Р Е Ц Е Н З И Я

от проф. дмн Галя Младенова Ангелова, Секция ЛМОЗ, ИИКТ-БАН
за дисертацията на Александър Николаев Попов
"*Моделиране на лексикалното знание с цел автоматична обработка на естествен език*",
представена за присъждане на образователната и научна степен "доктор"

Съгласно заповед 127/12.07.2018 г. на директора на ИИКТ-БАН съм определена за член на Научното жури за защита на представения дисертационен труд в професионално направление 4.6 "Информатика и компютърни науки". Дисертацията е свързана с някои от най-актуалните проблеми в информатиката днес: намиране на нови модели и методи за ефективна обработка на големи обеми от текстове, основани върху векторно представяне на думите (лексикално знание); създаване на публично достъпни масиви от представяния на думите като вектори и тестване на сценарии за сравняване и оценка на качеството на тези вектори; усъвършенстване на подходите за автоматично разрешаване на семантичната многозначност. Тези въпроси са част от доминиращото напоследък изследователското направление "deep learning" и с тях се свързват оптимистичните прогнози за развитието на изкуствения интелект.

Формално, следвайки Правилника за прилагане на Закона за развитието на академичния състав в Република България от 6 юли 2018 г., по група показатели Г за трудовете на дисертанта се събират 156 точки от индексирани в Scopus статии и 59 точки от неиндексирани публикации - общо 215 точки, при минимално изискване от 30 точки за група показатели Г. С това изискванията на ЗРАСРБ за получаване на образователната и научна степен "доктор" са удовлетворени и многократно надхвърлени. Резултатите от дисертацията са представени в 14 публикации (4 от тях самостоятелни), като:

- 1 е в международно списание с SJR ранк,
- 4 са в Сборници трудове на престижни международни конференции, издадени от Шпрингер в различни поредици - Computational Intelligence и Lecture Notes on Artificial Intelligence,
- 9 са в Сборници трудове на престижни международни конференции по обработка на естествен език, между които са RANLP, Global WordNet Conference, TLT и др.

Забелязани са 8 цитирания на трудовете на автора. Освен представяне на статиите на 6 международни конференции са изнесени няколко лекции на семинара по проекта ДемоСем в на които са обсъждани резултатите на дисертацията.

Трудът е на английски език и съдържа 140 страници основен текст включително 13 страници списък от използвана литература. Библиографията включва 141 заглавия (127, ако изключим публикациите с резултатите от дисертацията). Седем страници приложения съдържат списъци на таблиците, фигурите и съкращенията, използвани в дисертацията. Текстът е организиран в увод, 7 глави и заключение. Главите 3-8, които представят мотивацията и разработките на автора, завършват с обобщение на представените резултати и/или изброяване на отворени въпроси за изследване в дисертацията (Глава 3) и бъдещи изследвания (Глави 4-8).

Глава 1 (Въведение) има уводна роля. Тя представя важността на темата и я позиционира като част от полето на автоматичната обработка на естествен език. Представени са целите и задачите на дисертационния труд и мястото на българския език в разработката. Обсъжда се също и структурата на дисертационния труд.

Глава 2 (Дефиниция на основните понятия) представя рамката, в която се провеждат изследванията на докторанта, с въвеждане на основните понятия и формални означения. Обяснена е задачата за разрешаване на лексикалната многозначност (автоматично определяне на частите на речта), както и централната за дисертацията задача за автоматично разпознаване на значенията на думите (снемане на семантичната многозначност). Дефинирани са понятията подобие и свързаност на думи. Обсъдени са множества данни, използвани като лингвистични ресурси за решаване на тези задачи.

Глава 3 (Състояние на изследванията и свързани разработки) описва сегашните решения и езикови ресурси за представяне на думите в системите за обработка на естествен език. Тъй като дисертацията има за цел подобряване на представянията на думите, основен обект на разглеждане е лексиконът по принцип и базата WordNet в частност, както и някои други ресурси. Представен е задълбочен обзор на подходите за решаване на задачата за снемане на семантичната многозначност, както и нейните приложения. Също така е направен обзор на системите, използващи невронни мрежи при решаване на задачи за обработка на текст, с фокус върху задачата за снемане на семантичната многозначност. В обзора са споменати и резултатите на автора, дискутирани накратко като част от съвременното състояние на разработките в областта. Тези указания за оригиналност са полезни за разбиране на мястото на представените резултати в сценариите за решаване на задачата за снемане на семантичната многозначност.

Глава 4 (Рекурентни невронни мрежи за автоматично определяне на частите на речта) представя използването на рекурентни невронни мрежи (РНМ) за решаване на тази задача. Обучението на РНМ за построяване на векторно представяне за думите е извършено върху български корпус от около 220 млн. думи, а обучението на векторите представящи наставките – върху подмножество от около 10 млн. думи. В резултат са получени векторни представяния: думите са вектори с размерност 200, а наставките им –вектори с размерност 50. Предложена е многослойна архитектура с двупосочни рекурентни невронни мрежи с клетки с дълга краткотрайна памет, която позволява конкатенация на вектори за думи и наставки. За тестване е използван корпусът BulTreeBank. При обучение със 10,000 итерации е получена най-добра точност 78.16% при определяне на частите на речта, а с добавяне и на векторите за наставките със скрит слой от 125 неврона е достигната точност от 94.47%. Тази точност е сравнима с най-добрите решения за английски текст и показва, че софтуерът за базисна обработка на българския език вече е със световно качество. Интересно е също така емпиричното потвърждение, че морфологичната информация за думите може да се представя като embeddings толкова успешно, колкото и синтактично-семантичната информация. Това е важно знание за езиците с богата морфология като българския. Считам, че представеният модел за решаване на задачата за автоматично определяне на частите на речта ще бъде полезен на учени,

разработващи подобни системи за други флективни езици, тъй като лесно може да се реализира за друг език ако са налице необходимите лингвистични ресурси.

Глава 5 (Моделиране на лексикалната семантика чрез графи) представя един подход за разрешаване на семантичната многозначност на думите чрез използване на лексикон, построен като графи. Идеята е, че интегрирането на знания (във вид на специфични семантични мрежи) в лексикона ще допринесе за по-успешно автоматично идентифициране на значенията на думите. Първата задача е автоматично извличане на знания от текстове и тя е решавана както за български, така и за английски език. Сравнителният анализ на поведението на предлаганите алгоритми за два езика (български и английски) прави много добро впечатление, защото показва желание за дълбоко и изчерпателно разбиране на свойствата на лингвистичните ресурси, базирани върху графовидни представяния. Направените детайлни експерименти показват кои релации са най-удачно подобрение на лексикона с оглед решаване на задачата за разпознаване на значенията на думите с използване на знания. Показаната точност от над 68% при автоматично разпознаване на значенията показва, че добавянето на знания – информация за релациите, извлечени от графични представяния – подобрява съществено потенциала за решаване на тази задача. В тази глава е демонстрирана широката култура на докторанта в областта на техниките за обработка на естествен език: например синтактичен анализ, както и обстойното познаване на ресурси създадени от Секцията по лингвистично моделиране и обработка на знания (BulTreeBank) и в по-широк план граф-базираните модели за представяне на семантиката на текста, популярни напоследък особено за английския език. Доколкото ми е известно, представената дисертация за първи път изследва граф-базирани представяния на български текст в такава дълбочина и обем. Освен това, поради обема на експериментите, след прочита на глави 4 и 5 читателят оценява уменията на докторанта като програмист и знанията му на професионален информатик, който владее разнообразни техники и ги прилага с лекота. Това впечатление се задълбочава със следващите глави и към него се добавя положително мнение относно обстойните знания на Александър Попов в областта на лингвистиката и съществуващите езикови ресурси за различни езици.

Глава 6 (Дистрибутивно представяне на думи, основни форми и значения с помощта на лексикални ресурси) изследва начини за представяне на думи в контекст, като основен обект са векторите embeddings. Използват се модели за векторно представяне, обучени върху изкуствени поредици от лексикални единици, генерирани от граф-базирани представяния на текст. Тук се използват графи със знание, генерирани от различни множества релации, върху които се правят случайни обхождания с различна дължина. Чрез word2vec е обучен модел за на сходство и свързаност между думи. Показано е как може да се постигне увеличане на гъстотата на графа със знание чрез включване на релации между предикати и аргументи при наличие на филтър, който да отделя смислените релации. В досегашната работа са използвани релациите "субект", "пряко допълнение" и "непряко допълнение". Оценката на получените представяния над задачата за измерване на сходство и свързаност между думи показва подобрение с над 1-3% в зависимост от входните корпуси. Това е указание за продуктивността на подхода и поставя въпросът как в конструкцията да се включват и други смислени

семантични релации, които по принцип могат да се извличат от различни текстове включително от дефинициите в WordNet или Simple English Wikipedia (в последната текстовете по принцип са опростени и подлежат на сравнително по-лесно автоматично анализиране).

Глава 7 (Рекурентни невронни мрежи за снемане на лексикалната многозначност) представя две невронни архитектури *A* и *B*, които подхождат към задачата по различен начин, и съответни експерименти за решаване на задачата за автоматично разпознаване на значението на думата в конкретен контекст. Архитектура *B* се различава от архитектура *A* в заключителната фаза за представяне на контекста. С архитектура *B* е изпробвана стратегия за обогатяване на контекста с повече семантични признаци, тъй като английската Уикипедия е сведена до основни форми и конкатенирана към учебния корпус, над който са създадени векторни представяния за думите с дължина по 300 позиции. С двете архитектури са обучени три модела на векторни представяния, които при използване се доближават плътно до най-добрите известни модели в областта. Конкатенирането на лематизираната версия на Уикипедия към псевдокорпуса съществено подобрява векторното представяне и са надминати показателите на два известни модела, представени през 2015 и 2016 год.

Глава 8 (Успоредно обучение на невронни мрежи върху няколко задачи) представя резултати, свързани с изследване на поведението на невронни мрежи обучени успоредно с цел да се споделят параметри между двете. Показано е, че комбинирането на класификатор за снемане на лексикалната многозначност и система за научаване на векторни представяния на контексти дава по-точен модел и за двете задачи в сравнение с мрежи, обучени за всяка задача поотделно. Направени са също експерименти за успоредно обучение на невронни мрежи за разпознаване на частите на речта и за снемане на лексикалната многозначност. Моделът, обучен върху двете задачи, позволява споделяне на принципи и параметри и се справя по-добре от индивидуално обучени модели. В края на главата авторът пише “едно унифицирано решение, което може да моделира естествения език по различни начини със споделяне на параметрите между различните видове анализи, които се извършват, би било сериозна стъпка напред към построяване на многофункционални и сложно структурирани линвистични и концептуални представяния, които наподобяват човешката мисъл“.

В Заключението (Глава 9) са обобщени получените резултати и се обсъжда важността им. Изброените научни приноси са свързани с усъвършенстваната визия за лексикона като съвкупност от вероятностно обучени векторни модели в смесено пространство от думи, основни форми, граматическа и семантична информация. Научно-приложните приноси са свързани с предложените архитектури на невронни мрежи за обработка на текст и експерименталните доказателства за техните полезни свойства и приложение към ключови задачи. Видно е (което личи и от незабавно появяващите се цитати в публикации на чуждестранни автори), че някои от предложенията за представяне на лексикона чрез векторни модели са оригинални и иновативни в международен мащаб: това са добавянето на морфологична информация (наставки) и граматически роли към модела, както и обучението на дистрибутивни представяния на основни форми и синсети чрез генериране на изкуствени корпуси. Полезно е включването на таблицата на стр. 111-115 с публикации по темата на

дисертацията, в която се дава кратко резюме на съдържанието на всяка статия – така се проследява развитието на труда във времето. Списъкът идеи за бъдеща работа показва амбиция за извършване на изследвания на световно ниво и сравняване с резултатите на най-известните специалисти в полето. Задачите касаят както технологичната страна на използваните ИТ-инструменти, така и подобряване на качеството на лингвистичните ресурси.

Като цяло текстът на дисертацията е стегнат и добре организиран, с ясно разделяне по глави и тематика, и с удобна за възприемане последователност на изложението. Като технически недостатък бих посочила номерирането на бележките под линия от "1" в отделните глави (което не е типично за цялостен текст). Авторефератът на български език отразява коректно съдържанието на дисертационния труд. При превода старателно са търсени преводни еквиваленти на английските термини и според мен сегашната версия на автореферата е много по-добра от предишната в това отношение. Отразени са и моите забележки към версията на дисертационния труд, която беше предадена за предварителата защита.

Според личните ми наблюдения от срещи и семинари, Александър Попов е автор на представените иновативни резултати, което се потвърждава и от декларацията за оригиналност. Свидетел съм на големия интерес, който предизвикват неговите презентации и постери на престижни международни конференции. През четирите години на докторантурата си Александър Попов е натрупал капацитета на зрял специалист по компютърна лингвистика и приложения на невронните мрежи за обработка на текст. Той е отличен пример за експерт с интердисциплинарна подготовка между лингвистиката и информатиката, а текстът на дисертационния труд потвърждава компетентността му в двете области.

Заключение. Считам, че получените резултати и публикуваните статии доказват наличието на експертиза и качества за извършване на самостоятелна научна и научно-приложна работа, които се изискват от ЗРАСРБ за присъждане на образователната и научна степен "доктор". Дисертацията прави впечатление с амбицията си да се позиционира на световно ниво и да се сравнява с най-добрите постижения в областта. На тези основания ще гласувам положително за присъждане на степента и убедено предлагам на уважаемото научно Жури **да присъди на Александър Попов образователната и научна степен "доктор".**

12 октомври 2018