Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/16BE40Cvejic

ANALIZA RAZLIČITIH MODELA ZA PREPOZNAVANJE IMENOVANIH ENTITETA NA SRPSKOM JEZIKU

COMPARISON OF DIFFERENT APPROACHES TO NAMED ENTITY RECOGNITION IN SERBIAN LANGUAGE

Aleksandar Cvejić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Zadatak ovog rada jeste analiza različitih pristupa za prepoznavanje imenovanih entiteta (Named Entity Recognition, NER) u srpskom jeziku. Rad poredi performanse Conditional Random Fields (CRF) modela i transformer modela na NER zadatku. Kao transformer modeli korišćeni su BERT, DistilBERT i Electra transformer modeli. CRF se direktno trenira za NER zadatku, dok se transformer modeli treniraju u 3 koraka: (1) treniranje tokenizera, (2) pretreniranje generalnog jezičkog modela i (3) dotreniranje na NER zadatku. U radu se prikazuju rezultati više konfiguracija CRF modela, treniranim na različitim karakteristikama i rezultati transformer modela.*

Ključne reči: *NLP, NER, Transformeri, CRF,* prepoznavanje tokena

Abstract – *This paper aims to analyze different approaches to Named Entity Recognition (NER) in the Serbian language. The paper compares the performance of the CRF model to newer transformer models on the NER task. Applied transformer models are BERT, DistilBERT, and Electra. In the performed experiments, multiple configurations of CRF, trained on different attributes, are compared to transformer models.*

Keywords: *NLP, NER, Transformers, CRF, token classfication*

1. UVOD

NLP (eng. *Natural Language Processing*) se kao oblast razvila zahvaljujući dostupnosti veće količine podataka, kao i unapređenim u tehnikama procesiranja teksta. Naročito se povećava dostupnost nestrukturiranih podataka, zbog toga što je sve veći broj sistema digitalizovan, i veći broj ljudi imaju pristup internetu.

Ovaj rad rešava problem prepoznavanja imenovanih entiteta (eng. *Named Entity Recognition* – NER). Glavni cilj NER je prepoznavanje reči (ili skupa reči) u tekstu koje predstavljaju imenovane entitete (osobe, lokacije, organizacije i druge). Iako ovaj zadatak samostalno nije od većeg komercijalnog značaja, on predstavlja važan deo širih sistema za obradu teksta. Jedno od mesta gde bi NER model mogao primeniti zajedno sa analizom sentimenta (eng. *sentiment analysis*) jeste indirektna predikcije berze [1].

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Jelena Slivka, vanr. prof.

Zadatak ovog rada jeste analiza različitih pristupa za NER u srpskom jeziku. Jedan prikazani pristup je primena CRF [2] (eng. *Conditional Random Fields*) modela koji se tradicionalno koristi za ovaj zadatak. Isprobano je više različitih konfiguracija CRF modela, gde svaka konfiguracija predstavlja drugačiju kombinacija karakteristika korišćenih za predstavljanje podataka. CRF se pokazao kao *state-of-the-art* [3] pri obradi jezika koji nemaju bogat skup podataka (eng. *dataset*), kao što je srpski jezik. Pored ovoga, za rešavanje NER zadatka korišćeni su i dotrenirani BERT [4], DistilBERT [5] i Electra [6] transformer modeli. CRF se direktno trenira za NER zadatak, dok se transformer modeli treniraju u tri koraka: treniranje tokenizera, pretreniranje generalnog jezičkog modela (eng. *pre-training*) i dotreniranje na NER zadatku. Modeli istrenirani u radu se koriste originalne parametre i načine obučavanja kao i originalni autori transformer modela [4]–[6] i CRF rada [3].

Modeli predstavljeni u ovom radu trenirani su na jednom neanotiranom skupu podataka *oscar* [7] i dva anotirana skupa podataka, *SETimes.SR* [8] i *wikiann* [9]. Poređenja se vrše na *SETimes.SR* skupu podataka zbog upotrebe istog skupa podataka u drugim radovima [10], [11]. *Wikiann* je korišćen kao dodatni skup podataka za transformer modele, radi provere hipoteze o količini potrebnih podataka za treniranje ovako ogromnih modela.

Transformer modeli implementirani u ovom radu su javno dostupni, zajedno sa tokenizerima, generalnim jezičkim modelima i specijalizovanim modelima za NER na Huggingface [12] platformi[1].

U poglavlju 2 predstavljena je postojeća literatura, koja se bavi NER problemom. U poglavlju 3 su predstavljeni korišćeni podaci i treniranje modela. U poglavlju 4 su predstavljeni rezultati i diskusija. Poglavlje 5 predstavlja zaključak rada.

2. RELEVANTNA LITERATURA

U ovom poglavlju biće predstavljeni radovi u kojima se koriste statističke modele mašinskog učenja i moderne transformer modele za rešavanja NER problema. Biće predstavljena tačnost modela, kao i skupovi podataka nad kojima su modeli trenirani. Cilj je da se stekne okvirna predstava o oblasti, jer direktno poređenje nije moguće na različitim skupovima podataka. Ovaj rad je baziran na dve grupe studija: one koje se bave standardnim metodama za NER [3], [13], i one koje se bave transformer modelima za NER [11], [14]. Pored ovih radova, koriste se tehnike

[1] https://huggingface.co/Aleksandar

treniranja opisane u originalnim radovima o transformerima [15].

Pristup korišćen u ovom radu za odabir skupa svojstava (eng. *feature set*) je sličan pristupu u radu [3], a takođe se razmatra i isti pristup uključivanja morfoloških svojstava reči (eng. *parts of speech tags*) kod CRF modela. Autori rada [3] razmatraju NER performanse pomoću CRF modela nad češkim, kao i nad drugim jezicima. Pored toga, autori razmatraju prednost predloženog CRF modela ukoliko se koriste nadklase kod imenovanih entiteta. Kako autori u rada [3] primećuju, dodavanje morfoloških svojstava nema velikog uticaja na performanse NER zadatka, a zahteva dodatne informacije od modela. Postignuta f1 mera rada za češki jezik po *CoNLL* standardu za evaluaciju iznosi 74.08% nad malim i ograničenim češkim korpusom. Slična je situacija sa obeženim podacima za srpski jezik i dosta jezika (mali broj jezika je u povoljnoj grupi da ima pristup velikoj količini labeliranih podataka) [16].

Transformer rad [11] koji ćemo razmatrati jeste i najbitniji, jer potiče od autora samog skupa podataka koji je korišćen za evaluaciju NER modela u ovom radu. U radu se obučava *WordPiece tokenizer* (sa 32 hiljade tokena u rečniku), prvobitno predstavljen u upotrebi sa transformerima u *Google*-ovom sistemu [17]. Autori treniraju BERT transformer model (BERTić) na NER zadatku sa novim zadatkom pretreniranja generalnog jezičkog modela (korak pre dotreniranja na NER zadatak). Pristup treniranja generalnog jezičkog modela koji koriste autori rada [11] će biti primenjen i u ovom radu.

U radu [11] je prvi put uvedeno korišćenje *Electra* transformera [6]. Autori objašnjavaju da je razlog za odabir ovog načina predtreniranja (treniranje generalnog jezičkog modela – eng. *pre-training*) sadržan u boljim rezultatima po pitanju brzine konvergencije i manjoj količini resursa zahtevanih prilikom obučavanja samog transformera. Pretreniranje se postiže učenjem drugog zadatka, gde je cilj da transformer odgovori da li reč pripada originalnom tekstu ili je generisana pomoću generatora. Autori biraju ovaj pristup zbog toga što koriste manji skup podataka od ukupno 8.4 milijardi tokena (reči i ostali karakteri), koji je dosta manji od originalnog BERT modela na engleskom jeziku sa 2500 milijardi tokena. Promenjen način obučavanja se opisuje u 3.4 potpoglavlju. Ukupan broj tokena u skupu podataka korišćenim za treniranje u radu se dobija kombinacijom više jezika i više skupova podataka za jedan jezik. Autori postižu rezultate od 92.02% u f1 meri, nakon pet epoha treniranja na *SETimes.SR* i 87.92% u f1 meri na *ReLDI-sr* [18] skupu podataka. Drugi skup podataka je javno dostupan, ali bez NER tagova u trenutku izrade rada, zbog čega nije uzet u obzir u ovom radu[2].

Rešenje u ovom radu se razlikuje od dosadašnjih radova jer ograničava skup podataka za učenje transformera na jedno-jezičke modele, bez inicijalizacije modela na engleskom jeziku, radi čistog poređenja trenutnog stanja jedno-jezičkog modela na srpskom jeziku. Transformer modeli se porede sa 23 različite konfiguracije CRF modela.

[2] Nakon izrade rada, NER postale su dostupne anotacije na https://huggingface.co/datasets/classla/reldi_sr

3. METODOLOGIJA

U narednim poglavljima izloženi su skup podataka, tokenizer, arhitekture modela i trening modela.

3.1. Skup podataka

U radu se koriste 3 skupa podataka, neanotirani *oscar* [7] skup podataka i dva labelirana skupa podataka, *SETimes.SR* [8] i *wikiann* [9]. *Oscar* sadrži 645,747 jedinstvenih rečenica sa ukupno 207.5 miliona jedinstvenih tokena (*uhshuffled_deduplicated_sr*). *SETimes.SR* skup podataka sadrži 86,726 manualno anotiranih tokena, dok ima 3177 rečenica u trening skupu, 395 rečenica u validacionom skupu i 319 u testom skupu. *Wikiann* skup podataka sadrži 30 hiljada u trening skupu i po 10 hiljada za validacioni i testni skup podataka.

Anotirani skupovi podataka poštuju unutra-spolja-početak (eng. *inside-outside-beginning* – IOB) princip anotiranja podataka. Moguće klase kod *SETimes.SR*, sa obe varijante početak (B) i unutra (I), su: Osoba, Organizacija, Lokacija, Ostalo (eng. *misc*), izvedena osoba (eng. *deriv-per*) i klasa izvan skupa tagova (*O*). *Wikiann* skup podataka ima smanjeni broj klasa u odnosu na *SETimes.SR*, nema *B-misc*, *I-misc* i *B-deriv-per*.

Treba napomenuti da su svi skupovi podataka korišćeni u ovom radu samo na sprskom jeziku. Generalno se očekuje da kombinovanje sličnih jezika dovodi do unapređenja performansi [16].

3.2. WordPiece tokenizer

Sva tri transformer modela proučena u ovom radu koriste *WordPiece* način reprezentacije podataka. Ovaj algoritam je predstavljen u radu [19], ali je suštinski veoma sličan GPT-2 upotrebi sa *Byte-Pair Encoding* (BPE).

WordPiece se inicijalizuje tako što prvobitno uključi sve prisutne karaktere u rečnik (vokabular) u trening skupu (u ovom slučaju *oscar*), a potom vremenom uči delove reči pomoću pravila spajanja prethodno postojećih delova. Ideja je da se izabere *WordPiece* model koji rezultuje minimalnim brojem delova reči (eng. *word pieces*) u celom trening skupu (uz dodatne specijalne simbole). U odnosu na BPE, *WordPiece* tokenizer povećava verovatnoću pojave podataka nakon što su dodati u rečnik, umesto frekvencije reči, koja se koristi u BPE-u. Ovo je ekvivalentno pronalaženju para simbola, čija je verovatnoća veća u spojenoj formi u poređenju sa zbirom individualnih verovatnoća. Ovo znači da algoritam proverava da li vredi spojiti dva simbola (npr. „a“ i „b“), na osnovu računanja gubitaka koji nastaju spajanjem („ab“). Pristup je sličan *Unigram* principu odlučivanja modela koje tokene treba da spoji, ali razlika je u tome što je *WordPiece* „pohlepan“ (eng. *greedy*) algoritam, koji pokušava da nađe najbolje parove za spajanje u svakom koraku iteracije.

3.3. Tradicionalni model - CRF algoritam

CRF algoritam je baziran na neusmerenim grafičkim modelima. Jednostavni CRF lanci su kao *state-of-the-art* za NER zadatak među statističkim modelima mašinskog učenja. Mogu se koristiti i kao poslednji sloj transformera [14], [20]. CRF je definisan i analiziran u [2], [3], [13], koje treba konsultovati za više detalja.

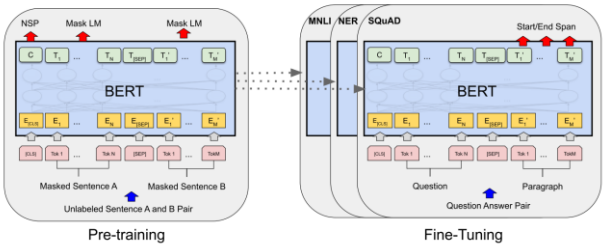
U ovom radu će se trenirati kombinacija različitih karakteristika po uzoru na [21], poput *word.lower()*, *word[-3:]*, *word.isupper()*, itd.

3.4. Transformer model – BERT

Bidirectional Encoder Representations from Transformers, odnosno BERT [4] je transformer model za modelovanje reprezentacije jezika. Postoje dva koraka u BERT implementaciji opisana u radu koji uvodi model: predtreniranje (eng. *pre-traning*) i dotreniranje (eng. *fine-tuning)*, kao što je prikazano na Slici 1.

Prvi korak podrazumeva treniranje modela uz pomoć nenadgledanog učenja na većim skupovima podataka. Ovo podrazumeva treniranje *WordPeice* tokenizera sa rečnikom od 30 000 tokena, nakon čega se trenira generalni jezički model na istom skupu nelabeliranih podataka.

Na kraju se koristi jedan od anotiranih skupova podataka da se istrenira na NER zadatku.



Slika 1. Prikaz dva koraka prisutna kod BERT transformera. Levo: prvi korak nenadgledanog obučavanja; desno: dotreniranje na specifičnom zadatku. Preuzeto iz [4].

Prilikom treniranje generalnog jezičkog modela pomoću nendagledanog učenja koristi se maskirani pristup, gde se umesto delova ulazne sekvence koristi [MASK] token za koji model treba da pretpostavi koji je token treba da bude.

3.5. Transformer model - DistilBERT

DistilBERT [5] transformer kao glavnu ideju koristi destilaciju znanja (eng. *knowledge distillation* [22]). Ovo podrazumeva specijalni režim obučavanja kompaktnog modela (studenta) u odnosu na veći (učitelj) model. Manji model je naučen da reprodukuje ponašanje već istreniranog celog modela. Rad pokazuje da se sa 40% manje parametara može održati 97% performansi mereno BLUE metrikom evaluacije (u odnosu na BERT model). Ovo znači da DistilBERT ima ukupno 66 miliona parametra, u odnosu na 110 miliona u BERT modelu.

3.6. Transformer model – Electra

ELECTRA [6] („*Efficiently Learning an Encoder that Classifies Token Representations Accurately*“) je transformer model koji menja način obučavanja od maskiranog pristupa do novog zadatka, gde se pogađa da li je token zamenjen ili pripada originalnom korpusu. Ovo je postignuto tako što se koristi jedna mala pomoćna mreža (generator). Prilikom treniranja transformer modela, menja se deo tokena tako što se uzimaju uzorci iz generator modela, dok se sam generator trenira istovremeno dok treniramo glavni model.

4. REZULTATI I DISKUSIJA

Prilikom izrade rada istrenirano sa konačnim konfiguracijama ukupno tri generalna jezička modela i šest specifičnih NER modela za transformere i 23 konfiguracije CRF modela.

Tabela 1 prikazuje rezultate treniranja specifičnog NER zadatka u 20 epoha dotreniranja. Bolji rezultati većeg skupa podataka *wikiann* ukazuju na manjak labeliranih podataka kod *SETimes.SR* skupa podataka.

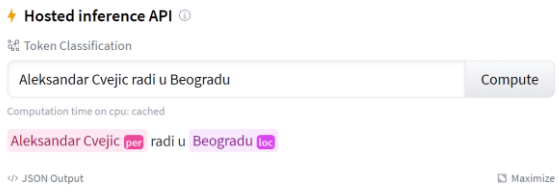
Model	**F1**	**Odziv**
BERT – *setimes.sr*	83.45%	84.65%
DistilBERT – *setimes.sr*	**84.92%**	**86.17%**
Electra – *setimes.sr*	83.46%	81.26%
BERT – *wikiann*	89.95%	90.82%
DistilBERT – *wikiann*	89.84%	**91.00%**
Electra – *wikiann*	**90.10%**	90.87%

Tabela 4.2 Rezultati dobijeni na dva skupa podataka sa različitim modelima.

Model	**Avg. F1**	**Makro F1**
BERT – *setimes.sr*	83.50%	-
DistilBERT – *setimes.sr*	84.44%	-
Electra – *setimes.sr*	81.13%	-
CRF – *setimes.sr* bez POS	84.2%	71.45%
CRF - *setimes.sr* bez isupper	88.58%	81.09%
BERTić [11] – *setimes.sr*	92.02%	-
CRF Janes [10] – *setimes*.sr	-	78.10%

Tabela 2 Rezultati dobijeni na dva skupa podataka sa različitim modelima.

Primer upotrebe modela je prikazan na slici 3. Rezultati konačnog CRF obučavanja, zajedno sa poređenjem sa rezultatima radova [10], [11] se mogu videti u tabeli 2. Najbolji CRF model je bolji od *Janes* modela [10], a lošiji od višejezičkog modela prikazanog u radu [11].



Slika 3 Primer upotrebe inference DistilBERT modela pomoću *HuggingFace API*-a, gde je zadatak pronalaženje imenovanih entiteta.

5. ZAKLJUČAK

U ovom radu predstavljeni su različiti modeli koji vrše prepoznavanje imenovanih entiteta (NER). Porede se tradicionalni pristupi NER problemu sa modernim transformer pristupima NER-u.

Rezultati dobijeni u radu imaju lošije performanse od *state-of-the-art* za engleski jezik i čak višejezičkog modela BERTić [11], ali treba imati u vidu količinu podataka korišćenu za treniranje. Razlike u transformer modelima su zanemarljive u odnosu na razlike između upotrebe većeg skupa podataka poput *wikiann*. Poznato je da veće mreže u NLP-u postižu bolje performanse, ali su potrebni podaci da se obuče takve velike mreže poput GPT-3.

6. LITERATURA

[1] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, 2009, doi: 10.1145/1462198.1462204.

[2] lafferty john, M. Andrew, and C. N. P. Fernando, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *ICML '01 Proc. Eighteenth Int. Conf. Mach. Learn.*, vol. 2001, no. June, pp. 282–289, 2001, doi: 10.29122/mipi.v11i1.2792.

[3] M. Konkol and M. Konopík, "CRF-based Czech named entity recognizer and consolidation of Czech NER research," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8082 LNAI, pp. 153–160, 2013, doi: 10.1007/978-3-642-40585-3_20.

[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[5] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," pp. 2–6, 2019, [Online]. Available: http://arxiv.org/abs/1910.01108.

[6] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," pp. 1–18, 2020, [Online]. Available: http://arxiv.org/abs/2003.10555.

[7] P. J. Ortiz Suárez, L. Romary, and B. Sagot, "A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages," pp. 1703–1714, 2020, doi: 10.18653/v1/2020.acl-main.156.

[8] V. Batanović, N. Ljubešić, T. Samardžić, and T. Erjavec, "Training corpus SETimes.SR 1.0." 2018, [Online]. Available: http://hdl.handle.net/11356/1200.

[9] A. Rahimi, Y. Li, and T. Cohn, "Massively multilingual transfer for NER," *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 151–164, 2020, doi: 10.18653/v1/p19-1015.

[10] D. Fišer, N. Ljubešić, and T. Erjavec, "The Janes project: language resources and tools for Slovene user generated content," *Lang. Resour. Eval.*, vol. 54, no. 1, pp. 223–246, 2020, doi: 10.1007/s10579-018-9425-z.

[11] N. Ljubešić and D. Lauc, "BERTić -- The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian," *Proc. 8th Work. Balto-Slavic Nat. Lang. Process. (BSNLP 2021)*, no. 1, pp. 37–42, 2021, [Online]. Available: https://www.aclweb.org/anthology/2021.bsnlp-1.5.pdf.

[12] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," pp. 38–45, 2020, doi: 10.18653/v1/2020.emnlp-demos.6.

[13] A. Cvejić, K. Grujić, A. Cvejić, M. Marković, and S. Gostojić, "Automatic Transformation of Plain-text Legislation into Machine-readable Format," *Proc. 11th Int. Conf. Inf. Soc. Technol.*, pp. 50–55, 2021.

[14] M. Arkhipov, M. Trofimova, Y. Kuratov, and A. Sorokin, "Tuning Multilingual Transformers for Language-Specific Named Entity Recognition," 2019, no. August, pp. 89–93, doi: 10.18653/v1/w19-3712.

[15] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[16] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," pp. 6282–6293, 2020, doi: 10.18653/v1/2020.acl-main.560.

[17] Y. Wu *et al.*, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," pp. 1–23, 2016, [Online]. Available: http://arxiv.org/abs/1609.08144.

[18] M. Miličević and N. Ljubešić, "Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets," *Slov. 2.0 empirical, Appl. Interdiscip. Res.*, vol. 4, no. 2, pp. 156–188, 2016, doi: 10.4312/slo2.0.2016.2.156-188.

[19] M. Schuster and K. Nakajima, "Japanese and Korean voice search," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 1, pp. 5149–5152, 2012, doi: 10.1109/ICASSP.2012.6289079.

[20] H. Yan, B. Deng, X. Li, and X. Qiu, "TENER: Adapting Transformer Encoder for Named Entity Recognition," 2019, [Online]. Available: http://arxiv.org/abs/1911.04474.

[21] M. Marcińczuk and M. Janicki, "Optimizing CRF-based model for proper name recognition in Polish texts," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7181 LNCS, no. PART 1, pp. 258–269, doi: 10.1007/978-3-642-28604-9_22.

[22] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," pp. 1–9, 2015, [Online]. Available: http://arxiv.org/abs/1503.02531.

Kratka biografija:

Aleksandar 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 – Računarstvo i automatika.
kontakt: cvejicaca@gmail.com