УНИВЕРЗИТЕТ У НОВОМ САДУ **ОБРАЗАЦ 6.**

ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА

ИЗВЕШТАЈ О ОЦЕНИ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

I ПОДАЦИ О КОМИСИЈИ

1. Датум и орган који је именовао комисију

   Решење декана Факултета техничких наука у Новом Саду, бр. 012-199/79-2017 од 19.07.2018.

2. Састав комисије са назнаком имена и презимена сваког члана, звања, назива уже научне области за коју је изабран у звање, датума избора у звање и назив факултета, установе у којој је члан комисије запослен:

   1. **Др Владо Делић**, редовни професор, **председник комисије**,
      УНО: Телекомуникације и обрада сигнала, (28.03.2013.),
      Универзитет у Новом Саду, Факултет техничких наука, Нови Сад
   2. **Др Драгана Бајић**, редовни професор, **члан комисије**,
      УНО: Телекомуникације и обрада сигнала, (15.06.2006.),
      Универзитет у Новом Саду, Факултет техничких наука, Нови Сад
   3. **Др Снежана Гудурић**, редовни професор, **члан комисије**,
      УНО: Романистика, (21.07.2009.),
      Универзитет у Новом Саду, Филозофски факултет, Нови Сад
   4. **Др Татјана Грбић**, ванредни професор, **члан комисије**,
      УНО: Теоријска и примењена математика, (19.02.2014.),
      Универзитет у Новом Саду, Факултет техничких наука, Нови Сад
   5. **Др Јелена Николић**, доцент, **члан комисије**,
      УНО: Телекомуникације, (08.12.2014.),
      Универзитет у Нишу, Електронски факултет, Ниш
   6. **Др Милан Сечујски**, ванредни професор, **члан комисије и ментор**,
      УНО: Телекомуникације и обрада сигнала, (03.03.2016.),
      Универзитет у Новом Саду, Факултет техничких наука, Нови Сад

II ПОДАЦИ О КАНДИДАТУ

1. Име, име једног родитеља, презиме:

   Стеван, Јоco, Острогонац

2. Датум рођења, општина, држава:

   15.10.1986, Суботица, Србија

3. Назив факултета, назив студијског програма дипломских академских студија – мастер и стечени стручни назив

   Факултет техничких наука, Енергетика, електроника и телекомуникације, Мастер инжењер електротехнике и рачунарства

4. Година уписа на докторске студије и назив студијског програма докторских студија

   2010. година, Енергетика, електроника и телекомуникације

5. Назив факултета, назив магистарске тезе, научна област и датум одбране:

   -

6. Научна област из које је стечено академско звање магистра наука:

-

III НАСЛОВ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

Модели српског језика и њихова примена у говорним и језичким технологијама

IV ПРЕГЛЕД ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

Навести кратак садржај са назнаком броја страна, поглавља, слика, шема, графикона и сл.

Докторска дисертација кандидата **Стевана Острогонца** написана је на 98 страница (са пратећом документацијом укупно 137 страница). Садржај је подељен у 6 поглавља. Дисертација садржи 15 слика, 10 графикона, 19 табела и 89 научних референци, као и 5 прилога. На почетку тезе су дати: наслов, кључна документацијска информација на српском и на енглеском језику, садржај рада, попис слика, попис табела, сажетак на српском и на енглеском језику, као и захвалница.

Кратак садржај дисертације обухвата следећих 6 поглавља:

1) Увод
2) Језички модели – стање у области
3) Ресурси за обуку језичких модела за српски језик
4) Морфолошке класе речи
5) Примене језичких модела
6) Закључак

На крају дисертације дати су списак научне литературе и прилози.

V ВРЕДНОВАЊЕ ПОЈЕДИНИХ ДЕЛОВА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

Наслов дисертације

Комисија сматра да наслов јасно назначава тематику која је предмет истраживања које је у докторској дисертацији приказано.

Поглавље 1 – Увод

У уводном поглављу објашњен је појам језичких модела и истакнут је њихов значај у области обраде природног језика. Образложена је потреба за развојем квалитетних језичких модела за српски језик, а наведени су и проблеми које је потребно решити. Међу овим проблемима највећи изазови су креирање адекватних ресурса за обуку модела, с обзиром на комплексну морфолошку структуру српског језика, као и проблем недовољне количине података за обуку, за чије је превазилажење кључан развој техника којима би се модели могли успешно обучавати и на релативно малим количинама података. Затим је дат је преглед садржаја остатка дисертације, уз истицање значаја језичких модела заснованих на морфолошким обележјима, који су за српски језик развијени у оквиру овог истраживања.

Мишљење Комисије је да је уводно поглавље пружило адекватну слику о мотивацији и потреби за истраживањем, као и о самом току и циљевима истраживања.

Поглавље 2 – Језички модели – стање у области

У оквиру овог поглавља дат је детаљан приказ најзначајнијих радова из области којој припада и истраживање које је предмет дисертације. Представљене су технике креирања статистичких *N*-грам модела уз адекватну математичку формулацију. Затим је дат преглед истраживања везаних за језичке моделе базиране на различитим структурама неуронских мрежа, такође уз адекватну математичку формулацију кључних алгоритама обуке. Истакнуте су предности и мане поменуте две парадигме језичког моделовања, а представљена су и истраживања у оквиру којих су различити типови модела комбиновани. За сва описана истраживања наведена је одговарајућа литература.

Комисија се слаже да је у оквиру овог поглавља дат јасан увид у стање истраживања у области језичког моделовања, као и да је коришћена литература актуелна и релевантна.

Поглавље 3 – Ресурси за обуку језичких модела за српски језик

Ово поглавље описује процес креирања текстуалних корпуса и алата за обуку језичких модела за српски језик. Поред тога, ово поглавље даје и преглед ресурса и алата који су настали у оквиру ранијих истраживања везаних за развој говорних технологија за српски језик, а који су релевантни за језичко моделовање. Представљени су и алати *SRILM* и

RNNLM, који су независни од језика и у свету су најчешће коришћени алати за обуку језичких модела. На крају, наведен је садржај текстуалних корпуса који су настали у оквиру овог истраживања, а који су креирани за различите функционалне стилове.

Комисија сматра да је у поглављу 3 јасно и концизно приказан део истраживања које је предмет дисертације, а које је резултовало постојањем драгоцених ресурса за српски језик, који се могу применити и у даљем развоју говорних и језичких технологија.

Поглавље 4 – Морфолошке класе речи

У поглављу 3 је наведено да су текстуални корпуси прикупљени за српски језик, значајно мањи од корпуса који се у моделовању језика користе за неке од светских језика, међу којима има и језика са значајно једноставнијом морфолошком структуром, као што је енглески. У дисертацији је, такође, још у уводном поглављу, речено да је стандардан начин за превазилажење проблема мањка података за обуку језичких модела груписање речи и креирање класних модела. Поглавље 4 даје детаљан приказ начина груписања речи на основу морфолошких и других обележја и моделе који се заснивају на оваквим класама. Описани су и модели који представљају комбинацију модела речи, основних облика речи – лема и морфолошких класа.

Оцена комисије је да је у оквиру поглавља 4 јасно и детаљно представљен концепт којим се третира проблем недовољне количине података за обуку језичких модела за српски језик.

Поглавље 5 – Примене језичких модела

У поглављу 5 систематично су представљени резултати сегмената истраживања са јасном назнаком о циљу сваког експеримента. У одељку 5.1 описана су истраживања на основу којих се дошло до закључака о томе које од постојећих техника су најпогодније за креирање модела, редукцију модела по потреби и друге врсте његове адаптације за потребе аутоматског препознавања говора. Представљено је и истраживање којим је показано да је груписање речи на основу морфолошких обележја погодније од аутоматског извођења класа помоћу Брауновог алгоритма, који се сматра стандардом за кластеризацију речи. Ово је потврђено већом тачношћу АлфаНум система за препознавање говора када су коришћени морфолошки модели у односу на тачност када су коришћени модели с аутоматски изведеним класама. У одељку 5.2 описано је истраживање које је показало да се помоћу језичких модела могу постићи бољи резултати у подели текста на реченице у оквиру АлфаНум система за синтезу говора, у односу на поделу на основу ручно имплементираних правила. У одељку 5.3 описано је истраживање везано за могућности примене морфолошких модела језика у детекцији и корекцији различитих врста грешака. Представљен је концепт паралелне анализе излаза модела речи и морфолошког модела с циљем детекције граматичких и семантичких грешака. Резултати постигнути у иницијалним експериментима указују да иницијални концепт представља добру основу за даљи развој. Одељак 5.4 описује друге примене језичких модела, од којих је, у оквиру ове дисертације, спроведено иницијално истраживање везано за класификацију текстуалних докумената на основу садржаја.

Комисија сматра да су експерименти описани спроведени коректно, а резултати интерпретирани објективно. Описани експерименти имају јасан допринос целокупном истраживању које представља значајан корак у развоју говорних и језичких технологија за српски.

Поглавље 6 – Закључак

У закључку су сумирани резултати истраживања и назначени су даљи правци рада.

Комисија сматра да су коректно описани проблеми који остају да се реше и да су логично представљени наредни кораци у развоју говорних и језичких технологија за спрски језик.

Литература

Комисија сматра да коришћена литература осликава систематичан приступ истраживању. Коришћена литература је актуелна и обухвата истраживања везана за језичко моделовање у оквиру различитих примена, што је у складу са циљевима ове дисертације.

Прилози

Дисертација садржи 5 прилога. У прилогу 1 приказана је структура статистичког *N*-грам модела. У прилогу 2 приказана је структура језичког модела базираног на рекурентној неуронској мрежи. У прилогу 3 приказани су изводи из корпуса речи, лема, морфолошких

класа и корпуса намењеног за обуку хибридних модела. Прилог 4 садржи примере садржаја једне морфолошке класе речи и једне класе која је добијена аутоматским извођењем. Прилог 5 приказује излазе различитих типова модела, када се они користе као генеративни модели.

Мишљење Комисије је да је садржај прилога адекватан и да доприноси лакшем разумевању детаља истраживања.

VI СПИСАК НАУЧНИХ И СТРУЧНИХ РАДОВА КОЈИ СУ ОБЈАВЉЕНИ ИЛИ ПРИХВАЋЕНИ ЗА ОБЈАВЉИВАЊЕ НА ОСНОВУ РЕЗУЛТАТА ИСТРАЖИВАЊА У ОКВИРУ РАДА НА ДОКТОРСКОЈ ДИСЕРТАЦИЈИ

Таксативно навести називе радова, где и када су објављени. Прво навести најмање један рад објављен или прихваћен за објављивање у часопису са ISI листе односно са листе министарства надлежног за науку када су у питању друштвено-хуманистичке науке или радове који могу заменити овај услов до 01. јануара 2012. године. У случају радова прихваћених за објављивање, таксативно навести називе радова, где и када ће бити објављени и приложити потврду о томе.

Рад у међународном часопису (M23)

1) Ostrogonac S., Pakoci E., Sečujski M., Mišković D., 2018. *Morphology-based vs Unsupervised Word Clustering for Training Language Models for Serbian*. Acta Polytechnica Hungarica, ISSN:1785-88601785-9599, accepted for publication.

Рад у часопису националног значаја (M52)

1) Ostrogonac S., Popović B, Mak R., Sečujski M., 2015. *Automatic Word Clustering Based on Semantics - an Approach for Serbian*. 3rd International Acoustics and Audio Engineering Conference, TAKTONS 2015, Novi Sad, Serbia, pp. 36-37, ISBN: 978-86-7892-758-4.

Саопштење са међународног скупа штампано у целини (M33)

1) Ostrogonac S., Popović B., Mak R., 2015. *The Use of Statistical Language Models for Grammar and Semantic Error Handling in Spell Checking Applications for Serbian*. 12th Int. Conf. on Electronics, Telecommunications, Automation and Informatics, ETAI 2015, Ohrid, Macedonia, ISBN: 978-9989-630-76-7.
2) Ostrogonac S., Popović B., Sečujski M., Mak R., Pekar D., 2013. *Language Model Reduction for Practical Implementation in LVCSR Systems*. In Proc. Int. Scientific-Professional Symposium INFOTEH, Jahorina, Bosnia and Herzegovina, pp. 391-394.
3) Ostrogonac S., Mišković D., Sečujski M., Pekar D., Delić V., 2012. *A Language Model for Highly Inflective Non-Agglutinative Languages*. 10. SISY – Int. Symp. on Intelligent Systems and Informatics, Subotica, Serbia: IEEEXplore, pp. 177-181, ISBN 978-1-4673-4749-5.
4) Ostrogonac S., Sečujski M., Mišković D., 2012. *Impact of training corpus size on the quality of different types of language models for Serbian*. 20. TELFOR, Belgrade, Serbia.

Саопштење са међународног скупа штампано у изводу (M34)

1) Ostrogonac S., Popović B., Sečujski M., 2016. *The Use of Semantic Classes in Document Classification*. Language Technologies & Digital Humanities, Ljubljana: Slovenian Language Technology Society, pp. 216-217, ISBN 978-961-237-862-2.
2) Ostrogonac S., Popović B, Mak R., Sečujski M., 2015. *Automatic Word Clustering Based on Semantics - an Approach for Serbian*. 3rd International Acoustics and Audio Engineering Conference, TAKTONS 2015, Novi Sad, Serbia, pp. 36-37, ISBN: 978-86-7892-758-4.

Саопштење са скупа националног значаја штампано у целини (M63)

1) Ostrogonac S., Mišković D., Sečujski M., Pekar D., 2012. *Discriminative Potential of a Language Model Based on the Class N-gram Concept*. DOGS, Kovačica, Serbia.

Прототип, нова метода, софтвер, стандардизован или атестиран инструмент (M85)

1) Bojanić M., Ostrogonac S., Sečujski M., Vujnović-Sedlar N., Suzić S., 2012. *A detector of spelling errors for Serbian - anSpellChecker*. Technical Solution: Prototype, Faculty of Technical Sciences and AlfaNum, Novi Sad, Serbia.

VII ЗАКЉУЧЦИ ОДНОСНО РЕЗУЛТАТИ ИСТРАЖИВАЊА

Ова дисертација преставља истраживање које обухвата комплетан процес креирања оптималних језичких модела за различите примене за српски језик. Закључци до којих се дошло односе се управо на то какве технике, обележја и алате је погодно користити за специфичне

примене, али и у општем случају, за српски језик. Најважнији од тих закључака су:

1) За решавање проблема недовољне количине података за обуку модела, груписање речи на основу морфолошких обележја даје боље резултате од аутоматске кластеризације речи, при креирању класних језичких модела за српски језик.
2) Постојећи корпуси за српски језик довољни су за обучавање морфолошких модела, али не и за обуку модела речи. Међутим, морфолошки модели су ограниченог потенцијала, када је у питању репрезентација језика, те их је најбоље користити у комбинацији са моделима речи или моделима лема, осим ако постоје екстремна ограничења по питању меморијских или процесорских ресурса за неку конкретну примену.
3) Уколико је потребно редуковати број параметара језичког модела, погодније је користити технику минималног пораста ентропије, него технику постављања прага минималног броја појављивања секвенци речи у корпусу за обуку.
4) Језички модели представљају адекватно решење за поделу текста на реченице, јер и са постојећим корпусима за српски језик постижу боље резултате од постојећег система базираног на ручно имплементираним правилима.
5) Морфолошки модели у комбинацији са моделима речи погодни су за детекцију различитих врста грешака у текстовима на српском језику.
6) Ресурси и модели развијени у оквиру овог истраживања представљају основу за почетак развоја разних технологија које још увек нису реализоване за српски језик, попут класификације докумената на основу садржаја и сличног.

VIII ОЦЕНА НАЧИНА ПРИКАЗА И ТУМАЧЕЊА РЕЗУЛТАТА ИСТРАЖИВАЊА

Експлицитно навести позитивну или негативну оцену начина приказа и тумачења резултата истраживања.

На основу детаљног увида у садржај докторске дисертације од стране чланова Комисије, закључено је да је истраживање пажљиво испланирано и систематично извршено, да су експерименти адекватно спроведени и да су резултати истраживања интерпретирани коректно и објективно. Оцена комисије је, дакле, позитивна.

Рад је проверен у софтверу за детекцију плагијаризма *iThenticate*.

IX КОНАЧНА ОЦЕНА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

Експлицитно навести да ли дисертација јесте или није написана у складу са наведеним образложењем, као и да ли она садржи или не садржи све битне елементе. Дати јасне, прецизне и концизне одговоре на 3. и 4. питање:

1. Да ли је дисертација написана у складу са образложењем наведеним у пријави теме

   Дисертација је у целини написана у складу са образложењем наведеним у пријави теме.

2. Да ли дисертација садржи све битне елементе

   Дисертација садржи све битне елементе који се захтевају по Статуту Факултета техничких наука и Универзитета у Новом Саду, као и Закона о високом образовању.

3. По чему је дисертација оригиналан допринос науци

   У оквиру дисертације предложен је начин груписања речи који подразумева коришћење морфолошких информација и који представља основу за креирање језичких модела за српски језик, са циљем превазилажења проблема мањка података за обуку. Експериментално је утврђено да су овакви модели погоднији од модела добијених аутоматском кластеризацијом речи, с аспекта примене у аутоматском препознавању говора.

   Језички модели засновани на морфолошким класама речи показани су експериментално као адекватни за коришћење заједно са моделима речи ради детекције и корекције различитих врста грешака у текстовима на српском језику.

4. Недостаци дисертације и њихов утицај на резултат истраживања

   Истраживање није обухватило детаљно испитивање различитих структура неуронских мрежа у језичком моделовању. Разлог овог недостатка је везан за величину корпуса прикупљеног за српски језик, а на основу којих су модели базирани на рекурентним неуронским мрежама дали лошије резултате него статистички модели. Међутим, резултати добијени

помоћу модела базираних на неуронским мрежама, иако лошији, указивали су на исте закључке на које су указивали и резултати добијени помоћу статистичких *N*-грам модела.

Стандардни начини евалуације језичких модела подразумевају рачунање перплексности и, ако је у питању примена у препознавању говора, грешке препознавања на нивоу речи. Ове величине у извесној мери зависе од података коришћених за евалуацију. Ипак, селекцијом репрезентативног узорка вишеструким понављањем свих експеримената описаних у дисертацији на мноштву различитих скупова података за тестирање, утврђена је конзистентност резултата, која је својеврстан индикатор поузданости.

X ПРЕДЛОГ:

На основу укупне оцене дисертације, комисија предлаже да се докторска дисертација „Модели српског језика и њихова примена у говорним и језичким технологијама“ кандидата Стевана Острогонца прихвати, а кандидату одобри јавна одбрана.

ЧЛАНОВИ КОМИСИЈЕ

_____

др Владо Делић, редовни професор
Факултет техничких наука, Нови Сад

_____

др Драгана Бајић, редовни професор
Факултет техничких наука, Нови Сад

_____

др Снежана Гудурић, редовни професор
Филозофски факултет, Нови Сад

_____

др Татјана Грбић, ванредни професор
Факултет техничких наука, Нови Сад

_____

др Јелена Николић, доцент
Електронски факултет, Ниш

_____

др Милан Сечујски, ванредни професор
Факултет техничких наука, Нови Сад

НАПОМЕНА: Члан комисије који не жели да потпише извештај јер се не слаже са мишљењем већине чланова комисије, дужан је да унесе у извештај образложење односно разлоге због којих не жели да потпише извештај.