УНИВЕРЗИТЕТ У БЕОГРАДУ
ФИЛОЛОШКИ ФАКУЛТЕТ

ИЗВЕШТАЈ О ОЦЕНИ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

I ПОДАЦИ О КОМИСИЈИ

1. Датум и орган који је именовао комисију

27. IV 2016. Научно-наставно веће Филолошког факултета

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 Цветана Крстев, редовни професор, библиотечка информатика, 20. V 2014, Филолошки факултет Универзитета у Београду
2. др Милош Утвић, доцент, библиотечка информатика, 18. XI 2014, Филолошки факултет Универзитета у Београду
3. др Рајна Драгићевић, редовни професор, српски језик, 15. V 2013, Филолошки факултет Универзитета у Београду
4. др Ђурица Грга, ванредни професор, клиничке стоматолошке науке, 1. VI 2012, Стоматолошки факултет Универзитета у Београду
5. др Душко Витас, ванредни професор, рачунарство и информатика, 8. VII 2011, Математички факултет Универзитета у Београду

II ПОДАЦИ О КАНДИДАТУ

1. Име, име једног родитеља, презиме:

Јелена Бранислав Јаћимовић

2. Датум рођења, општина, република:

1. I 1976. Пожаревац, Србија

3. Датум одбране, место и назив мастер рада:

22. I 2009. Београд. Информациона писменост и високошколско образовање

4. Научна област из које је стечено академско звање мастера:

Библиотекарство и информатика

III НАСЛОВ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

Аутоматско препознавање и нормализација временских израза у неструктурираним новинским и медицинским текстовима на српском језику

IV ПРЕГЛЕД ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:
Навести кратак садржај са назнаком броја страна поглавља, слика, шема, графикона и сл.

Докторска дисертација проучава временске изразе у новинским текстовима с аспекта њиховог аутоматског препознавања, обележавања и нормализовања у складу са важећим стандардима. Истраживање посебно обухвата:

- анализу постојећих стандарда и приступа у аутоматском препознавању временских израза;
- припрему корпуса текстова на српском језику за обуку и евалуацију система за препознавање временских израза;
- анализу структуре временских израза у текстовима на српском језику и израду формалног модела:
- израду и евалуацију софтверског алата за препознавање и нормализацију временских израза у текстовима на српском језику.

У истраживању се користе методе засноване на правилима, насупрот статистичким методама које се такође

користе за препознавање временских израза. Коришћена је методологија локалних граматика представљених коначним трансдукторима у форми синтаксичких графова организованих у каскаде и уз подршку електронских речника српског језика.

Као технолошка основа за подршку наведеној методологији коришћен је систем *Unitex* који је развијен у *LADL*-у (*Laboratoire d'Automatique Documantaire et Linguistique*), у оквиру Универзитета Марн-ла-Вале, Француска, као и електронски речници развијени у оквиру Групе за језичке технологије Универзитата у Београду.

Евалуација система за препознавање и нормализацију временских израза је показала да систем показује изузетно добре резултате на непознатим новинским текстовима, као и на текстовима из другог домена (медицинским наративним текстовима).

Дисертација обухвата 294 стране, а у оквиру тога 8 поглавља (214 страна), списак коришћене литературе (19 страна, 212 библиографских јединица), 5 прилога (57 страна), списак табела (2 стране) и списак слика (2 стране). У дисертацији укупно има 39 слика и 31 табела. Поглавља дисертације су:

1. Увод (10 страна, 1 слика).
2. Време у природном језику (9 страна).
3. Преглед постојећих ресурса и рачунарских приступа за обраду временских израза (54 стране, 2 табеле).
4. Препознавање временских израза (49 страна, 15 слика, 4 табеле).
5. Нормализација временских израза (38 страна, 23 слике, 5 табела).
6. Евалуација успешности система за аутоматску обраду временских израза српског језика (18 страна, 10 табела).
7. Препознавање и нормализација временских израза медицинских наративних текстова (33 стране, 10 табела).
8. Закључак (3 стране).

Прилози дисертације су:

A. Примери обележавања препознатих временских израза лексичким етикетама (18 страна).
B. Примери обележавања препознатих временских израза XML етикетама (18 страна).
C. Примери обележавања препознатих временских израза <TIMEX3> етикетама (7 страна).
D. Примери обележавања препознатих временских израза приликом процене успешности система (3 стране).
E. Примери обележавања препознатих временских израза медицинских наративних текстова (10 страна).

V ВРЕДНОВАЊЕ ПОЈЕДИНИХ ДЕЛОВА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

У уводном поглављу дисертације мастера Јелене Јаћимовић представљени су предмет, циљ и методологија истраживања. Предмет истраживања ове дисертације су временски изрази који се јављају у наративним текстовима, то јест у текстовима чији делови нису посебно истакнути. Како је основни циљ истраживања Јелене Јаћимовић аутоматска идентификација свих израза који указују на време, као и идентификација и успостављање хронологије догађаја неког текста, потребно је идентификоване временске изразе, догађаје и временске релације трансформисати у информације структурираног облика како би касније могле да се користе у оквиру неких сложенијих апликација обраде природног језика. У уводном поглављу кандидаткиња указује на неке од ових апликација чија ефикасност у многоме зависи од успешне идентификације временских израза, као што су системи за одговарање на питања, аутоматско резимирање текста, проналажење информација и екстракцију информација. Основни циљ истраживања кандидаткиње мастер Јелене Јаћимовић је аутоматско обележавање временских израза у неструктурираним текстовима српског језика са постизањем високог нивоа одзива и прецизности. Додатни циљ истраживања је процена ефикасности методе за препознавање временских израза у домену медицинских наративних текстова. Као резултат рада на овој дисертацији изграђен је систем који, без додатне припреме, успешно обележава временске изразе у текстовима на српском језику и корпус анотираних временских израза. Систем који је кандидаткиња развила ће омогућити изградњу додатних анотираних временских корпуса који би се могли користити као ресурси за развој система заснованих на методама машинског учења.

У другом поглављу „Време у природном језику", кандидаткиња се бави изражавањем времена у

природном језику, те истиче да се категорија времена везује искључиво за предикацију. Сложеност времена се одражава и у разноликим начинима изрицања временских односа који могу бити исказани на морфолошком, синтаксичком, лексичком нивоу, као и на нивоу текста. У природном језику, па и у српском, се временске информације изражавају путем средстава која се могу груписати у три групе: (а) догађаји (стања, активности, остварења и достигнућа); (б) временски изрази и (в) тип временских односа (сукцесивност, инклузија и поклапање). У овом поглављу кандидаткиња прецизно позиционира предмет свог рада који се односи на временске изразе помоћу којих се изражава позиција у времену, трајање у времену и учесталост понављања у времену.

У трећем поглављу „Преглед постојећих ресурса и рачунарских приступа за обраду временских израза“ кандидаткиња мастер Јелена Јаћимовић даје исцрпан приказ постојећих ресурса и рачунарских приступа који се баве аутоматском обрадом временских израза, пре свега у новинским текстовима. У овом прегледу кандидаткиња покрива три значајне теме: шеме и стандарде кодирања временских израза, најчешће коришћене анотиране временске корпусе за обуку и евалуацију и системе за обраду временских израза. Усвајање стандардних шема кодирања временских израза је предуслов за широку употребу система за обраду временских израза као и за могућност њиховог поређења. Кандидаткиња представља развој ових шема, почев од шеме TIMEX, преко шема TIDES и STAG, па све то актуелне шеме TimeML која је 2009. године усвојена као ISO стандард. Анотирани временски корпуси за обуку и евалуацију могу се припремати ручно или аутоматски са циљем припреме „златног стандарда“ у односу на који би се различити системи за обраду временских израза могли вредновати. Припрема ових корпуса у складу са стандардним шемама за кодирање је обавезан услов. Кандидаткиња представља корпус TERN, кодиран TIDES TIMEX2 шемом, као и више корпуса кодираних TimeML шемом. Системи за обраду временских израза се могу груписати у (а) системе засноване на правилима; (б) системе засноване на статистичким законитостима и (в) хибридне системе који комбинују најбоље особине претодна два приступа. Кандидаткиња одређује правац свог рада у односу на три теме представљене у овом поглављу: (а) систем који кандидаткиња развија ће поштовати стандард кодирања TimeML, (б) с обзиром да за српски језик не постоји „златни стандард“, један од резултата кандидаткињиног рада биће његово формирање и (в) кандидаткиња ће развијати систем заснован на правилима.

Најважнији резултати дисертације мастера Јелене Јаћимовић изложени су у поглављима 4-7. У четвртом поглављу, „Препознавање временских израза“, кандидаткиња даје, пре свега, класификацију временских израза, а потом и опис уобичајених типова временских израза који се могу срести у природном језику. Систем који кандидаткиња развија ће аутоматски препознавати временске изразе који одговарају на питања КАДА се нешто догодило, КОЛИКО ДУГО је нешто трајало и КОЛИКО ЧЕСТО се нешто дешавало, а који се могу изразити системом падежних конструкција и прилога. Класификација временских израза које развијани систем препознаје је учињена по две основе: с једне стране су класе израза који указују на позицију у времену (календарски датуми и времена дана), трајање и учесталост, а с друге стране су класе временских израза који су прецизни и независни од контекста и израза који су недовољно прецизни и зависе од контекста. За препознавање и даљу обраду временских израза потребно је одредити грануларност димензије времена на коју указују, а кандидаткиња се бави временским изразима чији је ниво грануларности миленијум, век, деценија, година, месец, недеља, дан, сат, минут и секунд. Од значаја је и одређивање коректног опсега препознатих временских израза и ту се кандидаткиња придржава препорука TimeML стандарда: тако у опсег временских израза улазе именице и именичке синтагме, придеви, прилози и придевско/прилошке синтагме који махом представљају окидаче препознавања тих израза, док предлози и везници који им претходе нису у њиховом саставу. Сви препознати временски изрази биће кодирани коришћењем стандардног језика за обележавање XML, етикетом TIMEX3 и обавезним атрибутима type (за прецизирање типа временског израза) и temporalFunction (за указивање на његову прецизност). Као корпус на основу кога је методом „пробај, па исправи“ (која се у англосаксонској литератури назива *bootstrapping*) кандидаткиња развијала свој систем, коришћен је велики узорак новинских текстова из више извора, који се састоји од више од пола милиона текућих речи и скоро 50 хиљада реченица. На основу свега овога кандидаткиња је развила систем заснован на правилима који користи *Unitex*, програмски алат за рад са корпусима, и морфолошке електронске речнике српског језика. Цео систем је развијен у облику каскаде коначних трансдуктора у којој сваки трансдуктор препознаје одређене временске изразе чиме припрема терен за успешан рад наредних трансдуктора. Структура ове каскаде је објашњена текстом, дијаграмима и илустративним примерима препознавања.

У петом поглављу, „Нормализација временских израза“, кандидаткиња, мастер Јелена Јаћимовић,

представља други део свог система чији је задатак нормализација, тј. свођење на нормализовани облик свих препознатих временских израза. И у овом случају следе се препоруке стандарда TimeML. Овај корак је од изузетног значаја за коректно анотирање текста јер он обезбеђује јединствено кодирање временских информација које могу у тексту да буду забележене различитим језичким средствима, на пример „3. IV 1999. године“ и „трећег априла 1999.“ У овом кораку, етикета TIMEX3 уметнута у претходном кораку обогаћује се додатним атрибутима: value (садржи нормализовану вредност временског израза у складу са TimeML стандардном шемом), mod (садржи евентуалну квалификацију временског израза), anchorTimeID (садржи идентификациони број временског израза), valueFromFunction (садржи потребне информације за рачунање нормализоване вредности недовољно прецизних временских израза), functionInDocument (даје функцију временског израза у документу), beginPoint и endPoint (одређују почетну и крајњу тачку временског израза који указује на трајање) и quant и freq (указују на учесталост и временски период понављања временских израза учестаности). За ову фазу рада кандидаткиња се опет одлучила за коришћење система *Unitex* и развој коначног трансдуктора који се примењује на већ анотирани текст системом за препознавање. Овакав избор је у потпуности у складу са светским искуствима која показују да су за нормализацију временских израза много успешнији системи засновани на правилима.

У шестом поглављу, „Евалуација успешности система за аутоматску обраду временских израза српског језика“, кандидаткиња Јелена Јаћимовић представља резултате исцрпне евалуације развијеног система. Ниједан систем за проналажење и екстракцију информација не може се сматрати завршеним уколико није процењена његова успешност измерена стандардним мерама одзив – колико је тражених израза пронађено – и прецизност – колико је међу пронађеним изразима оних који се траже. За евалуацију система кандидаткиња је користила новинске текстове који нису коришћени за развој система, али је домен текстова остао исти. Овај корпус се састојао од преко 100.000 текућих речи и преко 5.000 реченица. Евалуација је утврдила да је у овом корпусу било 2.050 временских израза, од којих 83 указују на временски период, а 34 на учесталост понављања у времену. Скоро 60% временских израза је прецизно изражено и не зависи од контекста, док је 35% израза непрецизно и захтева контекст за одређивање прецизне вредности. При процени препознавања временских израза водило се рачуна о грешкама пропуштања (непрепознати изрази), уљезима (препознато као временски израз нешто што то није), непоклапања (опсег временског израза није добро одређен), као и о пропустима до којих је дошло услед грешака у самом тексту. Осим процена препознавања временских израза, процењивани су и сви њима придружени атрибути. Резултати евалуације показују да изграђени систем има изузетну прецизност, p=0,99 (99%), док је одзив нешто слабији, r=0,80 (80%). Комбинована F_1 мера је 0,88 (88%), што овај систем сврстава у ред веома успешних система за препознавање и прецизирање временских израза. Како су сви препознати изрази и нормализовани, кандидаткиња је евалуирала и успешност одређивања нормализоване вредности пропознатих израза: овде су резултати више него успешни, јер је прецизност доделе нормализоване вредности 0,997 (99,7%), што значи да су практично сви успешно препознати временски изрази добили исправну нормализовану вредност.

У седмом поглављу „Препознавање и нормализација временских израза медицинских наративних текстова“ кандидаткиња Јелена Јаћимовић представља резултате примене система за аутоматско препознавање временских израза без икаквог прилагођавања на нови домен – медицинске наративне текстове. Време је изузетно значајно за репрезентовање садржаја медицинских текстова: у клиничким белешкама лекара изузетно је значајан хронолошки ред одређених чињеница. За процену рада система изабран је корпус од 100 насумично изабраних отпусних листи и 50 извештаја лекара који потичу из две наставне јединице Стоматолошког факултета Универзитета у Београду. Овај корпус се састојао од преко 35.000 текућих речи и нешто мање од 3.500 реченица. Медицински наративни текстови представљају посебан изазов за аутоматску обраду јер их пишу стручњаци за друге стручњаке, па обилују латинским називима, као и другом стручном терминологијом и великим бројем специфичних скраћеница, а пошто су готово увек писани у великој журби, и већим бројем типографских грешака но што је уобичајено. Евалуација је спроведена по истој методологији као и за новинске текстове. Утврђено је да овај корпус има укупно 884 временска израза, од којих 112 указује на временски период, а 60 на учесталост понављања у времену – процентуално знатно више него у новинским текстовима, што је и очекивано. Резултати евалуације показују да је прецизност p=0,94 нешто нижа док је одзив r=0,90 виши него у случају новинских текстова. Комбинована F_1 мера је 0,92, а коректност нормализације успешно препознатих израза опет скоро стопостотна, што све говори да је примена развијеног система, без икакве адаптације, на нови домен била изузетно успешна.

У деветом поглављу „Закључак“ Јелена Јаћимовић пре свега сажето излаже постављене задатке, коришћену методологију и технике као и остварене резултате. У завршном делу, кандидаткиња указује на правце даљег рада. Иако је успешност рада система за препознавање временских израза показана, још увек има простора за његово побољшање. То побољшање се може односити и на проширивање опсега његове примене (методом „пробај па исправи“) на текстове из разних домена, као и на дораду система у смислу допуне информација непрецизних препознатих временских израза. Нови правац рада је проширивање система на препознавање других временских ентитета, пре свега догађаја и временских релација.

На крају дисертације мастер Јелена Јаћимовић је приложила пет додатака:
A. Примери обележавања препознатих временских израза лексичким етикетама. У овом додатку су дате конкорданце препознатих временских израза обележених лексичким етикетама *Unitex*-а, које су разврстане према трансдукторима у каскади који те лексичке етикете производе.
B. Примери обележавања препознатих временских израза XML етикетама. У овом додатку су дате конкорданце препознатих временских израза обележених XML етикетама, које су разврстане према трансдукторима у каскади који те XML етикете производе.
C. Примери обележавања препознатих временских израза <TIMEX3> етикетама. У овом додатку су дате конкорданце препознатих и нормализованих временских израза обележених <TIMEX3> етикетама, које су разврстане према типовима временских израза: прецизни и непрецизни датуми, времена дана, трајање и учесталост.
D. Примери обележавања препознатих временских израза приликом процене успешности система. У овом додатку су дати примери обележавања успешно препознатих временских израза, делимично успешног и погрешног препознавања, као и примери обележених пропуштених (непрепознатих) временских израза.
E. Примери обележавања препознатих временских израза медицинских наративних текстова. У овом додатку је дат један наративни медицински текст, његова деидентификована варијанта, пример наративног текста обележен препознатим временским изразима, као и примери успешно препознатих временских израза, делимично успешног и погрешног препознавања, као и примери обележених пропуштених временских израза.

VI СПИСАК НАУЧНИХ И СТРУЧНИХ РАДОВА КОЈИ СУ ОБЈАВЉЕНИ ИЛИ ПРИХВАЋЕНИ ЗА ОБЈАВЉИВАЊЕ НА ОСНОВУ РЕЗУЛТАТА ИСТРАЖИВАЊА У ОКВИРУ РАДА НА ДОКТОРСКОЈ ДИСЕРТАЦИЈИ, уз напомену: Навести називе радова, где и када су објављени.

J. Jaćimović, "Automatic Processing of Temporal Expressions in Serbian Natural Language Texts," in *Natural Language Processing for Serbian - Resources and Applications*, eds. G. Pavlović-Lažetić et al., University of Belgrade, Faculty of Mathematics, Belgrade, 2014, pp. 57-69.

J. Jaćimović, C. Krstev, and D. Jelovac. 2015. “A Rule-Based System for Automatic De-identification of Medical Narrative Texts”. *Informatica* 39 (1): 45–53.

J. Jaćimović, "Recognition and Normalization of Temporal Expressions in Serbian Texts," in *Workshop on Computational Linguistics and Natural Language Processing of Balkan Languages - CLoBL 2012*, Novi Sad, 2012, pp. 97-100.

У случају радова прихваћених за објављивање, таксативно навести називе радова, где и када ће бити објављени и приложити потврду о томе.

VII ЗАКЉУЧЦИ ОДНОСНО РЕЗУЛТАТИ ИСТРАЖИВАЊА

Резултати изложени у овој дисертацији говоре да је кандидаткиња мастер Јелена Јаћимовић остварила циљеве зацртане у пријави дисертације. Кандидаткиња је детаљно и на формалан начин описала један важан тип именованих ентитета у српском језику – временске изразе. На основу резултата овог истраживања кандидаткиња је изградила потпуно функционални систем за препознавање и нормализацију временских израза. Као додатни резултат рада на овој дисертацији добијен је корпус новинских текстова као и корпус медицинских наративних текстова на српском језику са прецизно обележеним временским изразима, који будући истраживачи могу користити за развој система заснованих на другим методама, нпр.

машинском учењу. Све детаље овог система кандидаткиња је у дисертацији ставила на располагање чиме је омогућила репродукцију резултата истраживања као и будуће надградње.

Сам текст дисертације, као и списак литературе наведен на крају рада, говоре да је мастер Јелена Јаћимовић користила релевантну и савремену литературу, те да је постављене проблеме обрадила детаљно и сагледавајући их из разних углова. Овим радом Јелена Јаћимовић је отворила једно ново поље истраживања у области обраде српског језика, а будућим истраживачима ставила на располагање изузетно значајне ресурсе и алате за даљи рад.

VIII ОЦЕНА НАЧИНА ПРИКАЗА И ТУМАЧЕЊА РЕЗУЛТАТА ИСТРАЖИВАЊА
НАПОМЕНА: Навести позитивну или негативну оцену начина приказа и тумачења резултата истраживања.

Комисија сматра да је кандидаткиња мастер Јелена Јаћимовић у својој дисертацији *Аутоматско препознавање и нормализација временских израза у неструктурираним новинским и медицинским текстовима на српском језику* успешно обрадила 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.

X ПРЕДЛОГ:

На основу укупне оцене дисертације, комисија предлаже: Научно-наставном већу Филолошког факултета Универзитета у Београду да прихвати извештај о дисертацији *Аутоматско препознавање и нормализација временских израза у неструктурираним новинским и медицинским текстовима на српском језику* кандидаткиње мастер Јелене Јаћимовић и упути га Већу за друштвено-хуманистичке науке Универзитета у Београду, како би кандидаткиња била позвана на усмену одбрану рада.

ПОТПИСИ ЧЛАНОВА КОМИСИЈЕ

1. др Цветана Крстев, редовни професор
   Филолошки факултет Универзитета у Београду

2. др Милош Утвић, доцент
   Филолошки факултет Универзитета у Београду

3. др Рајна Драгићевић, редовни професор
   Филолошки факултет Универзитета у Београду

4. др Ђурица Грга, ванредни професор
   Стоматолошки факултет Универзитета у Београду

5. др Душко Витас, ванредни професор
   Математички факултет Универзитета у Београду