



УНИВЕРЗИТЕТ У БЕОГРАДУ - ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ

Булевар краља Александра 73, 11000 Београд, Србија

Тел. 011/324-8464, Факс: 011/324-8681

КОМИСИЈИ ЗА СТУДИЈЕ II СТЕПЕНА ЕЛЕКТРОТЕХНИЧКОГ ФАКУЛТЕТА У БЕОГРАДУ

Комисија за студије II степена, Електротехничког факултета у Београду, на својој седници одржаној 07.07.2020. године именovalo нас је у Комисију за преглед и оцену мастер рада дипл. инж. Ђурђина Симић под насловом „Методe за учење ниско-димензионалних векторских репрезентација речи у обради природних језика“. Након прегледа материјала Комисија подноси следећи

ИЗВЕШТАЈ

1. Биографски подаци кандидата

Ђурђина Симић је рођена 16.11.1996. године у Аранђеловцу. Гимназију „Милош Савковић“ у Аранђеловцу је завршила као вуковац. Електротехнички факултет у Београду уписала је 2015. године. Дипломирала је на одсеку за Сигнале и системе у септембру 2019. године са просечном оценом 9,20. Дипломски рад одбранила је у септембру 2019. године са оценом 10. Мастер студије на Електротехничком факултету у Београду је уписала у октобру 2019. године, на Модулу за сигнале и системе. Положила је све испите са просечном оценом 9,40.

2. Опис мастер рада

Мастер рад обухвата 48 страна, са укупно 19 слика, 14 табела и 12 референци. Рад садржи увод, 7 поглавља и закључак (укупно 9 поглавља), као и списак коришћене литературе, списак слика и списак табела.

Прво поглавље представља увод у коме су описани предмет и циљ рада. Дат је кратак осврт на значај обраде природних језика, као и на значај увођења дистрибуираних семантичких модела, који уче векторске репрезентације речи на основу контекста у којима се речи јављају.

У другом поглављу размотрене су разлике између *one-hot* и дистрибуиране репрезентације речи. Изнета је мотивација за увођење техника за добијање ниско-димензионалних векторских репрезентација речи, које мапирају речи у векторе реалних бројева фиксне димензије и које су позанте под заједничким називом *word embeddings*.

У трећем поглављу је детаљније размотрен *word2vec* модел који подразумева две различите архитектуре - *CBOW* и *Skip-gram*, и припада групи модела заснованих на предикцији речи. Такође, објашњени су и различити начини оптимизације и унапређења основних формулација *word2vec* модела.

Четврто поглавље описује *GloVe* модел, који се заснива на матрици заједничког појављивања речи и припада групи модела заснованих на пребројавању заједничких појављивања парова речи.

У оквиру петог поглавља описује се *fastText* модел, изграђен на ограничењима претходна два модела, који има способност да представи и оне речи које се не налазе у корпусу за обучавање.

Шесто поглавље садржи опис коришћеног корпуса српског језика *srWaC*. Такође, описују се методе за предобраду текста, које имају за циљ редуковање скупа на оне речи које су потенцијално најзначајније за корпус, а све то у циљу ефикаснијег обучавања.

У оквиру седмог поглавља описани су различити начини евалуације квалитета добијених векторских репрезентација речи, који се могу поделити у две основне групе –

посредни (*extrinsic*) и непосредни (*intrinsic*). Споменути модели тестирани су квантитативно на задацима анализе сентимента, сличности реченица, као и на скупу *SimLex999*, али и квалитативно, испитивањем најближих суседа одређених речи, аналогичности речи и помоћу *t-SNE* визуелизације.

У осмом поглављу приказани су резултати добијени на различитим методама евалуације, који омогућавају да се процени квалитет научених векторских репрезентација речи. Анализира се утицај различитих вредности хипер-параметара *word2vec*, *GloVe* и *fastText* модела и врши њихова упоредна анализа.

Девето поглавље је закључак у оквиру кога су сумирани добијени резултати и размотрена могућа унапређења и правци даљег развоја и истраживања.

3. Анализа рада са кључним резултатима

Мастер рад дипл. инж. Ђурђине Симић бави се проблематиком учења дистрибуираних векторских репрезентација речи на основу великих корпуса текста. Традиционалан начин представљања речи јесте помоћ индекса који одговара позицији те речи у неком фиксном, унапред задатом речнику. Како би се боље описале семантичке и синтаксичке сличности међу речима, последњих година су развијани методи машинског учења који речи представљају векторима знатно мањих димензија, и то тако да геометријски односи међу овим векторима одсликавају семантичке или синтаксичке сличности међу представљеним речима. Ове методе су биле катализатор значајног побољшања перформанси модела у разним задацима из домена обраде природних језика протеклих година. Ђурђина је анализирала данас три најпопуларније технике за учење дистрибуираних векторских репрезентација речи: *word2vec*, *GloVe* и *fastText*. Методе су у раду описане, имплементиране, обучене на корпусу текстова на српском језику, и темељно анализиране.

Основни доприноси рада су: 1) приказ наведених метода за учење дистрибуираних векторских репрезентација речи природног језика; 2) имплементација и примена метода на јавно доступном корпусу текстова српског језика *srWaC*, 3) критичка анализа и поређење перформанси трију метода по неколико различитих критеријума, као и анализа утицаја хипер-параметара ових алгоритама на њихове перформансе, 4) уочени могући правци за даљи рад на овој теми.

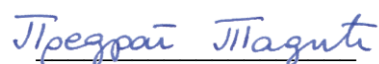
4. Закључак и предлог

Кандидат Ђурђина Симић је у свом мастер раду успешно описала, имплементирала и применила методе машинског учења намењене изналажењу дистрибуираних ниско-димензионалних векторских репрезентација речи природног језика. Критички је анализирала њихове перформансе, међусобне предности и недостатке, и уочила могуће правце за даље истраживање. Исказала је самосталност и систематичност у свом раду, способност коришћења савремене литературе и популарних библиотека намењених развоју софтверских решења из ове области.

На основу изложеног, Комисија предлаже Комисији за студије II степена Електротехничког факултета у Београду да рад дипл. инж. Ђурђине Симић прихвати као мастер рад и кандидаткињи одобри јавну усмену одбрану.

Београд, 20. 11. 2020. године

Чланови комисије:



Др Предраг Тадић, доцент.



Др Дражен Драшковћ, доцент.