

СТАНОВИЩЕ

от проф. д-р Людмила Петрова Димитрова-Рашкова,
Институт по математика и информатика, Българска академия на науките,
за дисертацията на Александър Николаев Попов “**Modeling Lexical Knowledge for Natural Language Processing**” („Моделиране на лексикалното знание с цел автоматична обработка на естествен език“) за придобиване на образователна и научна степен „доктор“ в област на висше образование 4. Природни науки, математика и информатика, професионално направление 4.6. Информатика и компютърни науки, докторска програма Информатика

Становището представям в качеството ми на член на Научното жури, назначено със Заповед № 127 от 12.07.2018 г. на Директора на ИИКТ-БАН.

Становището е изготвено в съответствие с изискванията на ЗРАСРБ (Закон за развитие на академичния състав в Република България) (ДВ бр. 30 от 3 април 2018 г.) и Правилника за неговото прилагане (ДВ бр. 56 от 6 юли 2018 г.), действащи към момента на откриване на процедурата.

1) Методиката на изследването

Методиката на изследването, използвана от докторанта, е ясно описана. Ясно са очертани целта на изследването, задачите и обхватът на изследването. Докторантът демонстрира грамотност по методологията на научните изследвания и показва способност да я използва практически за провеждане на собствено изследване с последващ анализ и обобщение на получените резултати.

2) Познаване на състоянието на проблема

Докторантът демонстрира разбиране на изследваната проблематика, систематичност и творчески подход, използва съвременни методи и техники за изследване и решаване на поставените задачи. Дисертацията е написана ясно, изложението е стегнато и последователно, изводите са обосновани.

3) Общо описание на дисертационния труд

Представеният дисертационен труд е написан на английски език и се състои от 147 страници, разпределени във Въведение, 7 глави (2 – 8), Обобщение и перспективи, Декларация за оригиналност, Списък с литературни източници, вкл. 14 публикации на дисертанта (4 самостоятелни, 10 в съавторство, като той е първи автор на една от тях), 23 таблици, 8 фигури и списък на съкращения. Дисертацията се фокусира върху модели за представяне на лексикалното знание с цел приложение в автоматичната обработка на естествен език. Лексикалното знание се разглежда в широк смисъл като включва традиционните аспекти на смисъла от лексикалната семантика, допълнени с морфо-синтактична информация, модели на колокация, знание за света и т.н. Представената работа изследва представяне на лексикалното знание със символни, статистически или хибридни средства за специфични задачи на естественоезиковата обработка (ЕЕО).

Разработени са и са оценени две системи за снемане на лексикалната многозначност, основани на знания (т.нар. СЛМОЗ). Използваните алгоритми за решаване на задачата разчитат не толкова на статистическо знание, научено директно от данни, колкото на информация, кодирана в моделиращите лексикона ресурси. Голямото предимство на СЛМОЗ е, че алгоритмите имат пълно покритие над значенията в използвания лексикон. Работата е иницирана във връзка с проект за изграждане на българска версия на лексикалния ресурс на WordNet и за неговото използване за снемане на многозначност на смисъла на думи в текст. Работата е използвана и в проект за изграждане на системи за машинен превод (МП). Тъй като разработваните МП системи обхващат както превода от български на английски, така и от английски на български, изследването е насочено успоредно към двата езика.

4) Анализ на научните и научно-приложните постижения в дисертационния труд

Най-напред (глава 2) са дадени дефиниции на разработваните проблеми. Всяка задача служи като тестова основа за хипотезите, изследвани в дисертацията. Проблемите са описани първо неформално (чрез примери за илюстрация), а след това са дадени формалните им определения. Подробно са обсъдени POS-маркирането, снемането на лексикалната многозначност, пресмятането на сходството и свързаността между думи. Предложен е обзор на определени модели за представяне на лексикони за ЕЕО, свързани с тезата (глава 3). Обзорът се основава на важни публикации, цитирани в Библиографията, включително шест работи на докторанта. Основните резултати са описани подробно в глави 4 – 8 на дисертацията. Рекурентни невронни мрежи за определяне частите на речта са описани в глава 4. Резултатите от експериментите с три модела невронни мрежи, които се различават по вида и броя на входните думи, са сравнени и обсъдени. Глава 5 изследва методите за снемане на лексикалната многозначност, основани на знание (СЛМОЗ). Някои от най-популярните варианти на СЛМОЗ са групирани под името *методи чрез графи*. В глава 6 авторът представя генериране на разпределени представяния на лексикални единици: думи, леми и значения. Глава 7 представя мрежови архитектури за снемане на лексикална многозначност. Последната 8 глава представя оригинални изследвания, проучващи възможностите за снемане на лексикална многозначност паралелно с други задачи на ЕЕО.

5) Оценка на приносите

Формулираните от дисертанта приноси в дисертацията и в автореферата правилно отразяват постигнатите резултати. Представените в дисертацията резултати са получени и използвани в два международни проекта: *QTLep: Quality Translation by Deep Language Engineering Approaches* и *EUCases - EUropean and National CASE Law and Legislation Linked in Open Data Stack*, финансирани от 7РП на ЕС, и в проект *Дълбоки модели на семантично знание (ДемоСем)*, финансиран от Фонд научни изследвания. Освен трите споменати проекта, резултатите от изследванията на дисертанта са представени на редица престижни конференции и семинари у нас и в чужбина.

6) Преценка на публикациите

Резултатите от дисертацията са публикувани в 14 статии. Всички оригинални резултати, описани в дисертацията, са представени на международни конференции и семинари и съответно публикувани в сборници с доклади от конференции, има 1 статия от 2018 г. в списание (SJR=0.204 за 2017). Съгласно базата данни Scopus са забелязани 5 цитирания (без автоцитати и цитирания на съавтори) на 4 от 8 видими статии. Съгласно Web-of-Science има 1 цитат на статия, който е отразен и в Scopus. Има забелязани 2 цитирания на 2 статии в БД, които не фигурират в Scopus / Web of Sciences.

Оценка по ЗРАСРБ (от 2018 г.) и Правилника на МС за прилагането му: *изискванията за получаване на образователната и научна и степен „доктор“ са изпълнени.*

7) Оценка на автореферата

Авторефератът на дисертацията е с обем от 46 страници. В него е обоснована актуалността на разглежданата тема, формулирани са целите и са поставени задачите за изследване. Представени са основните и най-важни резултати, получени в дисертацията, както и приносите и публикациите на автора, участието му с доклади в международни конференции и семинари у нас и в чужбина, имащи отношение към разработваната тематика. Считаю, че авторефератът, както и представеното според изискванията резюме, отразяват правилно съдържанието на дисертацията, резултатите и научните приноси на дисертанта.

8) Критични бележки

Нямам. Не познавам лично дисертанта, но представеният от него труд показва един ерудиран и перспективен млад учен-информатик.

9) Заключение

Представеният дисертационен труд на Александър Николаев Попов на тема „*Моделиране на лексикалното знание с цел автоматична обработка на естествен език*“ отговаря на всички изисквания на ЗРАСРБ за придобиване на образователната и научна степен „доктор“, както и на специфичните показатели, определени в Правилника на МС за прилагането му. Дисертантът е доказал компетентност, научна зрялост и способност за самостоятелна изследователска работа. Ето защо препоръчвам убедено на почитаемото научно жури да присъди на Александър Николаев Попов образователната и научна степен „доктор“ в област на висше образование 4. Природни науки, математика и информатика, професионално направление 4.6. Информатика и компютърни науки, докторска програма Информатика.

София, 12.10.2018 г.

