Универзитет у Београду
Студије при Универзитету

Веће за студије при Универзитету

Предмет: Реферат о урађеној докторској дисертацији кандидата Михаила Шкорића

Одлуком Већа за студије при универзитету бр. 06-4594/IV-1848/3-23 од 20. фебруара 2023. године, именовани смо за чланове Комисије за преглед, оцену и одбрану докторске дисертације кандидата **Михаила Шкорића** под насловом:

„Композитне псеудограматике засноване на паралелним језичким моделима српског језика“

После прегледа достављене дисертације и других пратећих материјала и разговора са кандидатом, Комисија је сачинила следећи:

Р Е Ф Е Р А Т

1 УВОД

1.1 Хронологија одобравања и израде докторске дисертације

Кандидат Михаило Шкорић је уписао интердисциплинарне докторске студије Универзитета у Београду, смер Интелигентни системи 2016 године. Током студија је положио све испите и испунио све обавезе у вези са студијским истраживачким радом предвиђене планом и програмом.

Кандидат је пријавио тему докторске дисертације под насловом: "Композитне псеудограматике засноване на паралелним језичким моделима српског језика", а Веће за студије при Универзитету је на седници одржаној 25. маја 2020. године донело одлуку број 06-4107/2.2-20 о именовању чланова комисије за оцену научне заснованости теме докторске дисертације под насловом: "Композитне псеудограматике засноване на паралелним језичким моделима српског језика". На седници Већа за студије при Универзитету одржаној 2. јула 2020. године, одлуком 06-148/1848/4-20 усвојен је извештај комисије за оцену научне заснованости предложене теме докторске дисертације под насловом: "Композитне псеудограматике засноване на паралелним језичким моделима српског језика", а за менторе ове докторске дисертације именовани су проф. др Ранка Станковић, ванредни професор и проф. др Цветана Крстев, редовни професор. Веће за интердисциплинарне, мултидисциплинарне и трансдисциплинарне студије, на седници одржаној 8. фебруара 2023. године, донело је одлуку број 06-355/I-4.1/2-23 о промени ментора докторске дисертације где је уместо проф. др Цветане Крстев, Филолошки факултет (због одласка у пензију) именована доц. др Јелена Граовац, Математички факултет.

Веће за студије при Универзитету, на седници одржаној 20. фебруара 2023. год. донело је одлуку број 06-4594/IV-1848/3-23 да се образује комисија за преглед и оцену докторске дисертације под насловом: „Композитне псеудограматике засноване на паралелним

језичким моделима српског језика“, кандидата Михаила Шкорића (докторске студије: Интелигентни системи) у саставу:

1. др Владан Девеџић, редовни професор,
   Универзитет у Београду, Факултет организационих наука
2. др Милош Утвић, доцент,
   Универзитет у Београду, Филолошки факултет
3. др Драган Станков, ванредни професор,
   Универзитет у Београду, Рударско-геолошки факултет

1.2 Научна област дисертације

Докторска дисертације припада научној области Интелигентни системи и ужим областима: Обрада природног језика и Рачунарска лингвистика, при чему тема има мултидисциплинарни карактер и поред рачунарских наука, конкретно вештачке интелигенције, укључује и лингвистичке науке, посебно област дигиталне хуманистике. Мултидисциплинарност дисертације се огледа у коришћеним методама, као и у примени спроведеног истраживања. Коришћене методе припадају различитим областима науке, пре свега рачунарства и лингвистике, јер је и сама област обраде природног језика мултидисциплинарна, а велику улогу су играле и методе вероватноће и статистике, као и машинског учења. Развијени ресурси и технологије већ налазе примену у рачунарству, као и у лингвистици.

Ментори докторске дисертације су проф. др Ранка Станковић, Рударско-геолошки факултет (ужа област математика и информатика) и доц. др Јелена Граовац, Математички факултет (ужа област рачунарство и информатика). Наведени ментори су аутори великог броја научних радова у истакнутим међународним часописима и испуњавају све формалне и законске услове за менторе ове дисертације. Релевантни радови ментора су наведени приликом пријаве теме докторске дисертације, односно приликом замене ментора.

1.3 Биографски подаци о кандидату

Михаило Шкорић рођен је 1992. године у Београду. Одрастао је у Обреновцу, где је похађао гимназију и матурирао 2011. године. Исте године уписује основне академске студије на Филолошком факултету Универзитета у Београду, модул Библиотекарство и Информатика, смер Језик, књижевност, култура. Дипломирао је 2015. године са просечном оценом 8.65 и уписао мастер академске студије истог профила. Завршни рад под називом „Сврставање појмова на скали позитивно-негативно, применом истраживања података над корпусом текстова који садрже емотиконе“ одбранио је 2016. године, и завршио мастер академске студије са просечном оценом 10. Исте године уписује интердисциплинарне докторске студије Универзитета у Београду, смер Интелигентни системи.

Од 2017. године ради у Рачунарском центру Рударског одсека на Рударско-геолошком факултету, у сфери системског и софтверског инжењерства, а исте године стиче и звање истраживач приправник при катедри за примењену математику и информатику. Звање истраживач-сарадник стиче 2020. године, када пријављује тему докторске дисертације „Композитне псеудограматике засноване на паралелним језичким моделима српског језика“. Аутор је и коаутор осамнаест научних радова из области обраде природног језика, интелигентних система и библиотечко-информационих наука, објављених у домаћим и страним часописима и монографијама, од којих су 3 са СЦИ листе.

У оквиру COST акције IC1302 – Keystone, учествује на летњој школи Keyword search in big linked data на Техничком Универзитету у Бечу. Био је активан учесник COST акције CA16204 – Distant Reading for European Literary History, у оквиру које учествује на четвородневном тренингу метода и техника удаљеног читања на Националном Универзитету Ирске у Галвеју 2018, као и на краткотрајној научној мисији на Институту за пољски језик у Кракову у марту 2020. године, са темом компаративне стилистичке и морфосинтаксичке анализе текстова. Активан је члан Друштва за језичке ресурсе и технологије – JePTex – где учествује у развоју система и алата за обраду српског језика.

1.4 Библиографија кандидата

M21a

Mihailo Škorić, Ranka Stanković, Milica Ikonić Nešić, Joanna Byszuk, Maciej Eder, "Parallel Stylometric Document Embeddings with Deep Learning Based Language Models in Literary Authorship Attribution", Mathematics (2022), ISSN: 2227-7390, MDPI AG. DOI 10.3390/math10050838 [IF=2.592 (2021)]

M22

Ranka Stanković, **Mihailo Škorić**, Branislava Šandrih Todorović, "Parallel Bidirectionally Pretrained Taggers as Feature Generators", Applied Sciences (2022), ISSN: 2076-3417, MDPI AG. DOI 10.3390/app11072892 [IF=2.838 (2021)]

Olivera Kitanović, Ranka Stanković, Aleksandra Tomašević, **Mihailo Škorić**, Ivan Babić, Ljiljana Kolonja, "A Data Driven Approach for Raw Material Terminology", Applied Sciences (2021), ISSN: 2076-3417, MDPI AG. DOI 10.3390/app12105028 [IF=2.838 (2021)]

M33

Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, Duško Vitas, **Mihailo Škorić**, Milica Ikonić Nešić, "Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection", Proceedings of the Language Resources and Evaluation Conference, June 2022, Marseille, France (2022), European Language Resources Association.

Milica Ikonić Nešić, Ranka Stanković, Christof Schöch and **Mihailo Škorić**, "From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back)", Proceedings of The 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference, June 2022, Marseille, France (2022), European Language Resources Association.

Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, **Mihailo Škorić**, "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian", Proceedings of the 12th Language Resources and Evaluation Conference, May Year: 2020, Marseille, France (2020), European Language Resources Association.

Ranka Stanković, Cvetana Krstev, Biljana Lazić, **Mihailo Škorić**, "Electronic Dictionaries - from File System to lemon Based Lexical Database", Proceedings of the 11th International Conference on Language Resources and Evaluation - W23 6th Workshop on Linked Data in Linguistics : Towards Linguistic Data Science (LDL-2018), LREC 2018, Miyazaki, Japan, May 7-12, 2018 (2018), European Language Resources Association (ELRA).

M51

Милош Утвић, Ранка Станковић, Александра Томашевић, **Михаило Шкорић**, Биљана Лазић, "Претрага корпуса заснована на употреби екстерних лексичких ресурса путем

веб-сервиса", Научни састанак слависта у Вукове дане - Vol. 48/3 Српски језик и његови ресурси (2019), Међународни славистички центар, Филолошки факултет, Универзитет у Београду. DOI 10.18485/msc.2019.48.3.ch12

M53

Ranka Stanković, **Mihailo Škorić**, Petar Popović, "SrpELTeC on Platforms: Udaljeno čitanje, Aurora, NoSketch", Infotheca 21/2 (2021), Faculty of Philology, University of Belgrade. DOI 10.18485/infotheca.2021.21.2.7

Petar Popović, **Mihailo Škorić**, Biljana Rujević, "The Use of the Omeka Semantic Platform for the Development of the University of Belgrade, Faculty of Mining and Geology Digital Repository", Infotheca 20/1 (2020), Faculty of Philology, University of Belgrade. DOI 10.18485/infotheca.2020.20.1_2.9

Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, **Mihailo Škorić**, "Annotation of the Serbian ELTeC Collection", Infotheca 21/2 (2021), Faculty of Philology, University of Belgrade. DOI 10.18485/infotheca.2021.21.2.3

Biljana Lazić, **Mihailo Škorić**, "From DELA Based Dictionary to Leximirka Lexical Database", Infotheca 19/2 (2019), Faculty of Philology, University of Belgrade. DOI 10.18485/infotheca.2019.19.2.4

Mihailo Škorić, Mauro Dragoni, "Medical Domain Document Classification via Extraction of Taxonomy Concepts from MeSH Ontology", Infotheca 19/1 (2019), Faculty of Philology, University of Belgrade. DOI 10.18485/infotheca.2019.19.1.3

Aleksandra Tomašević, Biljana Lazić, Dalibor Vorkapić, **Mihailo Škorić**, Ljiljana Kolonja, "The Use of the Omeka Platform for Digital Libraries in the Field of Mining", Infotheca 17/2 (2017), Faculty of Philology, University of Belgrade. DOI 10.18485/infotheca.2017.17.2.2

Mihailo Škorić, "Classification of Terms on a Positive-Negative Feelings Polarity Scale Based on Emoticons", Infotheca 17/1 (2017), Faculty of Philology, University of Belgrade. DOI 10.18485/infotheca.2017.17.1.4

Milena Obradović, Aleksandra Arsenijević, **Mihailo Škorić**, "Preparation of Multimedia Document "YU Rock Scene", Infotheca - Journal for Digital Humanities 16/1–2 (2017), Faculty of Philology, University of Belgrade. DOI 10.18485/infotheca.2016.16.1_2.6

M63

Милош Утвић, **Михаило Шкорић**, "Репозиторијум дигиталних идентификатора објеката – доиФил: изградња, стање и перспективе", Научна конференција Библиоинфо — 55 година од покретања наставе библиотекарства на високошколском нивоу, Београд 18. мај 2017. (2019), Филолошки факултет Универзитета у Београду. DOI 10.18485/biblioinfo.2017.ch7

Биљана Лазић, Александра Томашевић, **Михаило Шкорић**, "Дигиталне библиотеке у рударству и геологији са посебним освртом на представљање сиве литературе", Научна конференција Библиоинфо — 55 година од покретања наставе библиотекарства на високошколском нивоу, Београд 18. мај 2017. (2019), Филолошки факултет Универзитета у Београду. DOI 10.18485/biblioinfo.2017.ch13

2 ОПИС ДИСЕРТАЦИЈЕ

2.1 Садржај дисертације (обим, структура и основни садржај дисертације)

Докторска дисертација под насловом "Композитне псеудограматике засноване на паралелним језичким моделима српског језика" написана је на 149 страна, формата А4 (фонт: Book Antiqua 12, проред: single), садржи 48 илустрација, 40 табела, 14 примера и 145 библиографских референци. Дисертација је написана на српском језику коришћењем ћириличног писма. Текст дисертације је организован у три дела са укупно 10 поглавља.

Први је уводни део са 4 поглавља: 1. О језику и граматици (14 страна); 2. О псеудограматикама и језичким моделима (12 страна); 3. о генерисању квалитетног текста (13 страна); 4. О предмету и циљевима истраживања (5 страна). Други део је посвећен паралелним језичким моделима српског језика: 5. Паралелно процесирање у обради српског језика (15 страна); 6. паралелни језички модели у моделовању мини-језика (16 страна). Трећи део говори о композитним псеудограматикама српског језика: 7. Припрема језичких модела и других ресурса (14 страна), 8. Композитни модели засновани на паралелној перплексности (23 стране), 9. Евалуација (24 стране) и 10. Дискусија (9 страна). Поред тога, дисертација садржи: Насловну страну на српском и енглеском језику, Страну са подацима о менторима и члановима комисије, Сажетак на српском и енглеском језику са кључним речима, Садржај, Библиографију, Биографију докторанда, Изјаву о ауторству, Изјаву о истоветности штампане и електронске верзије докторског рада и Изјаву о коришћењу.

2.2 Кратак опис појединачних поглавља

У првом поглављу је дат кратак увод у теорију формалних језика, предочене су различите врсте граматика (по обухватности и начину генерисања), и потом је уведена проблематика регуларности природних језика.

Друго поглавље је посвећено псеудограматикама и језичким моделима. Описан је њихов историјски развој, са највећим акцентом на тренутно стање и најактуалније методе моделовања језика и језичке моделе, попут вештачких неуронских мрежа. Поглавље се наставља описом савременог моделовања језика, обрађеног кроз дубоко учење, врхунске моделе и генеративне предобучене трансформере.

Треће поглавље говори о проблематици евалуације квалитета текста и генерисања квалитетног текста. Описане су различите методе полуаутоматске и аутоматске евалуације. Даје се преглед популарних метрика за аутоматску евалуацију текста, конкретно екстринсичне и интринсичне метрике, а затим је описана хибридна евалуација квалитета текста, алтернативне методе у евалуацији текста и корпуси квалитетног текста.

Четврто поглавље говори о предмету и циљевима истраживања, постављају се истраживачка питања, описује ток истраживања и очекивани научни доприноси.

У петом поглављу се уводи паралелно процесирање у обради српског језика на примерима неколико тема: обележавање врстом речи, паралелно тагирање и обучавање паралелних тагера врстом речи за српски језик, где се описују анотирани ресурси, одабир тагера, алгоритми за обучавање и тагирање и коначно евалуација одређивања врсте речи.

Предмет шестог поглавља су паралелни језички модели у моделовању мини-језика, што је имплементирано на концепту удаљеног читања, анализе ауторства и моделовања стила као мини-језика. Уведени су појмови и технике везане за стилометрију, одређивање ауторства, векторизацију и угњеждавање докумената. Конкретно су описани векторизација

докумената заснована на фреквенцијама (н-грама) токена, потом векторизација докумената заснована на BERT моделима, паралелне матрице удаљености докумената. Студија случаја је представљена на угњеждавању старих српских романа, где су описани корпус и репрезентације докумената, потом произведене и евалуиране матрице удаљености докумената на задатку одређивања ауторства.

Седмо поглавље је посвећено припреми језичких ресурса за потребе дисертације: верзија Корпуса савременог српског језика – СрпКор2013 и проширења СрпКор2021, као и осталих корпуса коришћених у истраживању. Даље следи опис проширења корпуса семантичком и синтаксичком репрезентацијом текста. Поглавље се завршава описом обучавања језичких модела над различитим репрезентацијама корпуса.

Осмо поглавље се бави композитним моделима заснованим на паралелној перплексности, уводи се перплексност у служби композитне евалуације текста, потом основне композиције, вектори перплексности и потом композиције засноване на векторима перплексности (у служби детекције грешака у тексту и сажети вектори у служби евалуације перплексности). Даље следе композиције модела у служби генерисања текста и композиције засноване на вештачким неуронским мрежама.

Девето поглавље је посвећено евалуацији. Прво се описује припрема скупова за евалуацију, потом процес евалуације и резултати кроз: 1) детекцију уклоњене, уметнуте и замењене речи, 2) детекцију семантичких и синтаксичких неправилности, 3) моделовање мини-језика и 4) евалуацију композиција заснованих на вештачким неуронским мрежама. На крају поглавља се даје евалуација генеративних граматика.

Десето поглавље нуди дискусију постигнутих резултата, наводи се примена и значај, као и завршне напомене и обриси будућих планова за унапређење развијених модела.

3 ОЦЕНА ДИСЕРТАЦИЈЕ

3.1 Савременост и оригиналност

Разматрана докторска дисертација представља оригинални научно-истраживачки рад у научној области Интелигентни системи и ужим областима (Обрада природног језика и Рачунарска лингвистика) и обрађује врло актуелан проблем везан за моделирање језика. За развој композитних псеудограматика заснованих на паралелним језичким моделима српског језика коришћене се савремене методе вештачке интелигенције које се односе на обраду великих количина текстуалних података.

У оквиру ове докторске дисертације дефинисан је модел и развијено софтверско решење за имплементацију композитних интелигентних система заснованих на паралелним језичким моделима, који могу да се користе као псеудограматике природног језика. Систем заснован на спрези појединачних језичких модела, базиран на знању похрањеном у корпусима текстова и електронским речницима, решава специфичне задатке за стилометрију, одређивање ауторства, тагирање врстом речи, лемама и граматичким категоријама. Коришћени су савремени алати и технике, као што су обучавање језичких модела заснованих на принципима дубоког учења, и потом њихово обједињавање, коришћењем система заснованих на правилима, али и хеуристикама, у виду плитких и конволуционих неуронских мрежа, уз развој различитих електронских ресурса првенствено у виду великих колекција текста неопходних за изградњу језичких модела.

Савременост и оригиналност истраживања приказаног у овој докторској дисертацији потврђени су и публиковањем радова у међународним часописима и саопштењима на домаћим и међународним скуповима.

3.2 Осврт на референтну и коришћену литературу

У оквиру докторске дисертације цитирано је 149 литературних навода, који су углавном новији радови објављени у часописима међународног значаја. Преглед литературних података омогућио је да се прикаже стање у испитиваној научној области, као и да се сагледа актуелност проблематике предметне докторске тезе. Кандидат је првенствено прегледао обимну литературу која је везана за проучавање развоја разноврсних језичких модела, још од деведесетих година прошлог века, као и новију литературу о могућности спреге паралелних језичких модела и генеративних предобучених трансформера, што представља прилично актуелан правац истраживања последњих година. Из пописа литературе која је коришћена у истраживању, као и објављених радова кандидата може се запазити да кандидат на адекватном нивоу познаје област истраживања, као и актуелно стање истраживања у овој области у свету. Избор литературе и приступа који су коришћени за истраживање показује да се кандидат самостално бави научним радом али и дисеминацијом резултата.

3.3 Опис и адекватност примењених научних метода

У докторској дисертацији коришћене су опште и посебне методе истраживања: дескриптивна метода која укључује описивање, прикупљање и систематизацију података, компаративна и аналитичка истраживачка метода која подразумева упоређивање, вредновање и интерпретацију добијених резултата и анализу података из истраживања других аутора, као и методе статистичке обраде података.

У раду је разматран проблем моделирања језика, где је посебна пажња посвећена коришћењу већег броја расположивих језичких ресурса, укључујући корпусе текстова, и претходно припремљене лексичке ресурсе (попут морфолошких речника) или различите технологије моделирања. Имајући у виду растући број језичких ресурса, композитни модел је изразита прилика за њихово обједињавање и заједничку употребу. Примењена методологија укључује неколико корака у циљу изградње једног оваквог система. Предложени развој композитних псеудограматика је заснован на развоју паралелних језичких модела српског језика, што је захтевало коришћење различитих научних метода које наводимо кроз фазе истраживања.

У првој фази је преовладала дескриптивна метода која укључује описивање, прикупљање и систематизацију и хармонизацију података, у овом случају велике колекције неанотираних и анотираних текстова. Компаративна и аналитичка истраживачка метода је коришћена за упоређивање, вредновање и интерпретацију добијених резултата и анализу података из истраживања других аутора.

Најзначајније научне методе друге фазе су биле методе корпусне и рачунарске лингвистике, у комбинацији са методама вештачке интелигенције примењеним у аутоматском анотирању корпуса. Кандидат је потом применио најсавременије методе за развој језичких модела, попут дубоког учења у комбинацији са савременим софтверским решењима. За развој композитних система су примењене методе вероватноће и статистике као и методе машинског учења, претежно вештачке неуронске мреже.

У финалној фази, евалуацији развијених модела кандидат је користио методе аутоматске и ручне евалуације, као и статистичке методе. Резултати представљени у раду су праћени квантитативном и квалитативном анализом добијених резултата.

3.4 Применљивост остварених резултата

Резултати докторске дисертације кандидата Михаила Шкорића су значајни у научном смислу, али имају и велику практичну примену у рачунарској лингвистици и генерално у језичком инжењерству. Постављени задаци и остварени резултати у виду развоја модела и софтверског решења за имплементацију композитних интелигентних система заснованих на паралелним језичким моделима, применљиви као псеудограматике природног језика, усмерени су на решавање конкретних проблема из обраде текста на српском, дајући нарочито значајан допринос у техникама за генерисање текста на српском језику. На тај начин, дисертација пружа значајан научни допринос првенствено у области рачунарске лингвистике, систематизујући у спрегнуте језичке моделе информације екстраховане из прикупљеног корпуса текстова. Кандидат је у дисертацији препознао значај овог проблема, који је присутан и у обради српског језика, и у циљу његовог превазилажења развио је интегрални модел са информатичком подршком.

У практичном смислу развијени модел пружа ширу и потпунију слику од тренутно актуелних начина обраде, анотације и генерисања текста на српском, чиме олакшава креирање практичних решења у пословном процесу. Практична верификација развијеног модела и софтверског решења извршена је кроз евалуацију на задацима аутоматске детекције грешака у тексту, детекције синтаксички неисправних реченица (према облицима речи или према редоследу речи), семантички неисправних реченица, детекције неисправних реченица уопштено, као и разликовање (хуманих) експертских и машинских превода.

Развијене псеудограматике ће, дакле, наћи примену у решавању различитих проблема у обради природног језика, јер се, попут формалних граматика, и оне могу користити у морфосинтаксичкој анализи текста, за генерисање новог текста, као и за прецизније израчунавање сличности између текстова. Осим тога, представљаће и систем за евалуацију којим ће моћи да се одреди квалитет улазног текста. С обзиром на то да су машинско превођење, аутоматско генерисање текста и одговарање на упите већ умногоме аутоматизовани, одговарајућа аутоматска евалуација би додатно побољшала, а уз адекватну примену, и убрзала и олакшала те задатке. Осим тога, аутоматско проналажење (и потенцијално исправка) неквалитетног текста била би још једна могућа примена оваквог система, под условом да се он додатно усаврши.

3.5 Оцена достигнутих способности кандидата за самостални научни рад

Кандидат, Михаило Шкорић, је током израде докторске дисертације показао самосталност, систематичност и стручност у сагледавању проблема истраживања и критичке анализе добијених података. Током примене различитих аналитичких метода, обраде резултата и њихове презентације у објављеним радовима показао је да влада знањима везаним за област истраживања и методама научног рада. Осим тога, кандидат је успешно и квалитетно одговорио на циљеве и истраживачка питања постављене у предлогу ове дисертације, што указује на његову способност да објективно и у целини сагледа истраживачки процес и услове потребне за његову реализацију. Комисија сматра да кандидат поседује све квалитете који су неопходни за самосталан научни рад.

4 ОСТВАРЕНИ НАУЧНИ ДОПРИНОС

4.1 Приказ научних доприноса

Током истраживања остварени су очекивани научни доприноси:

- **Проширење Корпуса савременог српског језика**: групни рад на објављивању најновије допуне корпуса савременог српског језика, *СрпКор2021*, корпуса старих српских романа, *СрпЕЛТеК*.
- **Развој новог софтвера**[1] **отвореног кода за аутоматску анотацију корпуса, на основу прикупљених предобележених примера**, заснованог на паралелној употреби предобучених модела за анотацију који је приликом евалуације остварио најбоље резултате међу свим тестираним системима за српски језик.
- **Развој три савремена језичка модела српског језика на основу репрезентација Корпуса савременог српског језика** у виду предобучених генеративних трансформера друге генерације (ГПТ-2), обучена над трансформацијама текста заснованим на верзији корпуса проширеном врстама речи, ширим граматичким категоријама и лемама. Развијени модели су отворени за јавност и доступни на сајту заједнице *huggingface*[2], као и путем Пајтон пакета *transformers* под истим именима.
- **Развој детаљног модела композитног система за паралелно обједињавање креираних модела (укључујући и будуће моделе) и креирање псеудограматика српског језика које ће имати примену у задацима обраде природног језика, укључујући класификацију и евалуацију докумената, као и генерисање текста** (149 различитих композитних система: 122 за евалуацију квалитета, 20 за проналажење грешака у тексту и 7 за генерисање нових реченица).
- **Јавно доступна веб апликација**[3] **отвореног кода**[4] која демонстрира генерисање текста на основу креираних модела, те евалуацију унетог текста коришћењем припремљених граматика, као и његову трансформацију у векторе перплексности.
- **Тестирања и валидације на постојећим, претходно истраженим проблемима**. Све креиране композиције тестиране су на познатим проблемима попут детекције неисправних реченица и њиховој класификацији.

4.2 Критичка анализа резултата истраживања

Сагледавањем циљева и постављених хипотеза преточених у истраживачка питања, у односу на добијене резултате, може се констатовати да приказана истраживања у потпуности дају одговоре на постављена истраживачка питања и задовољавају критеријуме једне докторске дисертације. Увидом у доступну литературу из ове области, као и резултате изложене у дисертацији, може се констатовати да су коришћене методе у складу са

---

[1] https://github.com/procesaur/BEaSTagger, приступљено 2. марта 2023.

[2] https://huggingface.co/procesaur/gpt2-srlat, https://huggingface.co/procesaur/gpt2-srlat-sem, https://huggingface.co/procesaur/gpt2-srlat-synt, приступљено 2. марта 2023.

[3] https://plma.jerteh.rs, приступљено 2. марта 2023.

[4] https://github.com/procesaur/Parallel-language-models, приступљено 2. марта 2023.

савременим методама и да су резултати у овој докторској дисертацији значајни са научног аспекта.

Највећи допринос у смислу потпуно нових решења огледа се у обучавању синтаксичких и семантичких генеративних језичких модела, као и у употреби мере перплексности при њиховој комбинацији. Синтаксички језички модел је, пре свега, показао велику дискриминациону вредност на задатку разликовања синтаксички и семантички неисправних реченица, али и играо улогу у успеху композитних модела. Комбиновања на нивоу одаслате перплексности омогућава композицију не само модела исте бранше, већ било којих који производе перплексност или вероватноће, независно од типа и димензија, па би, на пример, омогућила комбинације *GPT* и *BERT* модела, или чак комбинације језичких модела са онима који нису језички.

4.3 Верификација научних доприноса

Научни допринос и резултати истраживања добијени током израде ове дисертације верификовани су у једном раду који је објављен у међународном часопису истакнутих вредности, са докторандом као првопотписаним, и другим радом у истакнутом међународном часопису, где је допринос докторанда једнак са првопотписаним аутором.

M21a - Рад у међународном часопису изузетних вредности

Mihailo Škorić, Ranka Stanković, Milica Ikonić Nešić, Joanna Byszuk, Maciej Eder, "Parallel Stylometric Document Embeddings with Deep Learning Based Language Models in Literary Authorship Attribution", Mathematics (2022), ISSN: 2227-7390, MDPI AG. DOI 10.3390/math10050838 [IF=2.592 (2021)]

M22 - Рад у истакнутом међународном часопису

Ranka Stanković, **Mihailo Škorić**, Branislava Šandrih Todorović, "Parallel Bidirectionally Pretrained Taggers as Feature Generators", Applied Sciences (2022), ISSN: 2076-3417, MDPI AG. DOI 10.3390/app11072892 [IF=2.838 (2021)]

M33 - Саопштење са међународног скупа штампано у целини

Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, Duško Vitas, **Mihailo Škorić**, Milica Ikonić Nešić, "Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection", Proceedings of the Language Resources and Evaluation Conference, June 2022, Marseille, France (2022), European Language Resources Association.

Milica Ikonić Nešić, Ranka Stanković, Christof Schöch and **Mihailo Škorić**, "From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back)", Proceedings of The 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference, June 2022, Marseille, France (2022), European Language Resources Association.

M53 - Рад у националном часопису

Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, **Mihailo Škorić**, "Annotation of the Serbian ELTeC Collection", Infotheca 21/2 (2021), Faculty of Philology, University of Belgrade. DOI 10.18485/infotheca.2021.21.2.3

4.4 Провера оригиналности докторске дисертације

Оригиналност докторске дисертације проверена је на начин прописан Правилником о поступку провере оригиналности докторских дисертација које се бране на Универзитету у Београду (Гласник Универзитета у Београду, бр. 204/22.06.2018). Помоћу програма "iThenticate" утврђено је да количина подударања текста по параметру Индекса сличности износи 3%. Овај степен подударности последица је: имена променљивих које се често користе у формулама, имена параметара за тренирање модела са платформе *HuggingFace*), цитата и библиографских података о коришћеној литератури, тзв. општих фраза: „као што је већ наведено", „с обзиром да то да су", „Универзитет у Београду". Стога сматрамо да је утврђено да је докторска дисертација Михаила Шкорића у потпуности оригинална, као и да су у потпуности испоштована академска правила цитирања.

5 ЗАКЉУЧАК И ПРЕДЛОГ КОМИСИЈЕ ЗА ОДБРАНУ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Докторска дисертација кандидата Михаила Шкорића, мастер професора језика и књижевности, под називом „Композитне псеудограматике засноване на паралелним језичким моделима српског језика", представља оригинални научни рад из ужих научних области Обрада природног језика и Рачунарска лингвистика, научне области Интелигентни системи и представља савремен, оригиналан и научно утемељен приступ решавању проблема моделирања српског језика. Комисија сматра да је кандидат Михаило Шкорић у својој дисертацији успешно обрадио ову комплексну и изузетно значајну тему, да је текст дисертације урађен према одобреној пријави дисертације, и да је реч о раду који представља оригинално и самостално научно дело. Скуп језичких модела, развијених у оквиру ове дисертације, представља значајан научни и практични допринос са становишта актуелних потреба језичког инжењерства за српски језик.

Резултати изложени у овој дисертацији показују да је кандидат Михаило Шкорић остварио циљеве постављене у пријави дисертације. У оквиру свог истраживачког рада на дисертацији, кандидат је допунио корпусе текстова на српском језику, који су коришћени у свим фазама истраживања и који остају као вредан ресурс за даља истраживања у области израде језичких модела. Уз то, изграђена софтверска платформа, као и језички модели развијени током истраживачког рада, представљају свакако значајан помоћни ресурс како за академску заједницу, тако и за индустрију која све више показује интересовање за језичке технологије. Кандидат је детаљно описао развијене језичке ресурсе и моделе, семантичка и синтаксичка проширења и начине комбиновања модела.

На основу резултата свог истраживања кандидат је доказао да је употреба композитних система заснованих на паралелним (језичким) моделима и њиховим пробабилистичким производима адекватна метода у области обраде природних језика, а поготово српског језика, што је показано на задацима евалуације, класификације и генерисања текста. Предложен је најподобнији метод комбиновања излаза језичких метода, наслагани класификатор заснован на израчунатим вредностима перплексности, векторима перплексности, те наслаганим перцептронима и конволуционим неуронским мрежама.

Кандидат је у дисертацији дао детаљан опис изграђених модела и софтвера, чиме је значајан део резултата својих истраживања, заједно са изузетно значајним ресурсима и алатима ставио на располагање другим истраживачима у овој области. Тиме је отворио ново поље

истраживања у области развоја и коришћења језичких модела за српски језик. Сам текст дисертације, као и списак литературе наведен на крају рада, показују да је Михаило Шкорић користио релевантну савремену литературу, те да је постављене проблеме обрадио детаљно, сагледавајући их из више углова.

Комисија за оцену и одбрану докторске дисертације са задовољством констатује да је докторска дисертација Михаила Шкорића од великог научног значаја, имајући нарочито у виду да се применом предложеног модела на практично применљив начин може остварити значајно унапређење у решавању различитих проблема у обради природног језика, а добијени модели се могу користити у морфосинтаксичкој анализи текста, за генерисање новог текста, као и за прецизније израчунавање сличности између текстова или евалуацију њиховог квалитета.

Комисија закључује да урађена докторска дисертација представља значајан и оригинални научни допринос у области рачунарске лингвистике и обраде природног језика, да је у свему израђена у складу са свим стандардима о научно-истраживачком раду, као и да испуњава све услове предвиђене Законом о високом образовању, Стандардима за акредитацију и критеријумима које је прописао Универзитет у Београду.

Комисија, на основу укупне оцене дисертације и горе наведеног, предлаже Већу за студије при Универзитету у Београду да докторску дисертацију под називом „Композитне псеудограматике засноване на паралелним језичким моделима српског језика“ кандидата Михаила Шкорића и овај извештај прихвати, дисертацију изложи на увид јавности и упути на коначно усвајање Већу научних области техничких наука Универзитета у Београду ради коначног усвајања, након чега би се приступило усменој одбрани дисертације пред комисијом у истом саставу.

У Београду 10. марта 2023. године

ЧЛАНОВИ КОМИСИЈЕ:

др Владан Девеџић, редовни професор
Универзитет у Београду, Факултет организационих наука

др Милош Утвић, доцент
Универзитет у Београду, Филолошки факултет

др Драган Станков, ванредни професор
Универзитет у Београду, Рударско-геолошки факултет