

Нови језички модели за српски језик

УДК 811.163.41'322.2

САЖЕТАК: У раду ће укратко бити приказан историјат развоја језичких модела за српски језик који су засновани на трансформерској архитектури. Биће представљено и неколико нових модела за генерисање и векторизацију текста, обучених на ресурсима Друштва за језичке ресурсе и технологије. Десет одабраних модела за векторизацију српског језика, међу којима су и два нова модела, биће упоређена на четири задатка обраде природног језика. Анализираћемо који су модели најбољи за изабране задатке, како величина модела и величина скупа за обучавање утичу на њихове перформансе на тим задацима и шта је потребно за обучавање најбољих модела за српски језик.

КЉУЧНЕ РЕЧИ: језички модели, српски језик, векторизација, обрада природног језика.

РАД ПРИМЉЕН: 27. јануар 2024.

РАД ПРИХВАЋЕН: 21. фебруар 2024.

Михаило Шкорић

mihailo.skoric@rgf.bg.ac.rs

ORCID: 0000-0003-4811-8692

Универзитет у Београду

Рударско-геолошки факултет
Београд, Србија

1. Увод

Почетком двадесет и првог века, дошло је најпре до наглог пораста количине доступних текстуалних података, а потом и до наглог раста рачунарске моћи, што је покренуло талас истраживања заснованих на идеји дубоког учења (*deep learning*) (LeCun, Bengio, and Hinton 2015). У случају обраде природних језика, истраживања кулминирају појавом архитектуре трансформера (Vaswani et al. 2017), заснованој на употреби енкодера, чија је главна намена анализа текста, и декодера, који су задужени за синтезу текста. Први изразито популаран модел овог типа био је *BERT*¹ (Devlin et al. 2018), заснован искључиво

1. *Bidirectional Encoder Representations from Transformers* – двосмерно кодирање репрезентације из трансформера

на трансформерском енкодеру. Овај модел направио је велики помак у обради природних језика, пре свега на задацима заснованим на векторизацији текста. Његове варијације, *RoBERTa*² (Liu et al. 2019) и *DeBERTa*³ (He et al. 2020) и данас постижу најбоље резултате на задацима угнежђивања речи (*word embedding*), анотације речи (нпр. обележавање врсте речи и препознавање именованих ентитета) и класификације реченица и докумената. Са друге стране, појављивање модела *GPT* (генеративни предобучени трансформер) (Radford et al. 2018) и *GPT-2* (Radford et al. 2019) популаризовало је језичке моделе засноване на траснформерском декодеру, а ова група модела се данас најбрже развија. Модели који комбинују употребу енкодера и декодера, као што су, на пример *BART* (Lewis et al. 2020) и *T5* (Raffel et al. 2020), остају недовољно запажени упркос изванредним резултатима које постижу на задацима трансформације текста, као што су машинско превођење, сумаризација и прилагођавање стила.

1.1 Преглед објављених модела за српски језик

Језички модели засновани на архитектури трансформера направили су продор у српски језик путем вишејезичних модела, најпре кроз *MBERT*⁴ (Devlin et al. 2018), а потом и кроз *XLM-RoBERTa* модел⁵ (Conneau et al. 2019), за чије је обучавање коришћено око четири милијарде токена из текстова писаних на српском или другом сродном језику (хрватски, босански). Потоњи модел објављен је у децембру 2019. године у две варијанте – *base* (279 милиона параметара) и *large* (561 милион параметара). И данас се, као један од највећих модела за векторизацију, употребљава у обради српског језика и притом остварује добре резултате, поготово након дообучавања.

Почетком 2021. године, на платформи *Huggingface*⁶ објављен је модел под називом *BEPTuћ* (*classla/bcms-bertic*) (Ljubešić and Lauc 2021), базиран на архитектури *ELECTRA* (Clark et al. 2020), са 110 милиона параметара, обучаван на корпусу од преко осам милијарди токена, босанског (800 милиона), хрватског (5,5 милијарди), црногорског (80 милиона) и српског језика (2 милијарде).

2. *Robustly Optimized BERT* – Робусно оптимизовани *BERT*

3. *Decoding-Enhanced BERT* – *BERT* проширен декодирањем

4. *Multilingual BERT* – вишејезични *BERT*

5. *Cross-lingual Language Model* – Међујезички језички модел

6. *Huggingface*, највеће веб чвориште за објављивање језичких модела.

Касније исте године направљени су и објављени први специфични модели за српски језик у оквиру једног ширег језичког истраживања за македонски језик (Dobrev et al. 2022). Прецизније, објављена је српска верзија *RoBERTa-base* модела, *macedonizer/sr-roberta-base* (120 милиона параметара) и српска верзија *GPT2-small* модела, *macedonizer/sr-gpt2* (130 милиона параметара). Оба модела су обучавана на корпусу српске Википедије и подржавају само ћирилично писмо.

Недуго потом предузет је сличан подухват, при којем је обучено пет *RoBERTa-base* модела за српски језик (Cvejić 2022). Првобитни модел *Andrija/SRoBERTa* имао је 120 милиона параметара и обучаван је на малом корпусу од 18 милиона токена познатом под именом *Leipzig* (Biemann et al. 2007), док су потоња четири модела имала по 80 милиона параметара, при чему је сваки обучаван на све већем корпусу. За модел *Andrija/SRoBERTa-base* додат је корпус *OSCAR* (Suárez, Sagot, and Romary 2019) (220 милиона токена), за модел *Andrija/SRoBERTa-L* је поред њега додат и *srWAc* (Ljubešić and Klubička 2014) (490 милиона токена), за модел *Andrija/SRoBERTa-XL* је уз претходне додат и део корпуса *cc100-hr* (21 милијарда токена) и *cc100-sr* (5,5 милијарди токена) (Wenzek et al. 2020), док су за модел *Andrija/SRoBERTa-F* сви поменути корпуси коришћени у целости.

Крајем 2022. године објављена су три експериментална генеративна модела за српски језик (Škorić 2023). Контролни модел *procesaur/gpt2-srlat* био је поново заснован на *GPT2-small* архитектури, имао је 138 милиона параметара и обучен је на исечку корпуса Друштва за језичке ресурсе и технологије (260 милиона токена) (Krstev and Stanković 2023). Друга два модела – *procesaur/gpt2-srlat-sem* и *procesaur/gpt2-srlat-synt*, настала су дообучавањем контролног модела коришћењем два специјално припремљена корпуса са циљем засебног моделовања семантике, односно, синтаксе текста. Три модела су потом употребљена за експеримент комбиновања језичких модела на задатку класификације реченица (Škorić, Utvić, and Stanković 2023).

Почетком наредне године, истраживачи са Универзитета у Нишу објавили су модел *JelenaTosic/SRBerta* (75 милиона параметара) који је такође заснован на *RoBERTa-base* архитектури, а обучаван је помоћу корпуса *OSCAR* (Suárez, Sagot, and Romary 2019). Занимљиво је да је овај модел, као и његова друга верзија (*nemanjaPetrovic/SRBerta*, 120 милиона параметара), пре објављивања дообучен над текстовима из домена права (Bogdanović, Kosić, and Stoimenov 2024). У периоду између објављивања ових модела, објављен је и *aleksahet/xlm-r-squad-sr-*

lat (Cvetanović and Tadić 2023), први модел за одговарање на питања на српском језику, настао прилагођавањем *XLM-RoBERTa* модела помоћу скупа података *SQuAD* (Rajpurkar, Jia, and Liang 2018), преведеног на српски језик.

Средином 2023. године објављена су још два генеративна модела заснована на ГПТ архитектури. Оба модела су обучавана над истим скупом података: над корпусима Друштва за језичке ресурсе и технологије (Krstev and Stanković 2023), докторским дисертацијама преузетим са платформе НАРДУС,⁷ корпусу јавног дискурса српског језика Института за српски језик САНУ под називом *PDRS* (Wasserscheidt 2023) и додатним јавно доступним корпусима са веба, као што су поменути *srWAc* (Ljubešić and Klubička 2014) и *cc100-sr* (Wenzek et al. 2020). Укупан број токена у овом скупу података броји око четири милијарде. Већи модел – *jerteh/gpt2-orao*⁸ броји 800 милиона параметара, заснован је на архитектури *GPT2-large* и представља тренутно највећи доступни модел предобучен за српски језик. Мањи модел – *jerteh/gpt2-vrabac*⁹ броји 136 милиона параметара и заснован је на архитектури *GPT2-small*. Оба модела су обучавана коришћењем рачунарских ресурса Националне платформе за вештачку интелигенцију Србије. Осим корпуса за обучавање, ова два модела деле и речник токена и токенизатор, специјално опремљен да упарује ћирилична и латинична слова, омогућујући им равноправну подршку.

Након објављивања генеративног модела од 800 милиона (*jerteh/gpt2-orao*) параметара, фокус се полако помера на дообучавање великих модела коришћењем текстова на српском језику. Тако су објављена два модела заснована на архитектури *Alpaca* (Taori et al. 2023), *datatab/alpaca-serbian-3b-base* (3 милијарде параметара) и *datatab/alpaca-serbian-7b-base* (7 милијарди параметара), а најављено је и објављивање још једног модела исте величине, заснованог на архитектури *Mistral-7b* (Jiang et al. 2023), који је обучаван на хрватским, босанским и српским текстовима (11,5 милијарди токена). Исти корпус коришћен је и за дообучавање *XLM-RoBERTa-large* модела у циљу поређења перформанси дообучаваних модела у односу на моделе обучаване од почетка (*od нуле*). Нови модел објављен је под именом *classla/xlm-r-*

7. НАРДУС – Национални репозиторијум докторских дисертација са свих универзитета у Србији.

8. [jerteh/gpt2-orao](#)

9. [jerteh/gpt2-vrabac](#)

bertic (Ljubešić et al. 2024) и броји 561 милион параметара, колико има и оригинални *XLM* модел.

Конечно, на скупу података над којим су обучавани *jerteh/gpt2-orao* и *jerteh/gpt2-vrabac*, обучена су још два модела за векторизацију текста. Већи модел, *jerteh/Jerteh-355*,¹⁰ заснован је на *RoBERTa-large* архитектури и броји 355 милиона параметара, док је мањи модел, *jerteh/Jerteh-81*,¹¹ заснован на *RoBERTa-base* архитектури и броји 81 милион параметара. Као и код модела *jerteh/gpt2-orao*, циљ је да се модели обуче на што квалитетнијем корпусу текстова. У овом раду биће представљено испитивање перформанси ових модела и њихово поређење са перформансама других одабраних модела како би се установило њихово место у хијерархији језичких модела за векторизацију текста на српском језику.

1.2 Поставка експеримента

У претходном одељку указано је на постојање већег броја вишејезичних модела који, у мањој или већој мери, подржавају обраду српског језика, као и на двадесетак модела који су припремљени специјално за обраду српског језика. Објављени модели се међусобно разликују према неколико особина: породици (архитектури) модела и броју његових параметара, речнику, односно токенизатору, на којем се заснивају, скупу коришћеном за њихово обучавање, задатку на којем је модел обучаван и дужини обучавања. Треба напоменути да неке од ових информација недостају за неке моделе, али и да су неке доступне информације (пре свега особине скупа коришћеног за обучавање) непроверљиве.

У наставку, рад ће се фокусирати на десет одабраних модела за векторизацију (општег типа). Основне информације о тим моделима биће представљене у одељку 2., експеримент поређења њихових перформанси на четири припремљена задатка приказаћемо у одељку 3., а резултати експеримента биће приказани и размотрени у одељку 4.. Напошетку, у одељку 5., биће предложен процес обучавања нових модела за српски језик. Овај рад неће разматрати генеративне моделе услед недостатка поузданог (али аутоматског) механизма за мерење њихових перформанси. Још није објављен ниједан енкодер-декодер модел специјално развијен за српски језик.

10. *jerteh/Jerteh-355*

11. *jerteh/Jerteh-81*

2. Одабрани модели за векторизацију текста

За потребе овог рада, од претходно поменутих модела (одељак 1.) одабрано је десет за детаљнију анализу. У тих десет улазе најпре четири *SRoBERTa* модела, који су врло погодни за овај експеримент јер се разликују искључиво по скупу података за обучавање. Даље, ту су најстарији модел, *classla/bcms-bertic* и најновији модел, *classla/xlm-r-bertic*, које је објавио центар за јужнословенске језике *CLASSLA*, као и два најпопуларнија вишејезична модела *xlm-roberta-base* и *xlm-roberta-large*. Коначно, ту су два модела која се први пут представљају у овом раду, модели *jerteh-81* и *jerteh-355*, који су обучени над ресурсима Друштва за језичке ресурсе и технологије.

Основне карактеристике ових десет модела приказане су у табели 1.

У приложеној табели, као и из описа модела у претходном одељку, види се да је најпопуларнија архитектура *RoBERTa* (6 од 10 одабраних модела), а додатна три модела заснована су на блиској, *XLM-RoBERTa* архитектури. Преостали модел *bcms-bertic*, заснива се на *ELECTRA* архитектури и једини је од одабраних који није обучаваан на задатку моделовања маскираног језика (предвиђања делова текста маскираних иза неке специјалне етикете).

Величина одабраних модела варира од 80 (за четири *SRoBERTa* модела) па до преко 560 милиона параметара (за моделе засноване на *XLM-RoBERTa-large*). Величина скупа за обучавање варира од 500 милиона за модел 1 (*SRoBERTa-base*), па до чак 11,5 милијарди токена за модел 6 (*classla/xlm-r-bertic*), при чему треба напоменути да није обучаваан од нуле, већ је у питању *xlm-roberta-large* модел дообучен за хрватски, босански и српски језик. Само четири од десет модела обучаваана су искључиво над корпусом српских текстова. У питању су модели 1, 2, 9 и 10, тј. прва два *SRoBERTa* модела, *jerteh/jerteh-81* и *jerteh/jerteh-355*.

Битно је напоменути и да десет приказаних модела користе само четири различита речника токена, односно токенизатора:

X_1 *SRoBERTa* токенизатор - прва 4 модела;

X_2 *bertic* токенизатор - модел број 5;

X_3 *XLM-R* токенизатор - модели 6 до 8;

X_4 *jerteh* токенизатор - последња 2 модела (9 и 10).

редни број	1	2	3	4	5	6	7	8	9	10
Идентификатор	Andrija/SRoBERTa-base	Andrija/SRoBERTa-L	Andrija/SRoBERTa-XL	Andrija/SRoBERTa-F	classla/bcns-bertic	classla/xlm-r-bertic	xlm-roberta-base	xlm-roberta-large	jerteh/jerteh-81	jerteh/jerteh-355
Речник токена	SRoBERTa				bertic	XLM-R			jerteh	
Архитектура	RoBERTa				ELE.	XLM-R			RoBERTa	
Величина модела	80				110	561	279	561	81	355
Величина скупа	500	1.000	3.750	5.700	8.400	11.500	4.000*		4.000	
Српски	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Хрватски			✓	✓	✓	✓	✓	✓		
Босански					✓	✓	✓	✓		
Црногорски					✓		✓	✓		

Табела 1. Десет одабраних модела за векторизацију текста на српском језику и њихове особине: речник токена на којем су засновани, архитектура модела, величина модела изражена у милионима параметара и величина скупа изражена у милионима токена. Подаци су преузети са платформе *HuggingFace*. *Величина скупа за обучавање код модела 7 и 8 (*xlm-roberta-base*, *xlm-roberta-large*) односи се на део скупа на српском, хрватском или другом сродном језику. Доњи део табеле приказује на којем од ових језика су обучавани модели.

3. Поставка евалуације перформанси модела

Десет одабраних модела је евалуирано на четири засебна задатка како би се упоредиле њихове перформансе:

- T_1 Моделовање маскираног језика (погађање недостајућих токена);
- T_2 Израчунавање (семантичке) сличности између реченица;
- T_3 Обележавање врстом речи;
- T_4 Препознавање именованих ентитета.

Прва два задатка припадају групи такозваних узводних задатака (*upstream*), тј. задатака који користе моделе у основном стању, док друга два задатка припадају групи низводних задатака (*downstream*) јер захтевају да се модели фино подесе (*fine-tune*) и тестирају на специјално припремљеном скупу података.

3.1 Евалуација модела на узводним задацима

Као што је већ поменуто, за узводне задатке није неопходно прилагођавање модела, те је потребно само припремити скупове за тестирање.

Како би се спровела евалуација модела на задатку моделовања маскираног језика (T_1) најпре је припремљен посебан скуп података у виду текстова у којима су у свакој реченици по један насумично изабрани токен маскирани – по један токен је сакривен, на пример иза маске <MASK>. За текстуалну грађу коришћена су четири извора:

- Y_1 *Дечко*, српски превод романа *Подросток* од Достојевског;
- Y_2 *Младић*, алтернативни превод *Дечка*;
- Y_3 *Пут око света за 80 дана*, српски превод романа Жила Верна;
- Y_4 *Пут око свијета у 80 дана*, хрватски превод романа Жила Верна.

Прва два извора нису коришћена за обучавање ниједног модела, док су друга два већ дуго доступни на вебу (Vitas et al. 2008) и стога су вероватно коришћени за обучавање већине, ако не и свих, наведених модела. Како неки модел не би био у посебној предности, текстови су токенизовани коришћењем сва четири токенизатора (X_1 до X_4), потом маскирани, а онда је пред сваким од модела био задатак да одмаскира свих шеснаест припремљених текстова (четири извора токенизована и маскирана на четири начина). У свакој реченици био је маскиран по

један токен, а модели су током евалуације за његово место нудили по три кандидата. За процену резултата теста узимана је мера тачности на овом задатку, при чему се као погодак рачунало свако одмаскирање код кога је маскирани, то јест тражени, токен био у скупу кандидата које је модел понудио за задату реченицу.

За потребе евалуације на задатку израчунавања сличности између реченица (T_2), коришћене су реченице из истих романа, то јест два пара паралелизованих романа (Y_1 и Y_2 , односно Y_3 и Y_4), али припремљене као триплети. С обзиром на то да су романи претходно паралелизовани на нивоу реченице, било је лако направити парове реченица које имају исто значење. Сваки триплет је употпуњавала додатна реченица из романа парњака, која дели што је више могуће токена са контролном реченицом и има сличну дужину, али је повучена из неког другог места у тексту. Пример триплета:

1. "Zaista, ko ne bi obišao svet i za manju cenu?" (контролна реченица, Y_3 : *Пут око света за 80 дана*)
2. "Doista, nije li i za manje od toga vrijedno izvršiti put oko svijeta?" (парњак, Y_4 : *Пут око свијета у 80 дана*)
3. "He! he! pa konačno zašto ne bi uspio?" (лажни парњак, Y_4 : *Пут око свијета у 80 дана*)

Задатак модела био је да у задатим триплетима препознају прави парњак (сличност између прве и друге реченице треба да буде већа него између прве и треће), а за процену резултата теста узимана је тачност на том задатку. Да би се израчунао реченични вектор, модел најпре додели векторске вредности сваком токenu у реченици, а потом се вредност тих вектора усредњава како би се добила векторска репрезентација реченице. Сличност између двеју реченица израчунава се као разлика броја 1 и косинусне удаљености израчунатих реченичних вектора.

3.2 Евалуација модела на низводним задацима

За потребе евалуације модела преостала два предвиђена задатка, модели су дообучени и тестирани на посебним скуповима података. За обележавање врстом речи (T_3) коришћен је јавно доступни скуп *Srp-Kor4Tagging* (Stanković et al. 2020) (триста педесет хиљада обележених токена), док је за препознавање именованих ентитета (T_4) коришћен други јавно доступни скуп *SrpELTeC-gold* (Šandrih Todorović et al. 2021). У оба случаја, модели су дообучени на 90% обележених реченица из

сваког скупа и тестирани на преосталих 10%. Како је реч о проблему вишекласне класификације, за процену резултата ова два задатка узете су F_1 -мере остварене приликом класификације над реченицама из скупа за тестирање.

4. Резултати евалуације

Резултати првог теста (T_1), тј. просечна тачност одабраних модела на задатку допуњавања недостајућег токена у сваком од шеснаест припремљених текстова, приказани су у табели 2.

У приложеним резултатима је може се уочити супериорност новог модела, *jerteh-355*, који остварује бољи резултат од осталих модела у тринаест од шеснаест случајева, односно бољи или исти резултат ($\pm 1\%$) у петнаест од шеснаест случајева. Осим тога, у девет од дванаест случајева модел *jerteh-355* надмашује друге моделе чак и када обрађује текст маскиран токенизатором тих модела. Једини модел који успева да га надмаши у два случаја је *SRoBERTa-F*, који се најчешће показује као најбољи у обради извора (Y_4) писаног на хрватском језику, који је у великом проценту био укључен у његов скуп за обучавање. Ипак, у просеку, његова тачност на овом тесту је нижа и од оне коју остварује други нови модел, *jerteh-81*. Модел 5 (*classla/bcms-bertic*), који, за разлику од осталих, није обучаван на задатку моделовања маскираног језика, није био укључен у евалуацију на овом задатку јер би био у неповољном положају.

Резултати теста израчунавања сличности између реченица (T_2) приказани су у табели 3. Вредности приказују тачност модела при препознавању реченица са истим/сличним значењем у триплетима екстрахованим из два српска превода истог романа, Y_1 и Y_2 (први ред вредности), из српског и хрватског превода истог романа, Y_3 и Y_4 (други ред), и просечну тачност (трећи ред).

Резултати за први низ триплета су веома добри за неколико модела: *SRoBERTa-L*, *SRoBERTa-XL*, *SRoBERTa-F*, *jerteh-81* и *jerteh-355* остварују сличну тачност од преко 95%. Модел *SRoBERTa-XL* остварује најбоље резултате, уз малу маргину, али и најбољи резултат за други низ триплета који садржи реченице на хрватском језику (92% тачности), па самим тим има и најбољи просечни резултат на овом задатку. Једини други модел који остварује тачност од преко 90% за други низ триплета је *SRoBERTa-F*, што је и очекивано јер су ова два модела обучавана на хрватским текстовима.

редни број	1	2	3	4	5	6	7	8	9	10
	Andrija/SRoBERTa-base	Andrija/SRoBERTa-L	Andrija/SRoBERTa-XL	Andrija/SRoBERTa-F	classla/bcms-bertic	classla/xlm-r-bertic	xlm-roberta-base	xlm-roberta-large	jerteh/jerteh-81	jerteh/jerteh-355
X_1-Y_1	0,43	0,63	0,66	0,70	/	0,43	0,46	0,51	0,70	0,75
X_1-Y_2	0,43	0,62	0,64	0,69	/	0,42	0,46	0,50	0,69	0,73
X_1-Y_3	0,37	0,56	0,59	0,63	/	0,34	0,38	0,43	0,66	0,72
X_1-Y_4	0,36	0,55	0,64	0,68	/	0,34	0,38	0,42	0,58	0,63
X_2-Y_1	0,36	0,47	0,51	0,54	/	0,47	0,50	0,55	0,57	0,60
X_2-Y_2	0,37	0,48	0,51	0,54	/	0,46	0,50	0,54	0,56	0,59
X_2-Y_3	0,31	0,41	0,45	0,48	/	0,42	0,45	0,50	0,50	0,54
X_2-Y_4	0,31	0,42	0,47	0,51	/	0,42	0,46	0,50	0,47	0,51
X_3-Y_1	0,37	0,49	0,52	0,54	/	0,48	0,50	0,55	0,57	0,60
X_3-Y_2	0,37	0,48	0,51	0,54	/	0,46	0,50	0,54	0,57	0,59
X_3-Y_3	0,30	0,41	0,44	0,47	/	0,41	0,45	0,49	0,50	0,54
X_3-Y_4	0,31	0,42	0,47	0,51	/	0,41	0,46	0,50	0,47	0,50
X_4-Y_1	0,42	0,60	0,63	0,67	/	0,43	0,47	0,51	0,73	0,78
X_4-Y_2	0,41	0,58	0,61	0,65	/	0,41	0,45	0,49	0,71	0,75
X_4-Y_3	0,35	0,53	0,55	0,60	/	0,33	0,38	0,42	0,69	0,76
X_4-Y_4	0,34	0,50	0,58	0,62	/	0,33	0,37	0,41	0,62	0,66
просек	0,36	0,51	0,55	0,59	/	0,41	0,45	0,49	0,60	0,64

Табела 2. Тачност модела у погађању токена (из три покушаја) на задатку моделовања маскираног језика над шеснаест припремљених маскираних текстова и њихова просечна тачност. Текстови су обележени (на почетку сваког реда) јединственим комбинацијама ознака токенизатора X и извора Y . У сваком реду најбољи резултат ($\pm 1\%$) обележен је подебљањем.

Резултати који су модели остварили на низводним задацима T_3 (обележавање врсте речи) и T_4 (препознавање именованих ентитета) приказани су у виду F_1 -мера у табели 4.

редни број	1	2	3	4	5	6	7	8	9	10
	Andrija/SRoBERTa-base	Andrija/SRoBERTa-L	Andrija/SRoBERTa-XL	Andrija/SRoBERTa-F	classla/bcms-bertic	classla/xlm-r-bertic	xlm-roberta-base	xlm-roberta-large	jerteh/jerteh-81	jerteh/jerteh-355
Y_1 - Y_2	0,93	0,95	0,96	0,96	0,92	0,76	0,90	0,87	0,95	0,95
Y_3 - Y_4	0,83	0,89	0,92	0,91	0,79	0,66	0,78	0,71	0,89	0,83
просек	0,88	0,92	0,94	0,93	0,85	0,71	0,84	0,79	0,92	0,89

Табела 3. Резултати одабраних модела на задатку препознавања реченица са истим значењем у триплетима екстрахованим из превода Достојевског (Y_1 - Y_2), Верна (Y_3 - Y_4), као и у просеку. У сваком реду најбољи резултат ($\pm 1\%$) обележен је подебљањем.

р.б.	1	2	3	4	5	6	7	8	9	10
	Andrija/SRoBERTa-base	Andrija/SRoBERTa-L	Andrija/SRoBERTa-XL	Andrija/SRoBERTa-F	classla/bcms-bertic	classla/xlm-r-bertic	xlm-roberta-base	xlm-roberta-large	jerteh/jerteh-81	jerteh/jerteh-355
T_3	0,974	0,980	0,982	0,982	0,986	0,987	0,984	0,986	0,985	0,986
T_4	0,908	0,922	0,929	0,935	0,942	0,942	0,933	0,935	0,928	0,928

Табела 4. F_1 -мера коју су модели остварили на задацима T_3 (обележавање врсте речи) и T_4 (препознавање именованих ентитета). У сваком реду најбољи резултат ($\pm 0.1\%$) обележен је подебљањем.

Из резултата приказаних у табели 4 види се да на задатку T_3 (обележавање врстом речи) девет од десет модела остварује јако добре резултате (преко 98%), при чему се резултати које остварују четири најбоља модела (*classla/bcms-bertic*, *classla/xlm-r-bertic*, *xlm-roberta-large* и *jerteh/jerteh-355*) разликују за мање од 0,02%, указујући да се модели полако приближавају горњим границама перформанси када је у питању овај задатак.

Када су у питању резултати остварени на последњем задатку T_4 (препознавање именованих ентитета), најбоље перформансе приказали су модели *classla/bcms-bertic* и *classla/xlm-r-bertic*, при чему су разлике у F_1 -мерама нешто више него на задатку T_3 ($\sim 4\%$ за задатак T_4 у односу на $\sim 1\%$ за задатак T_3), али и даље знатно ниже него што су разлике на узводним задацима (чак $\sim 28\%$ за задатак T_1).

У наредном одељку биће разматрани резултати које су модели остварили, као и разлози који су довели до тих резултата, са циљем да се установе најповољнији услови за обучавање модела за српски језик у будућности.

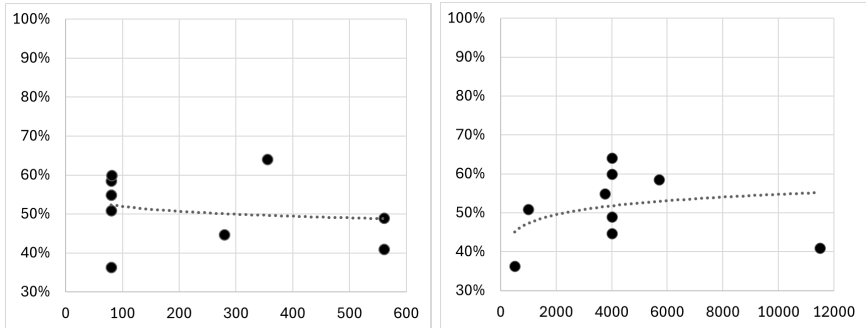
5. Дискусија

Претходно приказани резултати евалуације модела (табеле 2–4), показују да не постоји један модел, или група модела, који су најбољи у општем случају, већ су се различити модели показали као бољи (или лошији) на различитим задацима. У наставку, сваки од задатака биће посматран појединачно, пре свега у светлу односа остварених перформанси према величини модела, величини скупова за њихово обучавање и квалитету тих скупова.

5.1 Моделовање маскираног језика

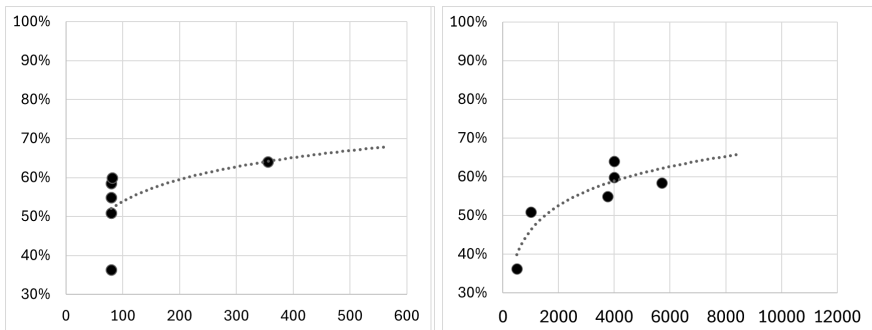
Резултати евалуације модела на задатку моделовања маскираног језика (табела 2) показују убедљиву предност модела *jerteh/jerteh-355*, са тим да добре резултате остварује и *jerteh/jerteh-81*, који је други најбољи модел за овај задатак када се посматрају просечне вредности. Пошто ова два модела користе исти скуп података за обучавање, то указује да би управо овај скуп могао бити разлог добрих резултата.

Просечна тачност модела на задатку T_1 према њиховој величини, односно према величини скупа који је коришћен за њихово обучавање приказан је на слици 1. На први поглед, приметни су неки истакнути



Слика 1. Тачност модела на задатку моделовања маскираног језика према величини тих модела (лево), као и према величини скупова за обучавање модела (десно). Приказана крива тренда одговара логаритамској функцији.

изузеци, пре свега модели засновани на *XLM-R* архитектури, који остварују неке од најлошијих резултата на овом задатку. Уколико се њихови резултати уклоне, појављују се нови трендови (слика 2). Дакле, када посматрамо само *RoBERTa* моделе, чини се да већи модел (не баш убедљиво) и већи скуп за његово обучавање (врло убедљиво) повољно утичу на перформансе модела.

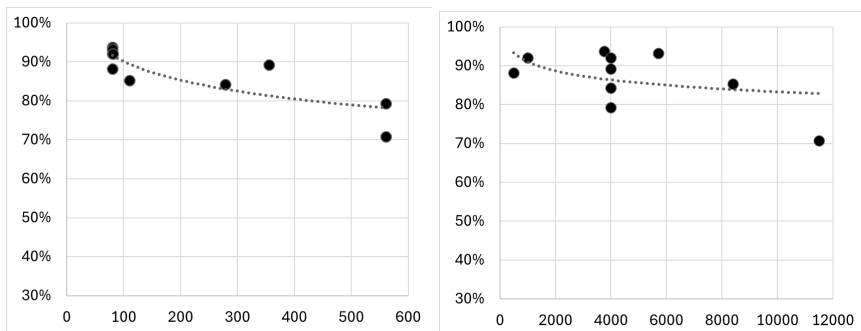


Слика 2. Тачност модела на задатку моделовања маскираног језика према величини тих модела (лево), као и према величини скупова за обучавање модела (десно), при чему су уклоњени резултати модела заснованих на *XLM-R* архитектури. Приказана крива тренда одговара логаритамској функцији.

5.2 Израчунавање сличности између реченица

Приликом евалуације задатка T_2 установљено је да најбоље резултате остварују модели *SRoBERTa-L*, *SRoBERTa-XL*, *SRoBERTa-F*, *jerteh-81* и *jerteh-355* када је у питању препознавање сличних реченица на српском језику, а *SRoBERTa-XL* и *SRoBERTa-F* када је у питању препознавање сличних реченица између српског и хрватског језика (табела 2). Ово указује да је кључ за добро угњежђивање реченица претходно обучавање модела за језике који се испитују.

Дакле, за угњежђивање реченица на српском језику најбољи су модели претходно обучени на довољно великом скупу реченица српског језика, али ако се обрађују и реченице на хрватском, модели који су обучавани и на српском и на хрватском имају предност. Са друге стране, модели засновани на *XLM-R* архитектури који су унапред обучени на сто светских језика поново показују најлошије резултате, вероватно због великог шума који разнолики скуп за обучавање производи.



Слика 3. Тачност модела на задатку угњежђивања према величини тих модела (лево), као и према величини скупова за обучавање модела (десно). Приказана крива тренда одговара логаритамској функцији.

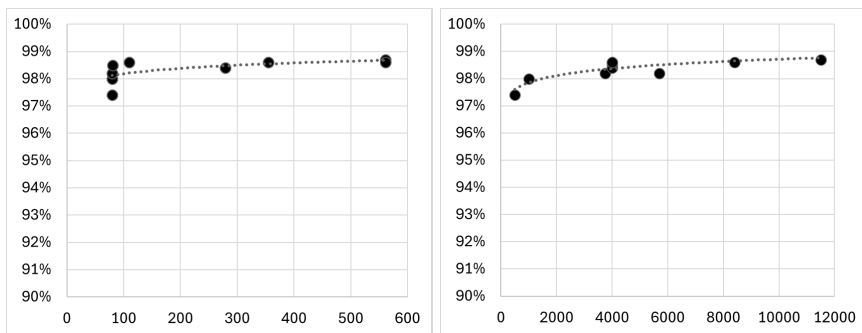
На слици 3 приказан је утицај величине модела и скупова за обучавање на перформансе на овом задатку, а линије тренда указују да се са повећањем и модела и скупа за обучавање перформансе смањују. Утицај величине скупа може донекле бити приписан претходно описаном феномену који погађа *XLM-R* архитектуру. Ипак, када је у питању утицај величине модела, постоје додатни индикатори да су мањи модели

бољи за овај задатак када је у питању обрада вишејезичних текстова. Тако *jerteh/jerteh-81* надмашује *jerteh/jerteh-355* на задатку утврђивања сличности између реченица на српском и хрватском језику. Разлог може бити то што је мањи модел, услед недостатка у величини, слабије прилагођен српском језику (*model underfit*), али зато има предност приликом генерализације.

5.3 Низводни задаци

За разлику од евалуације на узводним задацима, на низводним задацима резултати које постижу модели много су сличнији. На задатку обележавања врстом речи (T_3) готово сви модели остварују добре резултате (табела 4), укључујући и оне засноване на *XLM-R* архитектури. Штавише, *classla/xlm-r-bertic* и *xlm-roberta-large* два су од четири модела који остварују најбоље резултате (друга два модела су *classla/bcms-bertic* и *jerteh/jerteh-355*).

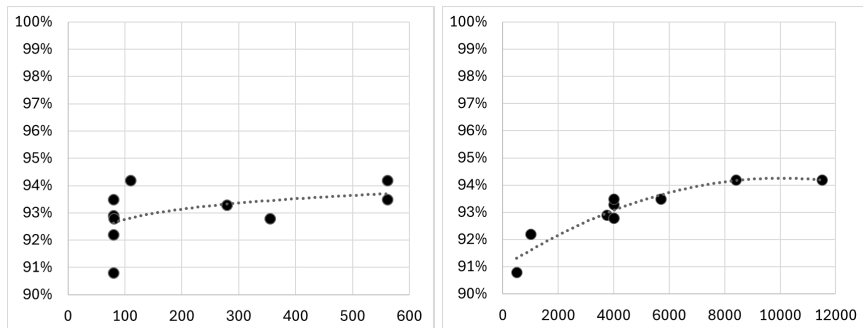
Заједничко за ова четири модела је да су или највећи модели или модели који су обучавани на највећим скуповима података. Позитивна корелација између перформанси и величине модела, као и између перформанси и величине скупова за обучавање уочава се и на слици 4.



Слика 4. Перформансе модела на задатку обележавања врстом речи према величини тих модела (лево), као и према величини скупова за обучавање модела (десно). Приказана крива тренда одговара логаритамској функцији.

Корелација између величине скупа за обучавање још је очигледнија у случају препознавања именованих ентитета (слика 5). Најбоље резултате

на овом задатку (T_4) остварила су два модела са највећим скуповима за обучавање, док је поново приметна и успешност *XLM-R* модела, као и код претходног задатака. Величина модела такође показује тек незнатну позитивну корелацију.



Слика 5. Перформансе модела на задатку препознавања именованих ентитета према величини тих модела (лево), као и према величини скупова за обучавање модела (десно). Приказана крива тренда одговара логаритамској функцији.

5.4 Закључак

Када је у питању моделирање маскираног језика, чини се да развој нових модела за српски језик иде у правом смеру. Модел *jerteh/jerteh-355* остварује убедљиво најбоље резултате, барем када је у питању рад са висококвалитетним текстовима, чак и када су маскирани токенизаторима других модела (табела 1). Премда повећање скупа података за обучавање повољно утиче на перформансе модела (слика 2), не треба занемарити ни квалитет скупа јер модели *jerteh/jerteh-355* и *jerteh/jerteh-81* надмашују моделе *Andrija/SRoBERTa-F* и *Andrija/SRoBERTa-XL* обучене на већим скуповима, указујући да веб-корпуси можда нису увек довољни за обучавање добрих модела. Овај закључак је у складу са закључком из једног другог недавног истраживања (Li et al. 2023); ипак, ново истраживање би требало да у скуп за евалуацију уврсти и нелитерарне изворе, како би се добила свеобухватнија слика стања.

На задатку израчунавања сличности између реченица, тј. угнежђивања реченица, истакли су се модели као што су *Andrija/SRoBERTa-F* и *Andrija/SRoBERTa-XL*, за којима сасвим мало заостају *Andrija/SRoBERTa-L*, *jerteh/jerteh-81* и *jerteh/jerteh-355*, када су у питању реченице на српском језику (табела 3). Оно што издваја ове моделе је да су мањи у односу на друге моделе и да су обучавани на већем скупу, па изгледа да је за овај задатак кључна генерализација, дакле, већи скупови података (или можда мањи модели). Такође, када је у питању обрада реченица ширег језичког спектра (нпр. јужнословенски језици) било би потребно да се реченице из комплетног спектра укључе у скуп за обучавање или, боље, да се речник прилагоди за мапирање ширег спектра токена, па самим тим и за правилну векторизацију ових реченица. Ново истраживање на ову тему требало би да истражи и модернији начин векторизације реченица, на пример, коришћењем архитектуре трансформера реченица (*sentence transformers*) (Reimers and Gurevych 2019).

У случају оба *узводна задатка*, тачност коју остварују модели засновани на *XLM-R* архитектури много је нижа у односу на тачност модела заснованих на *RoBERTa* архитектури. У случају првог задатка (T_1) то се може објаснити њиховим већим речником токена (па је самим тим и избор одговарајућег токена тежи). Ипак, такво објашњење не би било адекватно и за други задатак (T_2). Са друге стране модели засновани на *XLM-R* архитектури показали су се као најбољи (уз малу маргину) на низводним задацима, пре свега на задатку препознавања именованих ентитета (T_4). Чини се да је за успешно решавање овог задатка најбоље да се модел током обучавања сусретне са најразличитијим бројем токена, али побољшања доноси и додатно обучавање на српским текстовима. Изгледа да би за унапређење перформанси тренутно било оптимално дообучавање *XLM-RoBERTa-large* модела коришћењем што већег и квалитетнијег скупа текстова на српском језику. Када је у питању обележавање врстом речи, изгледа да би било који нови модел био адекватан за решавање овог задатка.

Захвалница

Најважније скупове података за обучавање модела *GPT2-orao*, *GPT2-vrabac*, *jerteh-81* и *jerteh-355* обезбедило је Друштво за језичке ресурсе и технологије.¹²

12. JeRTeh

Рачунарске ресурсе за обучавање модела *GPT2-orao* и *GPT2-vrabas* обезбедила је Национална платформа за вештачку интелигенцију Србије.

Рачунарске ресурсе за обучавање модела *jerteh-81* и *jerteh-355* обезбедио је Рударско-геолошки факултет Универзитета у Београду.

Истраживање је спроведено уз подршку Фонда за науку Републике Србије, #7276, *Text Embeddings – Serbian Language Applications – TESLA*.

Хвала!

Литература

Biemann, Chris, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. "The Leipzig corpora collection-monolingual corpora of standard size." *Proceedings of corpus linguistic 2007*.

Bogdanović, Miloš, Jelena Kocić, and Leonid Stoimenov. 2024. "SRBerta-A Transformer Language Model for Serbian Cyrillic Legal Texts." *Information* 15 (2). ISSN: 2078-2489. <https://doi.org/10.3390/info15020074>.

Clark, Kevin, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.10555*.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Unsupervised cross-lingual representation learning at scale." *arXiv preprint arXiv:1911.02116*.

Cvejić, Andrija. 2022. "Prepoznavanje imenovanih entiteta u srpskom jeziku pomoću transformer arhitekture." *Zbornik radova Fakulteta tehničkih nauka u Novom Sadu* 37 (02): 310–315.

Cvetanović, Aleksa, and Predrag Tadić. 2023. "Synthetic Dataset Creation and Fine-Tuning of Transformer Models for Question Answering in Serbian." In *2023 31st Telecommunications Forum (TELFOR)*, 1–4. IEEE.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*.

- Dobreva, Jovana, Tashko Pavlov, Kostadin Mishev, Monika Simjanoska, Stojancho Tudzarski, Dimitar Trajanov, and Ljupcho Kocarev. 2022. “MACEDONIZER-The Macedonian Transformer Language Model.” In *International Conference on ICT Innovations*, 51–62. Springer.
- He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. “DE-BERTA: Decoding-Enhanced BERT with Disentangled Attention.” In *International Conference on Learning Representations*.
- Jiang, Albert Q, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. “Mistral 7B.” *arXiv preprint arXiv:2310.06825*.
- Krstev, Cvetana, and Ranka Stanković. 2023. “Language Report Serbian.” In *European Language Equality: A Strategic Agenda for Digital Language Equality*, edited by Georg Rehm and Andy Way, 203–206. Cham: Springer International Publishing. ISBN: 978-3-031-28819-7. https://doi.org/10.1007/978-3-031-28819-7_32.
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep learning.” *nature* 521 (7553): 436–444. <https://doi.org/10.1038/nature14539>.
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.
- Li, Yuanzhi, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. “Textbooks are all you need ii: phi-1.5 technical report.” *arXiv preprint arXiv:2309.05463*.
- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. “Roberta: A robustly optimized BERT pretraining approach.” *arXiv preprint arXiv:1907.11692*.
- Ljubešić, Nikola, and Filip Klubička. 2014. “{bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian.” In *Proceedings of the 9th web as corpus workshop (WaC-9)*, 29–35.

- Ljubešić, Nikola, and Davor Lauc. 2021. “BERTić—The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian.” *arXiv preprint arXiv:2104.09243*.
- Ljubešić, Nikola, Vit Suchomel, Peter Rupnik, Taja Kuzman, and Rik van Noord. 2024. *Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining*. Submitted for review.
- Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. “Improving language understanding by generative pre-training.”
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. “Language models are unsupervised multitask learners.”
- Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. “Exploring the limits of transfer learning with a unified text-to-text transformer.” *The Journal of Machine Learning Research* 21 (1): 5485–5551.
- Rajpurkar, Pranav, Robin Jia, and Percy Liang. 2018. “Know what you don’t know: Unanswerable questions for SQuAD.” *arXiv preprint arXiv:1806.03822*.
- Reimers, Nils, and Iryna Gurevych. 2019. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.
- Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. “Serbian ner&beyond: The archaic and the modern intertwined.” In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 1252–1260.
- Škorić, Mihailo. 2023. “Композитне псеудограматике засноване на паралелним језичким моделима српског језика.” Докторска дисертација. PhD diss., Универзитет у Београду.
- Škorić, Mihailo, Miloš Utvić, and Ranka Stanković. 2023. “Transformer-Based Composite Language Models for Text Evaluation and Classification.” *Mathematics* 11 (22): 4660.

- Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. “Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian.” In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 3954–3962.
- Suárez, Pedro Javier Ortiz, Benoît Sagot, and Laurent Romary. 2019. “Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures.” In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.
- Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. “Alpaca: A strong, replicable instruction-following model.” *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html> 3 (6): 7.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention is all you need.” *Advances in neural information processing systems* 30.
- Vitas, Duško, Svetla Koeva, Cvetana Krstev, and Ivan Obradović. 2008. “Tour du monde through the dictionaries.” In *Actes du 27eme Colloque International sur le Lexique et la Gammaire*, 249–256.
- Wasserscheidt, Philipp. 2023. *Serbian Web Corpus PDRS 1.0*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1752>.
- Wenzek, Guillaume, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. “CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data” [in English]. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4003–4012. Marseille, France: European Language Resources Association, May. ISBN: 979-10-95546-34-4. <https://www.aclweb.org/anthology/2020.lrec-1.494>.