korisnik pregledava sadržaj Weba (Rasool, Jalil, 2020). Razlikuju se po tome što je web preglednik dio softvera koji dohvaća i prikazuje web stranice, dok pretraživač pomaže ljudima pronaći web stranice i ostale izvore na web ili datotečnim poslužiteljima temeljem postavljenih upita. Provedena istraživanja pokazala su da korisnici najčešće prvo posežu za web preglednikom kada imaju informacijsku potrebu (Hu, Jiang. Robertson i Wilson, 2019). Važna značajka pretraživača su „search snippets“. Snippets predstavlja kratki sadržaj članka prikazan ispod linka web stranice pomoću kojeg korisnik može dokučiti sadrži li taj članak informacije koje ga zanimaju ili ne. Snippets su jako važan dio današnjih pretraživača jer pružaju dodatni kontekst i upotpunjuje naslov linka web stranice (Hu, Jiang. Robertson i Wilson, 2019). Jedan od najpoznatijih primjera pretraživača koji koristi snippets je Google Search, gdje je prvi put uveden 2012 (Strzelecki i Rutecka, 2020). Snippets nerijetko kreiraju sažetke iz HTML metapodatkovnih oznaka (Hu, Jiang. Robertson i Wilson, 2019). Snippets se također mogu sastojati od nekoliko rečenica koje sadržavaju sve važne informacije web stranice koji je nastao automatskim postupkom sažimanja teksta. Algoritmi za stvaranje sažetaka ukomponirani su uz pojedine pretraživače kako bi svaki članak imao kvalitetan sažetak, u suprotnom bi korisnik mogao promašiti neki njemu bitan članak. Ti algoritmi koriste varijantu ekstrakcijskog sažimanja, iako stručnjaci preporučaju implementacije apstrakcijske metode uz ekstrakcijsku. Neprestana nadogradnja algoritma pretraživača neophodna je u stvaranju boljih i točnijih snippeta (Marcos, Gavin i Arapakis, 2015).

Još jedna od inačica Google pretraživača jest Google Scholar koji služi dohvaćanju akademskih članaka. Ispod svakog naslova članka također se nalazi snippet. Osim samog prikaza, Scholar nudi opciju preuzimanja dokumenta u PDF formatu ukoliko je raspoloživ, pregled citiranosti članka, mogućnosti odabira stila citiranja itd. Iako algoritam koji Scholar koristi nije poznat, mnoge knjižnice povezuju svoje korisnike s bazom Google Scholar kodirajući ga na naslovnicu knjižnice (Hartman i Bowering Mullen, 2008). U radu autora (Martín-Martín, Thelwall, Orduna-Malea, Delgado López-Cózar, 2021) u kojem su uspoređene baze podataka Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science i OpenCitations, ispostavilo se da je Google Scholar zasad najsveobuhvatniji jer nudi i pregled citiranosti radova.

Dodatni odličan primjer je SciSpaceov Copilot (2023) koji je ujedno i jedan od najnovijih alata, a dostupan na SciSpace web stranici od 2022. Copilot je AI asistent koji na korisnikov zahtjev može sažeti, pojasniti i pružiti kontekst rada. Služi kao pomoć tijekom čitanja i razumijevanja znanstvenih i drugih vrsta radova (Chandha, Sucheth, Ghosal, 2023). Bilo da korisnik istražuje neku temu ili čita za zabavu, Copilot nudi mnoge opcije kao npr. sažetak cijelog rada u dvije ili više rečenica, pregled ključnih točaka rada, pregled literature korištene u radu i sl. Prijavom na SciSpace korisnici mogu koristiti Copilota na svojim PDF dokumentima (Sucheth, 2022). Osim sažimanja Copilot nudi brojne mogućnosti po pitanju pojašnjenja teksta kao npr. pružanje objašnjenja na podcrtani pojam. Iako je nov, SciSpace programeri nastavljaju razvijati ovaj alat prikupljajući povratne informacije od korisnika kako bi poboljšali Copilot.

Paper Digest (2023) je također jedan od alata za sažimanje teksta ali umjesto staromodnog škrtog sažetka, Paper Digest automatski generira detaljan opis rada koji daje odgovor na pitanja o čemu se radi u tekstu, koji su glavni problemi, rješenja i diskusije i sl. Paper Digest radi na način da korisnik unese DOI članka, koji mora biti otvorenog pristupa, te AI generira sažetak. Jedna od specifičnosti Paper Digesta jest ta da je dostupan kao API (Paper Digest, 2019).



Slika 4 Taksonomija evaluacijskih mjera sažetaka (Izvor: Gambhir i Gupta 2017., str. 45)

4 Alati za sažimanje teksta

Postoje različiti alati za sažimanje koje možemo koristiti. Što se tiče vrste i formata podataka koji želimo sažeti postoje tekstualni dokumente, video, audio dokumenti i sl. Najšire korišteni alati su alati za sažimanje teksta koji su ujedno predmet interesa ovog rada.

Alati za sažimanje teksta su alati koji koriste metode i pristupe umjetne inteligencije (eng. Artificial Intelligence – AI) preciznije njezinog dijela koji se naziva obrada prirodnog jezika (eng. Natural Language Processing - NLP). Riječ je o tehnologiji kojom je moguće skratiti dužinu teksta što pridonosi boljem razumijevanju istog. Alati za sažimanje teksta analiziraju sadržaj i na temelju toga kreiraju sažetak koji sadržava sve glavne informacije i pokriva sve teme (Saggion, 2008). Drugim riječima pretvaraju dugačke rečenice u kratke ne mijenjajući pritom poantu rečenica.

Zanimanja koja iziskuju rad s velikim brojem tekstualnih datoteka, kao npr. blogeri, autori, znanstvenici, novinari i sl, koriste alate za sažimanje teksta kako bi si olakšali posao i učinili svoj rad privlačnijim. Alati za sažimanje mogu koristiti ekstrakcijsku , apstrakcijsku ili hibridnu metodu, međutim najčešće korištena metoda jest ekstrakcijsku metode zbog svoje jednostavne i jeftine primjene (El-Kassas, Salama, Rafea i Mohamed, 2021). Takvi sažetci sastoji se od rečenica koje je alat odredio kao ključne (key) rečenice. U sljedećem poglavlju prikazano je nekoliko primjera alata za automatsko sažimanje teksta.

4.1 Online alati za automatsko sažimanje teksta

U ovoj cjelini opisane su osnovne karakteristike i mogućnosti nekoliko alata za automatsko sažimanje teksta bez ulaženja u detalje njihovog principa rada te je izostavljena i njihova komparacija koja predstavlja zasebnu temu. Resoomer (2023) je jedan od popularnijih alata za sažimanje teksta, a baziran je na ekstrakcijskoj metodi što znači da izvlači ključne riječi i fraze iz originalnog teksta i od njih stvara sažetak. Korišten je diljem svijeta, a nudi mogućnosti sažimanja teksta na Engleskom, Francuskom, Arapskom, Španjolskom, Talijanskom, Poljskom i Njemačkom jeziku (Alliheibi, Omar, Al-Horais, 2021). Resoomer osim kreiranja sažetka odabranog dokumenta također nudi opciju podcrtavanja važnih rečenica. Autor Glenc (2021) u svom radu istražuje metode korištene odabranih alata za sažimanje teksta, među kojim je i Resoomer, te kvalitetu generiranih sažetaka. Ističe kako se u kvaliteti generiranih sažetaka na poljskom jeziku vidi malen pomak nabolje. Resoomer je praktičan online alat kojeg korisnici mogu koristiti za skraćivanje dužih tekstova, npr. poglavlja knjige (Srirejeki, Faturokhman, Praptapa, Irianto, 2022).

Scholarcy (2023) je alat koji koristi tehnologije umjetne inteligencije (eng. Artificial Intelligence – AI) i strojnog učenja (eng. Machine Learning – ML) u procesu automatskog generiranja sadržaja znanstvenih radova u oblik lako razumljivih sažetaka (Renn, 2021). Jedan od problema s kojim se znanstvenici susreću danas su poteškoće u izradi znanstvenih radova uslijed velikog broja publiciranih radova te česte potrebe za interdisciplinarnim uvidom u istraživačke teme. S obzirom na navedeno, moguća pomoć pruža se u obliku alata za sažimanje radova Scholarcy (Ardiansyah, Purwaningsih, Teruna, 2021). Tijekom provedbe korištenja Scholarcyja postignuti su zadovoljavajući rezultati po pitanju prihvaćanja implementacije Scholarcyja u procesu pisanja znanstvenih radova i povećanja entuzijazma istog. Scholarcy nudi proširenje preglednika kojim je moguće korištenje funkcije izdvajanja citata s API-jima CrossRefom i Unpaywalla (Gooch, 2023). Na taj način istraživači, dok čitaju rad, mogu pratiti trag znanja do verzija citiranih izvora s otvorenim pristupom. Iako ovaj alat nije u potpunosti besplatan alat (osim probnog razdoblja), najbolje značajke su uključene u plaćenoj verziji koja korisniku omogućuje pregled sažetka kompleksnih dokumenata.

Fitria (2021) je definirala QuillBot (2023) kao "online aplikaciju za parafraziranje, izbjegavanje plagijata, sažimanje dugih rečenica i poboljšanje gramatike kako bi bile preciznije i izgledale profesionalno". Kao jedan od najpopularnijih besplatnih online alata za parafraziranje i sažimanje, pogodan je za učenike, studente, profesore, pisce, blogere itd. QuillBot je praktičan alat za sažimanje dostupan na tržištu koji koristi umjetnu inteligenciju za sažimanje i parafraziranje bilo kojeg sadržaja. Glavni cilj alata je ispisivanje tekstualnog materijala promijenjene strukture rečenica i zamjenom riječi sinonimima uz zadržavanje izvornog značenja sadržaja (Fitria, 2021). Također, može sažeti rečenice i promijeniti strukturu deduktivnih rečenica na induktivne i obrnuto pritom ne mijenjajući sadržaj rečenice (Rakhmanina, Serasi, 2022).

Primjer potpuno besplatnog alata je Tools4Noobs (2023). Tools4Noobs obavlja obradu teksta u nekoliko faza: izvlači rečenice iz ponuđenog teksta, identificira ključne riječi i relevantnost svake riječi te potom identificira i odabire izraze s najvažnijim ključnim riječima (Costa, 2020). Tools4Noobs pruža neke alate pod licencom otvorenog koda (open source), a svi njegovi alati mogu se koristiti bez naknade zato što je on besplatan softver (Verma, Tiwari, 2014). Tools4Noobs dobro služi korisnicima koji žele jednostavan sažetak bez previše dodatnih mogućnosti (Ningrum, 2021). U istraživanju autora Christian, Agnus i Suhartono (2016) u kojem su uspoređeni razni alati, Tools4Noobs je postigao zadovoljavajuće rezultate po pitanju točnosti i dužine sažetaka. Provedeno istraživanje mjerilo je uspješnost alat za sažimanje pomoću kriterija Precision, Recall i F-mjere.

Smmry (2023) je alat stvoren za sažimanje svakog online članka i teksta. Njegov zadatak je pomoći da korisnik razumije teksta tako što ga svede na najvažnije rečenice (Roy, Mukhopadhyay, 2022). Autor Ningrum (2021) proveo je evaluaciju alata za sažimanje teksta Tools4Noobs i Smmry na temelju 5 karakteristika i 16 podkarakteristika koje su dio standarda ISO 9126. ISO 9126 jedan je od standarda koji se koristi za procjenu kvalitete softvera općenito iz razloga što su njegove karakteristike potpunije od ostalih modela procjene kvalitete softvera (Ningrum, 2021). Rezultati evaluacije pokazali su da je alatu sa sažimanje teksta Tools4noobs dodijeljena ocjena 3,816 od ukupne ocjene 5 dok je Smmry dobio ocjenu od 4,179 od ukupne ocjene 5. Vidljivo je da je Smmry dobio bolji konačni rezultat u odnosu na Tools4Noobs, što ukazuje da se radi o boljem sažetku.

Kod svih navedenih alata važnu ulogu ima evaluacija kreiranih sadržaja što je predmet sljedeće cjeline.

4.2 Evaluacija kreiranih sažetaka

Ključna točka u istraživanju alata za sažimanje teksta jest evaluacija kvalitete generiranog sažetka (Neto, Freitas i Kaestner, 2002). Kako bi se utvrdila uspješnost procesa sažimanja teksta potrebno je izvršiti kontrolu ili evaluaciju rezultata. Proces evaluacije izuzetno je bitan jer njime ne samo da ocjenjujemo rezultate sažetka već i radimo korak naprijed prema boljem i kvalitetnijem alatu za sažimanje teksta. Većina metoda i alata ocjenjuje prisutnost informacija u sažetku dok malo njih ocjenjuje kvalitetu samog sažetka. Novi pristupi nastoje uvesti automatizaciju procjene kvalitete sažetaka (Gambhir i Gupta, 2017).

Postoje dvije vrste evaluacija: subjektivna ili ručna i objektivna ili strojna evaluacija (Takeda i Takasu, 2009). Subjektivna evaluacija podrazumijeva da čovjek pročita rezultate sažetka, te dodijeli ocjenu prema određenom kriteriju. Objektivna evaluacija se vrši automatskim kalkuliranjem i usporedbom s drugim sustavima čiji sažetci imaju ručno pripremljene važne rečenice. Ručna evaluacija na velikom broju sažetaka često je neizvediva jer zahtjeva prisutnost čovjeka, stoga se češće koristi strojna evaluacija. Provođenje evaluacije sažetaka zahtjevan je i skup proces jer ne postoji definicija savršenog sažetka koja vrijedi za sve dokumente. Iz toga se razloga uzimaju u obzir mnogi različiti kriteriji za ocjenu sažetka. Neki od tih kriterija su precision, recall i mjera pokrivenosti (eng. Coverage). Precision i recall popularne su metrike kojima predviđamo pokrivenost između sažetaka koje su izradili ljudi i strojno generiranih sažetaka. Precision možemo definirati kao omjer broja odabranih točnih rečenica od ukupnog broja rečenica nekog teksta. Precision određuje točne udjele rečenica, a same rečenice odabrane su od strane ljudi i sustava. Recall je omjer broja odabranih točnih rečenica od ukupnog broja točnih rečenica (Neto, Freitas i Kaestner, 2002). Njime

procjenjujemo koliki udio rečenica, odabirnih od strane ljudi, sustav prepoznaje. Coverage je metrika procjene za mjerenje koliko dobro sustav proizvodi sažetak uzimajući u obzir redundanciju generiranog sažetka (Takeda i Takasu, 2009). Tijekom vremena uloženo je mnogo truda u razvoj automatskih metrika koje bi omogućile brzu i jeftinu procjenu modela (Kryściński, Keskar, McCann, Xiong i Socher, 2019). Autori Suleiman i Awajan (2020) u svom radu koriste sljedeće metrike za evaluaciju: ROUGE1, ROUGE2, i ROUGE-L. Još neke od metrika, pod leksičkom sličnošću, su BLEU, NIST, GTM, METEOR, TERp i Ol (Gambhir i Gupta, 2017). ROUGE (eng. Recall-Oriented Understudy for Gisting Evaluation) je najčešće korišteni alat za automatsku procjenu automatski generiranih sažetaka (Ganesan i sur., 2010). Sastoji se od skupa metrika i softverskog paketa kojim se provodi postupak evaluacije automatski generiranih sažetaka i strojno prevođenje u NPL-u (Gupta i Siddiqui, 2012). ROUGE se zadnjih nekoliko godina pokazao veoma učinkovitim u mjerenju kvaliteta sažetaka i u dobroj je korelaciji s ljudskim prosudbama. Međutim nedostatak mu je što povezuje samo nizove između sažetaka bez razmatranja značenja pojedinačnih riječi ili nizova riječi što zna rezultirati lošom kvalitetom sažetaka (Sun i sur., 2016).

Unatoč uloženom trudu u razvoj metrika, nije postignut željeni korak u smjeru sveobuhvatnijeg procesa evaluacije, različiti pristupi nisu postali popularniji, te je ROUGE ostao zadani alat za automatsku evaluacija sažimanja teksta (Kryściński i sur., 2019) unatoč svojim manama.

Metrike su važan dio procjene i evaluacije alata za sažimanje teksta jer se pomoću njih otkrivaju njihove mane i nedostaci. Metrike služe određivanju uspješnosti korištenih metoda kod stvaranja sažetaka time što ocjenjuju informativnost, tečnost, jezgrovitost i činjeničnost sažetka (Jangra, Mukherjee, Jatowt, Saha, Hasanuzzaman, 2021). Krajnji cilj metrika je stvoriti precizne sustave evaluacije koji će savršeno funkcionirati bez pomoći ljudske anotacije (Scialom, Lamprier, Piwowarski, Staiano, 2019). Kao i s pristupima, brojnost evaluacijskih metrika prelazi granice ovog rada stoga su nabrojane samo neke. Spomenute i još neke od metrika prikazane su na shemi na slici 4, a njihova objašnjenja mogu se naći u radu autora Gambhir i Gupta (2017).

5 Problemi kod automatskog sažimanja teksta

Postoje i određene prepreke koje se javljaju prilikom stvaranja i korištenja algoritma za automatsko sažimanje teksta. Strojni jezik ljudima je jednako nerazumljiv kao i računalima naš ljudski jezik, stoga je u komunikaciji između čovjeka i računala potrebno uložiti napor kako bi se postiglo razumijevanje s obje strane (Fattah, 2014). Kao što je već navedeno, nijednom se metodom još uvijek ne može ostvariti željeni cilj, a to je sažetak koji sadrži sve ključne informacije, informacije se ne ponavljaju, kraći je od originalnog teksta, koherentan je i sličan je sažetku kojeg je izradio čovjek (Petrović i Bušelić, 2020). Kreiranje sažetaka zahtjevan je posao, zbog toga je u interesu čovjeka da se razvije metoda koja bi pomogla u rješavanju kompleksnih zadataka kada je u pitanju sažimanje teksta, npr. sažimanje velikim dokumenata, dokumenti pisani na više jezika i sl.

5.1 Redundancija

Stanje u kojem isti podaci postoji na više mjesta u bazi podataka naziva se redundancija podataka (Lynch i sur., 2023) i ono je jedan od najvećih problema kod stvaranja sažetaka, pogotovo kod sažimanja više dokumenata (Gambhir i Gupta, 2017).

Redundancija je problem koji se pojavljuje u mnogim procesima sažimanja teksta, npr. u ekstrakcijskoj metodi (El-Kassas, Salama, Rafea i Mohamed, 2021). Redundancija utječe na kvalitetu sažetka, stoga je izuzetno bitno prije samog procesa sažimanja teksta izvršiti korak prethodne obrade kako bi se ona minimalizirala (Patel, Shah i Chhinkaniwala, 2019). Jedan od pristupa postupka obrade teksta prije procesa sažimanja autor Fattah (2014) opisuje ovako: rečenice se prikazuju grafom, kako bi se mogle zamijetiti i izbaciti irelevantne riječi, te primijeniti formule za rangiranje vrijednosti rečenica koje se temelje na baznom modelu. Svaka rečenica predstavlja čvor u pojedinačnom grafu. U slučaju ponavljanja četiri uzastopnih riječi u rečenici stvara se zajednička poveznica između čvorova. Snaga tih poveznice ovisi o omjeru zajedničkih riječi i dužini jedne ili dviju rečenica. U slučaju postojanja većeg čvora, manji čvorovi s istom poveznicom se eliminiraju iz procesa rangiranja rečenica. Stoga se rečenice koje se ponavljaju ili su gotovo identične uklanjaju (Fattah, 2014).

Tim se postupkom smanjuje redundancija, te se stvara kraći tekst s relevantnijim informacijama koje se ne ponavljaju. Tako se korisniku krati vrijeme čitanja dokumenta i smanjuje broj nepotrebnih ili ponovljenih informacija u tekstu (Gambhir i Gupta, 2017).

5.2 Financijska ulaganja

Automatsko sažimanje teksta kao alat ima velik potencijal, pogotovo kada ga primijenimo u različitim javnim ustanovama ili u znanstveno-istraživačkom radu općenito, gdje je ono iznimno potrebno. Ta potreba za njim raste s godinama ali postoje i problemi za koje ljudi još uvijek nisu našli rješenja, a to su velika ulaganja u sustave koja su neophodna za automatsko sažimanje koji su jedino isplativi dugoročno. Neki sustavi, kao što su npr. sustavi koji koriste apstrakcijsku metodu kreiranja sažetaka, zahtijevaju veća ulaganja u smislu vremena i financiranje za izgradnju i opremanje samog sustava, te više znanja u njegovom nadziranju (Abdaljalil, Bouamor, 2021). Takvi sustavi zahtijevaju stručnost u domeni, izvan svakog lingvističkog znanja. Zadnjih nekoliko godina bilježi se povećanje uloženih sredstva u aplikacije umjetne inteligencije (AI) od strane bankarskih industrija i osiguravajućih društva industrija, što je dobar vijest za automatsko ulaganje (Morgan, 2020). U slučaju da država nema dovoljno sredstva, ili ih ne želi odvojiti, ustanova će morati sama financirati sustave za sažimanje teksta. Uz ulaganja u sustav tu su i ulaganja u stručno osoblje koje će ga održavati. Sve to priječit će javne ustanove poput knjižnica i muzeja da svojim korisnicima pruži kvalitetniju uslugu po ovom pitanju.

5.3 Kohezija i koherentnost

Kohezija i koherentnost su načela strukture usmjerena na tekst (Mikić Čolić i Trtanj, 2019). Kohezija označava vezu između riječi i fraza u tekstu, pomoću nje se stvara dosljednost i skladnost teksta. Netočna kohezija u rečenici rezultira nejasnom porukom i obratno. Autori Brajković, Volarić i Vasić (2018) definiraju pojam kohezije kao povezanost između rečenica u bliskim segmentima koja je opisana u tekstu pomoću pragmatičnih i semantičkih odnosa između rečenica i rečenica. Koherentnost teksta podrazumijeva da je slijed riječi u rečenici organiziran, logičan, te da je dosljednost teme ispoštovana (Watson Todd i sur., 2004). Tekst koji je koherentan sastoji se od skladno sročenih i povezanih rečenica, logičan je i jasan za čitanje. Koherentan tekst čitatelj doživljava kao logičnu i povezanu cjelinu, te ga razumije i tumači na pravi način (Lakić, 2014).

Jedan od glavnih problema kod stvaranja teksta pomoću algoritama jesu kohezija i koherentnost. Huang i sur. (2010) opisali su četiri primarna zadatka koja bi se trebala uzeti u obzir prilikom generiranja sažetka za čitanje:

1. Pokrivenost informacija: sažetak treba sadržavati sve važne informacije ulaznih dokumenata.

2. Informacijski značaj: sažetak bi trebao pokrivati različite teme ulaznog dokumenta. Najvažnije teme ili teme koje preferiraju korisnici, zadane su kao glavne teme.
3. Redundancija informacija: minimizirajte suvišne ili duplicirane informacije u generiranom sažetku.
4. Koherentnost teksta: sažetak treba biti čitljiv i razumljiv tekst. Sve rečenice u sažetku su logične i jasno povezane

Prilikom stvaranja sažetaka pomoću sustava za automatsko sažimanje teksta, potrebno je postići ravnotežu između razine čitljivosti, omjera kompresije i kvalitete sažetka. Sažimanjem dugačkih dokumenata, kao što su knjige i romani, pojavljuju se problemi jer postojeći sustavi ne znaju kako ostvariti veću kompresiju sažetka (Wu i sur. 2017). Povećanje kvalitete generiranih sažetaka nije moguće sve dok sustavi ne nauče kako riješiti problem dvosmislenih riječi i sinonima. Bez pravilnog kodirana, sustav neće raspoznati da su dva sinonima ista riječ, te će ih interpretirati u sažetku kao dva različita pojma. Korištenjem znanja iz područja prirodne obrade jezika znatno se poboljšava kvaliteta generiranih sažetaka jer bez nje ekstrakcijski sažetci uvelike gube na koherentnosti, ravnoteži a time i na i značaju (Gupta i Lehal, 2010).

5.4 Dvosmislene riječi i sinonimi

Nastavno na temu spomenutu u prethodnom poglavlju, dvosmislene riječi i sinonimi predstavljaju problem ne samo u svakodnevnog govoru već i u programiranju alata za automatsko sažimanje teksta. Semantička sličnost bitan je dio prirodne obrade jezika (NLP) i može poboljšati njegovu izvedbu, kao što je razjašnjavanje smisla riječi, ekstrakcija informacija, sažimanje teksta i sustav odgovaranja na pitanja (Hasan, Mohd Noor, Rassem, Mohd Noah, Hasan, 2020). Dešifriranje tekstualne dvosmislenosti, uzimajući u obzir relevantnost između riječi ozbiljan je problem u tekstu na internetu. Autorica Šuman (2021) ističe da je problem i najveći izazov kod NLP-a jest dvosmislenost na više razina: značenju riječi, morfologiji, sintaktičkim svojstvima i ulogama i vezama između dijelova teksta. Problem dvosmislenost računala rješavaju učenjem iz grešaka, odnosno stvaranjem znanja iz prethodnih iskustva. Ako dvije različite riječi mogu zamijeniti u rečenici bez promijene značenja rečenice, znači da su sinonimi (Mohammed, 2020). Istraživanje sinonima je dugotrajan proces za jezikoslovce i leksikografe, a velik utjecaj ima i na područje NLP-a. Prilikom sažimanja teksta neophodno je da alat za sažimanje može raspoznati da dvije različite riječi imaju isto značenje, te ih tako i tretirati. U suprotnom može doći do ponavljanja informacija u sažetku. Također je važno da kada korisnik pretražuje određeni pojam, računalo prikaže i rezultate sinonima, npr.

pretraživanjem pojma „dijabetes“ u rezultatima bi se trebao pojaviti pojam „šećerna bolest“ i obratno.

6 Kratki osvrt na trendove u automatskom sažimanju teksta

Iz analize časopisnih izvora i konferencija koje objavljuju publikacije iz područja istraživanja sažimanja teksta, autori Widyassari, Rustad, Shidik, Noersasongko, Syukur i Affandy (2020) ističu kako se razina interesa za sažimanje teksta mijenjala kroz godine. Na tu temu 85 znanstvenih radova objavljeno je između 2008. i 2019. U tom preglednom radu istraživači su uzeli 80% časopisnih radova i 20% radova konferencija Časopis Expert System with the Application je u promatranom istraživanju objavio najviše tema o sažimanju teksta. U periodu od 2008. do 2012. vidljivo je da je interes za spomenuto područje nizak, svega jedan do maksimalno dva rada su objavljena po godini. 2015. interes je naglo porastao, što potvrđuje 15 napisanih radova. Interes je bio najviši 2018. kada je objavljeno 18.

Visok interes za područja pokrenuo je i razvoj različitih tema kao što su hibridne metode i pristupi. Najpopularnija područja interesa su sažimanje teksta više dokumenata i sažimanje pomoću ekstrakcijske metode. Sažimanje više dokumenata popularno je područje zbog svoje kompleksnosti u odnosu na sažimanje samo jednog dokumenta. Iako je ekstrakcijska metoda manje kompleksna od apstrakcijske, postoje mnogi izazovi koje istraživači žele riješiti, npr. kako maksimizirati prednosti ekstrakcijske metode i unaprijediti dobivene sažetke (Widyassari, Rustad, Shidik, Noersasongko, Syukur i Affandy, 2020). Razvojem ekstrakcijske metode dobili bi idealne sažetke uz minimalno truda jer ekstrakcijska je metoda jednostavnije i brža od apstrakcijske i hibridne metode, a njeni troškovi i potrebna znanja za njenu realizaciju znatno su manji. Dok se brojni znanstvenici bave istraživanjem i izučavanjem metoda za sažimanje, autori Widyassari i sur. (2020) u svome radu analiziraju i identificiraju teme u polju automatskog sažimanja teksta, daju uvid u pristupe i metode, diskutiraju o problemima sažimanja teksta i predlažu smjer budućeg razvoja. Jedan od zaključaka rada jest da je ekstrakcijska metoda još uvijek u centru istraživanja, međutim napominje se da ta metoda doseže svoje granice mogućnosti, te da se znanstveno područje sažimanja teksta treba posvetiti proučavanjem apstrakcijske i hibridne metode.

Napredak u korištenju apstrakcijske metode, pokazali su Facebookovi AI istraživači koji su korištenjem algoritma za generiranje u kombinaciji s modelom sažimanja rečenica proizveli sažetke visoke razine točnosti (Syed, Gaol, Matsuo, 2021). Googlovi istraživači predložili su pristup apstrakcijskoj metodi nazvanoj PEGA-SUS (Pre-training with Extracted Gap-sentences Abstractive Summarization Sequence to sequence models) koji je nadmašio prethodne modele boljim rezultatima. PEGA.SUS se koristi posebnim samonadziranim ciljem prije obuke koji se zove generiranje praznih rečenica (eng. Gap-Sentences Generation). GSG je takozvana metoda

samonadziranog učenja koja odabire nekoliko rečenica iz dokumenata i spaja ih u pseudosažetak. GSG zatim koristi ove pseudo-sažetke kao oznake za trening modela. Općenito možemo zaključiti kako je trend razvoja metoda i alata za automatsko sažimanje teksta u uskoj vezi s razvojem umjetne inteligencije (AI) općenito i posebno područja obrade prirodnog jezika (NLP).

7 Zaključak

Važnost i potreba za automatskim sažimanjem teksta neprestano raste iz razloga što mnogi autori gledaju na njega kao na rješenje pretrpanosti informacija. Područje automatskog sažimanja teksta izrazito je rašireno i privlači ne samo stručnjake već i sve širu publiku, kao npr. studente, profesore, medicinsko osoblje, korisnike e-maila i sl.

Najčešći medij prijenosa informacije je Internet. Preko njega je moguće doći do milijardi informacija od kojih su samo neke od koristi. Samo jedno pretraživanje kao što je npr. „evolucija čovjeka“ daje nam preko 150 tisuća rezultata. Kako bi korisnici pronašli ono što žele potrebno je bolje organizirati informacije na webu. Glavna svrha organizacije informacija jest smanjiti redundanciju, odnosno riješiti se dupliciranih i neorganiziranih informacija, te učiniti nešto s dinamičnim informacijskim izvorima. Alati za sažimanje teksta stvaraju sažetke koji pružaju korisniku sve važne informacije tekstualnog dokumenta koji bi trebao služiti kao zamjena za čitanje istog. Čitanjem sažetka korisnik može uštedjeti vrijeme, zaključiti da ga dokument ne zanima ili da ne sadrži informacije koje će zadovoljiti njegovu informacijsku potrebu. Međutim alati za sažimanje teksta imaju i negativnih strana. Osim što su skupi, alati za sažimanje još nisu dosegli očekivanu kvalitetu u kreiranju sažetaka koji su koherentni, sažeti i koji obuhvaćaju sve važne informacije. Kreirani sažetci trebali bi dostići kvalitetu ravnopravnu onim sažetcima koje je kreirao čovjek kako bi se alati za sažimanja teksta mogli primijeniti umjesto čovjeka.

Kako bi se ta očekivanja ispunila u budućnosti, izuzetno je bitna evaluacija, odnosno procjena postojećih alata. Evaluacijom bi se ocijenilo zadovoljstvo rezultata, te ukazao smjer za napredak. U procesu evaluacije jako su bitne metrike kojima se ocjenjuje uspješnost kreiranog sažetka. Vidljiv problem u informacijskim ustanovama jest nedostatak kvalitete sažetka ili nedostatak sažetka općenito jer nije svaki autor iskusan u području pisanja sažetka. Stoga je sažimanje teksta u području informacijskih ustanova korisno prilikom stvaranja sažetaka knjiga u digitalnim repozitorijima ili web stranica knjižnica kao i u bilo kojim drugim većim bazama podataka.

Zahvaljujući sve većim razvojem alata za sažimanje teksta rastući broj informacija i tekstualnih dokumenata više ne predstavlja problem jer se oni više ne obrađuju ručno. Iako još nismo tamo gdje bismo htjeli biti sa sažimanjem teksta, budući trendovi ukazuju na veliki potencijal koji bi se uskoro mogao ostvariti.

8 Popis literature

Abualigah, L., Bashabsheh, M. Q., Alabool, H., & Shehab, M. (2020). Text summarization: a brief review.

Recent Advances in NLP: the case of Arabic language, 1-15.

Abdaljalil, S., & Bouamor, H. (2021). An exploration of automatic text summarization of financial reports. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing* (pp. 1-7).

Andhale, N., & Bewoor, L. A. (2016, August). An overview of text summarization techniques. In 2016 international conference on computing communication control and automation (ICCUBEA) (pp. 1-7). IEEE.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017).

Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268.

Alliheibi, F. M., Omar, A., & Al-Horais, N. (2021). An Evaluation of Automatic Text Summarization of News Articles: The Case of Three Online Arabic Text Summary Generators. *International Journal of Advanced Computer Science and Applications*, *12*(5).

Ardiansyah, T., Purwaningsih, D., & Teruna, D. (2021). Pelatihan Menggunakan Aplikasi Scholarcy Mempermudah Pembuatan Jurnal. *Jurnal PKM: Pengabdian kepada Masyarakat Vol*, *4*(06).

Bagarić, B., & Jokić-Begić, N. (2020). Pretraživanje zdravstvenih informacija na internetu–implikacije za zdravstvenu anksioznost kod starijih osoba. Psychological Topics, 29(2), 401-425.

Bhat, I. K., Mohd, M., & Hashmy, R. (2018). Sumitup: A hybrid single-document text summarizer. In *Soft*

computing: Theories and applications (pp. 619-634). Springer, Singapore.

Bird, C., Ford, D., Zimmermann, T., Forsgren, N., Kalliamvakou, E., Lowdermilk, T., & Gazit, I. (2022).

Taking Flight with Copilot: Early insights and opportunities of AI-powered pair-programming tools.

Queue, *20*(6), 35-57.

Brajković, E., Volarić, T., & Vasić, D. (2018.) Pregled tehnika automatskog sažimanja teksta. *Fakultet prirodoslovnomatematičkih i odgojnih znanosti, Matice hrvatske bb*, *5*.

Chandha, S., Sucheth, R., & Ghosal, T. (2023). Setting the Scene: How Artificial Intelligence is reshaping how we consume and deliver research. Upstream.

Pristupljeno: 10.3.2023. https://upstream.force11.org/setting-the-scene-ai/

Chowdhury, G. G. (1999). The Internet and information retrieval research: A brief review. Journal of Documentation.

Christian, H., Agus, M. P., & Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, *7*(4), 285-294.

Collen, M. F. (2011). Computer medical databases: the first six decades (1950–2010).

Costa, B. D. S. (2020). Avaliação de sumarizadores automáticos de textos em inglês a partir da ferramenta computacional Rouge.

Čizmić, J. (2009). Pravo na pristup podacima u medicinskoj dokumentaciji. *Zbornik Pravnog fakulteta Sveučilišta u Rijeci*, *30*(1), 91-134.

Dernoncourt, F., Ghassemi, M., & Chang, W. (2018, May). A repository of corpora for summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization:
A comprehensive survey. *Expert Systems with Applications*, *165*, 113679.

Eric Conrad, ... Joshua Feldman, 2016. „Hardware Redundancy“ ScienceDirect.com. Objavljeno 2016 Pristupljeno: 12.9.2022. https://www.sciencedirect.com/topics/computer-science/hardware-redundancy

Fattah, M. A. (2014). A hybrid machine learning model for multi-document summarization. *Applied intelligence*, *40*(4), 592-600.

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, 47(1), 1-66.

Gong, Y., & Liu, X. (2001, September). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 19-25).

Gooch, P. How Scholarcy contributes to and makes use of open citations.

Pristupljeno: 13.2.2023. https://www.scholarcy.com/how-scholarcy-contributes-to-and-makes-use-of-open-citations/

Gupta, V., & Lehal, G. S. (2010). A survey of text summarization extractive techniques. Journal of emerging technologies in web intelligence, 2(3), 258-268.

Hartman, K. A., & Bowering Mullen, L. (2008). Google Scholar and academic libraries: an update. New library world, 109(5/6), 211-222.

Hasan, A. M., Mohd Noor, N., Rassem, T. H., Mohd Noah, S. A., & Hasan, A. M. (2020). A proposed method using the semantic similarity of WordNet 3.1 to handle the ambiguity to apply in social media text. In *Information Science and Applications: ICISA 2019* (pp. 471-483). Springer Singapore.

Huang, L., He, Y., Wei, F., & Li, W. (2010, April). Modeling document summarization as multi-objective optimization. In 2010 Third International Symposium on Intelligent Information Technology and Security Informatics (pp. 382-386). IEEE.

Hoang, M. (2022). Developing a Video Summarizing Tool using Machine Learning.

Hsu, T. Y., Suhara, Y., & Wang, X. (2022). Summarizing Community-based Question-Answer Pairs. arXiv preprint arXiv:2211.09892.

Jangra, A., Mukherjee, S., Jatowt, A., Saha, S., & Hasanuzzaman, M. (2021). A survey on multi-modal summarization. *ACM Computing Surveys*.

Kinga, S., & Gupta, G. S. (2021). Platforms as foundation of sharing economy. Delhi Business Review, 22(1), 1–13. https://doi.org/10.51768/dbr.v22i1.221202101

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019). Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Lakić, I. (2014). Analiza pisanog diskursa. Analiza diskursa: teorije i metode, 57-77.

Li, C., Xu, W., Li, S., & Gao, S. (2018, June). Guiding generation for abstractive text summarization based on key information guide network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 55-60).

Liu, Y., Li, S., Cao, Y., Lin, C. Y., Han, D., & Yu, Y. (2008, August). Understanding and summarizing answers in community-based question answering services. In Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008) (pp. 497-504).

Lynch, A. M., Kilroy, S., McKee, H., Sheerin, F., Epstein, M., Girault, A., ... & McKee, G. (2023). Active older adults goal setting outcomes for engaging in a physical activity app and the motivation characteristics of these goals (MOVEAGE-ACT). Preventive Medicine Reports, 31, 102084.

Mahajani, A., Pandya, V., Maria, I., & Sharma, D. (2019). A comprehensive survey on extractive and abstractive techniques for text summarization. *Ambient communications and computer systems*, 339-351.

Majstorović, M. (2020). RDF model podataka (Doctoral dissertation, Josip Juraj Strossmayer University of Osijek. Faculty of Humanities and Social Sciences. Department of Information Sciences).

Marcos, M. C., Gavin, F., & Arapakis, I. (2015, September). Effect of snippets on user experience in web search. In *Proceedings of the XVI International Conference on Human Computer Interaction* (pp. 1-8).

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*(1), 871-906.

Mikić Čolić, A., & Trtanj, I. (2019). O koheziji i koherenciji teksta. *Suvremena lingvistika*, *45*(88), 247-264.

Mohammed, N. (2020). Extracting word synonyms from text using neural approaches. *Int. Arab J. Inf.*
Technol., *17*(1), 45-51.

Nazari, N., & Mahdavi, M. A. (2019). A survey on automatic text summarization. *Journal of AI and Data Mining*, *7*(1), 121-135.

Nenkova, A., & McKeown, K. (2011). Automatic summarization. *Foundations and Trends® in Information Retrieval*, *5*(2–3), 103-233.

Nenkova, A., & McKeown, K. (2012) „A survey of text summarization techniques“. Mining text data, 43-76.

Neto, J. L., Freitas, A. A., & Kaestner, C. A. (2002, November). Automatic text summarization using a machine learning approach. In *Brazilian symposium on artificial intelligence* (pp. 205-215). Springer, Berlin, Heidelberg.

Ningrum, W. L. (2021). Analisis Perbandingan Kualitas Website Peringkas Teks Otomatis Menggunakan Model ISO 9126. *ICIT Journal*, *7*(2), 222-232. https://doi.org/https://doi.org/10.33050/icit.v7i2.1651

Paper digest n.d. „Paper digest.“ Pristupljeno: 12.3.2023. https://www.paper-digest.com/

Patel, D., Shah, S., & Chhinkaniwala, H. (2019). Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique. Expert Systems with Applications, 134, 167-177.

Petrović, Z., & Bušelić, V. (2020). Automatsko sažimanje hrvatskog teksta. *Polytechnic and design*, *8*(2), 121-128.

Prasasthy K B, 2021. „Brief history of Text Summarization“. Medium.com. Objavljeno lipanj 17, 2021 Pristupljeno: 25.8.2022. https://medium.com/@prasasthy.sanal/brief-history-of-text-summarization-9d1b3787a707

Putica, M. (2018). Semantički web. Hum: časopis Filozofskog fakulteta Sveučilišta u Mostaru, 13(19), 99-116.

Quillbot. n.d. „Quillbot.“ Pristupljeno: 12.3.2023. https://quillbot.com/

Radev, D., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization.
Computational linguistics, 28(4), 399-408.

Radić, B. (2017). Zaštita pisane baštine u zajednici AKM-a (arhivi-knjižnice-muzeji): Franjo Ksaver Kuhač (1834.-1911.) – utemeljitelj hrvatske etnomuzikologije i glazbene historiografije. *Vjesnik bibliotekara Hrvatske, 60* (2-3), 25-45. Preuzeto s https://hrcak.srce.hr/195864

Rakhmanina, L., & Serasi, R. (2022). UTILIZING QUILLBOT PARAPHRASER TO MINIMIZE PLAGIARISM IN STUDENTS'SCIENTIFIC WRITING. *Novateur Publications*, 26-33.

Rasool, A., & Jalil, Z. (2020). A review of web browser forensic analysis tools and techniques.
Researchpedia Journal of Computing, *1*(1), 15-21.

Renn, O. (2021). Science communication in crisis?: Can new technologies help and support?. *Bulletin VSH-AEU*, *3*(S).

Resoomer. n.d. „Resoomer.“ Pristupljeno: 12.3.2023. https://resoomer.com/en/
Roy, B. K., & Mukhopadhyay, P. (2022). Digital Access Brokers: Clustering and Comparison (Part II– from Summarization to Citation Map). SRELS Journal of Information Management, 59(6), 337-351.

Saggion, H. (2008). A robust and adaptable summarization tool. Traitement Automatique des Langues, 49(2), 68.

Saggion, H., & Poibeau, T. (2013). Automatic text summarization: Past, present and future. In *Multi-source, multilingual information extraction and summarization* (pp. 3-21). Springer, Berlin, Heidelberg.

Scialom, T., Lamprier, S., Piwowarski, B., & Staiano, J. (2019). Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Scholarcy. n.d. „Scholarcy.“ Pristupljeno: 12.3.2023. https://www.scholarcy.com/

Shannaq, B., & Adebiaye, R. (2015). Analytic-Synthetic Processing Of Information as Smart-Based Environment for Text Summarization. *International Journal of Innovative Research in Science,*
Engineering and Technology. Retrieved from https://www. researchgate. net/publication/271271153_Analytic-Synthetic_Processing_Of_Information_as_Smart-Based_Environment_for_ Text_Summarization [in English].

Slamet, C., Atmadja, A. R., Maylawati, D. S., Lestari, R. S., Darmalaksana, W., & Ramdhani, M. A. (2018). Automated text summarization for indonesian article using vector space model. In *IOP Conference Series: Materials Science and Engineering* (Vol. 288, No. 1, p. 012037). IOP Publishing.

SMMRY. n.d. „SMMRY“ Pristupljeno: 12.3.2023. https://smmry.com/

Srirejeki, K., Faturokhman, A., Praptapa, A., & Irianto, B. S. (2022). Academic misconduct: Evidence from online class. *Int J Eval & Res Educ*, *11*(4), 1893-1902.

Sucheth, 2022. „Introducing Copilot: Your AI assistant that helps explain papers“. Typeset.oi objavljeno 5.12.2022. Pristupljeno: 10.3.2023. https://typeset.io/resources/introducing-copilot-ai-assistant-explains-research-papers/

Sudeepthi, G., Anuradha, G., & Babu, M. S. P. (2012). A survey on semantic web search engine.
International Journal of Computer Science Issues (IJCSI), *9*(2), 241.

Suleiman, D., & Awajan, A. (2020). Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges. *Mathematical problems in engineering*, *2020*.
Svenonius, E., & Willer, M. (2005). *Intelektualne osnove organizacije informacija*. Benja.

Syed, A. A., Gaol, F. L., & Matsuo, T. (2021). A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access*, *9*, 13248-13265.

Špiranec, S. (2007). Model organizacije informacija u elektroničkoj obrazovnoj okolini (Doctoral dissertation).

Šuman, S. (2021). Pregled metoda obrade prirodnih jezika i strojnog prevođenja. *Zbornik Veleučilišta u Rijeci*, *9*(1), 371-384.

Takeda, T., & Takasu, A. (2009). Information Organization System for Duplicated Information Sources. In IADIS International Conference Information Systems 2009 (pp. 73-80).

Takeuchi, K. (2002). A Study on Operations used in Text Summarization. (PhD thesis), Nara Institute of Science and Technology.

Taylor, A. G., & Joudrey, D. N. (2004). The organization of information.

Todd, R. W., Thienpermpool, P., & Keyuravong, S. (2004). Measuring the coherence of writing using topic-based analysis. Assessing writing, 9(2), 85-104.

Tools 4 noobs. n.d. „Tools 4 noobs“ Pristupljeno: 12.3.2023. https://www.tools4noobs.com/

Van Berkel, N., Goncalves, J., Russo, D., Hosio, S., & Skov, M. B. (2021, May). Effect of information presentation on fairness perceptions of machine learning predictors. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

Verma, N., & Tiwari, A. (2014). A survey of automatic text summarization. *Int. J. Eng. Res. Technol*, *3*(6).

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., & Affandy, A. (2020). Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences*.

Woodsend, K., & Lapata, M. (2010, July). Automatic generation of story highlights. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 565-574).

Wu, Z., Lei, L., Li, G., Huang, H., Zheng, C., Chen, E., & Xu, G. (2017). A topic modeling based approach to novel document automatic summarization. Expert Systems with Applications, 84, 12-23.

Xue, F., & Lu, W. (2020). A semantic differential transaction approach to minimizing information redundancy for BIM and blockchain integration. *Automation in construction*, *118*, 103270.

Yadav, D., Desai, J., & Yadav, A. K. (2022). Automatic Text Summarization Methods: A Comprehensive Review. *arXiv preprint arXiv:2204.01849*.

9 Prilozi

9.1 Popis slika

Slika 1 Konceptualni prikaz ekstrakcijske i apstrakcijske metode sažimanja teksta (Izvor: Yadav, D i Desai Yadav, A., 2022, str 3) _____ *3*

Slika 2 Prikaz podjele pristupa sažimanja teksta i njihovih metoda (Izvor: El-Kassas, Salama, Rafea i Mohamed 2021., str. 8) _____ *5*

Slika 3 Prikaz klasifikacije sustava automatskog sažimanja teksta (Izvor: El-Kassas, Salama, Rafea i Mohamed 2021., str. 5) _____ *8*

Slika 4 Taksonomija evaluacijskih mjera sažetaka (Izvor: Gambhir i Gupta 2017., str. 45) _____ *16*

Text summarization as a useful tool in the organization, presentation and information retrieval

Summary

The subject of this paper is automatic text summarization in the context of organization, presentation and search of information. Automatic text summarization is applicable in many fields; however, this paper examines the role of text summarization in the domain of informational sciences. This paper presents different outlooks on text summarization and the primary methods of performing said task. Apart from two main, extract and abstract methods, a hybrid method and present-day metrics are also mentioned. Furthermore, the paper deals with shortcomings of methods in text summarization and the process of the summarization itself, which are redundancy, financial investments, double meanings, synonyms, cohesion and coherence. Nevertheless, automatic text summarization is helpful in processing text documents, especially with today's exponential increase of such information. The final goal of the text summarization field is to create high-quality summaries similar to those produced by humans, but without the need to monitor the summarization process. The evaluation of text summarization is emphasized, as well as the possible advancements of summaries created by tools for text summarization during the evolution of NLP and artificial intelligence in general.

Keywords: text summarization, extractive summarization method, abstractive summarization method, information retrieval and presentation, data and information organization