

**Образац 1 – Пријава теме докторске дисертације
(Кандидат доставља пријаву Програмском савету)
Студијски програм: Интелигентни системи
(назив студијског програма)**

Кандидат: Саша Петалинкар, мастер математичар

Молим да ми се одобри тема за израду докторске дисертације под насловом:

**„Машинско учење и велики језички модели у развоју семантичких мрежа и
њиховој примени на аутоматско разумевање текста“**

1. Биографија кандидата

Саша Петалинкар рођен је 24. децембра 1979. године у Бору, где је и одрастао и завршио основну и средњу школу. Године 1998. уписао је основне академске студије на Математичком факултету Универзитета у Београду, при Катедри за Рачунарство информатику, где је дипломирао 2009. године. Мастер академске студије завршио је на истом факултету 2012. године одбраном тезе под насловом „Обрада елемената логичког изгледа текста под системом UNITEX“. 2018 године уписао је докторске студије при Универзитету у Београду, модул Интелигентни системи.

Активан је члан Друштва за језичке ресурсе и технологије – ЈеРТех – где учествује у развоју система и алата за обраду српског језика.

2. Библиографија кандидата (категорисано према категоризацији надлежног Министарства)

2.1. Објављени радови или прихваћени за штампу

I. Саопштења са међународног скупа штампана у целини (M33):

- Ikonić Nešić, M., Petalinkar, S., Škorić, M., Stanković, R., & Rujević, B. (2024, September). Advancing Sentiment Analysis in Serbian Literature: A Zero and Few-Shot Learning Approach Using the Mistral Model. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)* (pp. 58-70).
<https://dr.rgf.bg.ac.rs/s/repo/item/8805>

- Ikonić Nešić, M., **Petalinkar, S.**, Stanković, R., Utvić, M., & Kitanović, O. (2024, September). SrpCNNeL: Serbian Model for Named Entity Linking. In *Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)* (pp. 465-473). IEEE. <https://doi.org/10.15439/2024F8827>
- Ikonić Nešić, M., **Petalinkar, S.**, Škorić, M., & Stanković, R. (2024, March). BERT Downstream Task Analysis: Named Entity Recognition in Serbian. In *Conference on Information Technology and its Applications* (pp. 333-347). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-71419-1_29

II. Рад у врхунском часопису националног значаја (M51):

III. Радови у националном часопису (M53):

IV. Саопштење са скупа националног значаја штампано у целини (M63)

3. Предмет и циљеви докторске дисертације

3.1. Предмет докторске дисертације (максимално 1 страна)

Семантичке мреже, попут ворднета (Miller 1995), представљају значајне ресурсе у лексикографији и обради природног језика, а нарочито су погодне за аутоматско разумевање текста. EuroWordNet је поставио основу за мултијезичке семантичке мреже (Fellbaum, 2005; Vossen, 2004), са јединственим идентификатором концепта у свим језицима: Inter-Lingual Index (ILI) који је уједно као стожер усвојен у свим каснијим пројектима. Српски Ворднет (SrpWN) креиран је у оквиру BalkaNet пројекта (Tufis et al., 2004), који је усвојио шему пројекта EuroWordNet. SrpWN верзија из 2018. године је садржала 25,320 синсета, међу којима је осим мапираних са Принстонског ворднета (PWN) такође налазе и они за значењима карактеристичних за подручје који су развијени у оквиру BalkaNet пројекта (Krstev et al. 2004; Stanković et al. 2018) Вредности сентимента (Liu 2010) су такође мапиране са SentiWordNet-a (SWN) (Baccianella, Esuli, and Sebastiani 2010), једног од најпопуларнијих лексикона са вредностима сентимента. Пошто је SWN базиран на PWN, синсетима који су карактеристични за подручје нису додељене вредности сентимента. Вредности сентимента из SrpWN коришћене су за анализу сентимента и откривање скривених значења (Mladenović et al., 2015).

Иако SrpWN представља значајан ресурс у рачунарској лингвистици за решавање различитих задатака за српски језик, његова примена је ограничена због величине, јер је око 4 пута мањи од PWN. Проширење ове семантичке мреже није једноставан задатак и захтева значајне људске ресурсе, јер он спада у ред ресурса који се ручно развијају (Krstev et al. 2004). Било је и раније покушаја аутоматизације изградње (Stanković et al. 2018), али је предложени процес и даље захтеван, те је потреба за ефикаснијом аутоматизацијом и даље неопходна. Проблем који није до сада решаван за српски језик је разрешавање вишезначности, што је кључно за примену SrpWN у аутоматском разумевању текста.

Речи могу имати једно или више значења, а у ворднету је свако значење представљено посебним синсетом, па тако једној речи може одговарати више синсета. Зато је неопходно да се свакој речи у тексту придружи одговарајуће значење. За решавање такве врсте проблема често се користе модели векторског простора (eng. Vector Space Models) (Camacho-Collados and Pilehvar 2018). Речима у тексту и синсетима додаје се векторска репрезентација, тако да је вектор речи најближи вектору синсета који најбоље описује њено значење.

Ова докторска дисертација истражује могућности примене метода машинског учења и великих језичких модела (BJM) (Grattafiori et al. 2024; Devlin et al. 2019; Mihailo Škorić 2024) за аутоматизацију развоја и унапређења SrpWN-a, као и коришћене исте за аутоматско разумевање текста. Посебан фокус је на развоју алата за аутоматску анотацију текста који обележава речи синсетима SrpWN-a (Orkphol and Yang 2019), као и на истраживању потенцијала проширења ворднета кроз интеграцију информација из PWN и других лексичких извора. Такође се предлаже и поравнање са описним речницима према значењу (eng. word sense alignment) (Oliver 2020), што би омогућило даљу примену SrpWN. Циљ овог истраживања је да допринесе не само лингвистичкој науци већ и применама у интелигентним системима, обради текста и образовању.

3.2. Циљеви докторске дисертације (максимално 1 страна)

Општи циљ ове дисертације је развој методологије за аутоматизацију, примену и унапређење семантичких мрежа, са фокусом на Српски Ворднет. Ова методологија обухвата ревизију постојећих вредности сентимента, проширење синсета новим лексичким јединицама и дефиницијама, интеграцију спољних ресурса, као и препознавање значења речи у тексту.

Специфични

циљеви

укључују:

- Развој аутоматског анотатора текста заснованог на ВЈМ и методама машинског учења, који ће омогућити повезивање речи са одговарајућим значењем те речи представљено синсетом.
- Замена постојећих вредности сентимента, добијених мапирањем са SWN, вредностима добијених машинским учењем
- Примена савремених ВЈМ за аутоматизацију прилагођавања синсета са PWN.
- Испитивање могућности примене генеративних модела за допуну SrpWN
- Испитивање примене обogaћеног SrpWN-a у задацима анализе сентимента, класификације текста и машинског превођења.
- Лексичко поравнавање SrpWN са описним речником за српски језик.

Циљ је креирање ефикасног и надоградивог система који интегрише постојеће и будуће језичке моделе.

3.3. Хипотезе

Основна хипотеза је да се применом метода машинског учења и ВЈМ могу аутоматизовати и унапредити процеси изградње семантичких мрежа, као и омогућити аутоматско обележавање текста чворовима те мреже, при чему ће обogaћени SrpWN имати боље перформансе од тренутно доступне верзије у задацима повезивања, претраживања и класификације текста.

Испитане претпоставке:

- Применом метода машинског учења се могу добити вредности сентимента који више одговарају српском језику него мапиране вредности са SWN-a

Претпоставке које су у фази истраживања су:

- У моделу векторског простора постоји таква векторска репрезентација да се њом може одреди значење речи, то јест придружити одговарајући синсет, како за реч у тексту, тако и за појединачно значење у описном речнику
- ВЈМ се могу користити за допуну синсета (на пример, аутоматско креирање примера или проналажење синонима), као и за прилагођавање

са енглеских синсета (будући да директан машински превод неће увек захватити све језичке разлике).

- SrpWN унапређен овим методама даје боље резултате у низводним НЛП задацима (на пример одређивање сентимента, детекција лажних вести, повезивање именованих ентитета са базом знања) него постојећа верзија.

4. План рада

Истраживање у оквиру предложене докторске дисертације извршиће се у три фазе и састојаће се од следећих корака:

I. Фаза припреме

- а. Изучавање литературе претходних релевантних истраживања и утврђивање теоријског оквира
- б. Истраживање најсавременијих технологија примене метода машинског учења и ВЈМ на проблеме разрешавања вишезначности речи, одређивање сентимента и генерисања текста
- с. Преглед постојећих ресурса и технологија који користе SrpWN за тражене задатке

II. Фаза развоја

- а. Припрема или проналажене аотираних и паралелних корпуса за задатке анализе сентимента, класификације текста и машинског превођења.
- б. Развијање новог софтверског решења за програмски приступ SrpWN
- с. Избор подскупа синсета за обуку за моделирање сентимента
- д. Развијање новог софтверског решења који ће омогућити фино подешавање модела за одређивање сентимента у SrpWN, и његову примену;

- e. Развијање новог софтверског решења који ће омогућити повезивање речи са одговарајућим значењем те речи представљено синсетима
- f. Припрема скупа синсета на енглеском и српском за прилагођавање и допуну
- g. Испитивање могућности допуне и прилагођавања синсета помоћу генеративних модела
- h. Припрема скупа који садржи део описног речника и одговарајуће синсете из SrpWN за поравнање као и испитивање могућности и метода семантичког поравнања на том скупу
- i. Развијање додатних система по потреби заснованих на SrpWN-а који ће решавати специфичне, претходно поменуте, задатке.

III. Финална фаза

- a. Евалуација појединачних делова пројекта и избор решења за примену;
- b. Примена изабраног решења на SrpWN да би се добила нова верзија SrpWN;
- c. Компаративна анализа постојећег и добијеног
- d. Дискусија и извођење закључка на основу добијених резултата.

5. Методе које се користе у истраживању (максимално 1 страна)

За решавање постављених проблема у овом раду ће се користити:

I. У првој фази ће преовлађивати дескриптивна метода за прикупљање, систематизацију и хармонизацију литературе, ресурса и постојећих софтверских решења. Биће примењене компаративна и аналитичка метода за упоређивање и интерпретацију резултата из претходних истраживања, уз основну статистичку обраду података.

II. У другој фази ће се примењивати методе рачунарске лингвистике за анотацију корпуса, у комбинацији са методама машинског учења (укључујући дубоко учење), ради развоја и финог подешавања језичких модела. За развој композитних система ће се

користити статистичке методе и вероватноћа, посебно у задацима анализе сентимента и разрешавања вишезначности. Језички модели ће се затим користити за креирање векторских репрезентација — било узимањем задњег или више слојева модела, било коришћењем упита.

III. У финалној фази, евалуација развијених модела ће се вршити помоћу аутоматске и ручне методе евалуације, уз квантитативну и квалитативну анализу добијених резултата. Применом статистичких тестова значајности и компаративних метода ће се утврђивати побољшање у односу на претходна решења и вршиће се формулисање коначних закључака. За задатке који ће се сводити на класификацију (нпр. поравнање по значењу, обележавање синсета) ће се користити мере прецизност, одзив и хармонијска мера F1, а за задатке који ће се сводити на регресију (нпр. додела вредности сентимента) ће се користити метод максималне ентропије.

Мултидисциплинарност теме (максимално 1/2 стране)

Мултидисциплинарност теме огледа се како у различитим научним методама које ће се користити, тако и у ресурсима који ће се користити и доменама примене добијених резултата. Обрада природног језика представља спој више дисциплина—пре свега рачунарства, лингвистике и статистике—па ће и ово истраживање обухватити методе корпусне и рачунарске лингвистике, потом примену машинског учења (укључујући дубоко учење), као и статистичке приступе. На тај начин, рад доприноси различитим областима: лингвисти добијају проширене ресурсе и моделе за анализу, док истраживачи у рачунарству могу да користе развијене моделе и алате за изградњу интелигентних система. Стога ће резултати истраживања, изузев доприноса научној заједници, имати и практичну вредност у алатима за обраду, анализу и разумевање текстова на српском језику.

6. Очекивани научни доприноси

Истраживање ће, иако усмерено превасходно на развоју семантичких мрежа на примеру SrpWN и њиховој примени на аутоматско разумевање текста, донети методолошке и технолошке доприносе који су применљиви и на остале језике са недовољно развијеним лингвистичким ресурсима. У ширем смислу, то подразумева бољу интероперабилност постојећих семантичких мрежа, као и лакше проширење или ревизију синсета у складу са појавом нових појмова или промена у самим језицима. Поред директне надоградње SrpWN-а—кроз увођење нових дефиниција, корекцију

постојећих семантичких веза и додавање нових лексичких јединица—унапређења ће се огледати у развоју универзалног оквира за аутоматско повезивање текста са релевантним синсетима. Тиме ће се омогућити и лакше интегрисање у друге NLP компоненте (сентимент анализа, машинско превођење, класификација документа), што ће бити корисно како у научним, тако и у практичним апликацијама.

Истовремено, употреба генеративних модела за синтетичко проширење лексичких ресурса унапредиће наше разумевање потенцијала (и ограничења) оваквих модела у процесу обраде природног језика. Посебан значај имаће резултати који демонстрирају у којим случајевима је ово проширење корисно и где се могу јавити проблеми попут непрецизних дефиниција, погрешних контекста или културолошки неодговарајућих примера. Стварање система који аутоматски може да поравна SrpWN са другим лексичким ресурсима—како постојећим, тако и будућим—повећаће трајност и флексибилност резултата, омогућавајући да даљи развој језичких модела буде лако интегрисан у семантичку мрежу. Тако ће методе настале у овом раду служити као чврст темељ за унапређење семантичких мрежа и у другим језицима, нарочито онима који немају богату дигиталну лингвистичку инфраструктуру.

7. Библиографски подаци релевантни за докторску дисертацију (максимално 1 страна)

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. 'SENTIWORDNET 3.0: An Enhanced Lexical Resource'. In 7th LREC, 2200–2204. <http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf>.

Camacho-Collados, Jose, and Mohammad Taher Pilehvar. 2018. 'From Word to Sense Embeddings: A Survey on Vector Representations of Meaning'. *Journal of Artificial Intelligence Research* 63:743–88. <http://glottometrics.iqla.org/wp-content/uploads/2021/06/g45zeit.pdf>

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding'. arXiv. <https://doi.org/10.48550/arXiv.1810.04805>.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. 'The Llama 3 Herd of Models'. <https://arxiv.org/abs/2407.21783>.

Krstev, Cvetana, Duško Vitas, Gordana Pavlović-Lažetić, and Ivan Obradović. 2004. 'Using Textual and Lexical Resources in Developing Serbian Wordnet'. *ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY* 7 (1–2): 147–61.

Liu, Bing. 2010. 'Sentiment Analysis and Subjectivity'. In *Handbook of Natural Language Processing*, Second Edition, edited by Fred J. Damerau and Nitin Indurkha, 629–61. Chapman & Hall/CRC.

Mihailo Škorić. 2024. 'New Language Models for Serbian'. *Infotheca - Journal for Digital Humanities* 24 (1). <https://arxiv.org/abs/2402.14379>.

Miller, George A. 1995. 'WordNet: A Lexical Database for English'. *Communications of the ACM* 38 (11): 39–41. <https://doi.org/10.1145/219717.219748>.

Mladenović, Miljana, Jelena Mitrović, Cvetana Krstev, and Duško Vitas. 2015. 'Hybrid Sentiment Analysis Framework for a Morphologically Rich Language'. *Journal of Intelligent Information Systems* 46 (3): 599–620. <https://doi.org/10.1007/s10844-015-0372-5>.

Orkphol, Korawit, and Wu Yang. 2019. 'Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet'. *Future Internet* 11 (5): 114. <https://doi.org/10.3390/fi11050114>.

Stanković, Ranka, Miljana Mladenović, Ivan Obradović, Marko Vitas, and Cvetana Krstev. 2018. 'Resource-Based WordNet Augmentation and Enrichment'. In *CLIB 2018*, 104–14. Sofia, Bulgaria: The Institute for Bulgarian Language BAS.

Tufis, D., D. Cristea, and S. Stamou. 2004. 'BalkaNet: Aims, Methods, Results and Perspectives. A General Overview'. *Romanian Journal of Information Science and Technology* 7 (1–2): 9–43.

Vossen, Piek. 2004. 'EuroWordNet: A Multilingual Database of Autonomous and Language-Specific WordNets Connected via an Inter-Lingual Index'. *International Journal of Lexicography* 17 (2): 161–73. <https://doi.org/10.1093/ijl/17.2.161>.

8. Изјава да предложеној тему кандидат није пријављивао на другој високошколској установи у земљи или иностранству

Ја, Саша Петалинкар, ЈМБГ 0710992710133, изјављујем под пуном моралном и материјалном одговорношћу да предложеној тему „Аутоматизација развоја и примене семантичких мрежа на примеру Српског Ворднета применом метода машинског учења и великих језичких модела“ нисам пријављивао на другој високошколској установи у земљи или иностранству.

9. Предлог два ментора и комисије за оцену теме докторске дисертације (име, презиме, звање, институција, ужа научна област)

1. Проф. др Ранка Станковић, ванредни професор, Рударско-геолошки факултет Универзитета у Београду (математика и информатика)
2. Проф. др Јелена Граовац, ванредни професор, Математички факултет Универзитета у Београду (рачунарство и информатика).

Прилози:

1. Сагласност ментора (треба да садржи име, презиме, звање, институцију и потпис)
2. Подаци о ментору

У Београду, 10.02.2025.

Подносилац молбе (потпис)


Саша Петалинкар