Strojno prepoznavanje idiomov z globokimi nevronskimi mrežami

Tadej ŠKVORC
Fakulteta za računalništvo in informatiko Univerze v Ljubljani,
Institut »Jožef Stefan«, tadej.skvorc@fri.uni-lj.si

Polona GANTAR
Filozofska fakulteta Univerze v Ljubljani, apolonija.gantar@ff.uni-lj.si

Marko ROBNIK-ŠIKONJA
Fakulteta za računalništvo in informatiko Univerze v Ljubljani,
marko.robnik@fri.uni-lj.si

Abstract

Idiomatic expressions are difficult to detect with machine learning approaches due to a lack of sufficiently large datasets and because their meaning cannot be inferred from their constituting words. We present a novel approach, called MICE, that uses contextual embeddings for that purpose. Our neural approach is trained on a new dataset of multi-word expressions with literal and idiomatic meanings. We test two recent contextual word embeddings: ELMo and BERT. We show that deep neural networks using contextual embeddings perform much better than existing approaches, and are capable of detecting idiomatic word use for expressions present and absent from the training set. We observe that the recognition rate differs significantly between different idioms.

Ključne besede: strojno prepoznavanje idiomov, nevronske mreže, kontekstne vložitve

Keywords: automatic detection of idioms, neural networks, contextual embeddings

1 Uvod

Idiomi so sestavljeni iz skupine besed z določenim pomenom, ki ga ni mogoče razbrati iz dobesednega pomena posameznih besed, ki jih sestavljajo (npr. *dobiti zajeten kos pogače* ali *zakopati bojno sekiro*). Pravilno prepoznavanje in razumevanje idiomov je ključno za pravilno delovanje metod za obdelavo naravnega jezika, kot so strojno prevajanje, povzemanje in odgovarjanje na vprašanja. V tem prispevku predstavljamo strojno prepoznavanju idiomov v slovenščini.

Problem trenutnih pristopov avtomatskega prepoznavanja idiomov je neuporaba kontekstnih nevronskih pristopov in pomanjkanje dovolj velikih učnih množic z označenimi idiomi, kar velja za vse jezike, ne le za slovenščino. Zaradi velikega števila različnih idiomov trenutno ni korpusa, ki bi vseboval zadovoljivo število primerov za vse idiome, kar bi omogočilo pristopom strojnega učenja, da se naučijo njihove specifične rabe. Večina obstoječih učnih množic je v angleškem jeziku, kar otežuje razvoj pristopov za druge jezike. Obstoječi pristopi uporabljajo razmeroma majhne korpuse, kot so na primer podatki iz izzivov SemEval 2013, naloge 5B (Korkontzelos et al. 2013) in PARSEME (Savary et al. 2017) ali iz množice VNC Tokens (Cook et al. 2008). Naštete učne množice zajemajo le majhno število idiomov in za vsak vsebovan idiom le majhno število učnih primerov, zaradi česar je strojno učenje manj uspešno.

Zaradi pomanjkanja zadovoljivih učnih korpusov in orodij si uporabniki pogosto pomagajo z leksikoni idiomov. Ti so izdelani ročno ali z uporabo preprostih računalniških orodij, ki upoštevajo samo jezikovno dokaj neodvisne značilnosti sočasnega pojavljanja. Uporaba leksikonov idiomov je dokaj težavna. Veliki ročno ustvarjeni leksikoni idiomov so redki zaradi zamudnega ročnega dela, ki je potrebno za njihovo sestavo. Seznami idiomov, ki so bili ustvarjeni s preprostejšimi strojnimi pristopi, so nezanesljivi, saj ne upoštevajo možnih diskontinuitet z vrivanjem elementov sobesedila (*šlo mi je na živce*) in skladenjske spremenljivosti idiomov (*začarani krog – krog je začaran*). Prepoznavanje idiomov in odkrivanje novih tako večinoma

temelji na leksikografskem delu in slovarskih podatkih, ki navadno niso na voljo v obliki, ki bi bila primerna za strojno obdelavo.

Globoke nevronske mreže so trenutno najuspešnejši pristop strojnega učenja na besedilnih podatkih in presegajo vse druge pristope v praktično vseh nalogah obdelave in razumevanja naravnega jezika (LeCun et al. 2015, Zhang et al. 2015, Kim et al. 2016, Peters et al. 2018, Devlin et al. 2019). Nevronske mreže na vhodu pričakujejo številske podatke. Za njihovo rabo besedilo pretvorimo v številske vektorje s postopkom, imenovanim vektorska vložitev besedila. Postopek mora zagotoviti, da se semantični odnosi med besedami odražajo v razdaljah in smereh vektorjev v številčnem prostoru, ki ima običajno nekaj sto dimenzij. Sodobne vektorske vložitve pridobimo z nevronskimi mrežami, ki jih učimo posebnih učnih nalog, tipično napovedovanja besede na podlagi njenega konteksta (okolice), kar imenujemo učenje jezikovnega modela. Primeri znanih metod vektorskih vložitev besed so word2vec (Mikolov in Sutskever 2013), GloVe (Pennington et al. 2014) in fastText (Bojanowski et al. 2017). Za dobro delovanje algoritmi za sestavo vektorskih vložitev uporabljajo obsežne enojezične besedilne korpuse.

Težava prve generacije nevronskih vektorskih vložitev, kot je word2vec, je njihov neuspeh pri izražanju večpomenskih besed. Med učenjem vektorskih vložitev vsi pomeni določene besede (npr. *list* kot list papirja ali list drevesa) prispevajo informacije o svojem kontekstu v sorazmerju s pogostostjo nekega pomena v učnem korpusu. Zaradi tega se končni naučeni vektor postavi v uteženo sredino vseh pomenov besede. Redki pomeni besed (ki so mnogokrat tudi del idioma) so zaradi tega s temi vektorskimi vložitvami slabše izraženi. Na primer, v angleščini noben od 50 najbližjih vektorjev besede *paper* (dobesedno ‘papir’, v enem od pomenov tudi ‘prispevek’ ali ‘znanstveni članek’) ni povezan z znanostjo.

Novejše kontekstne vektorske vložitve za vsak kontekst besede sestavijo drugačen vektor in lahko tako bolje predstavijo tudi večpomenske in redke besede. Te vložitve izboljšajo uspešnost strojnega učenja pri številnih nalogah obdelave naravnega jezika (Devlin et al. 2019). Obstoječi pristopi prepoznavanja idiomov ne uporabljajo

kontekstnih vložitev za razlikovanje med idiomatično in dobesedno rabo besed.

V prispevku predstavimo pristope za strojno prepoznavanje idiomov na podlagi globokih nevronskih mrež, ki uporabljajo vektorske vložitve. Najprej opišemo pristopa za izgradnjo kontekstnih vložitev, ki temeljita na globokih nevronskih mrežah ELMo in BERT. Nato predlagamo pristop rudarjenja idiomov s kontekstnimi vložitvami, imenovan MICE (*Mining Idioms with Contextual Embeddings*), pri katerem uporabljamo vektorske vložitve tipa ELMo in BERT na vhodu v nevronske mreže. Naš pristop učimo na za ta namen izdelani učni množici ročno označenih stavkov, ki vključujejo idiome v idiomatičnem in dobesednem pomenu. To je prvi tak pristop, ki je bil naučen in evalviran na večji množici slovenskih besedil. V nadaljevanju analiziramo rezultate samodejnega zaznavanja idiomov z različnih vidikov. Ovrednotimo nevronsko metodo za prepoznavanje idiomov MUMULS, ki uporablja strojno učenje z nevronsko mrežo, vendar brez predhodno naučenih kontekstnih vektorskih vložitev, in predlagani pristop MICE, ki poleg nevronske mreže uporablja še kontekstne vektorske vložitve besed. Metodi ovrednotimo z vidika klasifikacije idiomov, ki so prisotni v učni množici, in idiomov, ki jih v učni množici ni. Na podlagi analize rezultatov sklenemo, da se med obravnavanimi modeli v obeh nalogah najbolje obnese pristop MICE, ki za prepoznavo idiomov uporablja kontekstne vložitve, zasnovane za obravnavo večpomenskih besed. Prispevek zaključimo z izhodišči za nadaljnje analize ter izpostavimo možnosti za izboljšavo učne množice in delovanja predlaganega modela.

2 Obstoječi pristopi

Pristope za prepoznavanje idiomov v besedilu lahko v splošnem razdelimo na take, ki uporabljajo nadzorovane in nenadzorovane metode. V nadzorovanih pristopih prepoznavanje idiomov predstavimo kot problem binarne klasifikacije, kjer za vsak idiom naučimo ločen klasifikator (Liu in Hwa 2017). Pomanjkljivost tega pristopa je, da

ni primeren za veliko število idiomov, saj zahteva učenje ločenega modela za vsak idiom.

Več avtorjev je predlagalo pristope z nevronskimi mrežami. Pristop MUMULS (Klyueva et al. 2017) uporablja dvosmerno nevronsko mrežo tipa GRU (Cho et al. 2014) v kombinaciji z vektorskimi vložitvami. Pristop je poleg idiomov sposoben zaznati različne vrste večbesednih izrazov, označenih v izzivu PARSEME za identifikacijo glagolskih večbesednih enot (Savary et al. 2017). MUMULS je na izzivu dosegel najboljše rezultate pri več jezikih, vendar so avtorji poročali o slabi točnosti klasifikacije pri jezikih z manj učnimi podatki. Poleg tega niso uspeli zaznati izrazov, ki se niso pojavili v učnem korpusu. V izzivu PARSEME za leto 2018 (Ramisch et al. 2018) je bilo predstavljenih še več sistemov, ki temeljijo na nevronskih omrežjih (Berk et al. 2018, Ehren et al. 2018, Boroş in Burtica 2018). Sistemi so dosegli podobne rezultate kot MUMULS in so dobro delovali na več jezikih, vendar so dosegli nizko klasifikacijsko točnost pri jezikih z majhnimi učnimi množicami ter niso zaznali izrazov, ki niso bili prisotni v učnem korpusu. Primer takšnega pristopa sta predstavila Boros in Burtica (2018), ki uporabljata dvosmerno rekurenčno nevronsko mrežo s kratkim dolgoročnim spominom (angl. *bidirectional long short-term memory network*; biLSTM) v kombinaciji s podatki na podlagi grafov. V nasprotju z našim pristopom MICE, našteti pristopi ne uporabljajo kontekstnih vektorskih vložitev in ne izkoristijo kontekstnih informacij, ki jih te vsebujejo.

Druga skupina metod za zaznavanje idiomatične rabe besed so nenadzorovani pristopi. Njihova prednost je, da ne potrebujejo ročno označenih učnih množic z lokacijami idiomov, vendar pa na splošno dosegajo slabše rezultate. Primer takšne rešitve (Sporleder in Li 2009) uporablja le leksikalno kohezijo brez označenih korpusov ali drugih jezikovnih virov, kot so slovarji ali leksikoni. Podoben pristop (Liu in Hwa 2018) primerja kontekst pojava neke besede z vnaprej določenim »dobesednim kontekstom uporabe« (tj. zbirko besed, ki se pogosto pojavljajo v bližini dobesedne uporabe besede). S tem dobimo hevristično mero, ki kaže, ali se beseda uporablja dobesedno ali idiomatično. Dobljene ocene avtorji uporabijo

v verjetnostnem modelu, ki napove, ali ima beseda dobesedni ali preneseni pomen. Pristop doseže povprečno oceno F1 med 0,72 do 0,75 na nalogi SemEval 2013 5B (Korkontzelos et al. 2013) in na naboru podatkov VNC Tokens (Cook et al. 2008).

Težava obstoječih pristopov je pomanjkanje dovolj velikih učnih množic z označenimi idiomi, ki bi jih lahko uporabili za učenje klasifikacijskih modelov. Liu et al. (2017) uporabljajo podatke iz SemEval 2013, naloga 5B (Korkontzelos et al. 2013), ki vsebuje le 10 različnih idiomov s 2371 primeri. Klyueva et al. (2017) klasifikacijski model naučijo na izzivu PARSEME (Savary et al. 2017), ki vsebuje le majhno število idiomov v 20 jezikih. Obstajajo sicer večje množice, kot sta podatkovna množica VNC Tokens (Cook et al. 2008), ki vsebuje 2984 primerov in 53 različnih idiomov, in korpus, ki so ga predstavili Fadaee et al. (2018) in vsebuje 6.846 stavkov z 235 različnimi idiomi v angleščini in nemščini. Takšnih korpusov je malo in večinoma obstajajo le za angleščino. Uspešni pristopi, ki se omejijo izključno na slovenščino, trenutno ne obstajajo. Nekateri večjezični pristopi so bili evalvirani tudi na slovenskih besedilih, vendar le na majhnem številu idiomov (npr. izziv PARSEME 1.1 vsebuje 727 povedi z idiomi), zaradi česar je težko oceniti njihovo točnost na slovenskih besedilih.

Eden izmed ključnih problemov trenutnih pristopov je, da za delovanje potrebujejo seznam idiomov in učno množico zanje, da lahko naučijo klasifikacijski model. Ti pristopi namenjajo le malo pozornosti odkrivanju idiomov, ki se ne pojavijo v učnem korpusu, kar je težji problem. Zaradi velikega števila idiomov bi bil sistem, ki bi omogočal takšno uporabo, zelo koristen in bi bolje služil pri dejanski rabi. Tudi trenutni nenadzorovani pristopi, npr. (Liu in Hwa 2018), najprej ročno oblikujejo različne uporabe za vsak idiom in zato niso primerni za odkrivanje idiomov, ki niso vnaprej znani. Možna rešitev tega problema, ki jo predlagamo v prispevku, so kontekstne vektorske vložitve, za katere pri konstrukciji zajemamo semantične informacije iz besedila, ne da bi za učenje potrebovali označene podatke. To načeloma omogoča kasnejše prepoznavanje idiomov, tudi če niso vnaprej definirani.

Za predstavitev prepoznavanja idiomov s kontekstnimi vektorskimi vložitvami najprej opišemo dva sodobna pristopa za izgradnjo kontekstnih vložitev, ki temeljita na globokih nevronskih mrežah: ELMo (Peters et al. 2018) in BERT (Devlin et al. 2019).

2.1 ELMo

Pristop ELMo (*Embeddings from Language Models*; vložitve na podlagi jezikovnih modelov) zgradi velik nevronski jezikovni model, ki ustvari kontekstne vektorske vložitve, s katerimi lahko izboljšamo delovanje številnih sistemov strojnega učenja za obdelavo naravnega jezika. Arhitektura modela ELMo je sestavljena iz treh plasti nevronov, izhod po vsaki plasti daje en vektor vložitev. Skupaj dobimo torej tri različne vložitve, ki jih združimo v končno vložitev. Ker ELMo uporablja vhod v obliki znakov, je še posebej primeren za morfološko bogate jezike, kot je slovenščina, saj je zmožen obravnavati tudi besede izven slovarja.

V prispevku uporabljamo model ELMo, ki je bil predhodno naučen na veliki zbirki slovenskih besedil (Ulčar in Robnik-Šikonja 2020b). Kot vhod v naše modele uporabljamo povprečje treh slojev ELMo brez posebnega prilagajanja učni nalogi. Kot kažejo naši rezultati, tudi brez posebnega prilagajanja kontekstnih vložitev te izboljšajo uspešnost prepoznavanja idiomov v primerjavi s podobnimi pristopi, ki ne uporabljajo kontekstnih vložitev.

2.2 BERT

BERT (*Bidirectional Encoder Representations from Transformers*; predstavitev iz dvosmernih transformer kodirnikov) posploši idejo jezikovnih modelov na maskirne jezikovne modele, ki napovedujejo skrito besedo kjerkoli v besedilu. Maskirni jezikovni model naključno maskira nekaj delov vhodnega besedila in jih poskuša napovedati na podlagi njihove okolice. Model BERT uporablja nevronsko arhitekturo transformer (Vaswani et al. 2017), uporablja tako levi kot desni kontekst pri napovedovanju maskirane besede, poleg tega pa se uči, ali sta dva vhodna stavka zaporedna ali ne. S temi nalogami lahko iz

obsežnih jezikovnih korpusov izlušči veliko količino jezikovnih podatkov. Vhod v model BERT so zaporedja jezikovnih delčkov – žetonov (angl. *tokens*), ki sestavljajo besede. Vnaprejšnje razbitje besedila na žetone nekatere pogoste besede ohrani v celoti, druge pa razdeli na dele (npr. korene, predpone in pripone – če je potrebno vse do posameznih črk). Originalno je bil BERT naučen v treh oblikah: za angleščino, kitajščino in večjezični model. Slednji, imenovan večjezični BERT (mBERT), so učili hkrati na besedilih v 104 jezikih, tudi slovenščini. BERT se je odlično izkazal na številnih nalogah obdelave naravnega jezika (Wang et al. 2018), npr. ugotavljanju jezikovne sprejemljivost besedil, klasifikaciji sentimenta filmskih recenzij, parafraziranju, določanju podobnosti besedil, odgovarjanju na več vrst vprašanj, prepoznavanju imenskih entitet in zdravorazumskem sklepanju.

Pri nalogah obdelave naravnega jezika razvijalci večinoma uporabljajo vnaprej naučene modele BERT, ki jih prilagodijo posameznim nalogam. Ta pristop izkorišča zmožnost velikih vnaprej naučenih jezikovnih modelov, da izluščijo številne jezikovne informacije brez izgradnje posebnih učnih množic. Pri naši uporabi modelov BERT ne prilagodimo vseh uteži nevronske mreže, ampak nadomestimo le izhodno plast in se učimo le njenih uteži. Ta poenostavitev znatno zmanjšuje računsko zahtevnost učenja, vendar vodi do potencialne izgube točnosti napovedi. Izboljšavo prepuščamo nadaljnjemu delu.

3 Metoda MICE

V predlaganem pristopu, imenovanem MICE (*Mining Idioms with Contextual Embeddings*; rudarjenje idiomov s kontekstnimi vložitvami), uporabljamo vektorske vložitve tipa ELMo in BERT na vhodu v nevronske mreže. Pokažemo, da njihova uporaba izboljša rezultate v primerjavi z obstoječimi pristopi. Naš pristop učimo na novi učni množici slovenskih idiomov. Analiziramo različne lastnosti predlaganih modelov, na primer količino označenih podatkov, potrebnih za pridobitev koristnih rezultatov, in več različic modela BERT.

Pokažemo, da kontekstne vložitve vsebujejo veliko količino leksikalnih in semantičnih informacij, ki jih lahko uporabimo za

zaznavanje idiomov. Naš pristop MICE pri uspešnosti prepoznave presega obstoječe pristope, ki ne uporabljajo kontekstnih vložitev, tako pri odkrivanju idiomov, prisotnih v učni množici, kakor tudi idiomov, ki jih v učni množici ni.

Naš pristop temelji na kontekstnih vložitvah besed, ki so bile zasnovane za obravnavo večpomenskih besed. Namesto da vsaki pojavitvi besede dodelijo isti vektor, vsaki pojavitvi besede dodelijo različen vektor na podlagi njenega konteksta, tipično stavka. Ker se konteksti dobesedne in idiomatične uporabe skupka besed zelo verjetno razlikujejo, so kontekstne vložitve primerne za zaznavanje idiomatične rabe. Uporabili smo dva najsodobnejša pristopa vložitev: ELMo in BERT. Za ELMo smo uporabili slovenski model, ki sta ga zgradila Ulčar in Robnik-Šikonja (2020b). Model je bil naučen na korpusu slovenskih besedil Gigafida 2.0 (Krek et al. 2016, Krek et al. 2020). Za vložitve tipa BERT smo uporabili dva različna modela. Prvi je večjezični model mBERT, ki so ga predstavili Devlin et al. (2019) in je bil naučen na besedilih Wikipedije v 104 jezikih, vključno s slovenskim. Drugi, trojezični model CroSloEngual BERT (Ulčar in Robik-Šikonja 2020a), je bil naučen na angleščini, slovenščini in hrvaščini z uporabo Wikipedije za angleščino, korpusom Gigafida 2.0 za slovenščino in kombinacijo korpusa hrWaC (Ljubešić et al. 2011), člankov medijske skupine Styria in korpusa Riznica (Ćavar in Rončević 2012) za hrvaščino. Model BERT je primernejši za klasifikacijske naloge v slovenščini in hrvaščini kot mBERT, saj je bil naučen na večjih zbirkah besedil v teh dveh jezikih. Avtorja poročata o izboljšanem medjezikovnem prenosu naučenih klasifikacijskih modelov med vključenimi tremi jeziki.

V naši arhitekturi napovednih modelov prvi sloj nevronske mreže predstavljajo vektorske vložitve (ELMo ali BERT). Temu sloju sledi dvosmerna rekurenčna mreža tipa GRU s 100 celicami. Rekurenčne nevronske mreže so zmožne iz zaporedja besed razbrati semantične in sintaktične informacije, ki so koristne pri prepoznavanju idiomov. Rekurenčni mreži sledi sloj softmax, ki na podlagi pridobljenih informacij izračuna končne napovedi. Arhitektura sledi modelu za odkrivanje večbesednih enot, ki so ga predstavili Klyueva et al. (2017), z razliko, da uporabljamo kontekstne vložitve. Namenoma

uporabljamo preprosto arhitekturo nevronske mreže, da pokažemo, da že kontekstne vložitve same po sebi zajamejo dovolj semantičnih informacij za pravilno prepoznavanje idiomov.

Arhitekturo uporabljamo na dveh vrstah klasifikacij: klasifikaciji na ravni besede oz. žetona, kjer napovedujemo, ali ima posamezna beseda idiomatični ali dobesedni pomen, in klasifikaciji na ravni stavkov, kjer za celoten stavek napovemo, ali vsebuje izraz z idiomatičnim pomenom.

Hiperparametre nevronskih mrež prilagodimo z uporabo razvojne množice, sestavljene iz 7 % stavkov, naključno izbranih iz našega nabora podatkov. Mrežo smo učili 10 epoh z RMSProp optimizatorjem s stopnjo učenja 0,001, rho = 0,9 in epsilon = 10-7. Kot funkcijo izgube uporabimo binarno navzkrižno entropijo.

3.1 Podatkovne množice idiomov

Za ocenjevanje samodejnega zaznavanja idiomov na slovenskih besedilih smo zgradili korpus slovenskih idiomov, imenovan SloIE, ki je prosto dostopen na repozitoriju CLARIN.SI.[1] Korpus vsebuje 29.400 stavkov, izluščenih iz korpusa Gigafida 2.0 (Krek et al. 2016, Krek et al. 2020), ki vsebujejo 75 različnih idiomov (Priloga), izbranih na podlagi Leksikalne baze za slovenščino (Gantar in Krek 2011), za katere je bilo predhodno ugotovljeno, da se v stavkih pojavljajo v svojem idiomatičnem in dobesednem pomenu. Primer idioma, ki ustreza tema pogojema, je npr. *imeti krompir* v Tabeli 1.

Za namen prepoznavanja idiomatičnih in dobesednih pomenov v korpusnih stavkih smo izvedli označevalno kampanjo, v kateri so celoten nabor 29.400 iz korpusa izluščenih stavkov z vsebovanimi idiomi označile štiri študentke jezikoslovja, in sicer vsak stavek dve različni označevalki. Kot kaže Tabela 1, so imele pri označevanju na voljo štiri možne izbire: DA (izraz v določenem stavku se uporablja v idiomatičnem pomenu), NE (izraz se uporablja v dobesednem pomenu), NE VEM (nisem prepričana, ali se izraz uporablja v dobesednem ali idiomatičnem pomenu) in NEJASEN ZGLED (iz stavka ni

1 Povezava do korpusa na repozitoriju: http://hdl.handle.net/11356/1335.

mogoče razbrati dobesedne ali idiomatične rabe). Študentke so bile vnaprej seznanjene s kratkimi navodili in z vzorcem dobrih primerov.

Tabela 1: Primer označenih stavkov z oceno idiomatičnosti pomena vsebovanega idioma.

idiom	stavek	ocena 1	ocena 2
imeti krompir	Za kosilo so imeli v skledi zabeljen krompir.	NE	NE
	Njim ni nikoli ničesar manjkalo, krompir so imeli, sekira jim je padla v med.	NE VEM	NEJASEN ZGLED
	Kdo že ima debel krompir?	NEJASEN ZGLED	NEJASEN ZGLED
	Nekdo ima krompir, nekdo drug ima pa smolo.	DA	DA
	Ti imaš pa res vedno krompir.	DA	DA
	»V Šenčurju imamo pa res krompir,« je na sobotni prireditvi Praznik krompirja ugotovil župan Miro Kozelj.	NE	NE VEM

Hitri pregled 10 naključno izbranih idiomov (Tabela 2) je pokazal, da se približno polovica idiomov, ki se sicer pojavljajo v idiomatičnem in dobesednem pomenu, pojavlja v 50 ali več odstotkih korpusnih primerov v svojem idiomatičnem pomenu (obarvano) in približno polovica je takih, ki se v 50 odstotkih ali več pojavljajo v dobesednem pomenu.

Tabela 2: Odstotek prepoznanih idiomatičnih, neidiomatičnih in dvoumnih stavkov za posamezni idiom pri obeh označevalkah.

idiom	označevalka 1			označevalka 2		
	DA %	NE %	NEJASNO %	DA %	NE %	NEJASNO %
barvati kaj s črnimi barvami	50	50	0	50	50	0
kdo nosi hlače	19	75	5	10	70	19
kdo nosi težak križ	41	50	8	41	50	8
kdo pade v naročje	77	13	9	63	13	22
kdo si oblizuje prste	84	4	11	51	17	31
kislo jabolko	37	31	31	25	56	18
kot bi odrezal	59	31	9	45	3	51
letati od cveta do cveta	30	60	10	30	40	25
med in mleko	46	6	46	40	6	53
oprati si roke	85	10	3	66	14	19

Visoka stopnja idiomatičnih interpretacij pomena kot tudi razmeroma majhen nabor idiomov z izkazanima obema pomenskima rabama nakazujeta zanimiva raziskovalna izhodišča, kot je npr. zakaj se večina idiomov pojavlja pogosteje ali celo izključno v svojem idiomatičnem pomenu, čeprav je dobesedna raba skladenjsko in semantično možna, npr. *narediti kaj za čigavim hrbtom*, *zlesti komu pod kožo*. Poleg tega bi bilo v prihodnje smiselno pristop preizkusiti še na enotah, ki kažejo tendenco bodisi dobesednega bodisi idiomatičnega pomena in takih, ki izkazujejo večjo stopnjo dvoumnosti.

Ker gre za pilotno raziskavo, ki na slovenskem gradivu še ni bila opravljena, smo se pri oblikovanju učne množice omejili le na idiome, ki izpolnjujejo pogoj, da se v korpusnih stavkih pojavijo tako v idiomatičnem kot dobesednem pomenu, pri čemer smo domnevali, da govorci lahko dobesedno in idiomatično interpretacijo izraza prepoznamo na podlagi konteksta. Za ocenjevanje samodejnega zaznavanja idiomov smo iz celotnega korpusa izbrali samo stavke, kjer sta obe označevalki primer ocenili z DA ali NE. To je veljalo za 95,2 % primerov. Iz analize smo izpustili primere, kjer se označevalki nista strinjali in dvoumne primere (NE VEM, NEJASEN ZGLED), v prihodnje pa bi bilo smiselno, kot rečeno, razmisliti tudi o vključitvi takih primerov v podatkovno množico.

Zaradi narave idiomatičnih izrazov je naš korpus SloIE v zastopanosti idiomatičnih in neidiomatičnih stavkov za posamezni idiom neuravnotežen. Za večino izrazov vsebuje manj kot 100 korpusnih primerov, vsebuje pa tudi izraze z več tisoč pojavitvami. Tabela 3 prikazuje pregled podatkov korpusa SloIE.

Tabela 3: Pregled podatkov v korpusu SloIE.

Povedi	29.400
Besede	693.795
Idiomatične povedi	24.349
Dobesedne povedi	5.051
Idiomatične besede	67.088
Dobesedne besede	626.707
Št. različnih idiomov	75

SloIE je po številu stavkov veliko večji od drugih obstoječih naborov podatkov. Za primerjavo, angleški korpus VNC Tokens vsebuje 2.984 primerov in 53 različnih idiomov. Podatkovne množice za druge jezike so še manjše. Korpus SloIE bo torej koristen za nadaljnjo raziskovalno delo pri prepoznavanju idiomov.

3.2 Ocenjevanje rezultatov samodejnega zaznavanja idiomov

Rezultate samodejnega zaznavanja idiomov lahko ocenimo z več različnih vidikov.

1. Klasifikacija idiomov, ki so prisotni v učni množici. V tem primeru ocenjujemo, ali je pristop sposoben zaznati idiome, ki so bili prisotni v učnem korpusu. To ocenimo z dveh vidikov:
   i) klasifikacija na ravni stavka, kjer model strojnega učenja vrne eno napoved za celoten stavek, pri čemer napove, ali ta stavek vsebuje izraz z idiomatičnim pomenom, in
   ii) klasifikacija na ravni besed, kjer model za vsako besedo napove, ali ima dobesedni ali idiomatični pomen.

   Klasifikacija na ravni stavka je lažja, vendar je naloga na ravni besed lahko bolj koristna, saj lahko z njo zaznamo, katere besede »sodelujejo« pri idiomatičnem pomenu. V prihodnje bi bilo zato smiselno vpeljati tudi evalvacijo na ravni besedne zveze, kjer bi preverjali, v koliko primerih sistem pravilno napove idiomatični pomen vsaj ene izmed besed idioma.
2. Klasifikacija idiomov, ki niso prisotni v učni množici. Zaradi velikega števila idiomov je časovno zamudno ročno označiti korpus, ki bi vseboval večino možnih idiomov. Zaradi tega želimo, da bi napovedni model lahko prepoznal tudi idiome, ki niso prisotni v učni množici. Tako kot pri prvi nalogi tudi tukaj uporabimo dva načina klasifikacije: na ravni stavka in na ravni besed. Ta naloga je težja od zaznavanja idiomov, prisotnih v naboru podatkov, in jo je mogoče uspešno rešiti le, če kontekstne vektorske vložitve vsebujejo ustrezne informacije o idiomatični rabi besed (npr. kot smeri v vektorskem prostoru).

3. Težavnost prepoznavanja različnih idiomov. Idiomi se lahko glede pomena razlikujejo. Nekateri se približujejo dobesednemu pomenu, medtem ko so nekateri od dobesednega pomena zelo oddaljeni. Zaradi tega se lahko uspešnost strojnih metod razlikuje glede na idiom, ki ga želimo prepoznati.

V sledečih razdelkih ocenimo delovanje različnih pristopov samodejnega prepoznavanja idiomov. Kot izhodišče uporabljamo metodo podpornih vektorjev (angl. *support vector machines*; SVM), ki kot vhod prejme stavek, pretvorjen v vektorsko obliko z metodo tf-idf. Vektorska oblika ne upošteva zaporedja besed, zaradi česar lahko SVM prepozna idiome le na podlagi števila pojavitev besed v povedi. Posledično deluje dobro le v primerih, ko se besedne zveze pojavijo skupaj z besedami, ki jasno nakazujejo idiomatsko ali dobesedno rabo (npr. besedna zveza *držati pokonci*, ki se v idiomatskem pomenu velikokrat pojavi skupaj z besedo *glavo*). Ovrednotimo tudi dve nevronski metodi za prepoznavanje idiomov. Prva je MUMULS, ki uporablja strojno učenje z nevronsko mrežo, vendar pri tem ne uporablja predhodno naučenih kontekstnih vektorskih vložitev. Namesto tega zgradi vložitve iz besede, leme, in oblikoskladenjske oznake vsake besede. MUMULS ne uporablja vnaprej naučenih vložitev. Namesto tega so vložitve na začetku naključno generirane in se jih nevronska mreža nauči med učenjem prepoznavanja idiomov. Druga nevronska metoda je novo predlagan pristop MICE, ki poleg nevronske mreže uporablja še kontekstne vektorske vložitve besed.

Metode ocenimo z vidika klasifikacijske točnosti in binarne mere F1 (harmonična sredina preciznosti in priklica). Klasifikacijska točnost nam pove delež točnih napovedi metode in se pogosto uporablja za ocenjevanje pristopov strojnega učenja. V našem primeru metode ocenjujemo na neuravnoteženi podatkovni množici, zaradi česar klasifikacijska točnost ni najboljša izbira (zaradi neuravnoteženosti bi tudi večinski klasifikator dosegel visoko klasifikacijsko točnost). Raje uporabimo mero F1, ki je v primeru neuravnoteženosti bolj ustrezna.

3.2.1 Klasifikacija idiomov, ki so prisotni v učni množici

Za klasifikacijo idiomov, ki so prisotni v učni množici, nabor podatkov SloIE naključno razdelimo na učno, testno in razvojno množico v razmerju 63:30:7 (18.522, 8.820 in 2.058 stavkov). Delitev izvedemo na ravni idiomov – povedi vsakega idioma razdelimo v navedenem razmerju in jih nato združimo v učno, testno in razvojno množico. S tem zagotovimo, da vse množice vsebujejo dovolj povedi vsakega idioma. Prav tako zagotovimo, da vsaka množica vsebuje vsaj en pozitiven in negativen primer vsakega idioma. Pristope ovrednotimo v dveh sklopih: prepoznavanje posameznih besed v stavku kot idiomatičnih ali neidiomatičnih (tj. klasifikacija na ravni besede oz. žetona) in prepoznavanje, ali celoten stavek vsebuje ali ne vsebuje idiomov (tj. klasifikacija na ravni stavka). Za žeton pri modelih ELMo, SVM in pri večinskem klasifikatorju vzamemo posamezne besede. BERT za pravilno delovanje zahteva tokenizacijo na podbesedne enote, ki nato pri klasifikaciji predstavljajo žetone. Pri klasifikaciji na ravni žetonov za vsak žeton napovemo, ali ima dobeseden ali idiomatski pomen. Pri tem vse žetone v idiomatski besedni zvezi smatramo kot idiomatske. Podrobni podatki o parametrih in postopku učenja nevronskih modelov so na voljo v Škvorc et al. (2020).

Rezultati za klasifikacijo na ravni žetonov so predstavljeni v Tabeli 4:

Tabela 4: Rezultati zaznavanja idiomov na ravni žetonov. Idiomi v testni množici so bili prisotni tudi v učni množici.

Klasifikator	Klasifikacijska točnost	Mera F1
Večinski klasifikator	0,903	0,176
SVM	0,875	0,3962
MUMULS	0,975	0,0659
MICE + Slovenski ELMo	0,889	0,9219
MICE + mBERT	0,814	0,4556
MICE + CroSloEngual BERT	0,972	0,837

Klasifikator SVM doseže boljši rezultat F1 kot MUMULS, vendar nižji rezultat v primerjavi z različicami MICE. Nabor podatkov je zelo

neuravnotežen, saj ima 96,7 % vseh žetonov dobesedni pomen. MUMULS iz povedi ne more razbrati dovolj pomenske informacije in skoraj vsak žeton napove kot dobeseden. Posledično doseže visoko klasifikacijsko točnost, vendar zelo nizko oceno F1. Zaradi neuravnotežene narave nabora podatkov ocena F1 bolje odraža uspešnost pristopov na realnih problemih. V tem pogledu so različice MICE bistveno uspešnejše od drugih metod.

Od treh pristopov MICE ima tisti s slovenskim modelom ELMo najvišjo vrednost mere F1. Različice MICE z vložitvami BERT dosegajo nižje vrednosti klasifikacijske točnosti in mere F1. To je verjetno posledica drugačne tokenizacije, ki jo uporablja model BERT. Pri vložitvah ELMo lahko tokenizacijo izvedemo na ravni besed, medtem ko moramo pri BERTu besede razdeliti na podbesedne enote. Klasifikacija na ravni žetonov z BERTom mora posledično prepoznavati podbesede namesto celotnih besed. Poleg tega smo pri vložitvah ELMo uporabili vložitve, ki smo jih predhodno naučili na veliki množici samo slovenskih besedil. V času evalvacije enojezičnega slovenskega modela BERT, ki bi bil naučen na velikem številu besedil, še nismo imeli na voljo. Zaradi tega smo uporabili večjezične vložitve mBERT, ki so bile naučene na množici besedil iz 104 različnih jezikov, v kateri je bilo vključenih le malo slovenskih besedil, in vložitve CroSloEngual BERT, ki so bile naučene na veliki količini slovenskih, angleških, in hrvaški besedil. Zaradi učenja na večji količini slovenskih besedil z vložitvami CroSloEngual BERT dosegamo boljše rezultate.

Pri ocenjevanju na ravni stavka namesto klasifikacije vsakega žetona za celoten stavek napovemo, ali vsebuje idiom ali ne. To zmanjša pomen različnih pristopov tokenizacije med vložitvami ELMo in BERT. Slabost tega pristopa je, da ne pokaže, ali so modeli sposobni zaznati določene besede v stavku kot idiome. Rezultati te evalvacije so predstavljeni v Tabeli 5. Večinski klasifikator ustreza deležu dobesednih povedi v korpusu (tj. v korpusu je 82 % povedi dobesednih).

Klasifikacija na ravni stavka je manj zahtevna, kar vodi do boljših rezultatov pri vseh modelih. Klasifikator SVM tukaj preseže model

Tabela 5: Rezultati zaznavanja idiomov na ravni stavkov. Idiomi v testni množici so bili prisotni v učni množici.

Klasifikator	Klasifikacijska točnost	Mera F1
Večinski klasifikator	0,828	0,906
SVM	0,900	0,942
MUMULS	0,915	0,948
MICE + Slovenski ELMo	0,951	0,980
MICE + mBERT	0,897	0,908
MICE + CroSloEngual BERT	0,921	0,954

MICE + mBERT. MUMULS doseže boljše rezultate kot SVM in pristop MICE + mBERT. MICE s CroSloEngual BERT je pri tej nalogi bližje modelu ELMo, čeprav slednji še vedno dosega najboljše rezultate. MICE z mBERT verjetno zato dosega nižje rezultate, ker vložitve mBERT predhodno niso bile naučene na dovolj veliki količini slovenskega besedila.

3.2.2 Klasifikacija idiomov izven učne množice

V prejšnjem razdelku smo pokazali, da lahko samodejni pristopi za prepoznavanje idiomov dosežejo dobre rezultate pri idiomih, ki so bili prisotni tako v učni kot testni množici, zlasti z uporabo kontekstnih vektorskih vložitev. V številnih jezikih žal nimamo velikih, ročno označenih učnih množic. Tudi če take množice obstajajo, verjetno ne bodo vsebovale vseh možnih idiomov, ki jih najdemo v besedilih. Zaradi tega ocene na idiomih, ki so bili prisotni v učni množici, ne odražajo najbolje praktične uporabnosti preizkušenih metod.

Da bi dosegli bolj reprezentativne rezultate, smo preizkusili, kako dobro delujejo pristopi pri prepoznavanju idiomov izven učne množice. Za poskus smo nabor podatkov razdelili na učno in testno množico tako, da idiomi iz testne množice niso bili prisotni v učni množici. Razen te spremembe postopek ostane enak kot v prejšnji klasifikaciji.

Ker idiomi v testni množici niso prisotni v učni množici, se klasifikacijski modeli ne morejo naučiti, kako jih zaznati samo na podlagi

okoliških besed. Ker se pomen idiomov razlikuje od dobesednega pomena besed, ki idiom sestavljajo, bi se morali pojavljati v drugačnih kontekstih kot dobesedne besede. Nevronske mreže s kontekstnimi vektorskimi vložitvami bi lahko takšne pojave zaznale tudi za idiome, ki niso prisotni v učni množici.

Naši rezultati za zaznavanje idiomov na ravni besed in stavkov kažejo, da pristopi, ki ne uporabljajo kontekstnih vektorskih vložitev, ne morejo uspešno zaznati idiomov, ki niso prisotni v učni množici, medtem ko pristopi MICE s kontekstnimi vložitvami pridobijo koristne informacije.

Pri rezultatih na ravni besed zaradi neuravnotežene porazdelitve razredov (večina besed ima dobesedni pomen), vsi pristopi dosežejo slabšo klasifikacijsko točnost kot večinski klasifikator. Za SVM in MUMULS to velja tudi pri oceni F1. Pristop MICE z modeli ELMo in mBERT uspe pravilno razvrstiti številne idiome, vendar s slabšo točnostjo kot v prejšnjem razdelku. MICE z vložitvami ELMo je spet najboljša metoda, CroSloEngual vložitve pa so presenetljivo neuspešne. Rezultati so prikazani v Tabeli 6.

Tabela 6: Rezultati zaznavanja idiomov na ravni žetonov. Idiomi v testni množici niso bili prisotni v učni množici.

Klasifikator	Klasifikacijska točnost	Mera F1
Večinski klasifikator	0,903	0,176
SVM	0,870	0,029
MUMULS	0,873	0,000
MICE + Slovenski ELMo	0,803	0,866
MICE + mBERT	0,733	0,803
MICE + CroSloEngual BERT	0,759	0,176

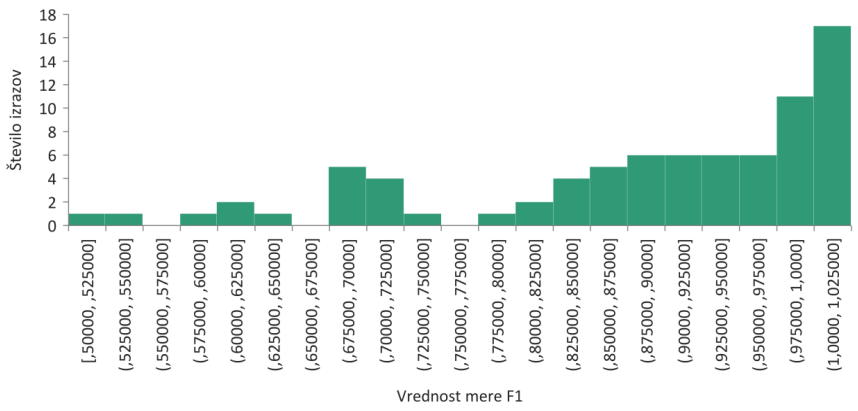
Na ravni stavka spet dosežemo boljše rezultate. Pristopa SVM in MUMULS še vedno zaostajata za privzetim klasifikatorjem glede klasifikacijske točnosti in mere F1. MICE pristopi so boljši, slovenska različica ELMo pa spet dosega najboljše rezultate. Rezultati so prikazani v Tabeli 7.

Tabela 7: Rezultati zaznavanja idiomov na ravni stavkov. Idiomi v testni množici niso bili prisotni v učni množici.

Klasifikator	Klasifikacijska točnost	Mera F1
Večinski klasifikator	0,828	0,906
SVM	0,783	0,689
MUMULS	0,520	0,672
MICE + Slovenski ELMo	0,842	0,907
MICE + mBERT	0,836	0,904
MICE + CroSloEngual BERT	0,771	0,837

3.2.3 Razlike pri zaznavanju različnih idiomov

Rezultati na celotni testni množici ne pokažejo polne slike delovanja samodejnega zaznavanja idiomov. Pristopi, ki jih obravnavamo v prispevku, delujejo na predpostavki, da se besede v idiomatičnem pomenu pojavljajo v drugačnih kontekstih kot v dobesednem pomenu. Zaradi tega je mogoče, da je nekatere idiome enostavno zaznati, druge pa težko. Ali to drži, preverimo tako, da modele naučimo na vseh razen enem idiomu (skupaj se učimo na 74 idiomih) in jih preizkusimo na izpuščenem idiomu. Postopek ponovimo za vse idiome in tako dobimo ločen model zaznavanja za vsak idiom. Za to nalogo uporabimo slovenski model MICE ELMo, saj je v prejšnjih testih presegel vse druge modele. Evalvacijo izvedemo s klasifikacijo na nivoju povedi.



Slika 1: Prikaz distribucije rezultatov zaznavanja različnih idiomov. MICE deluje dobro na večjem delu idiomov (F1 vrednosti > 0,8).

Slika 1 prikazuje porazdelitev ocen F1 med vsemi idiomi v korpusu SloIE. Porazdelitev kaže, da za večino idiomov model doseže visoke ocene F1 (nad 0,8), medtem ko nekaj idiomov prepoznava z nizko stopnjo F1 pod 0,6. Tabela 8 prikazuje pet najbolje in pet najslabše zaznanih idiomov.

Tabela 8: Prikaz rezultatov zaznavanja različnih idiomov. MICE nekatere idiome zazna v vseh primerih (F1 vrednost 1,0), nekatere pa precej slabše (F1 vrednost okoli 0,5).

Idiom	F1 vrednost	Število zaznanih idiomov
pospraviti kaj v arhive	1,0	4
kislo jabolko	1,0	9
pomešati jabolka in hruške	1,0	33
pristati v žepih koga	1,0	28
perje začne frčati	1,0	19
pospraviti kaj v arhiv	0,600	12
imeti krompir	0,597	162
gnilo jajce	0,571	11
kdo nosi hlače	0,525	218
želodec se obrne komu	0,487	10

Razlike v zaznavnosti idiomov z visoko in nizko F1 vrednostjo ni mogoče pojasniti s pogostnostjo njihove rabe oz. številom stavkov, v katerih se pojavljajo, saj so zlasti nižje pogostnosti prisotne v obeh setih idiomov, npr. *pospraviti kaj v arhive* (5 pojavitev), kjer so skoraj visi korpusni stavki prepoznani v idiomatičnem pomenu, in *pospraviti kaj v arhiv* (14 pojavitev), kjer nekoliko prevladuje dobesedna raba. Za dani primer je sicer mogoče sklepati, da je idiomatični pomen vezan na ustaljenost množinske oblike samostalnika. V obeh skupinah idiomov prevladuje bodisi idiomatična bodisi dobesedna raba, razlika med njima pa je tako v setu z večjo kot v setu z nižjo vrednostjo F1 bodisi minimalna (*kislo jabolko* 6-DA : 5-NE; *pospraviti kaj v arhiv* 6-DA : 8-NE), bodisi bolj opazna: *pristati v žepih koga* 27-DA: 2-NE; *gnilo jajce* 11-DA : 1-NE). Na podlagi tega ne moremo sklepati, da je mogoče idiome s prevladujočim deležem idiomatične ali dobesedne rabe bodisi lažje bodisi težje samodejno zaznati, bi

bilo pa v prihodnje smiselno analizirati tudi idiome, pri katerih izstopa število dvoumnih primerov (*imeti krompir, kdo nosi hlače*). Hkrati bi bilo vzroke mogoče iskati tudi v drugih lastnostih besed, kot je npr. večpomenskost, ustaljenost oblike ipd.

4 Zaključek

Predstavili smo nekaj novih načinov za strojno zaznavanje idiomov v besedilih. Predstavljeni pristopi temeljijo na globokih nevronskih mrežah in kontekstnih vložitvah besed. Pokazali smo, da lahko modeli strojnega učenja iz konteksta besede zaznajo, ali ima dobesedni ali idiomatični pomen. Modeli kontekst zajamejo s kontekstnimi vektorskimi vložitvami. Ko smo kot prvo plast nevronske mreže uporabili kontekstne vložitve (ELMo ali BERT) z enako arhitekturo kot obstoječi pristopi, ki takšnih vložitev ne uporabljajo, smo dosegli mnogo boljše rezultate. Pristopi za samodejno zaznavanje idiomov se dobro izkažejo pri klasifikaciji idiomov na ravni stavkov, na ravni žetonov pa delujejo nekoliko slabše.

Z uporabo kontekstnih vložitev so predlagani pristopi zmožni zaznati tudi idiome, ki niso prisotni v učni množici. To omogoča uspešno zaznavanje idiomov brez potrebe po velikih ročno označenih korpusih, kar odpira priložnost za samodejno zaznavanje idiomov v številnih aplikacijah ter v jezikih, kjer takšni korpusi niso na voljo. Pristope smo ovrednotili na novem slovenskem korpusu idiomov SloIE, ki je večji od večine obstoječih korpusov idiomov.

Ker lahko kontekstne vektorske vložitve pri idiomih zaznajo različne pomene besed, predpostavljamo, da bi podobne rešitve lahko delovale tudi pri prepoznavanju drugih figurativnih oblik jezika. V prihodnosti nameravamo podobne metode preizkusiti na metaforah in na drugih tipih večbesednih enot, kot so npr. stalne besedne zveze, ki imajo tako kot idiomi svoj pomen, vendar ta nima idiomatične vrednosti, čeprav je lahko nastal po idiomatični poti, npr. *črna skrinjica, taščin jezik*. Take stalne zveze je namreč težko ločevati od kolokacij, ki so tipične sopojavitve besed brez lastnega celostnega pomena. Prav tako bi bilo mogoče metodo preizkusiti na ravni besedne zveze

in v učni množici upoštevati tudi deleže idiomatičnih in dobesednih pomenov ter delež dvoumnih stavkov. V nadaljnjem delu nameravamo analizirati tudi prenos naučenih modelov v podobne jezike, kjer učna množica ne obstaja, npr. v hrvaščino.

Raziskava je pokazala tudi pomembnost izdelave (čim bolj obsežnih) korpusov z vključenimi semantičnimi podatki o večbesednih enotah, saj lahko njihova integracija v opisano arhitekturo smiselno pripomore k izboljšavi rezultatov in razvoju sistemov za strojno prepoznavanje idiomov.

Zahvala

Raziskovalna programa št. P6-0411 (Jezikovni viri in tehnologije za slovenski jezik) in št. P6-0215 (Slovenski jezik – bazične, kontrastivne in aplikativne raziskave), kakor tudi projekta J6-8256 (Nova slovnica sodobne standardne slovenščine: viri in metode) in J6-2581 (Računalniško podprta večjezična analiza novičarskega diskurza s kontekstualnimi besednimi vložitvami) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Delno je bilo delo sofinancirano tudi s strani okvirnega programa Evropske unije za raziskave in inovacije Obzorje 2020 projekt EMBEDDIA (št. proj. 825153, Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

Reference

Berk, G., Erden, B. in Güngör, T. (2018). Deep-BGT at PARSEME shared task 2018: Bidirectional LSTM-CRF model for verbal multiword expression identification. V A. Savary, Carlos R., J. D. Hwang, N. Schneider, M. Andresen, S. Pradhan in M. R. L. Petruck (ur.), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)* (str. 248–253). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W18-4927.pdf.

Bojanowski, P., Grave, E., Joulin, A. in Mikolov, T. (2017). Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, 5, 135–146. Dostopno prek: https://transacl.org/ojs/index.php/tacl/article/view/999.

Boroş, T. in Burtica, R. (2018). GBD-NER at PARSEME shared task 2018: Multi-word expression detection using bidirectional long-short-term memory net-works and graph-based decoding. V A. Savary, Carlos R., J. D. Hwang, N. Schneider, M. Andresen, S. Pradhan in M. R. L. Petruck (ur.), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)* (str. 254–260). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W18-4928.pdf.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. in Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. V A. Moschitti, B. Pang, W. Daelemans (ur.), *Proceedings ofthe 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (str. 1724–1734). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/D14-1179.pdf.

Cook P. in Fazly A., Stevenson S. (2008). The VNC-tokens dataset. V *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)* (str. 19–22). Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2008/workshops/W20_Proceedings.pdf

Ćavar, D. in Rončević, D. B. (2012). Riznica: the Croatian language corpus. *Prace filologiczne*, 63, 51–65.

Devlin, J., Chang, M.-W., Lee, K. in Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. V J. Burstein, C. Doran in T. Solorio (ur.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (str. 4171–4186). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/N19-1423.pdf.

Ehren, R., Lichte, T. in Samih, Y. (2018). Mumpitz at PARSEME shared task 2018: A bidirectional LSTM for the identification of verbal multiword expressions. V A. Savary, Carlos R., J. D. Hwang, N. Schneider, M. Andresen, S. Pradhan in M. R. L. Petruck (ur.), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)* (str. 261–267). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W18-4929.pdf.

Fadaee M., Bisazza A. in Monz C. (2018). *Examining the tip of the iceberg: A dataset for idiom translation*. Dostopno prek: https://arxiv.org/pdf/1802.04681.pdf.

Gantar P. in Krek, S. (2011). Slovene lexical database. V D. Majchraková in R. Garabík (ur.), *Natural language processing, multilinguality: 6th international conference* (str. 72–80). Brno: Tribun EU. Dostopno prek: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.396.1420&rep=rep1&type=pdf#page=72.

Kim, Y., Jernite, Y., Sontag, D. in Rush, A. M. (2016). Character-aware neural language models. V *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)* (str. 2741–2749). Dostopno prek: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12489.

Klyueva, N., Doucet, A. in Straka, M. (2017). Neural networks for multi-word expression detection. V S. Markantonatou, C. Ramisch, A. Savary in V. Vincze (ur.), *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (str. 60–65). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W17-1707.pdf.

Korkontzelos, I., Zesch, T., Zanzotto, F. M. in Biemann, C. (2013). Semeval-2013 task 5: Evaluating phrasal semantics. V S. Manandhar in D. Yuret (ur.), *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (str. 39–47). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/S13-2007.pdf.

Krek, S., Gantar, P., Arhar Holdt, Š. in Gorjanc, V. (2016). Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres. V T. Erjavec in D. Fišer (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (str. 200–202). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Krek-et-al_Nadgradnja-korpusov-Gigafida-Kres-ccGigafida-ccKres.pdf.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk in S. Piperidis (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*: *conference proceedings* (str. 3340–3345). European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. https://doi.org/10.1038/nature14539.

Liu, C. in Hwa, R. (2017). Representations of context in recognizing the figurative and literal usages of idioms. V *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-17)* (str. 3230–3236). Dostopno prek: https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14939.

Liu, C. in Hwa, R. (2018). Heuristically informed unsupervised idiom usage recognition V E. Riloff, D. Chiang, J. Hockenmaier in J. Tsujii (ur.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (str. 1723–1731). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/D18-1199.pdf.

Ljubešić, N. in Erjavec, T. (2011). hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. V I. Habernal in V. Matoušek (ur.), *Text, Speech and Dialogue: proceedings* (Lecture Notes in Computer Science, vol. 6836) (str. 395–402). Berlin; Heidelberg: Springer. https://doi.org/10.1007/978-3-642-23538-2_50.

Mikolov T., Le Q. V. in Sutskever I. (2013). *Exploiting similarities among languages for machine translation*. Dostopno prek: https://arxiv.org/pdf/1309.4168.pdf.

Pennington, J., Socher, R. in Manning, C. (2014). GloVe: Global vectors for word representation. V A. Moschitti, B. Pang in W. Daelemans (ur.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (str. 1532–1543). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/D14-1162.pdf.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. in Zettlemoyer, L. (2018). Deep contextualized word representations. V M. Walker, H. Ji in A. Stent (ur.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (str. 2227–2237). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/N18-1202.pdf.

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaite, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C. ... Walsh, A. (2018). Edition 1.1 of the

PARSEME shared task on automatic identification of verbal multiword expressions. V A. Savary, Carlos R., J. D. Hwang, N. Schneider, M. Andresen, S. Pradhan in M. R. L. Petruck (ur.), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)* (str. 222–240). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W18-4925.pdf.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., B., QasemiZadeh, Candito, M., Cap, F., Giouli, V., Stoyanova, I. in Doucet, A. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. V S. Markantonatou, C. Ramisch, A. Savary in V. Vincze (ur.), *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (str. 31–47). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W17-1704.pdf.

Sporleder, C. in Li, L. (2009). Unsupervised recognition of literal and non-literal use of idiomatic expressions. V A. Lascarides, C. Gardent in J. Nivre (ur.), *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (str. 754–762). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/E09-1086.pdf.

Škvorc, T., Gantar, P. in Robnik-Šikonja, M. (2020). *MICE: Mining Idioms with Contextual Embeddings*. Dostopno prek: https://arxiv.org/pdf/2008.05759.pdf.

Ulčar, M. in Robnik-Šikonja, M. (2020a). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. V P. Sojka, I. Kopeček, K. Pala in A. Horák (ur.), *Text, Speech and Dialogue: proceedings* (Lecture Notes in Computer Science, vol. 12284) (str. 104–111). Cham: Springer. https://doi.org/10.1007/978-3-030-58323-1_11.

Ulčar, M. in Robnik-Šikonja, M. (2020b). High quality ELMo embeddings for seven less-resourced languages. V N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk in S. Piperidis (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*: *conference proceedings* (str. 4731–4738). European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. in Polosukhin, I. (2017). Attention is all you need. V I. Guyon,

U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan in R. Garnett (ur.), *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (str. 5998–6008).

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. in Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. V T. Linzen, G. Chrupała in A. Alishahi (ur.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (str. 353–355). Association for Computational Linguistics. Dostopno prek: https://aclanthology.org/W18-5446.pdf.

Zhang, X., Zhao, J., LeCun, Y. (2015). Character-level convolutional networks for text classification, V C. Cortes, N. Lawrence, D. Lee, M. Sugiyama in R. Garnett (ur.), *Advances in Neural Information Processing Systems 28 (NIPS 2015)* (str. 649–657).

Priloga: Seznam idiomov z idiomatičnim in dobesednim pomenom, ki so bili uporabljeni v učni množici.

1	barvati kaj s črnimi barvami	39	ohladiti vroče glave
2	brusiti zobe	40	oprati si roke
3	dobiti debelo kožo	41	oprati svoje umazano perilo
4	dobiti ošpice	42	oprati umazano perilo
5	držati kaj pokonci	43	ovijati koga okoli prsta
6	držati vrečo	44	pade na plodna tla
7	dvigniti oblak prahu	45	pade v naročje komu
8	glava boli koga	46	pajčevina se nabira
9	gnilo jajce	47	paradni konj
10	igrati vlogo	48	perje frči
11	imeti debelo kožo	49	perje začne frčati
12	imeti jajca	50	plesati po taktih koga
13	imeti krompir	51	pobirati drobtine
14	imeti močan želodec	52	pobirati sadove česa
15	iskati kaj s povečevalnim steklom	53	pobrati smetano
16	jemati dih	54	pokaditi pipo miru
17	jemati kaj z veliko žlico	55	pokazati mišice

18	jemati komu sapo	56	polagati komu kaj na jezik
19	juha se ohladi	57	položiti komu kaj na jezik
20	kaj ima glavo in rep	58	pomešati hruške in jabolka
21	kaj pade na glavo	59	pomešati jabolka in hruške
22	kaj pade v vodo	60	posneti smetano
23	kaj rodi sadove	61	pospraviti kaj v arhiv
24	kdo bi si obliznil prste	62	pospraviti kaj v arhive
25	kdo drži skupaj	63	postaviti koga pokonci
26	kdo nosi hlače	64	pranje umazanega perila
27	kdo nosi težak križ	65	prati umazano perilo
28	kdo si oblizuje prste	66	prebiti led
29	kdo/kaj pasti v naročje komu	67	prestopiti prag
30	kislo jabolko	68	pristati na realnih tleh
31	kot bi odrezal	69	pristati na trdih tleh
32	letati od cveta do cveta	70	pristati na trdnih tleh
33	letati s cveta na cvet	71	pristati v naročju česa
34	med in mleko	72	pristati v žepih koga
35	mešati jabolka in hruške	73	pristati v žepu koga
36	odreti komu kožo	74	prižgati rdečo luč
37	ohladiti si glavo	75	želodec se obrne
38	ohladiti si pregreto glavo		