REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI SKLAD ZA REGIONALNI RAZVOJ
NALOŽBA V VAŠO PRIHODNOST

Učne množice za razumevanje naravnega jezika: zbirka nalog SuperGLUE

Poročilo projekta Razvoj slovenščine v digitalnem okolju

Aktivnost DS1.2

Avtorji: Aleš Žagar[°], Marko Robnik-Šikonja[°], Špela Arhar Holdt[°,□], Teja Goli[°,□]

[°] Fakulteta za računalništvo in informatiko Univerze v Ljubljani
[□] Filozofska fakulteta Univerze v Ljubljani

Ljubljana: Center za jezikovne vire in tehnologije, Univerza v Ljubljani, 2023

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Vsebina

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

1 Projektni okvir

Poročilo oz. kazalnik *Učne množice za razumevanje naravnega jezika: zbirka nalog SuperGLUE* je nastalo pod okriljem projekta Razvoj slovenščine v digitalnem okolju, ki sta ga med leti 2020 in 2023 sofinancirali Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj (Operacija se je izvajala v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020). Cilj projekta je bil zadovoljiti potrebe po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik za raziskovalne organizacije, podjetja in širšo javnost. Spletna stran projekta z dostopom do rezultatov: https://slovenscina.eu/.

Kazalnik se umešča v prvi projektni delovni sklop z naslovom *Jezikovni viri*. Namen delovnega sklopa je bil nadgraditi slovenske besedilne korpuse in leksikon besednih oblik. Prenovili smo učne množice in postopke za strojno označevanje sodobne slovenščine. Rezultat so osveženi in povečani jezikovni viri, ki so na voljo tako uporabniški skupnosti kot za strojno rabo. Z razvitimi postopki in orodji bo posodabljanje slovenskih korpusov v prihodnosti hitrejše in preprostejše.

Cilj aktivnosti 1.2 je bil nadgraditi oz. razviti napredno označeni učni množici za procesiranje in razumevanje naravnega jezika: učni korpus ssj500k in zbirko nalog SuperGLUE. V tem poročilu obravnavamo prevod in prilagoditev zbirke nalog SuperGLUE slovenskemu jeziku. V zaključku navedemo tudi načrte za nadaljnji razvoj te zbirke in prilagoditev še naprednejših evalvacijskih zbirk za slovenščino.

Razvoj metod za razumevanje naravnega jezika zahteva vrsto zahtevnih evalvacijskih nalog, ki spodbujajo razvoj novih pristopov in primerjavo z obstoječimi metodami. V svetu sta se za angleščino uveljavila nabora evalvacijskih nalog GLUE (General Language Understanding Evaluation) (Wang et al 2019a) in še zahtevnejši SuperGLUE (Wang et al 2019b). Nabor GLUE sestavlja 9 nalog različnih velikosti in zahtevnosti. Naloge sestavlja od nekaj tisoč do več sto tisoč primerov besedil. Zbirko SuperGLUE sestavlja 8 podobnih, vendar težavnejših nalog (sklepanje, odgovori na vprašanja, razdvoumljanje, koreferenčnost). Zbirko nalog SuperGLUE smo delno prevedli in priredili za slovenščino, ki je tako eden redkih jezikov s tako zbirko. To bo omogočilo, da bo slovenščina postala eden od jezikov, na katerih se bodo razvijala nova orodja za razumevanje naravnega jezika. Delo je potekalo v obdobju M2-M24.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Zbirko nalog SuperGLUE smo skladno z razpoložljivimi sredstvi delno ročno in delno strojno prevedli (124.000 ročno prevedenih besed in več kot 2.000.000 strojno prevedenih besed), kot natančneje predstavimo v nadaljevanju. Prevedena besedila smo uporabili za razvoj drugih orodij znotraj projekta, npr. učne množico za odgovore na vprašanja, za razdvoumljanje in odkrivanje koreferenčnosti. Del nalog, ki so namenjene zdravorazumskemu sklepanju in so vezane na anglo-saksonsko kulturno okolje, smo prilagodili slovenskemu kulturnemu okolju. Prevedeno in prilagojeno zbirko smo deponirali na repozitoriju CLARIN.SI (Žagar et al 2021). Podrobneje smo prevod, prilagoditev in analizo zbirke predstavili v publikaciji (Žagar in Robnik-Šikonja 2022).

V nadaljevanju na kratko predstavimo značilnosti slovenske zbirke SuperGLUE in njeno prvo analizo. V razdelku 2 predstavimo pomen zbirk nalog SuperGLUE za razumevanje naravnega jezika. V razdelku 3 opisujemo postopek prevajanja in končne učne množice iz zbirke SuperGLUE v slovenščini. V razdelku 4 opišemo tri postopke evalvacije (enojezikovno, medjezikovni prenos in večjezikovno), v razdelku 5 pa predstavimo rezultate. V razdelku 6 so predstavljeni zaključki, omejitve našega dela in načrti za nadaljnje izboljšave.

2 Pomen zbirke nalog SuperGLUE

Za merjenje napredka na hitro razvijajočem se področju obdelave naravnih jezikov (angl. *Natural language processing*, NLP) je bilo predlaganih več pomembnih zbirk nalog, kot so SentEval (Conneau et al. 2018), GLUE (Wang et al. 2019b) in SuperGLUE (Wang et al. 2019a). SentEval je namenjen ocenjevanju kodirnikov stavkov, GLUE in njegov zahtevnejši naslednik SuperGLUE pa ocenjujeta splošno razumevanje jezika. Ta dva nabora nalog vsebujeta naloge razumevanja naravnega jezika, ki jih je mogoče rešiti prek primerjalnega strežnika, ki daje ločene ocene za vsako od nalog in skupno združeno oceno. Naloge v SuperGLUE so raznolike in vključujejo odgovarjanje na vprašanja (QA), sklepanje v naravnem jeziku (NLI), razreševanje koreferenc in razločevanje besednih pomenov (WSD). Vse naloge so reševali nestrokovnjaki. Njihovi rezultati kažejo osnovno raven človeške uspešnosti in služijo za primerjavo s strojnimi pristopi.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Kljub nedavnim kritikam izračuna rezultatov znanih zbirk nalog (npr. aritmetična sredina ločenih metrik se uporablja za naloge različne zahtevnosti in različnih velikosti učnih množic) (Shavrina in Malykh 2021; Kasai et al. 2021) ni dvoma, da učne množice, ki jih vsebujejo zbirke nalog, pomembno prispevajo k napredku področja NLP. Žal se večina raziskav nanaša na angleščino, kar omejuje splošnost pristopov in ne naslavlja celotnega spektra kompleksnosti jezikov. Kljub temu lahko opazimo prizadevanja, da bi bile ustrezne učne množice večjezične ali prilagojene več jezikom. Večje večjezikovne podatkovne množice, kot sta XTREME (Hu et al. 2020) in XGLUE (Liang et al. 2020), zagotavljajo strojno prevedene nabore podatkov za več nalog in jezikov. Kot dopolnilo temu trendu predstavljamo zbirko v slovenščini, jeziku z manj viri, in sicer kombinirani strojno-človeški prevod zbirke nalog SuperGLUE. Opišemo težave, ki se pojavljajo pri prevajanju, in podajamo prvo vrednotenje učnih množic z uporabo velikih vnaprej naučenih jezikovnih modelov, podobnih modelu BERT (Devlin et al. 2019). Naše vrednotenje zajema enojezične, medjezikovne in večjezične pristope, primerjavo človeških in strojno prevedenih delov ter primerjavo dveh najsodobnejših sistemov strojnega prevajanja.

3 Slovenski prevod zbirke SuperGLUE

Da bi omogočili razvoj naprednih jezikovnih tehnologij v slovenščini, npr. medjezikovnega prenosa in posebnosti morfološko bogatih jezikov, smo učne primere iz zbirke SuperGLUE prevedli v slovenščino. Zaradi omejenih sredstev smo delno uporabili človeški prevod (HT), delno pa strojni prevod (MT). Skupaj je bilo prevedenih približno 120.000 besed. Nekatere učne množice so prevelike (BoolQ, MultiRC, ReCoRD, RTE), da bi jih lahko v celoti prevedli. V Tabeli 1 navajamo razmerja med človeško prevedenim delom množic in celotnimi angleškimi množicami. Za MT iz angleščine v slovenščino smo uporabili storitev GoogleMT Cloud. V naši analizi smo uporabili šest od izvornih osmih nalog. Kot je pojasnjeno v nadaljevanju, smo izključili nalogi ReCoRD in WiC.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Tabela 1: Število primerov v izvirnih angleških in prevedenih slovenskih nalogah SuperGLUE. HT pomeni človeški prevod, MT pa strojni prevod. "Razmerje" označuje razmerje med številom človeško prevedenih primerov in številom vseh primerov.

Množica	razbitje	angleško	HT	razmerje	MT
BoolQ	train	9427	92	0.0098	yes
	val	3270	18	0.0055	yes
	test	3245	30	0.0092	yes
CB	train	250	250	1.0000	yes
	val	56	56	1.0000	yes
	test	250	250	1.0000	yes
COPA	train	400	400	1.0000	yes
	val	100	100	1.0000	yes
	test	500	500	1.0000	yes
MultiRC	train	5100	15	0.0029	yes
	val	953	3	0.0031	yes
	test	1800	30	0.0167	yes
ReCoRD	train	101000	60	0.0006	/
	val	10000	6	0.0006	/
	test	10000	30	0.0030	/
RTE	train	2500	232	0.0928	yes
	val	278	29	0.1043	yes
	test	300	29	0.0967	yes
WiC	train	6000	/	/	/
	val	638	/	/	/
	test	1400	/	/	/
WSC	train	554	554	1.0000	/
	val	104	104	1.0000	/
	test	146	146	1.0000	/

Odločili smo se, da bomo pri testiranju kakovosti modelov strojnega učenja uporabili samo človeško prevedeni del (HT) naših testnih množic, da bi se izognili šumu zaradi prevodov. Zaradi tega so nekatere testne množice precej manjše v primerjavi z angleškimi. Množice ReCoRD nismo vključili v slovensko primerjavo zaradi nizke kakovosti prevedene testne množice, ki je sestavljena iz zmedenih in dvoumnih primerov. Med angleško in slovensko nalogo ReCoRD so tudi precejšnje razlike zaradi morfološkega bogastva slovenščine. V slovenščini namreč pravilna sklanjatev odgovora pogosto ni prisotna v besedilu, kar onemogoča iskanje povsem pravilnega odgovora. Nazadnje, podobno kot pri WSC

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

(obravnavano v nadaljevanju), tudi na ReCoRD vpliva problem prevajanja oznak HTML s strojnim prevajalnikom GoogleMT.

Naloge WiC ni mogoče prevesti in bi jo bilo treba zasnovati na novo, saj je nemogoče prenesti enak nabor pomenov določene besede iz angleščine v slovenščino.

Naloge WSC ni mogoče strojno prevesti, saj je potrebna človeška pomoč in preverjanje. Prvič, prevodi GoogleMT ne morejo obdelati pravilne postavitve oznak HTML, ki označujejo koreferenco. Drugi razlog je, da so v slovenščini koreference lahko izražene tudi z glagoli, medtem ko so koreference v angleščini predvsem samostalniki, lastna imena in zaimki. Zaradi tega je naloga v slovenščini v primerjavi z angleščino drugačna. Po eni strani je naloga v slovenščini težja, saj rešitve zajemajo več vrst besed, po drugi strani pa lahko slovenski glagoli v nekaterih primerih razkrijejo informacije o koreferenci.

4 Načrt evalvacije

Zbirka SuperGLUE se pogosto uporablja za primerjavo velikih vnaprej naučenih jezikovnih modelov v angleščini. V nasprotju s tem se v tem poročilu osredotočamo na slovenski prevod nalog SuperGLUE.

Primerjamo štiri modele BERT, ki so na voljo za slovenski jezik: enojezični slovenski model SloBERTa (Ulčar in Robnik-Šikonja 2021), trojezični (hrvaško-slovensko-angleški) model CroSloEngual BERT (Ulčar in Robnik-Šikonja 2020), masovno večjezični model mBERT (Devlin et al. 2019) (bert-base-multilingual-cased) in model XLM-R (Conneau et al. 2019) (xlm-roberta-base).

Skupna povprečna ocena (Avg v npr. drugem stolpcu Tabele 2) obsega povprečje ovrednotenih nalog. Za angleščino to pomeni povprečje osmih nalog, za slovenščino pa vseh šestih prevedenih nalog: BoolQ, CB, COPA, MultiRC, RTE in WSC. Pri nalogah z več metrikami smo te metrike povprečili, da smo dobili enotno skupno oceno naloge. Podrobnosti o tem, kako se izračuna ocena za posamezno nalogo, so na voljo v (Wang et al. 2019a). Modele smo preizkusili v treh nastavitvah. V enojezični nastavitvi se za učenje in testiranje uporablja isti jezik (slovenščina ali angleščina). V medjezikovnem načinu smo testirali medjezikovne modele (CroSloEngual, mBERT in XLM-R) ter prenos med angleškimi in

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

slovenskimi učnimi množicami v obe smeri. V večjezičnem načinu smo modele učili na združenih angleških in slovenskih podatkih polne velikosti.

Razpoložljive slovenske modele BERT smo prilagajali na nalogah SuperGLUE z orodjem Jiant (Phang et al. 2020). Za vsako nalogo posebej smo prilagajali parameter in modele prilagajali 100 epoh, pri čemer je bila začetna stopnja učenja 10-5. Vsak model je prilagojen z uporabo podatkov, za vsako od treh nastavitev.

5 Rezultati

V tem razdelku poročamo o rezultatih za vsako od treh nastavitev (enojezično, medjezikovne in večjezično) posebej.

5.1 Rezultati enojezikovnih modelov

V naši enojezični analizi primerjamo različne modele učene na celotnih učnih množicah, sestavljenih iz vseh človeško prevedenih primerov, medtem ko so preostali primeri strojno prevedeni. V Tabeli 2 so prikazani rezultati skupaj z več osnovnimi modeli, ki so bili učeni na izvornih angleških množicah. Nekatere primerjave z angleškimi modeli so vprašljive, saj so slovenski modeli učeni le na majhnem delu kakovostnejših človeško prevedenih množic (HT) (BoolQ, MultiRC) in tudi testirani na manjšem naboru teh testnih podatkov. V primeru modela BERT++ je bil angleški model dodatno naučen s prenosom znanja na nalogah, ki so podobne ciljnim (CB, RTE, BoolQ, COPA). Tako je glede na učne množice edina poštena primerjava mogoča s CB, COPA in WSC.

Glede na ocene Avg v Tabeli 2 je enojezični model SloBERTa najuspešnejši slovenski model. V povprečju so vsi slovenski modeli BERT uspešnejši od izhodiščnega modela, ki vrača najpogostejšo oznako (Most Frequent). Kar zadeva posamezne naloge, pri nalogi MultiRC nobeden od slovenskih modelov ne presega izhodiščne vrednosti Most Frequent. SloBERTa je bil bistveno boljši od preostalih modelov v nalogah CB, COPA in WSC, medtem ko je bil XML-R najboljši v nalogi BoolQ.

V primerjavi z angleškimi modeli je najboljši slovenski model (SloBERTa) dosegel boljše rezultate pri WSC. Zdi se, da se nobeden od angleških modelov ni ničesar naučil iz WSC (so pod izhodiščem Most

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Frequent), vendar je model SloBERTa dosegel rezultat 73,3 (izhodišče Most Frequent daje 65,8). Uspeh modela SloBERTa pri WSC morda izhaja iz morfologije slovenskih glagolov, ki vključujejo informacijo o spolu; ta informacija je koristna pri razreševanju koreferenc in v slovenščini poenostavi nekatere primere v primerjavi z angleščino. Kljub temu pa vidimo še velik razkorak v primerjavi s človeško uspešnostjo. Vsi slovenski modeli so pokazali dobro uspešnost pri CB in se uvrstili med angleškim modelom CBoW in BERT.

5.2 Rezultati medjezikovnega prenosa

V medjezikovnem scenariju smo preizkusili tri večjezične modele BERT (mBERT, CroSloEngual, XLMR) ter prenos med angleškimi in slovenskimi učnimi množicami (v obe smeri). Za slovenščino kot izvorni jezik smo uporabili razpoložljive človeško prevedene primere. Da bi bila primerjava uravnotežena, smo uporabili le iste primere iz angleških množic. Preizkusili smo prenos brez dodatnih primerov v ciljnem jeziku in prenos z nekaj dodatnimi primeri. Pri prilagajanju modelov z nekaj dodatnimi primeri smo za vsako nalogo uporabili 10 dodatnih primerov iz ciljnega jezika. Teh 10 primerov smo petkrat naključno vzorčili. Poročamo dosežena povprečja. Hiper parametri za prilagajanja modelov so enaki kot pri enojezičnem testiranju.

Rezultati so predstavljeni v Tabeli 3. V povprečju čez vse naloge so nekateri modeli izboljšali rezultate glede na osnovno različico, ki vrača najpogostejšo oznako (Most frequent). Na splošno so bili modeli precej neuspešni pri BoolQ, MultiRC in WSC, vendar vidimo nekaj obetavnih rezultatov pri COPA, RTE in zlasti CB. Dodatni učni primeri v scenariju z nekaj posnetki so prinesli nekaj vidnih izboljšav. Zdi se, da se modeli bolje obnesejo v angleško-slovenski smeri kot obratno. Najuspešnejši model je XLM-R, sledita mu CroSloEngual BERT in mBERT.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Tabela 2: Primerjava enojezikovnih modelov na nalogah SuperGLUE v angleščini (zgoraj) in slovenščini (spodaj). Vsi rezultati v angleščini so povzeti iz (Wang et al. 2019a). Najboljši rezultat za vsako nalogo in jezik je v krepkem tisku. Najboljši povprečni rezultati (Avg) za vsak jezik so podčrtani.

Task Models/Metrics	Avg	BoolQ Acc.	CB F1/Acc.	COPA Acc.	MultiRC $F1_a$/EM	ReCoRD F1/EM	RTE Acc.	WiC Acc.	WSC Acc.
Most Frequent	45.7	62.3	21.7/48.4	50.0	61.1/0.3	33.4/32.5	50.3	50.0	**65.1**
CBoW	44.7	62.1	49.0/71.2	51.6	0.0/0.4	14.0/13.6	49.7	53.0	**65.1**
BERT	69.3	77.4	75.7/83.6	70.6	70.0/24.0	**72.0/71.3**	71.6	**69.5**	64.3
BERT++	<u>73.3</u>	**79.0**	**84.7/90.4**	**73.8**	**70.0/24.1**	**72.0/71.3**	**79.0**	**69.5**	64.3
Human (est.)	89.8	89.0	95.8/98.9	100.0	81.8*/51.9*	91.7/91.3	93.6	80.0	100.0
Most Frequent	49.1	63.3	21.7/48.4	50.0	**76.4/0.6**	-	58.6	-	65.8
SloBERTa	<u>63.9</u>	66.6	**74.0/76.8**	**61.8**	62.7/21.9	-	**62.1**	-	**73.3**
CroSloEngual	57.8	66.6	62.1/72.4	58.2	56.7/15.6	-	**62.1**	-	56.2
mBERT	59.1	70.0	66.6/73.6	54.2	57.4/16.3	-	**62.1**	-	61.6
XML-R	58.7	**76.7**	66.2/73.2	50.0	55.3/13.9	-	55.2	-	65.8

*MultiRC ima več testnih množic, objavljenih v različnih obdobjih; označeni rezultati so izračunani na testni množici, ki je podmnožica naše.

Tabela 3: Rezultati modelov z medjezikovnim prenosom na nalogah SuperGLUE. Upoštevani so le človeško prevedeni primeri. Najboljši rezultati za medjezikovni prenos brez dodatnih primerov (Zero-shot) in z malo dodatnimi primeri (Few-shot) so v krepkem tisku.

Evaluation	Model	source	target	Avg	BoolQ acc.	CB F1/acc.	COPA Acc.	MultiRC $F1_a$/EM	RTE Acc.	WSC Acc.
Zero-shot	CroSloEngual	english	slovene	49.8	56.7	43.7/60.0	54.6	48.0/6.6	58.6	50.7
		slovene	english	52.6	60.0	53.8/70	**59.6**	**56.7/9.6**	48.3	58.2
	mBERT	english	slovene	47.4	56.7	36.2/57.2	50.2	47.3/8.7	55.2	64.4
		slovene	english	48.3	60.0	44.6/50.4	49.8	56.2/8.7	51.7	57.5
	XLM-R	english	slovene	**53.8**	**63.3**	**62.9/68.4**	53.6	48.5/0.3	**62.1**	56.2
		slovene	english	51.7	**63.3**	59.1/67.2	47.2	52.9/12.9	51.7	**65.8**
Few-shot	CroSloEngual	english	slovene	54.4	60.0	52.4/68.6	55.0	52.8/9.72	**65.5**	54.1
		slovene	english	53.0	60.0	53.8/70.0	**59.5**	56.0/12.1	49.7	58.2
	mBERT	english	slovene	50.9	60.1	53.1/66.2	50.4	50.8/9.8	53.8	64.4
		slovene	english	51.3	60.7	51.8/58.2	50.3	57.2/11.1	56.5	56.8
	XLM-R	english	slovene	**57.0**	**63.3**	**65.8/69.8**	53.3	**76.4/0.6**	62.1	57.4
		slovene	english	53.0	**63.3**	63.0/69.6	48.3	51.4/10.6	55.8	**65.8**
	Most frequent			52.4	63.3	23.0/52.7	50.0	77.3/0.3	58.6	**65.8**

Na splošno dokaj nizko uspešnost je mogoče pojasniti z majhnim številom učnih primerov v izvornem jeziku. Če si podrobneje ogledamo posamezne modele, lahko opazimo, da XLM-R kaže zelo dobre rezultate na CB v obeh smereh. CroSloEngual BERT je dosegel podoben dober rezultat pri COPA. Je edini model, ki se je tudi na tem naboru podatkov nekaj naučil.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Zaključimo lahko, da je za zahtevno zbirko nalog SuperGLUE medjezikovni prenos zahteven, vendar ne nemogoč. V prihodnosti nameravamo poskuse razširiti v več smeri. Najprej bomo angleške modele učili na celotnih učnih množicah SuperGLUE in jih prenesli na slovenske učne množice s človeškim in strojnim prevodom. Drugič, združili bomo prilagajanje za več nalog in preizkusili scenarije učenja s prenosom znanja.

5.3 Rezultati večjezikovnih modelov

V večjezikovnem načinu smo primerjali tri večjezikovne modele BERT (CroSloEngual BERT, mBERT in XLM-R), ki so bili prilagajani na združenih angleških in slovenskih množicah polne velikosti. Slovenske učne množice so sestavljali vsi človeško prevedeni in preostali strojno prevedeni primeri. Za slovenščino smo modele testirali samo na človeško prevedenih podatkih. Rezultati so prikazani v Tabeli 4.

Zanimivo je, da so bili najboljši rezultati za vse naloge doseženi pri testiranju na angleških testnih množicah (učni podatki pa so bili enaki in sestavljeni iz obeh jezikov). To je morda posledica dejstva, da so bili slovenski modeli BERT vnaprej naučeni na manjši količini podatkov, in pa nižje kakovosti slovenskega prevoda. Na splošno je najboljši model z najvišjo oceno Avg mBERT (najboljši v BoolQ, MultiRC in RTE), sledita XLM-R (najboljši v CB in WSC) in CroSloEngual BERT (najboljši v COPA). Pri MultiRC in WSC vsi modeli zaostajajo za osnovnim modelom Most Frequent.

Tabela 4: Rezultati večjezikovnih modelov na testnih množicah v angleščini in slovenščini (v slovenščini sestavljeni le iz človeško prevedenih primerov). Modeli so bili prilagojeni na kombiniranih angleških in slovenskih učnih množicah polne velikosti. Slovenske učne množice so sestavljene iz vseh človeško prevedenih učnih primerov in preostalih strojno prevedenih učnih primerov. Najboljši rezultati za vsako nalogo so označeni v krepkem tisku.

Evaluation on	Model	Avg	BoolQ Acc.	CB F1/acc.	COPA Acc	MultiRC F1a/EM	RTE Acc.	WSC Acc.
Slovene	CroSloEngual	59.8	70.0	67.7/74.7	59.4	58.4/15.6	51.7	58.2
	mBERT	60.2	73.0	66.5/71.9	51.6	57.5/17.0	62.1	58.9
	XML-R	59.9	63.3	69.9/74.7	52.8	58.8/18.8	58.6	61.0
English	CroSloEngual	59.9	63.3	67.3/75.5	**62.4**	59.5/16.7	55.2	57.5
	mBERT	**64.2**	**76.7**	69.9/74.7	58.6	**60.4/21.5**	**65.5**	63.0
	XML-R	61.4	70.0	**74.1/79.9**	51.8	60.1/19.4	48.3	**65.8**
	Most frequent	52.4	63.3	23.0/52.7	50	77.3/0.3	58.6	65.8

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

6 Ocena uspešnosti in prioritete za nadaljnji razvoj

Pripravili smo slovenski prevod zbirke nalog za razumevanje naravnega jezika SuperGLUE in ga objavili pod odprtokodno licenco (Žagar et al. 2021). Opisali smo postopek prevajanja in ovire pri prenosu v morfološko bogat slovenski jezik. Delno strojno in delno človeško prevedene nabore podatkov smo uporabili pri ocenjevanju štirih velikih vnaprej naučenih jezikovnih modelov na osnovi modela BERT, ki so na voljo za slovenščino. Rezultati kažejo, da je enojezični model SloBERTa trenutno najuspešnejši model za slovenščino. Uspešnost slovenskih modelov v primerjavi z najsodobnejšimi angleškimi modeli je vseeno še vedno bistveno slabša, kar kaže na precejšnje možnosti za izboljšanje pristopov NLP za jezike z manj viri.

Naše analize kažejo, da se je uspešnost modelov izboljšala z nabori človeško prevedenih podatkov, zato je treba v prihodnje povečati delež teh podatkov. Pri dveh angleških podatkovnih množicah prevod s strojnim prevajalnikom ni mogoč. Zato je treba ustvariti novo slovensko različico naloge WiC za razdvoumljanje besednih pomenov in ročno prilagoditi nalogo ReCoRD, da bomo pokrili celoten nabor nalog iz zbirke SuperGLUE. Pri nalogah WSC slovenska glagolska morfologija drugače kot v angleščini obravnava koreference. To vprašanje je potrebno analizirati in oblikovati zahtevnejšo slovensko množico WSC. Potencialna možnost za izboljšavo modelov je tudi izkoriščanje nekaterih prevedenih nalog iz zbirke SuperGLUE v druge jezike, npr. v ruščino. V nadaljnjih večjezikovnih in medjezikovnih poskusih bi bilo koristno vzpostaviti delotoke za prenos virov iz sorodnih, predvsem slovanskih jezikov, in za združitev in povečanje enakovrednih učnih množic.

Ker smo ohranili format izvornih nalog iz SuperGLUE, lahko slovenske rezultate ocenjujemo z izvorno lestvico SuperGLUE. To sicer omogoča primerjavo z angleškimi modeli, vendar dobljeni rezultati niso javno dostopni. Naloge iz zbirke SuperGLUE smo zato vključili v evalvacijski sistem SloBench (https://slobench.cjvt.si/), ki vsebuje ločeno slovensko lestvico. Na ta način spodbujamo skupnost NLP, da slovenščini nameni več pozornosti.

SupeGLUE je bila ob svoji predstavitvi najnaprednejša zbirka nalog za vrednotenje NLP-sistemov. Pred kratkim pa se je pojavila nova, še bistveno večja in bolj raznolika zbirka nalog s področja obdelave in razumevanja naravnega jezika, z imenom BIG-bench, ki vsebuje 204 naloge. V prihodnje bo potrebno za slovenščino prilagoditi vsaj del nalog, ki jih vsebuje BIG-bench, in jih vključiti v SloBench.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

7 Literatura

Conneau et al. 2018 = Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jegou, H. (2018). Word translation without parallel data. In *Proceedings of International Conference on Learning Representation ICLR*.

Conneau et al. 2019 = Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Devlin et al. 2019 = Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *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.

Hu et al. 2020 = Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating crosslingual generalisation. In: *International Conference on Machine Learning*, str. 4411–4421.

Kasai et al. 2021 = Kasai, J., Sakaguchi, K., Bras, R. L., Dunagan, L., Morrison, J., Fabbri, A. R., Choi, Y., and Smith, N. A. (2021). Bidimensional leaderboards: Generate and evaluate language hand in hand. ArXiv preprint 2112.04139.

Liang et al. 2020 = Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., et al. (2020). Xglue: A new benchmark dataset for crosslingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, str. 6008–6018.

Phang et al. 2020 = Phang, J., Yeres, P., Swanson, J., Liu, H., Tenney, I. F., Htut, P. M., Vania, C., Wang, A., and Bowman, S. R. (2020). Jiant 2.0: A software toolkit for research on general-purpose text understanding models. http://jiant.info/.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Shavrina in Malykh 2021 = Shavrina, T. in Malykh, V. (2021). How not to lie with a benchmark: Rearranging NLP leaderboards. ArXiv preprint 2112.01342.

Shavrina et al. 2020 = Shavrina, T., Fenogenova, A., Anton, E., Shevelev, D., Artemova, E., Malykh, V., Mikhailov, V., Tikhonova, M., Chertok, A., and Evlampiev, A. (2020). Russian SuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, str. 4717–4726.

Ulčar in Robnik-Šikonja 2021 = Ulčar, M. in Robnik-Šikonja, M. (2021). SloBERTa: Slovene monolingual large pretrained masked language model. In *Proceedings of Data Mining and Data Warehousing, SiKDD*.

Ulčar in Robnik-Šikonja 2020 = Ulčar, M. in Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Proceedings of Text, Speech, and Dialogue, TSD 2020*, str. 104–111.

Wang et al. 2019a = Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019a). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

Wang et al. 2019b = Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019b). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In International Conference on Learning Representations.Language Resource References.

Žagar et al. 2020 = Žagar, Aleš; Robnik-Šikonja, Marko; Goli, Teja and Arhar Holdt, Špela, (2020). Slovene translation of SuperGLUE, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1380.

Žagar in Robnik-Šikonja 2022 = Žagar, A. in Robnik-Šikonja, M. (2022) Slovene SuperGLUE benchmark: translation and evaluation. In CALZOLARI, Nicoletta (ur.). LREC 2022: Language Resources and Evaluation Conference, 2022. str. 2058-2065.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.