СТАНОВИЩЕ

за дисертацията на Ивелина Мирчева Николова
„Приложение на обработката на естествен език при изграждане на семантични системи“
представена за присъждане на образователната и научна степен „доктор“
от проф. дмн Галя Ангелова, ИИКТ-БАН

Дисертацията е свързана с някои от най-актуалните тенденции в компютърната лингвистика: *(i)* да се обработват все по-големи обеми от реални текстове; *(ii)* да се анализират текстове, написани на естествени езици, за които са налични ограничени количества от лингвистични ресурси; *(iii)* да се обработват текстове в областта на био-медицината (в случая, клинични пациентски записи) в рамките на идеята за вторично използване на електронните пациентски записи и *(iv)* софтуерните компоненти, анализиращи текст, да се усъвършенстват до степен, позволяваща успешното им влагане в интелигентни системи с по-широко предназначение. Отбелязвам, че дисертантката постепенно се изправи пред горните предизвикателства без колебание (и бих искала с радост да допълня, че това е 3-та дисертация от подобен тип, предадена за защита у нас през 2014).

Съгласно Правилника за специфичните условия за придобиване на научни степени и за заемане на академични длъжности в Института по информационни и комуникационни технологии (ИИКТ) при Българската академия на науките (БАН), кандидатът за получаване на образователната и научна степен "доктор" трябва да има "*поне 3 научни публикации, поне една от които да е в списание с импакт фактор или в специализирано международно издание*". Резултатите от дисертацията са представени в 10 публикации, като 4 от тях са в серии Lecture Notes на Шпрингер и 3 са в издания на международни професионални организации – ACL (Асоциация по компютърна лингвистика) и ELRA (European Language Resources Association). Изброени са и 4 цитирания на трудовете от чуждестранни автори в международни издания. Така изискванията на Правилника на ИИКТ-БАН към кандидатите за получаване на образователната и научна степен "доктор" са изпълнени.

Трудът съдържа 129 страници и е организиран в увод, 4 глави и заключение, списък на използваните термини, съкращения и означения, списък на фигурите, списък на таблиците и библиография с 67 заглавия.

Уводът представя актуалността на темата и мотивира целите и задачите на дисертацията. В Глава 1 се въвежда принципната архитектура на семантични системи, които съдържат компоненти за обработка на естествен език, и накратко се представя състоянието на изследванията по релеванти под-теми в компютърната лингвистика: извличане на понятия и релации от текст, разпознаване на парафрази и автоматично структуриране на описания от пациентски записи.

В Глава 2 "Приложение на OEE за създаване на концептуални модели на предметна област" се представят авторски резултати, свързани с извличане на понятия от текст чрез разпознаване на парафрази (на основните наименования) и извличане от текст на понятийни отношения /релации/ между понятията. Специално за релациите идеята е да се създадат процедури, които използват извънредно богатия английски източник UMLS (Unified Medical

Language System, натрупван в Националната библиотека по медицина на САЩ в последните 30 години). Макар и несистемно декларирани, в UMLS са дефинирани имена на множество отношения между наличните стотици хиляди понятия. Дисертацията показва как тези релации могат да се извличат от оригиналния източник на английски език и да се внасят експлицитно в декларативен понятиен модел, откъдето сравнително лесно да се пренасят в модел с етикети на български език. Според мен едно от важните постижения на дисертацията е, че показва как публичен софтуер за автоматичен анализ на английския език (система RelEx) може да се интегрира в работно място за решаване на споменатата задача. В крайна сметка, едва ли можем да преведем на български език цялото богатство от знания в ресурсите на UMLS, но бихме могли да създаваме внимателно извлечени техни понятийни под-модели, които да се допълват с онтологична терминология на български език. Дисертацията дава едно интересно решение как може да се направи това.

Глава 3 "Структуриране на текстови описания в биомедицината" разглежда извличане на величини (симптоми на диабет и стойности от изследвания) от неструктуриран текст на епикризи на български език, както и информация за събития и темпорални маркери от английски текстове в учебен корпус, предоставен от организаторите на състезанието i2b2 2012. При работа с българските текстове е използван хибриден метод, съчетаващ машинно самообучение и анализ чрез правила. За анализ на английските текстове, освен споменатите подходи, са използвани и платформите MetaMap и GATE. Анализът на българските епикризи е представен в самостоятелна публикация, приета за Докторантския семинар на конференцията на Северо-американското подразделение на Асоциацията по компютърна лингвистика през 2012. Получените резултати съответстват на успеваемостта на подобни задачи, решавани за английски език. Участието в състезанието i2b2 2012 с получен сравнително добър резултат, наред със системи представени от научни групи, работещи в областта от десетилетия, е още едно потвърждение за качеството на постигнатите резултати.

В Глава 4 "Интеграция в прототипи" е представен класификатор, който по думи, налични в пациентски запис, разпознава с висока точност дали се изказва предположение, че пациентът има диабет или фамилна обремененост и рискови фактори относно диабет. Класификаторът е обучен над корпус с ръчно анотирани положителни и отрицателни примери. Направени са редица експерименти за установяване на подходящ метод за машинно самообучение. Класификаторът е интегриран в система, разпознаваща потенциално-застрашени от диабет граждани чрез автоматичен анализ на текстове с техни записи.

Заключението резюмира коректно научните и научно-приложни приноси на дисертацията, като изброява оригиналните резултати заедно с оценка за качеството на обработката на текста.

Дисертационният труд е прегледно подреден и ясно написан. Би следвало да се въведат повече определения на основните понятия, които сега са обяснени в неформален вид в речника на преводните съответствия на термините. Отстраняването на някои технически пропуски (напр. номерация на главите) ще подобри качеството на Автореферата.

Относно речника на преводните съответствия, бих препоръчала прецизно уточнение на термините и поставяне на речника в Интернет на видимо място. Това ще бъде от полза за нарастващото множество експерти по семантични системи и Data Science, които – поради

липса на уточнена терминология на български – използват английски термини дори в текст на кирилица.

Заключение

Считам, че получените резултати и публикуваните статии доказват експертизата на кандидатката и нейния потенциал за извършване на самостоятелна научна и научно-приложна работа, които се изискват от ЗРАСРБ за присъждане на образователната и научна степен "доктор". **Подкрепям с положително заключение присъждането на образователната и научна степен "доктор" на г-ца Ивелина Николова и предлагам на членовете на Научното жури единодушно да гласуват в подкрепа на такова решение.**

31 декември 2014 г.
София