УНИВЕРЗИТЕТ У БЕОГРАДУ

МАТЕМАТИЧКИ ФАКУЛТЕТ

Наставно-научном већу

На 314. седници Наставно-научног већа Математичког факултета, која је одржана 9.5.2014. године, одређени смо за чланове комисије за писање извештаја о докторској дисертацији

Екстракција информација вођена онтологијама

(Модел за српски језик)

кандидата Сташе Вујичић Станковић. После прегледа рукописа који је кандидат предао комисији, подносимо Наставно-научном већу Математичког факултета следећи

И З В Е Ш Т А Ј

БИОГРАФИЈА КАНДИДАТА

Сташа Вујичић Станковић рођена је у Београду, 1982. године. Завршила је Основну школу „Бранко Ћопић“ као носилац диплома „Вук Караџић“ и „Ђак генерације“. После тога је завршила Математичку гимназију у Београду и Математички факултет у Београду, где је дипломирала на смеру „Рачунарство и информатика“ са просечном оценом 9,39. Током школовања је била носилац стипендије Фонда за младе таленте Министарства просвете Републике Србије.

Од 2007. године запослена је на Математичком факултету Универзитета у Београду, као сарадник у настави на Катедри за рачунарство и информатику, а од 2009. године као асистент у настави. Држала је вежбе из низа предмета на основним и мастер студијама и то: Програмирање 1, Програмирање 2, Објектно

оријентисано програмирање, Информациони системи, Увод у организацију рачунара, Основи управљања и Управљање пројектима у индустрији и науци.

Основне области интересовања су јој обрада природних језика, екстракција информација, базе података и претраживање информација и истраживање веба. Учесник је научних пројеката „Српски језик и његови ресурси: теорија, опис и примене“ и „Инфраструктура за електронски подржано учење у Србији“ које финансира Министарство науке Републике Србије.

Као аутор или коаутор је објавила 21 рад и то: 1 рад у часопису са SCI листе на којем је први аутор (категорија М23), 4 рада у монографијама и тематским зборницима (категорија М14), 9 радова у зборницима међународних скупова од којих је један самосталан (категорија М33), 4 рада у научним часописима од којих је један самосталан (категорија М53) и 3 рада са скупова националног значаја (категорија М63). Присуствовала је зимској школи *PARSEME 1st Training School*, од 19. до 23. јануара 2015, на Универзитету за формалну и примењену лингвистику у Прагу у Чешкој, на летњој школи *EUROLAN 2011 „Natural Language Processing Goes Industrial“*, од 28. августа до 4. септембра 2011, у граду Клуж у Румунији и летњој школи *GATE (General Architecture for Text Engineering)*, одржаној у јулу 2009. године на Универзитету Шефилд у Енглеској.

Од 2014. године члан је и један од оснивача Друштва за језичке ресурсе и технологије. Обавља дужност секретара Семинара Друштва за језичке ресурсе и технологије.

РАДОВИ КОЈИ КВАЛИФИКУЈУ КАНДИДАТА ЗА СТИЦАЊЕ АКАДЕМСКЕ ТИТУЛЕ И НАУЧНОГ ЗВАЊА ДОКТОРА НАУКА

Рад у међународном часопису (категорија М23)

[M23.1] **Vujičić Stanković, S.**, Kojić, N., Rakočević, G., Vitas, D., Milutinović, V. (2013). A Classification of Data Mining Algorithms for Wireless Sensor Networks, and Classification Extension to Concept Modeling in System of Wireless Sensor Networks Based on Natural Language Processing. *Advances in Computers: Connected Computing Environment, 90*, 223-283. ISBN: 978-0-12-408091-1, IF(2013) = 0.489.

Саопштење са међународног скупа штампано у целини (категорија М33)

[M33.1] **Vujičić Stanković, S.** (2012). Named Entity Recognition in the System for Information Extraction. In S. Halupka-Rešetar, M. Marković, T. Milićev, and N. Milićević, editors, Selected Papers from SinFonIJA 3 (pp. 206-223). Newcastle upon Tyne, UK: Cambridge Scholars Publishing. ISBN: 978-1443840804.

БИБЛИОГРАФИЈА ОСТАЛИХ РАДОВА КАНДИДАТА

Монографије и тематски зборници (категорија М14)

[M14.1] **Vujičić Stanković, S.**, Pajić, V. (2015). Upotreba vlastitih imena u kulinarskom domenu. *Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opisi primene, 44/3*, 137-142. ISBN: 978-86-6153-305-1 UDC: 811.163.41'322, 811.163.41'367.622.12.

[M14.2] Pajić, V., **Vujičić Stanković, S.** (2014). Finite State Transducers for Generating Texts of Meteorological Reports in Serbian. In G. Pavlović-Lažetić, C. Krstev, I. Obradović, and D. Vitas, editors, *Natural Language Processing for Serbian: Resources and Applications*, 79-86. ISBN: 978-86-7589-088-1.

[M14.3] **Vujičić Stanković, S.**, Pajić, V. (2014). Formiranje domenskog korpusa – kulinarska leksika. *Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opisi primene, 43*(3), 51-59. ISBN: 978-86-6153-203-0 UDC: 811.163.41'322, 811.163.41'373:641/642]: 004.738.5.

[M14.4] Zečević, A., **Vujičić Stanković, S.** (2014). Language Identification: The Case of Serbian. In G. Pavlović-Lažetić, C. Krstev, I. Obradović, and D. Vitas, editors, *Natural Language Processing for Serbian: Resources and Applications*, 101-112. ISBN: 978-86-7589-088-1.

Саопштење са међународног скупа штампано у целини (категорија М33)

[M33.2] Krstev, C., **Vujičić Stanković, S.**, Vitas, D. (2014). Approximate Measures in the Culinary Domain: Ontology and Lexical Resources. In T. Erjavec and J. Žganec Gros, editors, *Proceedings of the 9th Language Technologies Conference IS-LT 2014* (pp. 38-43). Ljubljana, Slovenia: Institut "Jožef Stefan". ISBN: 978-961-264-077-4.

[M33.3] **Vujičić Stanković, S.**, Krstev, C., Vitas, D. (2014). Enriching Serbian WordNet and Electronic Dictionaries with Terms from the Culinary Domain. *The Proceedings of Seventh Global WordNet Conference* (pp. 127-132). Tartu, Estonia: University of Tartu. ISBN: 978–9949–32–492–7.

[M33.4] Zečević, A., **Vujičić Stanković, S.** (2013). The Mysterious Letter J. *Proceedings of the Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants* (pp. 40-44). Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA. ISBN: 978-954-452-026-7.

[M33.5] **Vujičić Stanković, S.**, Pajić, V. (2012). Information Extraction from the Weather Reports in Serbian. *Proceedings of the Fifth Balkan Conference in Informatics* (pp. 105-108). Novi Sad: Faculty of Sciences, University of Novi Sad. ISBN: 978-86-7031-200-5.

[M33.6] **Vujičić Stanković, S.**, Rakočević, G., Kojić, N., Milićev, D. (2012). A Classification and Comparison of Data Mining Algorithms for Wireless Sensor Networks. *Proceedings of the 2012 IEEE International Conference on Industrial Technology* (pp. 265-270). Athens, Greece: IEEE. ISBN: 978-1-4673-0340-8, DOI:10.1109/ICIT.2012.6209949.

[M33.7] **Vujičić Stanković, S.**, Rakočević, G., Milutinović, V. (2011). A Metadata-Supported Distributed Approach for Data Mining Based Prediction in Wireless Sensor Networks. *10th International Conference on Telecommunication in Modern Satellite Cable and Broadcasting Services (TELSIKS, Volume 1)* (pp. 181-185). Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. ISBN: 978-1-4577-2018-5, DOI: 10.1109/TELSKS.2011.6112030.

[M33.8] **Vujičić, S.**, Vitas, D. (2010). Odonym Recognition in Serbian. In T. Váradi, J. Kuti, and M. Silberztein, editors, *Applications of Finite-State Language Processing: Selected Papers from the 2008 International NooJ Conference* (pp. 152-159). Newcastle upon Tyne, UK: Cambridge Scholars Publishing. ISBN: 978-1-4438-2573-3.

[M33.9] **Vujičić, S.**, Vitas, D., Utvić, M. (2010). Recognition of odonyms in Serbian language. In E. Tomaž, editor, *Proceedings of the Seventh Language Technologies Conference - Proceedings of the 13th International Multiconference INFORMATION SOCIETY - IS 2010. C* (pp. 74-77). Ljubljana, Slovenia: Jožef Stefan Institute. ISBN: 978-961-264-026-2.

Рад у научном часопису (категорија М53)

[M53.1] Pajić, V., **Vujičić Stanković, S.**, Pajić, M. (2015). An Algorithm for Sentence Recovery from PDF Files. *Infotheca: Journal for Digital Humanities, 15*(2), 42-55. UDC: 81'322.2:004.912.

[M53.2] **Vujičić Stanković, S.** (2013). Model sistema za ekstrakciju informacija iz tekstova pisanih na srpskom jeziku. *Info M*, 47/2013, 4–9. ISSN: 1451-4397, UDC: 004.822:519.76.

[M53.3] Pajić, V., Pajić, M., **Vujičić Stanković, S.** (2012). Nov metod ekstrakcije informacija baziran na transduktorima. *Info M, 11*(44), 33-40. ISSN: 1451-4397, UDC: 004.832.2:025.4.

[M53.4] Pajić, V., **Vujičić Stanković, S.**, Pajić, M. (2012). Transducers for Annotating Weather Information in Meteorological Texts in Serbian. *Infotheca: Journal for Digital Humanities*, *13*(2), 36-51. UDC: 811.163.41’322.2 , 004.9:811.111’374.

Саопштење са скупа националног значаја штампано у целини (категорија М63)

[M63.1] **Vujičić Stanković, S.** (2012). Using Natural Language Processing for Data Mining Algorithms in Wireless Sensor Networks: Planning a Weather Forecast Application. *Proceedings of the 18th Symposium on Computer Sciences and Information Technologies YU INFO 2012* (pp. 216–220). Beograd: Društvo za informacione sisteme i računarske mreže. ISBN: 978-86-85525-09-4.

[M63.2] **Vujičić Stanković, S.**, Vitas, D., Rakočević, G., Milutinović, V. (2011). Classification of Data Mining Algorithms and Concept Modeling Approaches in Wireless Sensor Networks. *Proceedings of the 17th Symposium on Computer Sciences and Information Technologies YU INFO 2011* (pp. 375-379). Beograd: Društvo za informacione sisteme i računarske mreže. ISBN: 978-8-6855-2508-7.

[M63.3] **Vujičić Stanković, S.**, Vitas, D., Rakočević, G., Milutinović, V. (2011). Simulator Strategy for Data Mining and Concept Modeling in Wireless Sensor Networks. *Proceedings of the 17th Symposium on Computer Sciences and Information Technologies YU INFO 2011* (pp. 444-447). Beograd: Društvo za informacione sisteme i računarske mreže. ISBN: 978-8-6855-2508-7.

ПРЕДМЕТ ДИСЕРТАЦИЈЕ

Тема докторске дисератције Сташе Вујичић Станковић припада области екстракције информација. Основни циљ њеног рада се састојао у развоју ресурса за српски језик који ће опремити неструктурирани или семи-структурирани текст информацијама за претраживање и класификовање текстова из одређеног домена према сложеним семантичким критеријумима и имплементирању развијених ресурса у систем који омогућава њихову експлоатацију.

Сама област екстракције информација, када је српски језик у питању, је неразвијена, а то је махом случај и са другим словенским језицима. Основни проблем у развоју оваквих система представља богати и високо развијени морфолошки систем словенских језика како на флективном, тако и на деривационом нивоу.

Рад је усмерен првенствено на методе екстракције и обележавања именованих ентитета, њихову нормализацију и њихово повезивање одговарајућим семантичким релацијама.

За домен на коме ће се вршити истраживање је изабран подјезик кулинарства који, као језички феномен, није истраживан у својој свакодневној и синхроној употреби. Овај подјезик обилује специфичним класама именованих ентитета, посебно када су у питању ентитети који описују мере, и посебном употребом специфичне лексике у описивању релација међу ентитетима. Напоменимо да традиционална српска лексикографија у највећој мери занемарује опис употребе лексике овог подјезика. Такође, уобичајене класе именованих ентитета су морале бити проширене на ентитете специфичне за овај подјезик, а посебно када су у питању мере и састојци.

У решавању постављеног проблема било је потребно извршити анализу и адаптацију програмских система који се могу применити на решавање проблема екстракције, пре свега, именованих ентитета, као и развити одговарајуће дигиталне ресурсе попут одговарајућих електронских речника, семантичких мрежа и локалних граматика. Кандидат се определио за надградњу система Unitex, дајући му предност над системом GATE, због високе флексибилности у управљању језичким ресурсима. Полазећи од овог опредељења, кандидат је извршио анализу лексике изабраног доменског корпуса, идентификовао лексику

коју је потребно описати за рад на екстракцији информација и опремио је семантичким маркерима који омогућавају да се аутоматски препознају одређене класе објеката као што су намирнице, полупроизводи, производи, итд. Сва лексика анализираног домена има генерички семантички маркер Culinary. За овако идентификовани лексички фонд, кандидат је саставио електронски морфолошки речник који исцрпно описује у тзв. LADL-формату флективна својства лексике. С друге стране, кандидат је у српску верзију семантичке мреже WordNet уградио синсетове који описују лексику исхране и кулинарства. Овај корак обезбеђује потенцијалну примену развијеног система у анализи вишејезичних садржаја. Користећи се овим резултатима, кандидат је развио више онтологија које ће користити у екстракцији информација.

Применом изграђених ресурса, кандидат је обележио лексику доменског корпуса додељујући његовим елементима нормализовани облик (лему) и семантичке етикете. Посебно је значајно решење за руковање приближним или неодређеним мерама.

У даљем раду, кандидат је изградио релациону базу података на основу доменског корпуса и платформу која омогућава претраживање његовог садржаја према сложеним критеријумима на основу изграђених ресурса за српски језик. Овај корак подразумева да ће се колекција неструктурираних или семи-структурираних текстова конвертовати у претраживи корпус, у коме су именовани ентитети обележени и додељени им семантички атрибути, а претраживање се ослања на екстраховане информације и вођено је уграђеним онтологијама.

ПРИКАЗ ДИСЕРТАЦИЈЕ

Докторска дисертација Сташе Вујичић Станковић „Екстракција информација вођена онтологијама (Модел за српски језик)“ у 236 страна текста садржи 5 поглавља, 6 прилога, 48 слика и 7 табела, као и библиографију са 251 библиографском јединицом.

У првом поглављу су дата уводна разматрања везана за проблем обраде улаза на природном језику, као и општа разматрања проблема екстракције информација и метода које се користе за њихово решавање. Назначени су мотиви

за избор кулинарског подјезика као домена на коме ће се извршити истраживање. Представљена је формализација основног објекта истраживања, текста у електронском облику, и описане апроксимације природних језика које се користе у његовој обради. У завршном делу увода посебна пажња посвећена је онтологијама и WordNet-у, и истакнут значај његове употребе као онтологије.

У другој глави су представљени језички ресурси и алати који су коришћени у дисертацији. Описани су морфолошки речници и локалне граматике који се користе за решавање проблема екстракције информација из текстова на српском језику. Дат је преглед система за екстракцију информација и описан ток обраде текстова на српском језику при решавању задатка екстракције информација у програмским системима Unitex и GATE.

У трећој глави је представљен главни резултат дисертације, модел за решавање проблема екстракције информација интегрисањем језичких ресурса и алата, који обухвата формирање корпуса текстова, дефинисање задатака екстракције информација, изградњу коначних модела за екстракцију информација, примену развијених коначних модела, итеративну доградњу морфолошких електронских речника, проширење WordNet-а и изградњу нових онтологија. Детаљно је описан сваки од ових корака. Иако је модел првенствено разматран из угла решавања проблема који се јављају при обради српског језика, може бити примењен и за обраду текстова на другим језицима уз развој адекватних језичких ресурса.

Имплементација описаних корака приказана је у четвртом поглављу кроз систем за екстракцију информација из текстова кулинарског домена на српском језику. Описана је спрега при развоју и међусобној допуни доменских лексичких ресурса кроз кораке формирања доменског корпуса, препознавања кулинарске лексике, проширења и доградње WordNet-а и морфолошких електронских речника. Такође је описан развој доменске онтологије хране, доменске онтологије састојака који могу да се употребе као међусобне замене у кулинарском домену и доменске онтологије приближних мера у кулинарском домену. Развијени систем за екстракцију информација је послужио за реализацију система за напредно претраживање рецепата. Још један од доприноса дисертације јесте примена развијених онтологија у задацима конвертовања приближних кулинарских мера у стандардне мере и утврђивања сличности између рецепата. Сличност рецепата је

дефинисана као сличност текстова који описују поступак припреме јела према одређеном рецепту.

У последњој глави су приказани закључци и правци даљег рада.

НАУЧНИ ДОПРИНОС КАНДИДАТА

Главни научни доприноси у приложеном раду су:

- Развој коначних трансдуктора и локалних граматика за екстракцију именованих ентитета који припадају кулинарском домену (назива намирница, кулинарских производа, јела, посебно оних у чијој структури учествују властита имена, оброка, пића, кухињског прибора, приближних мера које се користе у кулинарском домену)
- Креирање доменског корпуса текстова кулинарских рецепата
- Допуна електронских морфолошких речника кулинарском лексиком
- Доградња WordNet-а синсетима који припадају кулинарској лексици
- Нов модел за екстракцију информација вођену онтологијама који поред решавања задатака екстракције информација обухвата и изградњу нових и доградњу постојећих лексичких ресурса и онтологија
- Креирање доменских онтологија хране, састојака који могу да се употребе као међусобне замене у кулинарском домену и приближних мера које се користе у кулинарском домену
- Изградња система за екстракцију информација из кулинарског домена и система за напредно претраживање рецепата
- Примена развијених онтологија у задацима конвертовања приближних кулинарских мера у стандардне мере
- Развој метода за успостављање веза између текстова сродних и сличних рецепата

ЗАКЉУЧАК И ПРЕДЛОГ

У рукопису „Екстракција информација вођена онтологијама (Модел за српски језик)“ кандидат Сташа Вујичић Станковић је показала познавање области

и проблема екстракције информација, као и истраживачку способност да своје резултате стави у функцију решавања актуелних информатичких проблема. Развијен је модел екстракције информација на српском језику и дата његова имплементација. Модел је унапређен интеграцијом онтологија у процес екстракције. Кандидат је кроз овај рад дао теоријски, методолошки и практични допринос решавању проблема екстракције информација на српском језику као и широј области истраживања текста.

Предлажемо стога Наставно-научном већу Математичког факултета да рукопис "Екстракција информација вођена онтологијама (Модел за српски језик)" прихвати као докторску дисертацију и одреди комисију за јавну одбрану.

У Београду, 8.6.2016.

Чланови комисије:

проф. др Душко Витас, ментор
ванредни професор Математичког факултета Универзитета у Београду

_____

проф. др Гордана Павловић-Лажетић
редовни професор Математичког факултета Универзитета у Београду

_____

доц. др Весна Пајић
доцент Пољопривредног факултета Универзитета у Београду

_____