Uporaba postopkov strojnega učenja pri samodejni slovenski grafemsko-fonemski pretvorbi

Janez Križaj*, Simon Dobrišek*, Aleš Mihelič†, Jerneja Žganec Gros†

*Laboratorij za strojno inteligenco, Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška cesta 25, 1000 Ljubljana, Slovenija
janez.krizaj@fe.uni-lj.si, simon.dobrisek@fe.uni-lj.si
†Alpineon razvoj in raziskave, d. o. o., Ulica Iga Grudna 15, 1000 Ljubljana
Tržaška cesta 25, 1000 Ljubljana, Slovenija
jerneja.gros@alpineon.si, ales.mihelic@alpineon.si

1 Uvod

Grafemsko-fonemska pretvorba se nanaša na pretvarjanje izvirno črkovno zapisanih besed danega jezika v njihove fonemske zapise oziroma predstavitve. Nabor osnovnih grafemskih enot, ki se jih razume kot osnovne enote pisave in se jih upošteva pri črkovnih zapisih besed, navadno določa pravopis danega jezika, in enako velja tudi za slovenski jezik (SAZU, 1990). Osnovnim grafemskim enotam pravimo tudi grafemi, njihovim vidno zaznavnim različnim pisnim simbolnim predstavitvam, kot so velike in male črke, pa pravimo alografi. Nabor fonemov je na drugi strani določen predvsem na osnovi glasoslovnega pomensko razločevalnega slušnega kriterija. Grafemi in fonemi so kot osnovne enote do določene mere sicer povezani, a se pri pretvorbi grafemov v foneme lahko tudi več zaporednih črk v zapisani besedi preslika v posamezne foneme. Pretvarjanje grafemskih zapisov besed v njihove fonemske zapise tudi ne temelji samo na nekem manjšem številu osnovnih pravil in pri slovenskem govorjenem jeziku obstaja veliko izjem, ki se ne podrejajo osnovnim pravilom (Toporišič, 2000).

Pri razvoju jezikovnih tehnologij se postopki samodejnega računalniškega pretvarjanja grafemskih zapisov besed v njihove fonemske zapise uporabljajo tako pri izgradnji samodejnih razpoznavalnikov govora kot tudi pri sistemih za tvorjenje umetnega govora (Žganec Gros et al., 2016). V okviru razvojnega in raziskovalnega projekta Razvoj slovenščine v digitalnem okolju (RSDO, 2020) smo izvedli in ovrednotili več različnih uveljavljenih postopkov samodejne grafemsko-fonemske pretvorbe, ki so bili uporabljeni za tovrstno pretvarjanje zapisov slovenskih besed. Preizkusili in ovrednotili smo tri izbrane postopke samodejne grafemsko-fonemske pretvorbe, ki so se uveljavili v zadnjih nekaj letih in so na kratko opisani v nadaljevanju. Za preizkus in ovrednotenje izbranih postopkov smo uporabili množico besed iz slovenskega leksikona Sloleks 2.0 (Dobrovoljc et al., 2019). Množico besed smo na različne načine razdelili na učno in testno množico, ki smo ju nato uporabili za strojno učenje in preizkus izbranih samodejnih grafemsko-fonemskih pretvornikov.

2 Obravnavani postopki

V literaturi je predstavljenih mnogo različnih postopkov za samodejno grafemsko-fonemsko pretvorbo zapisov besed. Starejši postopki praviloma izvajajo pretvorbo na podlagi predhodno definiranih slovničnih pravil (Black et al., 1998). Pomanjkljivost teh postopkov je predvsem v dolgotrajnem ročnem oblikovanju pravil, ki zahtevajo znanje s področja jezikoslovja in glasoslovja in morajo vključevati tudi seznam izjem z različnimi posebnostmi pri izgovorjavah besed. Pri kasneje predlaganih postopkih se je uveljavila pretvorba z modeli skupnih zaporedij (Bisani in Ney, 2008), ki s poravnavo grafemskega zaporedja s fonemskim zaporedjem tvorijo posebne skupne enote, imenovane grafoni. Za modeliranje grafonskih zaporedij nato uporabljajo jezikovne modele n-gramov, udejanjene v obliki uteženega končnega pretvornika (angl. weighted final state transducer), ki omogočajo predvidevanja grafemsko-fonemske pretvorbe za besede, ki niso bile del učne množice.

Avtorji Novak et al. (2015) so razvoj grafemsko-fonemskega pretvornika osnovali na modelih uteženih končnih pretvornikov in predlagali postopek grafemsko-fonemske pretvorbe, ki temelji na prilagojeni metodi maksimizacije upanja za poravnavo niza grafemov z nizom fonemov in več dekodirnih postopkov, med njimi tudi jezikovni model, ki temelji na modelih rekurenčnih nevronskih omrežij (angl. recurrent neural networks).

Yolchuyeva et al. (2019) so dosegli visoko uspešnost grafemsko-fonemske pretvorbe z uporabo globokega modela, ki je poznan pod imenom *transformer*. Ti modeli imajo zgradbo vrste kodirnik-dekodirnik z dodanim mehanizmom pozornosti, ki pomaga pri strojnem učenju soodvisnosti med učnimi pari nizov grafemov in fonemov, kar se odraža tako v hitrejšem strojnem učenju kot tudi pri bolj zanesljivi pretvorbi preizkusnih nizov grafemov v ustrezne nize fonemov.

3 Kvantitativno ovrednotenje

Pri kvantitativnem ovrednotenju obravnavanih postopkov grafemsko-fonemskih pretvorb smo uporabili njihove izvedbe v prosto dostopnih računalniških programskih knjižnicah. Postopek, predlagan v (Bisani in Hermann, 2008), smo udejanjili s programskim orodjem Sequitur[1], postopek avtorjev Novak et al. (2015) je implementiran z orodjem Phonetisaurus[2], za evalvacijo metode avtorjev Yolchuyeva et al. (2019) pa smo uporabili programsko orodje Deep Phonemizer[3].

Pri tvorjenju in preizkušanju vseh obravnavanih modelov in izvajanje postopkov njihovega strojnega učenja smo uporabili ročno validirani del slovenskega leksikona Sloleks 2.0 (Dobrovoljc et al., 2019), ki poleg posameznih besed vsebuje tudi informacijo o njihovih osnovnih besednih oblikah oziroma lemah ter tudi njihove fonemske oziroma fonetične prepise. Validirani del leksikona Sloleks 2.0, ki smo ga uporabili za naše eksperimente, tako vsebuje 646.994 posameznih besed oziroma 62.729 besednih lem. Pri preizkušanju smo opazili, da so rezultati precej odvisni od tega, kako se množico razpoložljivih grafemsko-fonemsko pretvorjenih besed razdeli na učni in testni del. Pri preizkusih smo zato izvedli dve različni razdelitvi množice vseh besed v učno množico, ki je vsebovala 90 % besed iz slovarja, in testno množico, ki je vsebovala preostalih 10 % besed. Pri naključni razdelitvi, v nadaljevanju označeni z oznako "RandomSplit", smo razdelitev izvedli povsem naključno z uporabo sistemskega naključnega generatorja. Pri razdelitvi, ki je temeljila na razvrščanju besed v učno oziroma testno množico glede na njihove leme, pa smo poskrbeli, da se v testni množici ne pojavljajo besede, ki se od besed v učni množici razlikujejo le po končnicah. To namreč pogosto velja za besede z istimi lemami. Polega tega smo poskrbeli, da se leme besed v testni množici razlikujejo za vsaj tri črke glede na njim najbolj podobne leme v učni množici besed. Ta razdelitev je v nadaljevanju označena z oznako "LemmaSplit".

Pri izvajanju poskusov smo ugotovili, da je rezultat po pričakovanjih tudi precej odvisen od upoštevanega nabora fonemskih enot pri grafemsko-fonemskih pretvorbah. Pri gradnji samodejnih razpoznavalnikov govora se tako navadno ne ločuje med dolgimi in kratkimi samoglasniki oziroma med naglašenimi in nenaglašenimi samoglasniki. To ločevanje pri razpoznavalnikih govora namreč ni pomembno po pomensko razločevalnem kriteriju določanja fonemskih enot. To ločevanje pa je pomembno pri gradnji sistemov za tvorjenje umetnega govora, kjer so prozodične značilnosti umetnega govora odvisne od informacije o naglašenih in nenaglašenih samoglasnikih v besedah. V skladu s temi predpostavkami smo učno in testno množico dodatno razdelili na različna načina, glede na to, katere osnovne fonemske enote se je upoštevalo. V nadaljevanju tako oznaka ASR označuje razdelitev, ki je bila primerna za samodejne razpoznavalnike govora in temelji na upoštevanju samo 34 osnovnih fonemskih enot oziroma fonemskih različic. Oznaka TTS pa označuje razdelitev, ki je primerna za sisteme za samodejno tvorjenje umetnega govora in temelji na upoštevanju 39 osnovnih fonemskih enot. Povečanje števila fonemskih enot je posledica upoštevanja ločevanja med dolgimi in kratkimi oziroma naglašenimi in nenaglašenimi samoglasniki. V nadaljevanju predstavljeni rezultati so potrdili predvidevanja, da je pri slovenskem jeziku najteže samodejno napovedovati naglasno mesto v besedah oziroma naglašene samoglasnike. Pri naglaševanju slovenskih besed je namreč zelo veliko izjem, ki se ne podrejajo nekemu bolj splošnemu manjšemu naboru osnovnih pravil naglaševanja besed.

Rezultati uspešnosti samodejnih grafemsko-fonemskih pretvorb so v nadaljevanju podani v obliki odstotnega deleža napačno pretvorjenih besed (angl. word error rate, WER) in deleža napačno pretvorjenih fonemskih enot (angl. phoneme error rate, PER). Kot je razvidno iz tabele so se glede na različne delitve množice besed in upoštevanja ločevanja med naglašenimi in nenaglašenimi samoglasniki pri rezultatih dejansko potrdila predvidevanja. Pri naključni razdelitvi so tako rezultati bistveno boljši kot pri razdelitvi po lemah, saj se pri naključni razdelitvi v testni množici lahko pojavljajo besede, ki se od najbolj podobnih besed v učni množici

[1] https://github.com/sequitur-g2p/sequitur-g2p
[2] https://github.com/AdolfVonKleist/Phonetisaurus
[3] https://github.com/as-ideas/DeepPhonemizer

razlikujejo samo po končnici ali predponi. Rezultati pri večjem naboru osnovnih fonemskih enot, ki vključuje ločevanje med dolgimi in kratkimi samoglasniki (oznaka TTS), pa so prav tako po pričakovanjih precej slabši, kot pri manjšem naboru, ki tega ločevanja ne upošteva (oznaka ASR). To potrjuje druge že obstoječe ugotovitve, da je pri slovenskem jeziku dejansko težko samodejno napovedovati naglasno mesto v besedah (Žganec Gros et al., 2016).

Orodje	Slovar	WER [%]	PER [%]
Sequitur (Bisani in Hermann, 2008)	ASR_RandomSplit	16,5	1,9
	ASR_LemmaSplit	25,4	2,9
	TTS_RandomSplit	17,3	2,2
	TTS_LemmaSplit	50,2	7,4
Phonetisaurus (Novak et al., 2015)	ASR_RandomSplit	1,0	0,1
	ASR_LemmaSplit	14,1	1,6
	TTS_RandomSplit	2,0	0,3
	TTS_LemmaSplit	29,1	4,1
Deep Phonemizer (Yolchuyeva et al., 2019)	ASR_RandomSplit	1,1	0,1
	ASR_LemmaSplit	8,6	0,9
	TTS_RandomSplit	1,7	0,3
	TTS_LemmaSplit	16,1	2,6

Tabela 1: Uspešnost grafemsko-fonemske pretvorbe obravnavanih postopkov.

4 Zaključek

V prispevku so predstavljeni rezultati izvedb in preizkusov različnih samodejnih grafemsko-fonemskih pretvornikov za slovenski jezik. Glede na ugotovitve lahko uporabniki tovrstnih pretvornikov za izgradnjo samodejnih razpoznavalnikov govora pričakujejo približno 91% pravilno pretvorbo besed, ki niso vključene v obstoječe slovenske leksikone. Pri izgradnji sistemov za tvorjenje umetnega govora, pri katerih je pomembno pravilno določanje naglasnega mesta, pa lahko pričakujejo samo približno 84% pravilno pretvorbo.

Zahvala

Predstavljeno delo je bilo delno financirano s strani Ministrstva za kulturo in Evropskega sklada za regionalni razvoj v okviru projekta RSDO (Razvoj slovenščine v digitalnem okolju), s strani Javne agencije za raziskovalno dejavnost Republike Slovenije v okviru aplikativnega raziskovalnega projekta L7-9406 OptiLEX in s strani ARRS v okviru raziskovalnega programa Metrologija in biometrični sistemi (P2-0250).

Literatura

Maximilian Bisani in Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication,* 50(5):434–451.

Alan W. Black, Kevin Lenzo in Vincent Pagel. 1998. Issues in Building General Letter to Sound Rules. V: *Zbornik 3rd ESCA Workshop on Speech Synthesis*, str. 77–80.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Jaka Čibej, Luka Krsnik in Marko Robnik-Šikonja. 2019. Morphological lexicon Sloleks 2.0. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1230.

Josef R. Novak, Nobuaki Minematsu in Keikichi Hirose. 2015. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, 22(6):907–938.

RSDO - Razvoj slovenščine v digitalnem okolju. 2020. https://www.slovenscina.eu/.

SAZU - Slovenska akademija znanosti in umetnosti. 1990. Slovenski pravopis 1: Pravila. Državna založba Slovenije, Ljubljana.

Jože Toporišič. 2000. *Slovenska slovnica*. Založba Obzorja, Maribor.

Sevinj Yolchuyeva, Géza Németh in Bálint Gyires-Tóth. 2019. Transformer Based Grapheme-to-Phoneme Conversion. V: *Zbornik konf. Interspeech 2019*, str. 2095–2099, Gradec, Avstrija.

Jerneja Žganec Gros, Boštjan Vesnicer, Simon Rozman, Peter Holozanin Tomaž Šef. 2016. Sintetizator govora za slovenščino eBralec. V: *Zbornik konf. Jezikovne tehnologije in digitalna humanistika*, str. 180–185, Ljubljana, Slovenija.