Slovenski meta-povzemalnik

Aleš ŽAGAR
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Marko ROBNIK-ŠIKONJA
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Povzetek

Povzemanje besedil je pomembna naloga obdelave naravnega jezika, zato so raziskovalci v zadnjih letih razvili različne pristope, od sistemov, ki temeljijo na pravilih, do nevronskih mrež, ki v zadnjem času prevladujejo. Žal ne obstaja en sam model ali pristop, ki bi dobro deloval na vseh vrstah besedil, zato predlagamo meta-model, ki priporoča najprimernejši povzemalnik za določeno besedilo. Predlagani meta-model uporablja polno povezano nevronsko mrežo, ki analizira vhodno vsebino in napove, kateri povzemalnik bi bil za dano vhodno besedilo najboljši v smislu ocene ROUGE. Meta-model izbira med štirimi različnimi modeli povzemanja, razvitimi za slovenščino, pri čemer uporabi različne lastnosti vhodnega besedila, zlasti njegovo dokumentno predstavitev Doc2Vec. Štirje uporabljeni slovenski povzemalniki naslavljajo različne izzive, povezane s povzemanjem besedil v jeziku z manj viri. Delovanje predlaganega modela SloMetaSum ovrednotimo avtomatsko, njegove dele pa tudi ročno. Rezultati kažejo, da sistem uspešno avtomatizira korak ročne izbire najboljšega modela za dano besedilo.

Ključne besede: povzemanje besedil, jeziki z manj viri, meta-model, slovenski jezik

Abstract

Text summarization is an essential task in natural language processing, and researchers have developed various approaches over the years, ranging from rule-based systems to neural networks. However, there is no single

model or approach that performs well on every type of text. We propose a system that recommends the most suitable summarization model for a given text. The proposed system employs a fully connected neural network that analyzes the input content and predicts which summarizer should score the best in terms of ROUGE score for a given input. The meta-model selects among four different summarization models, developed for the Slovene language, using different properties of the input, in particular its Doc2Vec document representation. The four Slovene summarization models deal with different challenges associated with text summarization in a less-resourced language. We evaluate the proposed SloMetaSum model performance automatically and parts of it manually. The results show that the system successfully automates the step of manually selecting the best model.

Keywords: text summarization, low-resource languages, meta-model, Slovene language

1 Uvod

Povzemanje besedila izbere bistvene informacije v dokumentu ali zbirki dokumentov ter jih predstavi na kratek in koherenten način. Kljub dolgotrajnim prizadevanjem raziskovalcev na področju obdelave naravnega jezika (NLP) je povzemanje besedil še vedno zahtevna naloga. Z eksplozivno rastjo digitalnih informacij postaja povzemanje velikih količin besedil v krajšo in bolj obvladljivo obliko vse pomembnejše.

Obstajata dva glavna pristopa k povzemanju besedil: ekstraktivni in abstraktivni. Ekstraktivno povzemanje izbere podmnožico stavkov ali besednih zvez iz izvornega besedila, ki najbolje predstavljajo vsebino. Izbrani stavki se združijo v povzetek. V nasprotju s tem abstraktivno povzemanje ustvarja nove stavke, ki zajamejo pomen izvirnega besedila. Ekstraktivno povzemanje je preprostejše in hitrejše od abstraktivnega povzemanja, vendar lahko privede do povzetkov, ki vsebujejo odvečno in ponavljajočo se vsebino. Abstraktivno povzemanje je zahtevnejše in zahteva naprednejše tehnike

obdelave naravnega jezika, vendar lahko ustvari podobne povzetke kot človek.

Tehnologija povzemanja besedil je v zadnjih letih doživela velik razvoj z arhitekturo nevronskih mrež transformer in na njej osnovanih velikih vnaprej naučenih jezikovnih modelih, kot sta T5 (Raffel idr., 2020) in GPT-3 (Brown idr., 2020). Tako so nastali modeli za povzemanje, katerih povzetki so zelo podobni tistim, ki jih je napisal človek, z malo ponovitvami in netočnostmi, predvsem pri povzemanju novic. Ti modeli so sposobni obdelovati vedno daljše vsebine, kar omogoča izdelavo povzetkov za daljša besedila. Posledično so lahko najsodobnejši samodejni povzetki jasni in enostavni za razumevanje.

V morfološko bogati slovenščini je povzemanje besedil zaradi omejene razpoložljivosti virov in podatkov ter raziskav še večji izziv kot v angleščini. Naučili smo štiri slovenske modele za povzemanje besedil z različnimi lastnostmi na različnih učnih podatkih.[1] Naši štirje modeli zajemajo dva ekstraktivna povzemalnika (eden temelji na enostavnem izboru stavkov s pogostostjo besed, drugi pa na grafu), abstraktivni povzemalnik, ki temelji na modelu T5, in hibridni ekstraktivno-abstraktivni model. Na splošno se najbolje obnese slovenski transformerski model, ki temelji na modelu T5, vendar se ne nujno dobro posploši za vse vrste vhodnih besedil. Zato se ukvarjamo s problemom meta-povzemalnika, ki ugotavlja, kateri model povzemanja je najprimernejši za določeno besedilo glede na dolžino in žanr besedila.

Rezultat raziskave je nov slovenski sistem za povzemanje (imenovan SloMetaSum), ki ga sestavljajo ekstraktivni, abstraktivni in hibridni povzemalniki ter meta-model, ki med njimi izbira. Predlagani meta-sistem je sestavljen iz polno povezane nevronske mreže, ki analizira vhodno vsebino in priporoča najprimernejši model povzemanja za določeno besedilo. V ta namen SloMetaSum uporablja vektorsko predstavitev dokumentov Doc2Vec (Le in Mikolov, 2014) in napoveduje ocene ROUGE za vsakega od povzemalnikov. S kombiniranjem večih pristopov lahko sistem učinkovito ustvari kakovostne

1 V sklopu projekta RSDO: https://www.cjvt.si/rsdo/.

povzetke, ki so informativni in lahko razumljivi za več vrst besedil, ne glede na njihovo dolžino in žanr.[2]

Prispevki naše raziskave so naslednji:

- Razvili smo štiri modele za povzemanje, ki lahko učinkovito povzemajo besedila različnih dolžin in žanrov, zaradi česar so vsestransko uporabni.
- Uspešno smo naslovili izzive slovenskega jezika z malo viri in ustvarili visoko učinkovite modele za povzemanje slovenskih besedil.
- Ustvarili smo meta-model, ki na podlagi parametrov, kot so dolžina, zapletenost, raven abstrakcije in predvideni primer uporabe, priporoči najprimernejši model povzemanja za določeno besedilo.

Preostali del prispevka je razdeljen na šest razdelkov. V Razdelku 2 predstavimo sorodne raziskave. V Razdelku 3 so opisani nabori podatkov. V Razdelku 4 opisujemo osnovne povzemalnike in meta-model. V oddelku 5 predstavljamo empirično ovrednotenje in predstavimo ugotovitve. Razdelek 6 vsebuje zaključke in priporočila za nadaljnjo delo.

2 Sorodna dela

Zgodnji pristopi k povzemanju besedil so temeljili na statističnih frekvencah besed, položaju stavkov in stavkih, ki vsebujejo ključne besede (Nenkova in Vanderwende, 2005). Cilj teh pristopov je bil iz besedila izluščiti pomembne stavke ali besedne zveze in z njihovim povezovanjem ustvariti povzetek. Abstraktivne metode so vključevale brisanje manj pomembnih besed iz besedila za oblikovanje povzetka (Knight in Marcu, 2002).

Metode, ki temeljijo na grafih, so priljubljen pristop k povzemanju besedil. Pri tem pristopu je dokument predstavljen kot graf, kjer so povedi vozlišča, povezave pa predstavljajo odnose med njimi.

2 Demonstracijska predstavitev je na voljo na spletni strani https://slovenscina.eu/en/povzemanje. Repozitoriji kode so na voljo na naslovih https://github.com/azagsam/metamodel in https://github.com/clarinsi/SloSummarizer.

Graf se nato uporabi za izdelavo povzetka z izbiro najpomembnejših stavkov. Ta metoda je bila raziskana v več delih (Mihalcea in Tarau, 2004; Erkan in Radev, 2004).

S pojavom nevronskih mrež se je povečalo zanimanje za razvoj tehnik abstraktivnega povzemanja. Prvi nevronski abstraktivni sistemi so uporabljali arhitekturo rekurentnih nevronskih mrež, kot je LSTM (See idr., 2017; Nallapati idr., 2016). Danes najsodobnejši modeli za abstraktivno povzemanje uporabljajo arhitekturo transformer (Zhang idr., 2020; Lewis idr., 2020). Ta arhitektura intenzivno uporablja mehanizem samopozornosti za selektivno osredotočanje na pomembne dele besedila. Modeli s to arhitekturo lahko v primerjavi s prejšnjimi metodami ustvarijo bolj tekoče in koherentne povzetke.

Čeprav je bilo za povzemanje besedil predlaganih več pristopov, so mnogi omejeni na določene žanre besedil. V tem delu je naš cilj zgraditi sistem za povzemanje, ki lahko obravnava raznovrstna besedila različnih žanrov, ki nastopajo v realnosti. To vključuje besedila različnih dolžin, tematik in slogov, s ciljem izdelati povzetke, ki zajamejo najpomembnejše informacije v besedilu. Želimo razviti robusten in prilagodljiv model, ki se lahko nauči kakovostno povzemati besedila različnih vrst.

3 Učne množice

V tem razdelku opisujemo nabore podatkov, ki smo jih uporabili v naši raziskavi. V nadaljevanju podajamo kratek opis podatkovnih množic, v Tabeli 1 pa njihove statistične lastnosti.

Množica STA (splošni novinarski članki Slovenske tiskovne agencije) obsega 366.126 dokumentov. Kot približek povzetka smo uporabili prvi odstavek vsakega članka, saj ta množica ne vsebuje ročno napisanih človeških povzetkov. Naša izbira povzetka sledi pogosti praksi pri povzemanju besedil, zlasti v jezikih, ki nimajo namenskih učnih množic za povzemanje novic.

AutoSentiNews (Bučar, 2017) je podoben nabor besedil kot STA; sestavljen iz 256.567 člankov iz slovenskih novičarskih portalov

24ur, Dnevnik, Finance, RTVSlo in Žurnal24. Povzetki so izdelani iz prvega odstavka na enak način kot v podatkovni zbirki STA.

Učna množica SURS je manjša zbirka finančnih novic Statističnega urada Slovenije in obsega 4.073 dokumentov.

Korpus slovenskih akademskih besedil KAS (Žagar idr., 2022) je sestavljen iz diplomskih, magistrskih in doktorskih del, napisanih med letoma 2000 in 2018, ki so zbrana v digitalnih knjižnicah slovenskih visokošolskih zavodov ter dostopna na portalu odprte znanosti.[3] Korpus vsebuje človeške povzetke akademskih besedil.

Nabor podatkov CNN/Daily Mail (Hermann idr., 2015) je namenjen povzemanju besedil. Vsebuje s strani človeka ustvarjene izvlečke povzetkov iz novic na spletnih straneh CNN in Daily Mail. Korpus ima 286.817 učnih parov, 13.368 validacijskih parov in 11.487 testnih parov. Izvorni dokumenti imajo v povprečju 766 besed, povzetki pa 53 besed. Zbirko podatkov smo prevedli v slovenščino z uporabo strojnega prevajanja (Lebar Bajec idr., 2022).

Tabela 1: Korpusi in učne množice, uporabljene za učenje modela za predstavitev dokumentov Doc2vec in meta-modela.

Množica	Število dokumentov
STA	334.696
AutoSentiNews	256.567
SURS	4.073
CNN/Daily Mail	311.672
KAS	82.308
Skupno	**677.644**

4 Povzemalni modeli in meta-model

V tem razdelku opisujemo sestavne dele našega sistema SloMetaSum, ki ga sestavljajo štirje povzemalniki, sistem za predstavitev dokumentov in meta-model.

3 http://openscience.si/

4.1 Povzemalni modeli

Izdelali smo štiri povzemalnike, ki jih v nadaljevanju na kratko opišemo.

Osnovni povzemalnik (Nenkova in Vanderwende, 2005) za izbiro najbolj informativnih stavkov uporablja preprost pristop s frekvenco besed. Model **Grafi**, zasnovan na grafih (Žagar in Robnik-Šikonja, 2021), se zgleduje po algoritmu TextRank (Mihalcea in Tarau, 2004) in za razvrščanje stavkov uporablja ocene centralnosti stavkov. Oba modela spadata med ekstraktivne metode in se lahko uporabljata na dokumentih poljubne velikosti. V nasprotju z izvornim pristopom TextRank smo za numerično predstavitev stavkov uporabili kodirnik stavkov LaBSE (Feng idr., 2022), ki temelji na arhitekturi transformer. Model abstraktivnega povzemanja **T5-članki** uporablja vnaprej naučen slovenski model SloT5 (Ulčar in Robnik-Šikonja, 2023) in je prilagojen na strojno prevedenem naboru člankov CNN/Daily Mail (Hermann idr., 2015) z uporabo slovenskega sistema strojnega prevajanja (Lebar Bajec idr., 2022). Model **Hibrid-dolga** je kombinacija modela, ki temelji na grafu, in modela T5-članki. Najprej sestavi kratko besedilo z združevanjem najbolj informativnih stavkov (ekstraktivni korak). V naslednjem, abstraktivnem koraku pa se ti stavki povzamejo s povzemalnikom T5-članki.

4.2 Predstavitev dokumentov z modelom Doc2Vec

Za izbiro najprimernejše metode povzemanja za določeno besedilo mora meta-model pridobiti informacije o različnih lastnostih besedila. Za predstavitev dokumentov uporabimo model Doc2Vec in ga naučimo na slovenskih besedilih iz Tabele 1 (brez povzetkov). V koraku predobdelave smo odstranili visokofrekvenčne besede, ki ne prispevajo k pomenu dokumentov, kot so zaimki, vezniki itd.; da bi dodatno zmanjšali število različnih besed, smo celotno zbirko besedil lematizirali.

4.3 Meta-model

Naš meta-model je sestavljen iz polno povezane nevronske mreže, naučene za napovedovanje ocen ROUGE povzetkov. Za učno

množico smo naključno izbrali 93.419 primerov iz celotne zbirke surovih besedil. Najprej je vsak od naših štirih povzemalnikov pripravil povzetek za vse primere. Izračunali smo ocene ROUGE med referenčnimi in izdelanimi povzetki. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) je metrika, ki se najpogosteje uporablja za ocenjevanje samodejno ustvarjenih povzetkov. Kakovost povzetka meri s številom prekrivajočih se enot (n-gramov, zaporedij besedil itd.) med povzetki, ki jih je ustvaril človek, in povzetki, ki so jih ustvarili sistemi za povzemanje. ROUGE ni ena sama metrika, ampak družina metrik. Najpogosteje se uporabljata ROUGE-N in ROUGE-L. Prva meri prekrivanje n-gramov (običajno unigramov in bigramov), druga pa meri najdaljšo skupno zaporedje besed v obeh povzetkih. Kot vhodni podatek za naš meta-model uporabljamo štiri ocene F1 ROUGE (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-LSum), ki kažejo, kako dobri so ustvarjeni povzetki. Podatke smo razdelili na učno, validacijsko in testno množico v razmerju 90 : 5 : 5.

Velikosti obeh podatkovnih množic sta predstavljeni v Tabeli 2. V Tabeli 3 so predstavljene povprečne vrednosti ROUGE naših povzetkov za dolga in kratka besedila. Povzemalniki, ki so specializirani za kratka besedila, dosegajo boljše rezultate na kratkih besedilih in obratno.

Tabela 2: Število učnih primerov za model predstavitve besedil in meta-povzemalnik.

Model	Velikost učne množice
Doc2Vec	677.644
Meta-model	93.419

Tabela 3: Ocene povzetkov, izražene z mero ROUGE, za dolga in kratka besedila. Najboljši rezultati za kratka in dolga besedila so v krepkem tisku.

	T5-članki	Osnovni	Grafi	Hibrid-dolga
Kratka	**14,01**	13,11	13,15	12,55
Dolga	10,51	13,12	**17,71**	17,59

5 Rezultati

V tem razdelku predstavimo načrt evalvacije modelov in rezultate. Rezultate predstavitvene metode Doc2Vec in povzemalnikov podamo v ločenih podrazdelkih.

5.1 Doc2Vec

Za učenje modela za predstavitev dokumentov Doc2Vec smo uporabili naslednje hiperparametre: največja dovoljena velikost slovarja je 100 000, velikost vektorja za predstavitev besed je 256, velikost okna konteksta je 5, najmanjša frekvenca besede, ki se vključi v slovar, je 1, skupno število epoh za učenje modela pa je 5.

Model Doc2Vec smo ocenili ročno in z avtomatskimi merami. Pri ročni analizi smo za vsakega od nekaj naključno izbranih vzorcev z uporabo kosinusne podobnosti pregledali 3 najbolj podobne vrnjene dokumente in opazovali, ali se teme dokumentov prekrivajo. Teme dokumentov so si bile v večini primerov podobne. Na podlagi tega smo sklepali, da model deluje v skladu s pričakovanji. Avtomatsko ocenjevanje je bilo del celotnega cevovoda, v katerem so bili hiperparametri modela prilagojeni za optimizacijo funkcije izgube meta-modela.

5.2 Meta-model

Naši končni rezultati so predstavljeni v Tabeli 4. Predlagani mehanizem za izbiro povzemalnih modelov smo primerjali s tremi osnovnimi mehanizmi. Mehanizem *Povprečje* vzame napovedi za vsak model povzemanja in jih povpreči. Vedno izbere model z najvišjim številom točk. Mehanizem *Drevo* uporablja regresijsko drevo; z uporabo iskanja po mreži hiperparametrov je najmanjše število vzorcev, potrebnih za razdelitev notranjega vozlišča, 100. Mehanizem *Gozd* uporablja naključni gozd; eksperimentirali smo s podobnimi vrednostmi kot pri mehanizmu Drevo in določili število dreves na 300.

Naš najboljši mehanizem za izbor povzemalnikov je nevronska mreža z dvema skritima nivojema. Skrita nivoja vsebujeta po 1024

nevronov, med postopkom učenja pa smo uporabili 10 % primerov iz učne množice za validacijo. Aktivacijska funkcija, uporabljena za ta model, je popravljena linearna enota (ReLU). Uporabili smo strategijo zgodnje ustavitve učenja s parametrom potrpežljivosti 2. Funkcija izgube, uporabljena za ta model, je povprečna kvadratna napaka.

Meta-model se je prenehal učiti po 7 epohah in na testni množici dosegel skoraj 15 točk nad srednjo osnovno vrednostjo. Opazili smo, da izbira različnih hiperparametrov ne vpliva bistveno na rezultate. Eksperimentirali smo z različnimi velikostmi skritih plasti, številom enot in aktivacijskimi funkcijami. Preizkusili smo tudi različne velikosti največjega besedišča in oken modela Doc2Vec. Navajamo samo vrednosti najboljšega modela.

Na splošno se je ta model izkazal za najučinkovitejšega med preizkušenimi mehanizmi za izbor najprimernejšega povzemalnika. Veliko število nevronov v skritem sloju je verjetno prispevalo k njegovi boljši učinkovitosti, saj omogoča večjo stopnjo kompleksnosti pri predstavitvi podatkov v modelu.

Tabela 4: Rezultati štirih mehanizmov za izbor povzemalnikov na testni množici. Meta-model-osnovni je pokazal znatno izboljšanje v primerjavi z metodami Povprečje in Drevo. Izrecno kodiranje dolžine besedil ali uravnoteženje podatkovne množice nista izboljšali rezultatov.

Model	Srednja kvadratna napaka
Povprečje	84,493
Drevo	81,631
Gozd	74,975
Meta-model-osnovni	**70,066**
Meta-model+dolžina	70,146
Meta-model+uravnoteženje	79,044

Nadalje smo preizkusili dve različici meta-modela. Meta-model+dolžina doda še en vhodni nevron, ki izrecno kodira vhodno dolžino. Ugotovili smo, da to ne izboljša modela; domnevamo, da so akademska besedila različnih žanrov, kar naša predstavitev dokumentov že dobro pokriva. Poskušali smo tudi uravnotežiti podatke, saj prvotni nabor podatkov vsebuje razmerje 1 : 5 med dolgimi in

kratkimi besedili, kar povečuje morebitno težavo pretiranega prilagajanja kratkim besedilom. Zmanjšali smo število kratkih besedil v učni množici, da smo dobili uravnoteženo učno množico s 16,932 povzetki za naš uravnoteženi meta-model. Rezultat tega uravnoteženja je slabši model, vendar še vedno boljši od osnovnega modela Povprečje.

Tabela 5 prikazuje frekvence, kolikokrat je meta-model priporočil vsak model od 1.000 vzorcev iz testne množice. Vidimo lahko, da je bil model T5-članki priporočen največkrat, in sicer 595-krat od 1.000 vzorcev. Model Hibrid-dolgi je bil priporočen 254-krat, sledi mu model Osnovni, ki je bil priporočen 80-krat. Model Graf je bil priporočen najmanjkrat, in sicer 71-krat od 1000 vzorcev.

Tabela 5: Pogostost priporočil meta-modela za vsakega od osnovnih povzemalnikov na vzorcu 1.000 povzetkov iz testne množice.

Model	Frekvenca
T5-članki	595
Hibrid-dolga	254
Osnovni	80
Grafi	71
Total	1.000

Tabela 6: Klasifikacijski rezultat meta-modela, izražen z natančnostjo, priklicem in mero F1 za vsako metodo ter število primerov v testni množici (podpora). Primerjamo uspešnost meta-modela pri izbiri modelov T5-članki, Hibrid-dolga, Osnovni in Grafi.

Povzemalnik	Natančnost	Priklic	Mera F1	Podpora
T5-članki	0,33	0,11	0,16	1.069
Hibrid-dolga	0,25	0,34	0,29	817
Osnovni	0,28	0,10	0,15	1.196
Grafi	0,38	0,67	0,48	1.589

Glede na Tabelo 6 je meta-model najbolj učinkovito priporočal povzemalnik, ki temelji na grafu, z rezultatom F1 0,48, natančnostjo 0,38 in priklicem 0,67. Za model Hibrid-dolga je dosegel oceno

F1 0,29, z natančnostjo 0,25 in priklicem 0,34, pri modelu T5-članki pa F1 0,16, natančnost 0,33 in priklic 0,11. Najnižji rezultat je dosegel pri metodi Osnovni z F1 0,15, natančnostjo 0,28 in priklicem 0,15. Večinski klasifikator ima na testni množici klasifikacijsko točnost 0,34. Klasifikacijska točnost meta-modela pa je bila 0,34.

5.3 Meta-model proti ostalim

V Tabeli 7 so predstavljeni končni rezultati ocenjevanja, pridobljeni s poskusi na testni množici. Omeniti velja, da je predlagani meta-model presegel vse druge modele pri vseh ocenah ROUGE. Ta rezultat poudarja učinkovitost in superiornost meta-modela pri izbiri najprimernejšega povzemalnika za določeno besedilo. Rezultat tudi kaže, da je avtomatizacija postopka izbire najboljšega modela za povzemanje koristna in odpravlja potrebo po ročni izbiri povzemalnika.

Tabela 7: Uspešnost na testni množici za vse modele. Meta-model doseže najboljše rezultate v vseh treh kategorijah.

Povzemalnik	ROUGE-1	ROUGE-2	ROUGE-L
T5-članki	19,01	5,61	13,52
Grafi	19,47	5,52	12,50
Hibrid-dolga	18,55	5,42	11,73
Osnovni	18,86	5,04	12,25
Meta-model	**20,38**	**5,85**	**13,67**

6 Zaključki

V sestavku predlagamo nov povzemalni sistem, ki vsebuje dva ekstraktivna, abstraktivni, hibridni in meta-model povzemanja. Novost je meta-model, ki je sestavljen iz polno povezane nevronske mreže, ki analizira vhodno vsebino in priporoča najprimernejši model povzemanja zanjo. Pristop deluje na kratkih in dolgih besedilih različnih žanrov in omogoča učinkovito in uspešno izdelavo kakovostnih povzetkov za slovenska besedila.

Čeprav predlagani model SloMetaSum predstavlja inovativno rešitev problema izbire najprimernejšega povzemalnika za dano besedilo, ni brez slabosti. Ena glavnih pomanjkljivosti je zanašanje na oceno ROUGE kot edino merilo za izbiro modelov. Čeprav je ROUGE pogosto uporabljena metrika na področju povzemanja besedil, ne odraža vedno dobro kakovosti povzetka ter ne zajame njegove koherentnosti in berljivosti. Druga morebitna slabost je omejen obseg študije, ki se osredotoča izključno na slovenski jezik. Čeprav so štirje modeli povzemanja, razviti za slovenščino, pomemben prispevek k temu področju, se morda ne bodo v enaki meri posplošili na druge jezike z manj viri, saj je za to potreben dober sistem za strojno prevajanje.

V prihodnjem delu bi bilo smiselno sistem razširiti na druge jezike. Sistem bi bilo smiselno primerjati z najnovejšimi povzemalnimi pristopi za velike jezike, predvsem angleščino. Poleg samodejnega ocenjevanja kakovosti sistema bi bilo koristno izvesti tudi uporabniške študije, da bi ocenili njegovo uporabnost in učinkovitost v realnih scenarijih. Podrobneje bi lahko analizirali delovanje sistema pri povzemanju novic, akademskih člankov in drugih vrst realnih vsebin, kjer se povzemanje pogosto uporablja.

Zahvala

Delo sta podprla Ministrstvo za kulturo Republike Slovenije preko projekta Razvoj slovenščine v digitalnem okolju (RSDO) in Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS) preko raziskovalnega programa P6-0411 in projektov J6-2581, J7-3159 in CRP V5-2297. Projekt Razvoj slovenščine v digitalnem okolju sta med leti 2020 in 2023 sofinancirali Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj (Operacija se je izvajala v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020).

Literatura

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., idr. (2020). *Language models are few-shot learners*. arXiv. https://arxiv.org/abs/2005.14165

Bučar, J. (2017). *Automatically sentiment annotated Slovenian news corpus AutoSentiNews 1.0* [Data set]. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1109

Erkan, G. in Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22,* 457–479.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N. in Wang, W. (2022). Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 878–891.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., in Blunsom, P. (2015). *Teaching machines to read and comprehend*. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, 1693-1701.

Knight, K. in Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence, 139(1),* 91–107.

Le, Q. in Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, 1188–1196.

Lebar Bajec, I., Repar, A., Bajec, M., Bajec, Ž. in Rizvič, M. (2022). *NeMo neural machine translation service RSDO-DS4-NMT-API 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1739

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. in Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

Mihalcea, R. in Tarau, P. (2004). TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411.

Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, Ç. in Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 280–290.

Nenkova, A. in Vanderwende, L. (2005). *The impact of frequency on summarization*. Tech. rep., Microsoft Research. https://www.academia.edu/21603307/The_impact_of_frequency_on_summarization

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. in Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21(140),* 1–67.

See, A., Liu, P. J. in Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 1073–1083.

Ulčar, M. in Robnik-Šikonja, M. (2023). Sequence to sequence pretraining for a less-resourced Slovenian language. *Frontiers in Artificial Intelligence, 6*.

Žagar, A., Kavaš, M., Robnik-Šikonja, M., Erjavec, T., Fišer, D., Ljubešić, N., Ferme, M., Borović, M., Boškovič, B., Ojsteršek, M. in Hrovat, G. (2022). *Corpus of academic Slovene KAS 2.0* [Dataset]. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1448

Žagar, A. in Robnik-Šikonja, M. (2021). Unsupervised approach to multilingual user comments summarization. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 89–98.

Zhang, J., Zhao, Y., Saleh, M. in Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, 11328–11339.