Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/16BE39Cvejic

PREPOZNAVANJE IMENOVANIH ENTITETA U SRPSKOM JEZIKU POMOĆU TRANSFORMER ARHITEKTURE

NAMED ENTITY RECOGNITION FOR SERBIAN LANGUAGE WITH TRANSFORMER ARHITECTURE

Andrija Cvejić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Za treniranje neuronskih mreži za obradu prirodnog jezika već postoje ustaljeni šabloni i principi isprobani nad engleskim jezikom. Prirodni sled događaja jeste istraživanje i razvijanje oblasti za druge jezike. U ovom radu predstavljena je arhitektura modela za prepoznavanje imenovanih entiteta u srpskom jeziku. Ulaz model je prirodno pisan jezik. Istrenirani model daje kao izlaz verovatnoće pripadnosti reči imenovanoj kategoriji. Predloženi su koraci za poboljšanje i dalji razvoj oblasti.*

Ključne reči: *NLP, NER, Obrada prirodnog jezika, Prepoznavanje imenovanih entiteta, BERT, RoBERTa*

Abstract – *When training neural networks for natural language processing, there are already common methods and practises tried and validated on the English language. Current research is the natural sequence of development and is focused on improving models for non English languages. In this paper, we present a model architecture used for named entity recognition of the Serbian language. The model's input is natural text. The trained model's outputs are category probability of each word for each named entity.*

Keywords: *NLP, NER, natural language processing, named entity recognition, BERT, RoBERTa*

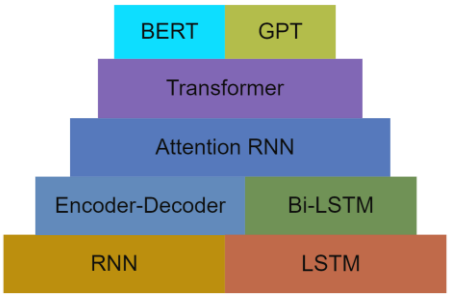
1. UVOD

Znatan deo komunikacije u savremenom društvu odvija se putem interneta, u obliku tekstualnog, video ili audio zapisa. Zahvaljujući tome, bitna odlika modernog sveta postala je povećana količina informacija, koje se prenose preko interneta.

Zbog toga javljaju se prvi pokušaji da se računarske tehnologije upotrebe u svrhe obrade teksta. Grana nauke koja se bavi takvim i sličnim problemima naziva se obrada prirodnog jezika (eng. *Natural Language Processing - NLP*).

Postoje različite tehnike za pronalaženje imenovanih entiteta; među prvima pojavile su se heurističke metode i ručno napravljena pravila (eng. *rule based*). Međutim, iakosu u izvesnim instancama pružale visoku tačnost, ove tehnike su bile preskupe i zahtevale su previše vremena za implementaciju, pri čemu i dalje nisu bile efikasne u opštim, već samo u pojedinačnim slučajevima.

Heurističke metode nasledili su kasnije statistički modeli, poput skrivenih Markovljevih lanaca (eng. *Hidden Markov Model – HMM*), koji su davali dobre rezultate, ali su zahtevali stručnjake koji bi mogli da izračunaju odgovarajuće verovatnoće za svaki jezik.



Slika 1. Hijerarhija nadogradnji NLP modela

Ovakvi pristupi imaju nekoliko ograničenja koja sprečavaju njihovo korišćenje u praksi. Iz tog razloga se prešlo na metode mašinskog učenja (eng. *Machine Learning - ML*) koje su se pokazale kao efikasnije u pronalaženju imenovanih entiteta [1]. Na slici 1. se može videti istorija napredka i nadogradnje za *ML* arhitekture. Do značajnog unapređenja u obradi prirodnog jezika dolazi sa pojavom Transformer arhitekture [2], koja se zasniva na pretpostavci da nam rekurentne neuronske mreže (RNN) nisu potrebne, već da nam je dovoljan samo aspect „pažnje“ (eng. attention). Ova nova arhitektura dala je izvanredne rezultate na planu prevođenja jezika.

U odnosu na RNN, ova arhitektura unapređuje razumevanje konteksta, ubrzava obučavanje i podržava rad sa neograničenim sekvencama, pri čemu očuvava međuzavisnost i poznavanje prethodne rečenice pomoću pažnje. To znači da se u ovom modelu može pozivati na reči u dalekoj prošlosti ili uticati na odabir trenutne reči. Istorijska nadogradnja modela može se videti na slici 1.

S obzirom na znatan napredak u oblasti obrade prirodnog jezika, pristup koji je korišćen u radu zasniva se na arhitekturi sa trenutno najboljim rešenjima kao *BERT* (eng. *Bidirectional Encoder Representations from Transformers - BERT*) [3] arhitekturi za reprezentaciju jezika [4], s tim da su uzeta u obzir i sitna unapređenja osnovnog modela, koja daju nešto bolje rezultate kao RoBERTa (*Robustly Optimized BERT Pretraining Approach*) [4]. Pored toga, predstavljena rešenja u ovome radu proizilaze iz očiglednog nedostatka obučenih modela za obradu imenovanih entiteta u srpskom jeziku [5]. Odabrana je HuggingFace infrastrukture za veb pristup modelima i

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Milan Segedinac, vanr. prof.

skupovima podataka (eng. dataset) iz navedenih razloga, u radu je korišćena BERT arhitektura, primenjena na srpski jezik, pomoću biblioteke HuggingFace specijalizovane za NLP [6][14]. Takođe se koriste algoritmi za tokenizaciju koji vrši transformaciju prosleđenih reči u vektore brojeva, konkretno o tokenizaciji pod-reči zvanoj šifrovanja bajtparova (eng. Byte Pair Encoding – BPE). Za reprezentaciju teksta u vektorskom prostoru, u radu je korišćena kontekstualna RoBERTa arhitektura.

Evaluacija prepoznavanje imenovanih entiteta izvršena je nad skupom podataka HR500k, koji je predstavljen kod [7] i korišćen kod [8], sa uporednom analizom performansi primenjenog BERT modela u istoj studiji. Pored toga, u ovom radu se primenjuje nadogradnja BERT modela zvana RoBERTa predstavljena kod [4].

U narednom poglavlju detaljnije je predstavljena postojeća literatura, korišćene tehnike opisane su u poglavlju 3, a poglavlje 4 sadrži analizu dobijenih rezultata i mogućih poboljšanja. Poglavlje 5 predstavlja sumarizaciju celog rada.

2. PRETHODNA REŠENJA

U ovom poglavlju biće predstavljeni radovi koji su se bavili prepoznavanjem imenovanih entiteta u prirodnom jeziku, sa akcentom na primenu u srpskom jeziku. U tom smislu, ukazaćemo na skupove podataka i tehnike koje su korišćene, kao i na njihove rezultate dobijene evaluacijom. Pored toga, biće objašnjen način na koji naš rad izvlači inspiraciju iz navedenih radova. Opisani radovi su dati hronološkim redosledom, gde je poseban akcenat stavljen na studiju koji su sačinili [9], a koja pružadobre smernice za celo polje i uputstva za rešavanje ovog problema. Takođe, od velikog značaja za našu problematiku su istraživanja koja su sproveli [11], zatim [10], kao i [12], rešavajući slične probleme u domenima sopstvenih jezika. Poseban značaj imaju studije koje su sačinili [13][8], jer su se autori bavili prepoznavanjem imenovanih entiteta u srpskom jeziku.

U radu [6] prezentuju HuggingFace biblioteku i platformu, čija je glavna prednost mogućnost deljenja obučenih modela i skupova podataka široj javnosti. Zbog toga, HuggingFace je postalo jedno od najpopularnijih okruženja u razvoju modela za obradu prirodnog jezika, kao i za mnoge specifične zadatke, poput prepoznavanja imenovanih entiteta. *HuggingFace* se prvenstveno bavi *Transformer* arhitekturama.

Međutim, dok je u ovoj oblasti zabeležen znatan napredak po pitanju dostupnosti obučenih modela, to ipak nije ostao slučaj za jezike koji nisu engleski, kako su pokazali Wolf, Debut, Sanh et al [11]. U svom radu autori obučavaju *BERT* model nazvan *CamemBERT*, primenjen na francuski jezik. Kao model, odabrana je nadograđena arhitektura za obučavanje, RoBERTa, uz pomoć HuggingFace biblioteke. Autori porede rezultate obučavanjana malom skupu podataka – primenjenog samo na francuski jezik, to jest jednojezični model (eng. mono-lingual model) – sa M-BERT (eng. BERT-base-multilingual-cased) više jezičnim modelom (eng. Multi-language model) obučavanja na većem skupu podataka. Zaključak ovoga rada sastoji se u tome da jednojezični model daje bolje rezultate od višejezičnog. Takođe, autori pokazuju da podaci preuzeti sa internet stranica daju bolje rezultate u odnosu na Wikipedia skupove podataka, i to sa unapređenjem od 1.5% na planu preciznosti.

Takođe rad [12] poredi višejezične modele M-BERT sa jednojezičnim modelima, u njihovom slučaju na nemačkom, švedskom, finskom, danskom i norveškom. Autori poređenje vrše na rasponu više zadataka, od generisanja teksta do specifičnih problema kao što je NER. Njihov rad takođe pokazujeda jednojezični modeli znatno bolje generišu tekst od M-BERT-a. Štaviše, autori smatraju da je M-BERT praktično beskoristan za generisanje teksta, posmatrano sa aspekta gramatičke i sintaksičke ispravnosti generisanih reči. Na pojedinačnom slučaju, NER performanse su bolje oko jedan do dva procenta kod jednojezičnih modela. Takođe, autori pokazuju da predikcija reči kod M-BERT jezičkog modela ima poteškoća sa skandinavskim jezicima zbog njihove kompleksnosti (morfološkog bogatstva) i manjeg skupa podataka za obučavanje.

U radu [8] su radu izdvojili tri ključna koraka za implementaciju svog rešenja: prvi se odnosi na odabir *BERT* modela; drugi na samo obučavanje modela da razume kontekst i gramatiku srpskog jezika, što je izvršeno na velikom skupu neobeleženih podataka; a treći na obučavanje obrade imenovanih entiteta (*NER*). Za *BERT* model autori koriste velike skupove podataka preuzete sa interneta: *srWaC[15]*, *hrWaC[15]*, *bsWaC[15]*, *cc100-hr[16] i cc100-sr[16]*. . Za obučavanje *NER* koristi se skup podataka *HR500k*. Tačnost koju su autori postigli u F1 meri varira od 0.90 do 0.96, u zavisnosti od test-skupa. Time su prikazani odlični rezultati u obradi imenovanih entiteta na srpskom jeziku, gde postoje mali skupovi podataka zbog *Transformer* arhitekture. Autori svoje rešenje porede sa drugim *M-BERT* modelom istreniranima na višejezičnim skupovima podataka.

3. METOD

U ovom poglavlju biće detaljno objašnjena postavka implementacije sistema za obučavanje modela na srpskom jeziku, kao i konkretan zadatak modela vezan za pronalaženje imenovanih entiteta

U prvom delu opisana je arhitektura rešenja. Drugi deo detaljno sagledava skupove podataka. Treći deo razjašnjava proces pretprocesiranja skupova podataka. Četvrti deo govori o obučavanju modela. Peti deo razmatra pitanja preuzimanja i ponovnog korišćenja obučenih modela u ovom radu.

3.1. Skup podataka

Skup podataka je podeljen u dva dela – (1) skup podataka za treniranje jezičkog modela i (2) skup podataka za prepoznavanje imenovanih entiteta.

Za treniranje jezičkog modela jezički model RoBERTa preuzeta su šest javno dostupna skupa podataka — Leipzig, OSCAR, srWac, hrWac, cc100-hr i cc100-sr. Prvi preuzeti srpsko-hrvatski Leipzig skup podataka je na latiničnom pismu, a podaci dolaze sa sajta Wikipedia, kao i sa .rs i .hr domena iz 2014. godine.

Drugi skup podataka je zvan OSCAR-sr, koji obuhvata samo srpski jezikna ćiriličnom pismu, dalje u radu obeležen kao OSCAR. On predstavlja prečišćenu verziju podataka preuzetih sa svih .rs domena od strane Common Crawl skupa podataka.

Treći skup podataka srWac, većinom je na latinici (16.7% ćirilica), i preuzet je sa .ba, .rs i .hr domena iz 2014. godine. Ovaj skup podataka izvorno ima veličinu 17 GB, međutim, nakon pretprocesiranja, sveden je na 3 GB.

Četvrti skup podataka hrWac, potpuno je na latinici, i preuzet je sa .hr domena iz 2014. godine. Ovaj skup podataka nakon pretprocesiranja je sveden na 6 GB.

Peti skup podataka cc100-sr, preuzet i prečišćen podaci iz 2018. godine sa Common Crawl. Koji samo sadrži ćirilicu.

Šesti skup podataka cc100-hr, preuzet i prečišćen podaci iz 2018. godine sa Common Crawl. Ovaj skup podataka je samo na latinici.

Ukupna veličina skupa podataka iznosi 28.848 milijarde tokena i 17.19 miliona rečenica, odnosno oko 41.5 GB podataka. Distibucija skupova podataka se može videti u tabeli 3.1.

Skup podataka za NER nazvan je HR500k[7] . On je spoj više skupova podataka – srpskog i hrvatskog skupa

Skup podataka	Prosečna dužina rečenice	broj tokena	broj rečenica
OSCAR	98	220 mil.	4 mil.
Leipzig	75	18 mil.	0.19 mil.
srWaC	38	490 mil.	13 mil.
hrWaC	21	1.4 mlrd.	67 mil.
cc100-sr	152	5.42 mlrd.	36 mil.
cc100-hr	148	21.3 mlrd.	143 mil.

Tabela 3.1 Skupovi podataka

podataka SETimes.RS i SETimes.HR. Skup sadrži 24800 reči i 500 hiljada tokena.

Obeležavanje imenovanih entiteta je urađeno prema IOB [17] (eng. Inside-Outside-Beginning) šemi označavanja. Opšti oblik takvog označavanja izgleda kao "prefiks-oznaka". Prefiks "I" u oznaci označava da se radi ocelini imenovanog entiteta, dok "B" prefiks indikuje da je reč o početku celine imenovanog entiteta. "B" prefiks je korišćen samo kada ga prati ista oznaka bez pojavljivanja "O" oznake između. Oznaka "O" označava da reč nije deo celine imenovanih entiteta. Ovaj način označavanja poznat je i pod nazivom BIO (eng. Beginning-Inside-Outside). Podržane oznake u korpusu su Osoba (PER), Organizacija (ORG), Lokacija (LOC) i Ostalo (MISC). Takvih oznaka u skupu podataka ima 11:'I-org', 'B-misc', 'B-per', 'B-deriv-per', 'B-org', 'B-loc', 'I-deriv-per', 'I-misc', 'I-loc', 'I-per' i 'O'. Oznaka koja zahteva dodatno objašnjenje je "deriv-per" - označava izvedenicu od imena osobe. U Tabeli 3.2 je distribucija skupa podataka za imenovane entitete HR500k prema kategorijama. U Tabeli 3.3 prikazana je distribucija rečenica i tokena za obuku, validaciju i testiranje.

Imenovani entitet	Broj tokena	Procenat skupa	Oznaka za kategoriju
Osoba	10,241	2.02%	PER
Izvedenica od osobe	319	0.06%	DERIV-PER
Lokacija	7,445	1.47%	LOC
Ogranizacija	11,216	2.21%	ORG
Razno	7,514	1.48%	MISC
Ukupno	36,735	7.25%	

Tabela 3.2 Distribucija imenovanih entiteta u skupu podataka *hr500k*

Veličina validacionog skupa iznosi 10%, a test-skupa 15% od obučavajućeg skupa, pri čemu svaka podela sadrži klase koje se prepoznaju za imenovane entitete

	Obučavanje	Validacija	Testiranje
Rečenica	20K	2K	2.7K
Tokena	550K	52K	68K

Tabela 3.3 Distribucija rečenica i tokena na skupovima podataka za obučavanje, validaciju i testiranje

3.2.Arhitektura

Arhitektura sistema je podeljena u dva dela – (1) obučavanje jezičkog modela za srpski jezik RoBERTa, u daljem toku rada nazvanim SRoBERTa, i (2) obučavanje glave modela za prepoznavanje imenovanih entiteta, u daljem toku rada nazvanim SRoBERTa-NER.

Jezički model poseduje RoBERTa arhitekturu, pa se, iz tog razloga, nećemo ponovo osvrtati na njene opšte odlike, već ćemo samo ukazati na parametre značajni za naš model, kao što su broj slojeva, broj glava za pažnju.

	M-BERT	SB-L	SB-base
Jezik	104 jezika	Sr + Hr	Srpski
Skup pod.	Wikipedia	WOL*	OL**
Obučavanja	MLM+NSP	MLM	MLM
Tokenizacija	WordPiece	BPE	BPE
Vokabular	110 hilj.	48 hilj.	48 hilj.
Dužina ulaza	512	514	514
Slojevi	12	6	6

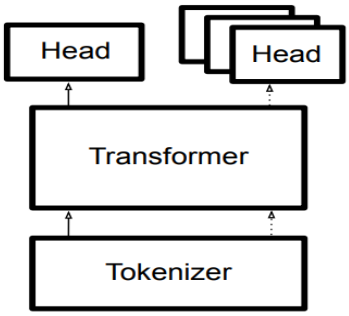
Tabela 3.4 Razlika između modela *M-BERT*, *SB-L* i *SB-base*

* WOL – podrazumeva spojena 3 izvora skupove podataka — srWac, OSCAR, Leipzig [20].

** OL- podrazumeva spojena 2 izvora skupa podataka — Oscar i Leipzig

Sam RoBERTa model je bitno uticao na odabir parametara, koji su dosta slični parametrima za BERT mrežu. Međutim, takvi parametri se i dalje znatno razlikuju od parametaramreže sa kojom vršimo poređenje, kakva je M-BERT. Sve razlike između parametara u navedenim modelima date su u Tabeli 3.4 (svi modeli SRoBERTa u tabeli su označeni sa SB). SB-XL ima istu konfiguraciju kao SB-L, ali koristi sve skupove podataka navedeni iz tabele 3.1. Takođe SB modeli se razlikuju po zadacima koje izvršavaju samo obučavanje modela maskiranog jezika (eng. Masked Language Model - MLM), dok BERT koristi i predikciju sledećih rečenica (eng. Next Sentence Prediction - NSP).

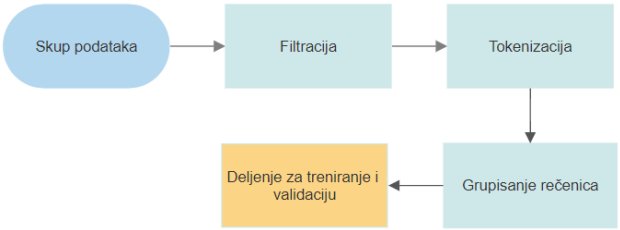
Naš model za prepoznavanje imenovanih entiteta je zasnovan na SRoBERTa modelu, gde se vrši dodavanje "glave" prema HuggingFace terminologiji, koja će da izvršava specifičan zadatak, kao što je ilustrovano na Slici 3.1. Ova glava je zadužena za NER i naziva se Token Classification, odnosno Roberta For Token Classification. Kao ulaz, model prima veličine embedovane dimenzionalnosti od E=768 vektora realnih brojeva. Glava NER modela se zasniva na jednom potpuno povezanom sloju koji za izlaz daje klase imenovanih entiteta. Klase imenovanih entitetase određuju pomoću softmax funkcije, koja utvrđuje najveću verovantnoću kojoj klasi dati entitet pripada.



Slika 3.1 Prikaz ponovne upotrebe jezičkog modela za više specifičnih zadataka [6]

3.3. Pretprocesiranje podataka za jezički model

Na Slici 3.2 predstavljeni su koraci u pretprocesiranju skupova podataka - filtracija, tokenizacija, grupisanje rečenica i deljenje za treniranje i validaciju.



Slika 3.2 Pretprocesiranje skupova podataka za jezički model

Prvi korak je podrazumevao filtraciju podataka, u ovom koraku ulaz je predstavljao originalan skup podataka, a izlaz su činile filtrirane rečenice u razdvojenim redovima. Filtracija podrazumeva izbacivanjem suvišnjih informacija (veb tagove i enkodovane veb tagove) i transformaciju u odgovarajući format. Drugi korak, tokenizacija, predstavlja proces prebacivanja podaka iz rečenica u nizove tokenizovanih informacija — niz tokena, niz ulazne pažnje i niz segmenata. Niz tokena predstavlja prebacivanje reči u niz tokena podreči dužine t; svaki token podreči je ceo broj. Reč će biti jedan token u slučaju da cela reč postoji u vokabularu. U slučaju da cela reč ne postoji u vokabularu, ona postaje niz tokena koji čine tu reč. Niz ulazne pažnje je neophodan da bi se token popunjavanja "[PAD]" razlikovao u odnosu na druge tokene prilikom ulaza, pri čemu evaluacija modela može da zanemari izlaz ovih tokena. Niz segmenata predstavlja niz celih brojeva, a vrednost se dodeljuje u odnosu na to kojoj rečenici token pripada. Na primer, ako token pripada prvoj rečenici, onda će niz na t poziciji imati vrednost 1.

Treći korak podrazumevao je grupisanje rečenica, što je bilo neophodno iz tri razloga. Prvo, da bi ulaz odgovarao predefinisanoj veličini ulaza u model, jer u prethodnom koraku tokenizacije nije vršeno odsecanje preko dužine ulaza niti popunjavanje do dužine ulaza. Drugo, uvidi do kojih su došli Kosec, Fu i Krell [18], pokazuju da je moguće udvostručiti brzinu obučavanja ukoliko se ne troši snaga računanja na „[PAD]" token. Takođe, Sun, Qiu, Xu et al. [19] su pokazali da se bolji rezultati mogu dobiti ako se pri klasfikaciji teksta ne vrši odsecanje na kraja, već je bolje sredinu. Autori su takođe pokazali da tekst pri kraju ima veći značaj po odlučivanje modela. Treće, da bi se moglo propustiti više reči u procesu obučavanja, pošto skupovi podataka sadrže i veoma duge rečenice.

Skup podataka hr500k ima 3 koraka za pretprocesiranje. Prvi korak se odnosi na transformaciju iz conllu formata u csv format, pri čemu se čuvaju informacije o rečenicama, njihovim odgovarajućim imenovanim entitetima i POS oznakama. Drugi se odnosi na tokenizaciju reči u skupu podataka. On podrazumeva da svi tokeni (pošto reči mogu da postanu više tokena) dobiju istu oznaku NER iz skupa podataka.

3.3. Obučavanje

U predloženom rešenju obučavana su četiri RoBERTa modela — *SRoBERTa* , *SRoBERTa-base* , *SRoBERTa-L* i *SRoBERTa-XL*.

Na Tabeli 3.5 su prikazani različiti parametri obučavanaja jezičkog modela.Najbolji *m*odel *SB-XL* je treniran 135h

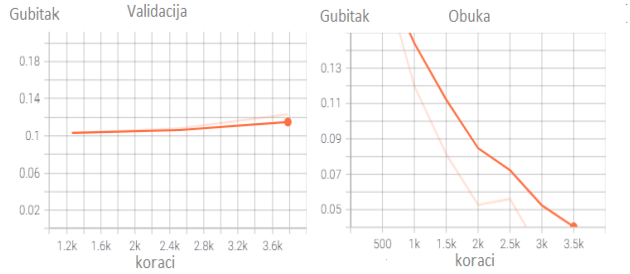
	SB	SB-base	SB-L	SB-XL
Batch size	94	180	48	50
Vreme	12 h	40h	30h	135h
Epoha	3	2	2	1
Skup Podataka	Leipzig	OL	WOL	WOLX
Broj parametra	120mil.	80mil.	80 mil.	80 mil.
Gubitak	3.2	2.9	2.6	2.2

Tabela 3.5 Razlike između obučavanih modela

na RTX 3070 sa *batch size* 50 sa gubitkom 2.2. U Tabeli 3.5 su prikazani različito obučavanje SB modela. Za ovo obučavanje su korišćeni svih 6 skupova podataka iz tabele 3.1. zajedno zvani WOLX.

U sklopu predloženog rešenja obučavana su četiri *NER* modela — *SB-NER*, *SB-base-NER*, *SB-L-NER*, *SB-XL-NER* [1] i *M-BERT-NER*. Svi modeli su obučavani na odgovarajućim *Transformer* modelima na istom skupu podataka.

Kada je reč o obučavanju modela za pojedinačni zadatak prepoznavanja imenovanih entiteta, situacija je nešto drugačija. Posle tri epohe gubitak na validacionom skupu uglavnom počinje da raste. Tada model kreće da se previše istrenira na skup podataka iz obučavajućeg seta (eng. Overfitting). Na Slici 5.1 možemo uočiti nizak obučavajući gubitak (desno), jer se kriva približava nuli, dok validacioni gubitak (levo) kreće polako da raste sa dužim obučavanjem. Dodatno obučavanje nije doprinelo boljim rezultatima.



Slika 3.3 Izgled obuke modela *SRoBERTa-L-NER*

4. REZULTATI I DISKUSIJA

U Tabeli 4.1 prikazani su rezultati obučavanja u predloženom rešenju za sve NER SB modele i M-BERT. Za evaluaciju rezultatasu korišćene prethodno opisane mere: f-mera, tačnost i odziv. F1 mera se povećava od *SRoBERTa-NER* do *M-BERT-NER* modela u rasponu od 0.66 do 0.86. Iz obučavanih modela se može primetit veliki uticaj povećanje skupa podataka sa Leipzig (18 mil. Tokena) na WOLX (28 mlrd.). Povećanje na SB-XL-NER sa jednom epohom treniranja postiže povećanu tačnost od 7%. Takođe, vidi se neophodna veličina skup podataka da se dostigne jednaka tačnost M-BERT-u. Tabele 4.2 i 4.3 prikazuju tačnost. Mada za dalje napredovanje, trebaju druge tehnike da se primene, kao što je povećavanje batch size,

Možemo zaključiti da su predložena rešenja u radu uspešno reprodukovana tačnost modela NER M-BERT čije je obučavanje opisano kod Ljubešića i Lauca [8] na hr500k skupu podataka. Međutim, naše rešenje nije uspelo poboljšati tačnost za pojedinačan zadatak sa jednojezičnim modelom, u meri u kojoj su to pomenuti autori postigli na drugim jezicima [10] [11][12].

Tokom treninga NER modela menjan je korak učenja (eng. learning rate), a najbolje rezultate ostvaruje korak između 1e-5 i 1.5e-5 za NER mrežu. Za evaluaciju NER modela korišćen je skup podataka koji su predstavili Ljubešić, Agić, Klubickaet al [7].

[1] https://huggingface.co/Andrija/SRoBERTa-XL-NER

Ime Modela	Preciznost	Odziv	F1 mera
SB-NER	0.65	0.67	0.66
SB-base-NER	0.72	0.73	0.73
SB-L-NER	0.79	0.80	0.80
SB-XL-NER	0.85	0.86	0.86
M-BERT-NER	0.87	0.86	0.86

Tabela 4.1 Rezultati metrika prilikom obučavanja NER modela

Imenovani entitet	SB	SB-base	SB-L	SB-XL	M-BERT
LOC	0.69	0.75	0.82	0.83	0.86
ORG	0.53	0.65	0.72	0.80	0.84
MISC	0.45	0.42	0.50	0.60	0.66
PER	0.61	0.72	0.80	0.85	0.86
D-PER	0.32	0.48	0.70	0.78	0.83
Ukupno	**0.56**	**0.66**	**0.73**	**0.80**	**0.83**

Tabela 4.2 F1 mera *NER* modela na test-delu skupa podataka

Imenovani entitet	SB	SB-base	SB-L	SB-XL	M-BERT
LOC	0.69	0.75	0.82	0.83	0.86
ORG	0.53	0.65	0.72	0.80	0.84
MISC	0.45	0.42	0.50	0.60	0.66
PER	0.61	0.72	0.80	0.85	0.86
D-PER	0.32	0.48	0.70	0.78	0.83
Ukupno	**0.56**	**0.66**	**0.73**	**0.80**	**0.83**

Tabela 4.3 F1 mera *NER* modela na test-delu skupa podataka

5. ZAKLJUČAK

U ovom radu predstavljen je model za prepoznavanje imenovanih entiteta u srpskom jeziku. Dobijeni rezultati nisu neposredno napredovali prepoznavanja imenovanih entiteta, ali daju zadovoljavajuće NER rezultate za jezički model u odnosu na M-BERT. Premda zabeležen je uspeh na planu razvoja jednog jezički modela u odnosu na M-BERT, koji bolje rešava problem skrivenih reči. U tom smislu, rezultati ovog rada nadovezuju se na rezultate koje su izveli Rönnqvist, Kanerva, Salakoski et al. [12],

tvrdeći da je višejezični model M-BERT praktično beskoristan za generisanje reči sa aspekta njihove semantičke i gramatičke ispravnosti.

U procesu izrade rešenja, identifikovano je nekoliko problema koji se javljaju u prepoznavanju imenovanih entiteta. Jedan od najvažnijih predstavlja nedovoljno radne memorije na grafičkim karticama, što se kod jezičkog modela manifestuje na tri plana - batch size, veličinu arhitekture (dubinu slojeva u mreži) i veličinu vokabulara. Smernice za dalje istraživanje svode se na povećanje vremena obuke jezičkog model, što je neophodno jer se tekstovi na srpskom jeziku javljaju na dva pisma, latinici i ćirlici, pa time zahtevaju više vremena za obuku kroz adekvatan skup podataka. Jednom kada se to postigne, biće moguće ostvariti bolju reprezentacija konteksta i samim time bolje rezultate NER zadatka. Istovremeno, ovo polazište potvrđuje pretpostavku da skupovi podataka srodnih jezika značajno doprinose izgradnji sveobuhvatnijeg konteksta. U daljem istraživanju ove smernice mogu biti korisne za opšti razvoj obrade prirodnog jezika, kao i, konretno, za razvoj jezičkih i NER modela.

6. LITERATURA

[1] Mansouri, A., Affendey, L. S., & Mamat, A. (2008). Named entity recognition approaches. International Journal of Computer Science and Network Security, 8(2), 339-344.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[5] Krstev, C., Vitas, D., Obradović, I., & Utvić, M. (2011, July). E-dictionaries and Finite-state automata for the recognition of named entities. In Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing (pp. 48-56).

[6] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). HuggingFace's Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

[7] Ljubešić, N., Agić, Ž., Klubicka, F., Batanović, V., & Erjavec, T. (2018). hr500k–A Reference Training Corpus of Croatian.

[8] Ljubešić, N., & Lauc, D. (2021, April). BERTić-The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing (pp. 37-42).

[9] Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavvaf, N., & Fox, E. A. (2020). Natural language processing advancements by deep learning: A survey. arXiv preprint arXiv:2003.01200.

[10] Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., ... & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076.

[11] Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de La Clergerie, É. V., ... & Sagot, B. (2019). Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894.

[12] Rönnqvist, S., Kanerva, J., Salakoski, T., & Ginter, F. (2019). Is multilingual BERT fluent in language generation?. arXiv preprint arXiv:1910.03806.

[13] Šandrih, B., Krstev, C., & Stanković, R. (2019, September). Development and evaluation of three named entity recognition systems for serbian-the case of personal names. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019) (pp. 1060-1068).

[14] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). HuggingFace's Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

[15] Ljubešić, N., & Klubička, F. (2014, April). {bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian. In Proceedings of the 9th Web as Corpus Workshop (WaC-9) (pp. 29-35).

[16] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

[17] Ramshaw, L., & Marcus, M. P. (1995). Text chunking using transformation-based learn.

[18] Kosec, M., Fu, S., & Krell, M. M. (2021). Packing: Towards 2x NLP BERT Acceleration. arXiv preprint arXiv:2107.02027.

[19] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, October). How to fine-tune bert for text classification?. In China National Conference on Chinese Computational Linguistics (pp. 194-206). Springer, Cham.

[20] Goldhahn, D., Eckart, T., & Quasthoff, U. (2012, May). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In LREC (Vol. 29, pp. 31-43).

Kratka biografija:

Andrija Cvejić rođen je 1996. godine u Novom Sadu. Osnovne akademske studije završio je 2019. godine na Fakultetu tehničkih nauka, na kom brani i master rad 2021. godine iz oblasti Elektrotehnike i računarstva – Primenjene računarske nauke i informatika.
kontakt: andrijacvejic@gmail.com