Pristopi k denormalizaciji besedil: pregled področja

Melanija Vezočnik in Marko Bajec
Univerza v Ljubljani Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana, Slovenija
melanija.vezocnik@fri.uni-lj.si, marko.bajec@fri.uni-lj.si

Izvleček

Sodobni sistemi za samodejno razpoznavo govora učinkovito pretvorijo govorjeni jezik v pisno obliko, vendar pogosto ustvarijo zgolj surov prepis brez ustrezno oblikovanih števil, datumov in časovnih izrazov, kar zmanjšuje njegovo berljivost in uporabnost. Denormalizacija je postopek, ki odpravlja te pomanjkljivosti tako, da preoblikuje prepis v standardizirano pisno obliko. Članek podaja sistematičen pregled in analizo glavnih pristopov k denormalizaciji, ki jih je mogoče razvrstiti v tri skupine: pristopi, ki temeljijo na pravilih, nevronski pristopi ter hibridni pristopi. Pristopi, ki temeljijo na pravilih, tipično izhajajo iz končnih avtomatov, nevronski pristopi uporabljajo nevronske mreže, hibridni pristopi pa združujejo elemente obeh pristopov. Pristopi, ki temeljijo na pravilih, dosežejo visoko natančnost, a ne upoštevajo konteksta besedila. Nasprotno nevronski pristopi upoštevajo kontekst besedila, vendar pa zahtevajo obsežne količine podatkov za učenje. Hibridni pristopi predstavljajo kompromisno rešitev, ki združuje prednosti obeh pristopov. Delo prispeva k razumevanju izzivov ter izboljšanju učinkovitosti denormalizacijskih sistemov.

Ključne besede: denormalizacija besedila, inverzna normalizacija besedila, pregled področja, samodejna razpoznava govora

Approaches to Inverse Text Normalization: A Review

Abstract

Modern automatic speech recognition systems effectively convert spoken language into written text. However, they often produce only a raw transcript without properly formatted numbers, dates, and time expressions, which can reduce readability and usability. Denormalization is a process that addresses these issues by transforming the transcript into a standardized written form. This article provides a systematic review and analysis of the main approaches to denormalization, which can be classified into three categories: rule-based, neural, and hybrid approaches. Rule-based approaches typically rely on finite-state machines, while neural approaches utilize neural networks. Hybrid approaches combine elements of both approaches. Rule-based approaches achieve high accuracy but tend to overlook the context of the text. In contrast, neural approaches consider the context but require large amounts of training data. Hybrid approaches offer a balanced solution that harnesses the strengths of both approaches. This work contributes to understanding the challenges and improving the efficiency of denormalization systems.

Keywords: automatic speech recognition, inverse text normalization, review, text denormalization

1 UVOD

S hitrim razvojem tehnologije in umetne inteligence so računalniški sistemi vse bolj sposobni obdelovati jezik na načine, ki posnemajo človekovo razumevanje. Sistemi za samodejno razpoznavo govora, ki omogočajo pretvorbo govorjenega jezika v pisno obliko, predstavljajo pomemben del tega napredka [1], [2], [3]. Generiranje naravnega, človeku berljivega besedila, je ključno tako za izboljšanje uporabniške izkušnje kot za zagotavljanje točnih in uporabnih informacij v različnih kontekstih. Med najpogostejšimi aplikacijami so pametni pomočniki, kot sta med drugim Siri in Google Assistant, samodejno generiranje podnapisov in sodobno iskanje informacij.

Čeprav so sodobni sistemi za samodejno razpoznavo govora zelo učinkoviti pri pretvorbi govorjenega jezika v pisno obliko, rezultat pogosto ostaja zgolj surov prepis, v katerem pisna znamenja, med drugim števila, datumi in čas, nimajo ustrezne oblike [4]. Na primer, številski izraz *dva tisoč petindvajset* se ne pretvori v številčno obliko 2025. Takšna odstopanja zmanjšujejo berljivost in uporabnost besedila, še posebej pri samodejnem generiranju podnapisov. Denormalizacija besedila je postopek, ki pretvori prepis v standardizirano pisno obliko, ki med drugim vključuje pravilne zapise števil, datumov, časovnih izrazov in okrajšav [5]. Učinkovita denormalizacija izboljša uporabniško izkušnjo, saj zagotavlja, da je izhodno besedilo usklajeno s pričakovanji končnega uporabnika. Zaradi jezikovnih dvoumnosti, raznolikih kontekstov in specifičnih oblikovnih pravil, vezanih na posamezna področja, denormalizacija še vedno predstavlja zahteven izziv [6].

Nekateri sodobni sistemi, kot je med drugim Whisper [7], omogočajo generiranje prepisov v standardizirani obliki in tako odpravljajo potrebo po ločenem koraku denormalizacije. Vendar pa takšne rešitve niso vedno dostopne ali ustrezno prilagojene specifičnim jezikom in aplikacijam. Poleg tega razvoj integriranih rešitev pogosto zahteva velike količine anotiranih podatkov, ki pri jezikih z omejenimi viri, kot je slovenščina, običajno ni na voljo. Nadalje zvočni signal na postopek denormalizacije ne vpliva, saj ne vsebuje indikacij o formatu zapisa [8]. Zato je denormalizacijo smiselno ločiti od sistema za samodejno razpoznavo govora, saj omogoča večjo prilagodljivost in boljšo uporabnost besedil za nadaljnjo obdelavo, na primer pri integraciji s sistemi, ki zahtevajo strukturirane podatke. Zato so samostojni pristopi k denormalizaciji še vedno široko razširjeni.

Kolikor nam je znano, celovitega preglednega članka, ki bi sistematično obravnaval področje denormalizacije besedil v kontekstu samodejne razpoznave govora, v obstoječi znanstveni literaturi še ni. V tem kontekstu članek podaja sistematičen pregled obstoječih pristopov k denormalizaciji, pri čemer analizira njihove značilnosti, prednosti in omejitve ter ocenjuje njihovo primernost za različne jezikovne in aplikativne kontekste. Na podlagi teh ugotovitev želimo v nadaljevanju razviti učinkovit denormalizator za slovenski jezik, ki je kot jezik z omejenimi viri pogosto spregledan v sodobnih rešitvah za obdelavo naravnega jezika. Takšno orodje lahko bistveno izboljša uporabnost govornih prepisov v različnih aplikacijah, kot so samodejno generiranje podnapisov avdio- in video vsebin, arhiviranje sodnih ali parlamentarnih sej, prepisovanje intervjujev v raziskovalne namene ter glasovno upravljanje v lokaliziranih pametnih asistentih. Nadalje tudi olajša dostop do informacij uporabnikom s posebnimi potrebami.

Struktura preostanka članka je naslednja. Drugi razdelek opisuje metodologijo izbire člankov, tretji obravnava pristope k denormalizaciji besedil in vsebuje primerjalno analizo identificiranih pristopov, četrti razdelek vključuje razpravo, peti pa sklepne ugotovitve.

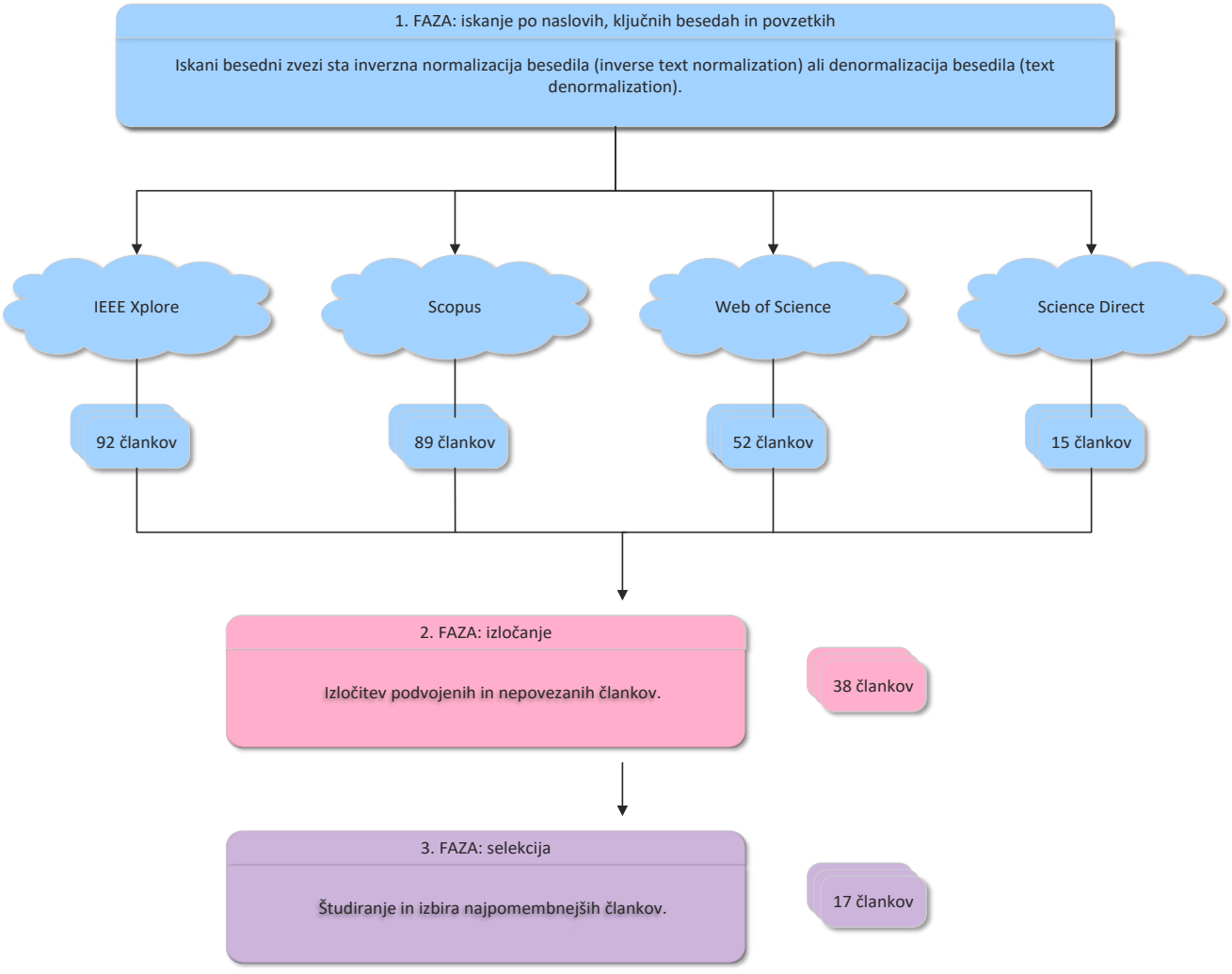
2 METODE

Postopek pregleda literature smo zasnovali sistematično in ga izvedli v treh fazah, kot prikazuje slika 1. Struktura članka dosledno sledi tej metodološki zasnovi. V prvi fazi smo skladno z zastavljenimi raziskovalnimi vprašanji identificirali in analizirali nabor relevantnih člankov na podlagi ključnih besed. V drugi fazi smo izvedli sistematični pregled pristopov k denormalizaciji besedil, kot jih predstavljajo relevantne raziskave v zadnjem desetletju. V tretji fazi smo izvedli poglobljeno primerjalno analizo izbranih rešitev, kjer smo se osredotočili na ključne značilnosti, prednosti in slabosti. Na podlagi ugotovitev, pridobljenih skozi celoten postopek, smo na koncu podali smernice in možnosti za nadaljnji razvoj raziskovalnega področja.

V prvi fazi pregleda področja smo najprej definirali raziskovalna vprašanja, ki so omogočila celovito obravnavo tematike denormalizacije besedil.

- Kakšna je zasnova obstoječih rešitev za denormalizacijo besedil?
- Katere kriterije je mogoče uporabiti za oceno učinkovitosti in natančnosti pristopov za denormalizacijo besedil?
- Na kakšen način denormalizacija besedil vpliva na praktične aplikacije in kakšne so njene širše posledice za področje obdelave naravnega jezika?
- Kakšni so odprti problemi in perspektivne usmeritve v raziskavah, povezanih z denormalizacijo besedil?

Literaturo smo pregledali z iskanjem po digitalnih bibliografskih bazah IEEE Xplore, Scopus, Web of Science in ScienceDirect. Za zajem ustreznega nabora relevantnih člankov smo definirali nabor ključnih besed, ki smo jih uporabili pri iskanju v naslovih,



Slika 1: **Diagram poteka postopka pregleda literature.**

povzetkih in ključnih besedah člankov. Uporabljena iskalna izraza sta bila:

- inverzna normalizacija besedila (inverse text normalization) ter
- denormalizacija besedila (text denormalization).

Ključna izraza smo iskali kot posamezne besedne zveze (in ne kot celotne fraze z navednicami), kar je omogočilo zajem širšega nabora zadetkov. Iskanje je bilo omejeno na naslove, povzetke in ključne besede člankov. Poleg tega smo uporabili filtra za angleški jezik ter časovno obdobje med leti 2015 in 2025. S tem smo zagotovili osredotočenost na aktualne in znanstveno relevantne prispevke.

V drugi fazi smo iz začetnega nabora 248 člankov najprej odstranili duplikate. Nato smo preučili povzetke in naslove preostalih člankov in izločili tiste članke, ki niso neposredno obravnavali problematike denormalizacije besedil ali so bili metodološko nepovezani z obravnavanim področjem. Po tej fazi je ostalo 38 člankov, ki smo jih podrobneje analizirali na osnovi celotnega besedila. Končni nabor je obsegal 17 člankov, ki so izpolnjevali vsebinske in metodološke kriterije ter ustrezali zastavljenim raziskovalnim vprašanjem. Izbrani članki celovito pokrivajo obravnavano temo ter predstavljajo osnovo za nadaljnjo analizo. Na podlagi vsebinske obravnave člankov smo pripravili sistematični pregled pristopov k denormalizaciji besedil, ki je predstavljen v naslednjem razdelku.

3 DENORMALIZACIJA BESEDILA

Denormalizacija besedila je postopek, ki prepis pretvori v standardizirano pisno obliko [5]. Ta običajno vključuje ustrezno zapisane glavne in vrstilne štev-

nike, decimalna števila, datume, časovne izraze, telefonske številke, elektronske naslove, kratice ter okrajšave [4]. Besede, ki sodijo v takšne specifične kategorije, uvrščamo v tako imenovane semiotične razrede.

Ker zvočni signal ne vsebuje informacij o tem, kako naj bodo izrazi iz različnih semiotičnih razredov zapisani v pisni obliki, denormalizacija običajno poteka kot ločen korak po fazi samodejne razpoznave govora [8]. Tako je vhod sistema za denormalizacijo besedila običajno prepis, ki ga je generiral sistem za samodejno razpoznavo govora. Izhod pa je konvencionalna, človeku bolj berljiva pisna različica prepisa. Učinkovitost sistemov za denormalizacijo besedil se običajno ocenjuje z deležem besednih napak, ki meri razlike med izhodnim in pričakovanim besedilom na ravni posameznih besed.

Pristope k denormalizaciji besedil lahko razvrstimo v tri glavne kategorije, in sicer v

- pristope, ki temeljijo na pravilih (razdelek 3.1),
- nevronske (razdelek 3.2) in
- hibridne pristope (razdelek 3.3).

Pristopi, ki temeljijo na pravilih, uporabljajo eksplicitna slovnična in formalna pravila za sistematično pretvorbo besedila v konvencionalno pisno obliko. Nevronski pristopi se učijo kompleksnih jezikovnih vzorcev iz velikih količin anotiranih podatkov z uporabo modelov globokega učenja. Hibridni pristopi združujejo elemente obeh pristopov ter vključujejo slovnična pravila in nevronske modele za obravnavo konteksta in variabilnih jezikovnih struktur.

V nadaljevanju podrobneje predstavimo značilnosti, prednosti in omejitve posameznih kategorij pristopov.

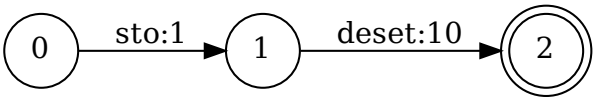
3.1 Pristopi, ki temeljijo na pravilih

Pristopi k denormalizaciji besedil, ki temeljijo na pravilih, tipično vključujejo uporabo končnih avtomatov ali sorodnih formalnih modelov, s katerimi definiramo gramatiko jezika [4]. Učinkovitost teh pristopov je v veliki meri odvisna od natančnosti gramatike, ki definira jezikovno specifična slovnična pravila in strukture. Razvoj tovrstnih sistemov zato zahteva poglobljeno jezikoslovno znanje ter natančno poznavanje strukture in značilnosti obravnavanega jezika.

Postopek denormalizacije besedil v sistemih je običajno dvofazni in sestoji iz klasifikacije in verbalizacije [4], [9], [10]. Prva faza zajema klasifikacijo enot, kot so besede ali fraze, v ustrezne semiotične razrede [4]. Semiotični razredi med drugim zajemajo glavne in vrstilne številnike, decimalna števila, datume, časovne izraze, telefonske številke, elektronske naslove, kratice, okrajšave in splošne besede. Za vsakega izmed identificiranih semiotičnih razredov je potrebno definirati ustrezno gramatiko, ki omogoča njihovo natančno prepoznavanje v fazi klasifikacije. Enote, ki jih ni mogoče uvrstiti v noben semiotični razred, se običajno razvrstijo v splošno kategorijo besed. Druga faza pa zajema verbalizacijo, ki pretvori enote v ustrezno pisno obliko [4].

Slika 2 prikazuje primer končnega avtomata, ki sprejme zaporedje slovenskih številskih besed »sto« in »deset« ter na podlagi prehodov preide med začetnim, vmesnim in končnim stanjem. Ta stanja so po vrsti označena s števili 0, 1, in 2. Prehoda med stanji sta dva. Iz stanja 0 v stanje 1 vodi prehod z oznako sto:1, kar pomeni, da ob vhodnem simbolu sto avtomat doda vrednost 1. Iz stanja 1 v stanje 2 vodi prehod z oznako deset:10, kar pomeni, da ob vhodnem simbolu deset avtomat doda vrednost 10. Skupaj avtomat ob zaporedju simbolov sto in deset doseže končno stanje in s tem sprejme vhod. Skupna številska vrednost prepoznanega zaporedja je 110.

Čeprav pristopi, ki temeljijo na pravilih, še zdaleč niso novi, še vedno ostajajo pomembni tudi v sodobnih sistemih. Večina tradicionalnih sistemov za denormalizacijo besedil uporablja gramatike, definirane z uporabo končnih avtomatov. Pogosto so ti sistemi uporabni tudi pri normalizaciji besedil, ki je obraten postopek od denormalizacije besedil. Primera takih sistemov sta Googlova Kestrel [10] in Sparrowhawk [9], ki predstavlja odprtokodno različico rešitve Kestrel [10]. Rešitev Kestrel [10] podpira angleščino in ruščino, njen povprečni delež besednih napak na naboru podatkov po meri pa je pod 10 %.



Slika 2: **Primer končnega avtomata.**

Zhang idr. [4] so predlagali rešitev NeMo Text Processing [11], ki temelji na ogrodju Sparrowhawk [9] ter podpira normalizacijo in denormalizacijo besedil. Omenjena rešitev uporablja dvofazni postopek, ki v prvi fazi razvrsti besede ali fraze v semiotične razrede, v drugi fazi pa jih pretvori v ustrezno obliko [4]. Zhang idr. [4] so predlagano rešitev za angleški jezik testirali na Googlovi testni množici za normalizacijo [12] in poročali povprečni delež besednih napak v višini 10,1 %. Knjižnica NeMo Text Processing [4] omogoča definiranje gramatik za druge jezike in izvoz definiranih gramatik v format za uporabo v ogrodjih Sparrowhawk [9] ali NVIDIA Riva [13]. Wang in Wang [14] sta prilagodila knjižnico NeMo Text Processing [4] specifični domeni dispečerstva v elektroenergetskem sektorju za kitajščino. Avtorja poročata o 10,5 % povprečnem deležu besednih napak na naboru podatkov po meri.

Kljub visoki učinkovitosti pa takšni pristopi kažejo omejeno prenosljivost med jeziki, saj vsak jezik zahteva definiranje posebej prilagojenih pravil in za slovnične pretvorbe. Poleg tega je razvoj tovrstnih sistemov časovno zahteven, saj vključuje ročno kodiranje kompleksnih jezikovnih pravil in struktur. Kljub navedenim omejitvam pa imajo pristopi, ki temeljijo na pravilih, še vedno pomembno vlogo v okoljih, kjer je ključna visoka stopnja nadzora nad vedenjem sistema, ali kjer primanjkuje podatkov za učenje modelov. V nadaljevanju predstavimo nevronske pristope, katerih razvoj je tesno povezan z napredkom na področju globokega učenja ter z vse večjo dostopnostjo obsežnih jezikovnih korpusov.

3.2 Nevronski pristopi

Poleg pristopov, ki temeljijo na pravilih in zahtevajo ročno definiranje slovničnih pravil, so se v zadnjem desetletju uveljavili nevronski pristopi. V primerjavi s pristopi, ki temeljijo na pravilih, izkazujejo visoko stopnjo prilagodljivosti in zmogljivosti pri denormalizaciji besedil, saj se učijo kompleksnih jezikovnih vzorcev iz velikih količin anotiranih podatkov z uporabo modelov globokega učenja, med katerimi prevladujejo nevronske mreže s povratno zanko, konvolucijske nevronske mreže, nevronske mreže z dolgim kratkoročnim spominom in transformerji [15], [16], [17].

Jedro večine nevronskih pristopov predstavlja t.i. model zaporedje v zaporedje (seq2seq) [15], [18], ki je učinkovit tudi pri denormalizaciji besedila. Rešitve, ki temeljijo na modelu seq2seq, običajno uporabljajo arhitekturno zasnovo kodirnik–dekodirnik, pri čemer kodirnik pretvori vhodno zaporedje v notranjo predstavitev, dekodirnik pa slednjo pretvori v izhodno ciljno zaporedje. Obe komponenti sta tipično realizirani z modeli globokega učenja [15], [18].

Suter in Novak [16] sta zasnovala rešitev na modelu seq2seq z uporabo transformerja. Njihovi modeli za angleščino, ruščino in nemščino so dosegli povprečni delež besednih napak 6,36 %, 4,88 % oziroma 5,23 %. Za angleščino ter ruščino so uporabili Sproutovo množico podatkov, ki je javno dostopna na platformi Kaggle [19], za nemščino pa so uporabili ParaCrawl [20]. Tudi Nayan in Haque [21] sta rešitev zasnovala na modelu seq2seq, pri čemer sta uporabila nevronsko mrežo z dolgim kratkoročnim spominom. Za izboljšanje učenja predlaganega modela za denormalizacijo in normalizacijo sta Nayan in Haque [21] uporabila tehniko CycleGAN, ki omogoča učenje brez popolnoma poravnanih parov podatkov. Za angleščino je predlagana rešitev dosegla povprečni delež besednih napak 33,6 % pri denormalizaciji na Googlovi testni množici [12]. Chen idr. [22] so rešitev prav tako zasnovali z uporabo modela seq2seq in uporabili nevronsko mrežo z dolgim kratkoročnim spominom in transformer. Predlagana rešitev uporablja bogatenje podatkov in nevronsko strojno prevajanje, ki sta zasnovana posebej za jezike z malo viri. Za angleščino je predlagana rešitev dosegla povprečni delež besednih napak 13,8 % na Googlovi testni množici [12] .

Pomembna omejitev modelov seq2seq je njihova dovzetnost za halucinacije, tj. tvorjenje jezikovno pravilnih, a vsebinsko napačnih izhodov. Za odpravo te težave so Antonova idr. [17] predlagali rešitev, ki denormalizacijo obravnava kot nalogo označevanja. Predlagana rešitev vsako besedno enoto bodisi zamenja, kopira brez spremembe ali pa izbriše, kar omogoča večji nadzor nad vedenjem sistema in zmanjšuje tveganje za halucinacije. Predlagana rešitev temelji na enoprehodnem klasifikatorju, ki je enostavnejši kot arhitekturni model seq2seq [17]. Predlagana rešitev je dosegla povprečni delež besednih napak 9,1 % za angleščino in 8,05 % za ruščino na Googlovi testni množici [12].

Ena ključnih prednosti nevronskih pristopov je sposobnost upoštevanja širšega konteksta pri odločanju. To je še posebej pomembno pri dvoumnih zapisih. Na primer, zaporedje *deset trideset* se lahko nanaša na časovni izraz 10.30 ali na glavna številka

10 in 30. Sistemi, ki temeljijo na pravilih, pogosto ne morejo zanesljivo razločiti med takimi primeri, medtem ko se nevronski modeli iz podatkov naučijo razlikovati med možnostmi na podlagi konteksta.

Učenje nevronskih modelov zahteva velike količine anotiranih podatkov, zato glavni izziv pogosto ni samo konstrukcija modela, temveč predvsem pridobivanje in označevanje podatkov. Na primer, Sproutova množica podatkov [19] vsebuje več kot milijardo besednih enot za angleški jezik. Zaradi zahtevnosti ročnega označevanja so bili številni podatkovni korpusi ustvarjeni s pomočjo sistemov, ki temeljijo na pravilih [23]. Kljub temu ostaja velik izziv pridobiti kakovostne učne pare, ki pokrivajo širok spekter semiotičnih razredov, kot so glavni in vrstilni števniki, datumi, časovni izrazi, denar, ulomki, decimalna števila, telefonske številke, mere, elektronski in spletni naslovi ter kratice. To je še posebej problematično za jezike z malo razpoložljivimi viri, kot je slovenščina. Način, kako se nevronski modeli naučijo povezav, ostaja neznan.

Kljub izjemnemu napredku nevronskih modelov pri denormalizaciji besedil se v praksi kaže potreba po večji zanesljivosti, interpretabilnosti in prilagodljivosti sistemov, zlasti v okoljih z omejenimi podatkovnimi viri ali tam, kjer so posledice napak lahko kritične. Nevronski pristopi sicer uspešno zajemajo kontekst in obvladujejo kompleksne jezikovne vzorce, vendar njihova odvisnost od obsežnih anotiranih korpusov, dovzetnost za halucinacije in zahtevnost ponovnega učenja ob spremembah predstavljajo pomembne omejitve. Hibridni pristopi kombinirajo nevronske pristope s pristopi, ki temeljijo na pravilih, in tako omogočijo boljšo obvladljivost sistemskega vedenja, večjo interpretabilnost ter modularno nadgradljivost, kar je še posebej pomembno v okoljih z visokimi zahtevami glede nadzora in zanesljivosti.

3.3 Hibridni pristopi

Hibridni pristopi združujejo nevronske pristope s pristopi, ki temeljijo na pravilih. Pri denormalizaciji besedila so še posebej učinkoviti, saj pravila omogočajo zanesljivo obravnavo sistematičnih vzorcev v besedilu, nevronski pristopi pa omogočajo ustrezno obravnavo kontekstualno pogojenih jezikovnih struktur [23]. Gramatike so tipično bistveno manjše in enostavnejše kot tiste v sistemih za denormalizacijo besedil, ki temeljijo izključno na pravilih. Njihova uporaba zajema klasifikacijo besedila ali korekcijo izhodov nevronskega modela. Modularna zasnova hibridnih sistemov omogoča selektivno posodabljanje posameznih komponent, kar bistveno olajša njihovo prenosljivost med jeziki, domenami in aplikacijskimi scenariji [6].

Pusateri idr. [23] so rešitev zasnovali z uporabo pravil in nevronske mreže z dolgim kratkoročnim spominom. Problem denormalizacije besedila so definirali kot problem označevanja in uporabili s pravili definirane gramatike za pripravo besedila za nevronsko mrežo. Predlagano rešitev so ovrednotili na podlagi pomenske enakovrednosti povedi, ki meri pravilnost celotne povedi. Predlagana rešitev je dosegla približno 99 % natančnost. Podobno so Alphonso idr. [24] predlagali hibridni pristop, ki temelji na nevronskih mrežah z dolgim kratkoročnim spominom, n-gramskem jezikovnem modelu in končnih avtomatih. Potencialne rešitve za pisno obliko so generirali z uporabo pravil, nato pa so jih rangirali z uporabo nevronskih pristopov. Povprečni delež besednih napak je znašal 5,6 %.

Tudi Tan idr. [25] so predlagali rešitev, ki združuje prednosti nevronskega učenja in deterministočnega označevanja z uporabo pravil. Predlagana rešitev generira kontekstualne predstavitve vhodnega besedila z uporabo transformerja, nato pa na podlagi pravil, ki temeljijo na končnih avtomatih, poskrbi za ustrezen format besedila. Podobno so Phan idr. [26] uporabili nevronski model seq2seq za klasifikacijo semiotičnih razredov, nakar so klasificirane segmente pretvorili z uporabo pravil. Glavna prednost njihovega pristopa je v jasni razmejitvi med komponento, ki realizira pravila, in nevronsko komponento, kar omogoča večjo modularnost. Podobnemu principu sledi tudi hibridni pristop, ki so ga predlagali Gaur idr. [8], kjer so uporabili nevronski model transformer za kategorizacijo besednih enot v semiotične razrede, nato pa na označenih segmentih uporabili pravila, specifična za vsak semiotični razred. Ta pretočna in računsko učinkovita arhitektura omogoča visoko natančnost ob znatno manjši kompleksnosti. Vse predlagane rešitve dosegajo primerljive rezultate z drugimi rešitvami na številnih množicah podatkov.

Uspešnost teh sistemov v veliki meri ostaja odvisna od pokritosti jezikovnih pravil, saj se v odsotnosti ustrezne gramatike sistem povrne k nevronskemu modelu. S tem se tveganje za halucinacije ponovno poveča. Nekateri pristopi vključujejo pravila že v fazi predobdelave, s čimer zmanjšajo odvisnost od

nevronskih modelov in izboljšajo robustnost sistema že pred dejanskim procesom denormalizacije. Na primer, Sunkara idr. [6] so uporabili pravila za obdelavo besedila v prvi fazi, nato pa v drugi fazi uporabili na transformerju osnovan model seq2seq. Njihova rešitev omogoča prenosljivost na nove jezike ne da bi potrebovali obsežno jezikovno znanje.

Guo idr. [27] so zasnovali rešitev, ki združuje transformer in pravila, implementirana s končnimi avtomati. Choi idr. [28] pa sistem za denormalizacijo besedil osnovali na arhitekturi, ki je uveljavljena pri strojnem prevajanju, in vključili velik jezikovni model ter transformer. Na ta način so rešili izzive, povezane s pomanjkanjem anotiranih podatkov. Na drugi strani so Wang idr. [14] predlagali hibridni pristop, ki temelji na medjezikovni destilaciji znanja. V sistem so vključili učiteljski model, naučen na jeziku z bogatimi viri, in študentski model, naučen na jeziku z omejenimi viri. Pravila, zasnovana s končnimi avtomati, so vključili kot dodatno obliko nadzora in tako zmanjšali napake in izboljšali kontekstualno interpretacijo.

Hibridni pristopi združujejo robustnost pravil in prilagodljivost nevronskih modelov. Rezultat združevanja različnih pristopov predstavljajo učinkovite rešitve, ki so natančne, razširljive in primerne za produkcijsko rabo. Kljub obetavnim rezultatom pa se ti pristopi razlikujejo glede na obseg in način uporabe pravil, nevronskih modelov, integracije komponent in aplikativnosti. V nadaljevanju podajamo primerjalno analizo obravnavanih pristopov, pri čemer izpostavimo njihove ključne značilnosti, prednosti in slabosti.

3.4 Primerjava pristopov

V nadaljevanju primerjamo tri glavne pristope k denormalizaciji besedil, in sicer pristope, ki temeljijo na pravilih, ter nevronske in hibridni pristope. Za njihovo sistematično primerjavo smo identificirali šest ključnih kriterijev, ki se nanašajo za učinkovitost, uporabnost in vzdrževanje rešitev v različnih jezikovnih in aplikativnih kontekstih. Poleg tega smo upoštevali tudi praktični razvoj, kot sta med drugim modularnost arhitekture in možnost nadgradnje. Izbrani kriteriji omogočajo celovito presojo prednosti in omejitev posameznih pristopov ter lahko predstavljajo osnovo za izbiro najprimernejšega pristopa glede na specifične zahteve aplikacije. Tabela 1 vsebuje primerjavo pristopov.

Tabela 1: **Primerjava pristopov za denormalizacijo besedil.**

kriterij	**pristopi, ki temeljijo na pravilih**	**nevronski pristopi**	**hibridni pristopi**
način delovanja	slovnična pravila, ki tipično temeljijo na končnih avtomatih	nevronski modeli, med drugim nevronske mreže z dolgim kratkoročnim spominom in transformerji, uporabljeni v modelu seq2seq	kombinacija pristopov, ki temeljijo na pravilih, in nevronskih pristopov
potreba po podatkih pri razvoju	ni potrebe po anotiranih podatkih	zahteva po veliki količini anotiranih podatkov	delna odvisnost od anotiranih podatkov
prenosljivost med jeziki	ne	zmerna	zmerna
robustnost	visoka za strukturirane podatke, nizka za odprte kontekste	visoka pri kontekstualni interpretaciji, nizka v primeru halucinacij	visoka, saj združuje prednosti uporabe pravil in upoštevanja konteksta
interpretabilnost	pravila so eksplicitna	delovanje modela težje razložljivo	pravila delno olajšajo razlago rezultatov
razvojna zahtevnost	visoka (zaradi ročnega kodiranja pravil)	srednja do visoka (zaradi zahteve po učenju modelov)	zmerna (zaradi modularne strukture je razvoj nekoliko olajšan)
natančnost	visoka za enostavne primere, nizka za nepodprte primere	visoka za kompleksne kontekste, obstaja možnost pojavljanja halucinacij	visoka zaradi združevanja natančnosti pravil in fleksibilnosti nevronskih modelov
odvisnost od konteksta	ne	da	da

Količina učnih podatkov lahko predstavlja ključen dejavnik pri izbiri pristopa. Pristopi, ki temeljijo na pravilih, so primerni za jezike, kjer ni na voljo večjih količin anotiranih korpusov. Nasprotno nevronski pristopi zahtevajo obsežne količine kakovostno anotiranih podatkov za učenje, kar je lahko omejitveni dejavnik pri manj raziskanih jezikih. Hibridni pristopi omogočajo zmanjšanje potrebe po podatkih, saj kombinacija pravil z učenjem omogoča, da se nevronski model nauči tudi na manjši količini podatkov, pri čemer pravila tipično opredelijo jezikovne strukture v besedilu.

Pristopi, ki temeljijo na pravilih, imajo nizko prenosljivost, saj pri denormalizaciji besedila izhajajo izključno iz definiranih gramatik. Vsak nov jezik zahteva novo definicijo zbirke pravil, kar pomeni visoke stroške razvoja. Nevronski pristopi imajo višjo stopnjo prenosljivosti, še posebej pri uporabi večjezikovnih modelov, ki omogočajo učenje na več jezikih hkrati. Vendar pa zahtevajo velike količine anotiranih podatkov pri učenju, ki pa za jezike z malo viri niso na voljo. Hibridni pristopi lahko z ustrezno zasnovo omogočajo boljšo prenosljivost, vendar še vedno zahtevajo nekaj prilagoditev pri prehodu na nov jezik. Stopnja prenosljivosti hibridnih pristopov je odvisna od razmerja med obsegom na pravilih osnovanih komponent in deležem nevronskih modelov ter načina njihove integracije.

Nevronski pristopi so zmožni upoštevati raznolike in kompleksne jezikovne vzorce, ki se pojavijo v učnih podatkih. Pristopi, ki temeljijo na pravilih, so v primerjavi z nevronskimi pristopi manj robustni, saj kontekstov, ki niso bili eksplicitno definirani s pravili, ne upoštevajo. Hibridni pristopi v tem pogledu ponujajo uravnoteženost, saj s pravili obravnavajo enostavne in ponovljive primere, medtem ko nevronske komponente izboljšajo učinkovitost sistema pri zahtevnejših vhodih.

Interpretacija rezultatov je ključna na primer v pravu ali medicini. Pristopi na osnovi pravil imajo visoko interpretabilnost, saj so pravila eksplicitno definirana in razumljiva. Način, kako se nevronski modeli naučijo povezav, ostaja neznan, zato pogosto ne moremo razložiti, zakaj je model sprejel določeno odločitev. Hibridni pristopi poskušajo ublažiti to pomanjkljivost tako, da za bolj predvidljive jezikovne pojave uporabijo pravila, kar prispeva k večji razložljivosti celotnega sistema.

Razvojna zahtevnost pristopov k denormalizaciji besedil se razlikuje glede na izbrano metodologijo. Pristopi, ki temeljijo na pravilih, zahtevajo ročno definicijo slovničnih pravil, kar zahteva poglobljeno jezikoslovno znanje. Nevronski pristopi zahtevajo obsežno količino anotiranih podatkov in poznavanje področja strojnega učenja. Postopek učenja modelov, optimizacija hiperparametrov in zagotavljanje kakovosti izhodov lahko predstavljajo izziv. Hibridni pristopi omogočajo ponovno uporabo komponent in hitrejšo prilagoditev aplikacijam.

Natančnost označuje stopnjo pravilnosti denormaliziranega besedila. Ena izmed najpogostejših metrik za določanje natančnosti predstavlja povprečni delež besednih napak. Za jezike, kjer so na voljo obsežne zbirke anotiranih podatkov, so nevronski pristopi pogosto najbolj natančni, saj lahko zajamejo kontekst in subtilnosti jezika, ki jih pravila ne upoštevajo. Pristopi, ki temeljijo na pravilih, dosegajo visoko natančnost pri omejenem in dobro definiranem domenskem besedilu, vendar pogosto dosegajo slabše rezultate pri variabilnejšem jeziku. Hibridni pristopi dosegajo primerljive rezultate, saj lahko kombinirajo prednosti obeh pristopov in tako zmanjšajo verjetnost sistematičnih napak.

Pristopi, ki temeljijo na pravilih, izhajajo iz vnaprej določenih pravil, ki niso občutljiva na širši kontekst povedi ali besedila. Zato lahko v primerih, kjer je interpretacija izrazov odvisna od pomena v konkretnem okolju, pride do napačnih pretvorb ali nedoslednosti. Po drugi strani zasnova nevronskih pristopov omogoča upoštevanje širšega konteksta besedila. Vendar pa so nevronski modeli nagnjeni k halucinacijam. Hibridni pristopi združujejo navedena pristopa ter pravila uporabljajo za sistematične in kontekstualno nevtralne primere, medtem ko nevronske komponente omogočajo interpretacijo izrazov v širšem besedilnem kontekstu. Takšna kombinacija omogoča večjo zanesljivost in prilagodljivost.

4 DISKUSIJA

Pregled literature je temeljil na sistematičnem iskanju znanstvenih prispevkov v recenziranih virih. Pri zasnovi sistematičnega pregleda smo se zavestno omejili na vključitev recenziranih znanstvenih virov, indeksiranih v uveljavljenih zbirkah, kot so IEEE Xplore, Science Direct, Scopus in Web of Science. Čeprav Google Scholar in Semantic Scholar predstavljata pomembni zbirki znanstvene literature, ki vključujeta tudi predobjave ter druge nerecenzirane prispevke, smo se odločili, da njune vsebine v ana-

lizo ne vključimo. Ključni razlog za to odločitev je zagotavljanje konsistentne ravni znanstvene verodostojnosti, saj zbirki ne zagotavljata preglednega nadzora nad kakovostjo objavljenih del. Posledično bi vključitev nerecenziranih virov lahko vplivala na zanesljivost ugotovitev ter otežila primerjalno vrednotenje obravnavanih pristopov. Zavedamo se, da so predobjave, zlasti na hitro razvijajočih se področjih, kot je obdelava naravnega jezika, pogosto nosilci najnovejših raziskovalnih trendov.

V pregledu pristopov k denormalizaciji besedil smo ugotovili, da vsak od treh glavnih identificiranih pristopov izkazuje specifične lastnosti, ki vplivajo na samo natančnost delovanja rešitve. Pristopi, ki temeljijo na pravilih, omogočajo visoko stopnjo nadzora in natančnosti nad predvidenim vhodnim besedilom, saj temeljijo na ročno definiranih jezikovnih pravilih in gramatikah. Njihovo glavno pomanjkljivost predstavljajo visok strošek razvoja, slaba prenosljivost med jeziki ter omejena sposobnost obravnave jezikovnih variacij, ki niso bile predhodno definirane s pravili. Nevronski pristopi predstavljajo sodobnejšo alternativo, saj omogočajo učenje kompleksnih jezikovnih vzorcev iz učnih podatkov brez potrebe po ročnem definiranju pravil. Njihova učinkovitost pa je močno pogojena z razpoložljivostjo obsežnih in raznolikih anotiranih besedil, kar predstavlja izziv predvsem pri jezikih z omejenimi viri. Poleg tega ti pristopi niso vedno zanesljivi, saj lahko generirajo halucinacije ali izhode, ki kljub formalni pravilnosti niso ustrezni v danem kontekstu. Hibridni pristopi, ki združujejo prednosti pristopov, ki temeljijo na pravilih, in nevronskih pristopov, so še posebej obetavni za produkcijska okolja, kjer je potrebna visoka natančnost ob ohranjanju fleksibilnosti. Vendar pa ostajajo izzivi pri zagotavljanju optimalne integracije obeh komponent, še posebej pri usklajevanju rezultatov iz obeh virov.

Kritičen dejavnik pri nevronskih in hibridnih pristopih ostaja dostopnost ustreznih podatkovnih množic. Zaradi omejene razpoložljivosti anotiranih besedil za večino jezikov se pogosto uporabljajo metode za bogatenje podatkov, kar lahko vpliva na kakovost in robustnost naučenih modelov. Pomanjkanje standardiziranih podatkovnih množic za različne jezike prav tako otežuje neposredno primerjavo rezultatov med različnimi študijami. V prihodnje bi bilo smiselno raziskati možnosti prenosa znanja iz virov z bogatimi podatki, kot je na primer angleščina, na jezike z manj viri, kot je slovenščina. Prav tako bi bilo koristno razviti bolj transparentne modele, ki bi omogočali boljšo razlago odločitev, s čimer bi povečali zaupanje uporabnikov v delovanje sistemov, še posebej v občutljivih domenah, kot sta medicina ali pravo.

Sklepamo lahko, da optimalna rešitev za denormalizacijo besedil trenutno še ni univerzalna. Učinkovitost posameznega pristopa je v veliki meri odvisna od razpoložljivih virov, zahtevane stopnje natančnosti ter ciljne domene. Nadaljnje raziskave bi morale stremeti k razvoju prilagodljivih, razširljivih in razložljivih sistemov, ki bodo kos kompleksnosti naravnega jezika v različnih kontekstih.

Denormalizacija besedil ostaja odprt raziskovalni izziv, ki zahteva usklajevanje jezikovno-specifičnih potreb, razpoložljivih virov in kontekstualnih zahtev. V prihodnje bo ključno razvijati prilagodljive in robustne rešitve, ki bodo omogočale učinkovito delovanje tudi v pogojih omejenih podatkovnih virov. Posebno pozornost bo treba nameniti tudi večjezičnosti, razložljivosti modelov ter etičnim in praktičnim vidikom uporabe v realnih aplikacijah.

Med jeziki z omejenimi viri izstopa tudi slovenščina, ki zahteva prilagojene pristope zaradi svojih morfosintaktičnih posebnosti in majhnega števila anotiranih podatkovnih množic. Pregledane usmeritve tako odpirajo možnosti za nadaljnje raziskave, usmerjene v razvoj zanesljivih in razložljivih rešitev za denormalizacijo slovenskega jezika, ki bodo lahko učinkovito podprle tako raziskovalne kot tudi praktične aplikacije v govornih in jezikovnih tehnologijah.

5 ZAKLJUČEK

V članku smo predstavili pregled pristopov k denormalizaciji besedil in jih umestili v širši kontekst jezikovnih tehnologij, s poudarkom na jezikih z omejenimi podatkovnimi viri, kot je slovenščina. Bistvene ugotovitve so sledeče. Pristopi, ki temeljijo na pravilih, so zanesljivi, vendar pogosto premalo prilagodljivi za zajem jezikovne raznolikosti. Nevronski pristopi ponujajo večjo fleksibilnost in sposobnost učenja iz podatkov, a zahtevajo obsežne in kakovostno anotirane korpuse, ki jih za slovenščino primanjkuje. Hibridni pristopi se zato kažejo kot najbolj obetavni za praktično rabo, saj združujejo prednosti obeh pristopov.

Zato izbira optimalnega pristopa ni univerzalna, temveč mora upoštevati ciljno aplikacijo, specifičnost jezika in razpoložljivost virov. V kontekstu slovenščine se kot ključni izzivi pojavljajo pomanjkanje javno

dostopnih anotiranih podatkov, odsotnost standardiziranih evalvacijskih okvirov in omejene zmogljivosti obstoječih orodij za razlago odločitev nevronskih modelov.

V prihodnje bo ključno usmeriti raziskave v razvoj jezikovno prilagojenih denormalizatorjev, zasnovanih posebej za značilnosti slovenskega jezika. Čeprav se v širšem kontekstu pogosto predlaga prenos znanja iz resursno bogatejših jezikov prek večjezičnih ali samonadzorovanih pristopov, se takšni prenosi izkažejo za omejeno uporabne pri slovenščini zaradi izrazitih morfoloških, skladenjskih in pravopisnih posebnosti. Zato je še toliko pomembneje razvijati namenske rešitve, ki upoštevajo slovenske jezikovne norme in rabo.

Poleg tega ostaja izziv tudi razložljivost modelov, ki je ključna za sprejemljivost in zanesljivost uporabe v realnih aplikacijah. Nadaljnji razvoj naj se zato osredotoči tudi na transparentnost sistemov in njihovo prilagodljivost specifičnim okoljem – od samodejnega generiranja podnapisov do glasovnega upravljanja ter digitalne dostopnosti za slovensko govoreče uporabnike. Na podlagi teh ugotovitev želimo v nadaljevanju razviti učinkovit denormalizator za slovenski jezik, ki je kot jezik z omejenimi viri pogosto spregledan v sodobnih rešitvah za obdelavo naravnega jezika.

LITERATURA

[1] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, »End-to-End Speech Recognition: A Survey,« *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 325–351, 2024, doi: 10.1109/TASLP.2023.3328283.

[2] D. O'Shaughnessy, »Trends and developments in automatic speech recognition research,« *Computer Speech & Language*, vol. 83, p. 101538, 2024, doi: https://doi.org/10.1016/j.csl.2023.101538.

[3] S. Alharbi *et al.*, »Automatic Speech Recognition: Systematic Literature Review,« *IEEE Access*, vol. 9, pp. 131858–131876, 2021, doi: 10.1109/ACCESS.2021.3112535.

[4] Y. Zhang, E. Bakhturina, and B. Ginsburg, »NeMo (Inverse) Text Normalization: From Development To Production,« in *INTERSPEECH 2021*, in Interspeech. 2021, pp. 4857–4859.

[5] J. Liao, Y. Shi, and Y. Xu, »Automatic Speech Recognition Post-Processing for Readability: Task, Dataset and a Two-Stage Pre-Trained Approach,« *IEEE Access*, vol. 10, pp. 117053–117066, 2022, doi: 10.1109/ACCESS.2022.3219838.

[6] M. Sunkara, C. Shivade, S. Bodapati, and K. Kirchhoff, »Neural inverse text normalization,« in *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2021, pp. 7573–7577.

[7] H. Ahlawat, N. Aggarwal, and D. Gupta, »Automatic Speech Recognition: A survey of deep learning techniques and approaches,« *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 201–237, 2025, doi: https://doi.org/10.1016/j.ijcce.2024.12.007.

[8] Y. Gaur *et al.*, »Streaming, Fast and Accurate on-Device Inverse Text Normalization for Automatic Speech Recognition,« in *2022 IEEE spoken language technology workshop (SLT)*, 2023, pp. 237–244. doi: 10.1109/SLT54892.2023.10022543.

[9] »Sparrowhawk.« Na spletu: https://github.com/google/sparrowhawk, 2015.

[10] P. Ebden and R. Sproat, »The Kestrel TTS text normalization system,« *Natural Language Engineering*, vol. 21, no. 3, pp. 333–353, 2015, doi: 10.1017/S1351324914000175.

[11] »NeMo Text Processing.« Na spletu: https://github.com/NVIDIA/NeMo-text-processing, 2021.

[12] »Google text normalization dataset.« Na spletu: https://www.kaggle.com/datasets/google-nlu/text-normalization, 2016.

[13] »NVIDIA Riva.« Na spletu: https://docs.nvidia.com/deeplearning/riva/user-guide/docs/index.html, 2024.

[14] L. Wang *et al.*, »Zero-Shot Text Normalization via Cross-Lingual Knowledge Distillation,« *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4631–4646, 2024, doi: 10.1109/TASLP.2024.3407509.

[15] A. Vaswani *et al.*, »Attention is all you need,« in *Proceedings of the 31st international conference on neural information processing systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[16] B. Suter and J. Novak, »Neural Text Denormalization for Speech Transcripts,« in *Interspeech 2021*, 2021, pp. 981–985. doi: 10.21437/Interspeech.2021-1814.

[17] A. Antonova, E. Bakhturina, and B. Ginsburg, »Thutmose tagger: Single-pass neural model for inverse text normalization,« in *INTERSPEECH 2022*, in Interspeech. 2022. doi: 10.21437/Interspeech.2022-10864.

[18] C. Peyser, H. Zhang, T. N. Sainath, and Z. Wu, »Improving Performance of End-to-End ASR on Numeric Sequences,« vol. abs/1907.01372, pp. 2185–2189, 2019, doi: 10.21437/Interspeech.2019-1345.

[19] »Text Normalization for English, Russian and Polish.« Na spletu: https://www.kaggle.com/datasets/richardwilliamsproat/text-normalization-for-english-russian-and-polish, 2020.

[20] M. Esplà, M. Forcada, G. Ramírez-Sánchez, and H. Hoang, »ParaCrawl: Web-scale parallel corpora for the languages of the EU,« in *Proceedings of machine translation summit XVII: Translator, project and user tracks*, M. Forcada, A. Way, J. Tinsley, D. Shterionov, C. Rico, and F. Gaspari, Eds., Dublin, Ireland: European Association for Machine Translation, Aug. 2019, pp. 118–119. Available: https://aclanthology.org/W19-6721/

[21] Md. M. R. Nayan and M. A. Haque, »Parallel Training of TN and ITN Models Through CycleGAN for Improved Sequence to Sequence Learning Performance,« in *2022 asia-pacific signal and information processing association annual summit and conference (APSIPA ASC)*, 2022, pp. 1137–1141. doi: 10.23919/APSIPAASC55919.2022.9980278.

[22] S.-J. Chen, D. Paul, Y. Pang, P. Su, and X. Zhang, »Language Agnostic Data-Driven Inverse Text Normalization,« in *INTERSPEECH 2023*, in Interspeech. 2023, pp. 451–455. doi: 10.21437/Interspeech.2023-2066.

[23] E. Pusateri, B. R. Ambati, E. Brooks, O. Platek, D. McAllaster, and V. Nagesha, »A Mostly Data-Driven Approach to Inverse Text Normalization,« in *INTERSPEECH*, Stockholm, 2017, pp. 2784–2788.

[24] I. Alphonso, N. Kibre, and T. Anastasakos, »Ranking Approach to Compact Text Representation for Personal Digital Assi-

stants,« in *2018 IEEE spoken language technology workshop (SLT)*, 2018, pp. 664–669. doi: 10.1109/SLT.2018.8639542.
[25] S. Tan, P. Behre, N. Kibre, I. Alphonso, and S. Chang, »Four--in-One: a joint approach to inverse text normalization, punctuation, capitalization, and disfluency for automatic speech recognition,« in *2022 IEEE spoken language technology workshop (SLT)*, IEEE, 2023, pp. 677–684.
[26] T. A. Phan, N. D. Nguyen, H. L. Thanh, and K.-H. N. Bui, »Neural Inverse Text Normalization with Numerical Recognition for Low Resource Scenarios,« in *Intelligent information and database systems*, N. T. Nguyen, T. K. Tran, U. Tukayev, T.-P. Hong, B. Trawiński, and E. Szczerbicki, Eds., Cham: Springer International Publishing, 2022, pp. 582–594.
[27] K. Guo, S. S. Clarke, and K. Kalyanam, »Inverse text normalization of air traffic control system command center planning telecon transcriptions,« in *AIAA AVIATION FORUM AND ASCEND 2024*, 2024, pp. 4360–4365.
[28] H. Choi *et al.*, »Spoken-to-written text conversion with Large Language Model,« in *Proc. Interspeech 2024*, 2024, pp. 2410–2414.

▪

Melanija Vezočnik je asistentka v Laboratoriju za podatkovne tehnologije na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Njeno trenutno raziskovalno delo je usmerjeno v področje govornih in jezikovnih tehnologij. Leta 2023 je na isti fakulteti doktorirala iz računalništva in informatike. Med doktorskim študijem se je raziskovalno ukvarjala z analizo hoje z inercijskimi senzorji, še posebej z oceno dolžine koraka. Rezultate svojega raziskovalnega dela redno objavlja v znanstvenih revijah, za raziskovalne dosežke med doktorskih študijem pa je prejela priznanje dekanje.

▪

Marko Bajec je redni profesor na Fakulteti za računalništvo in informatiko (Univerza v Ljubljani) in vodja Laboratorija za podatkovne tehnologije. Ukvarja se z načrtovanjem in razvojem podatkovno intenzivnih sistemov. V zadnjih letih se posveča jezikovnim in govornim tehnologijam ter digitalizaciji slovenskega jezika. Svoje rezultate redno objavlja v domačih in tujih revijah ter konferencah. Je prejemnik več nagrad in priznanj za raziskovalno, pedagoško in aplikativno delo.