

## РЕЦЕНЗИЯ

от д-р **Иван Колев Койчев**, професор в ФМИ на СУ“Св. Климент Охридски“  
за дисертационния труд на тема

*„Моделиране на лексикалното знание с цел автоматична обработка на  
естествен език“*

с автор **Александър Николаев Попов**

с научен ръководител: доцент д-р Кирил Симов

за придобиване на образователната и научна **степен „доктор“** по:

област на висше образование 4. Природни науки, математика и информатика;  
професионално направление 4.6. Информатика и компютърни науки;

Тази рецензия е изготвена на основание заповед 127/12.07.2018 г. на директора на ИИКТ-БАН, с която съм включен в състава на научно жури за по-горе спомената дисертация и решението на първото заседание на журито, което ми възложи тази задача.

Моделирането на естествените езици е важна задача, както за дисциплините изучаващи принципите на естествените езици, така и за редица съвременни компютърни приложения. Една от задачите, която е важна за обработката на естествен език (ОЕЕ) е намирането на подходящо формално представяне на лексикалното знание. Много често, лексиконът (речник) бива разглеждан като средство за представяне на множество типове информация, като например: синтактична структура, семантични връзки, допустими съсрещания и други. Представената дисертация предлага съвременни подходи за автоматизирано създаване и подобряване на такива речници, които намират широко приложение в много съвременни софтуерни системи.

Следователно, може да заключим, че темата на представената дисертация е в много **актуалната област**.

Дисертацията е с общ обем от 147 страници, в които около 120 страници е самата дисертация. Написана е на английски език, който според мене е много добър. **Добре структурирана е**, като се състои от увод, седем глави и заключение.

В **Глава 1 (увод)** се дава кратка мотивация на темата, ясно се определят целта на дисертацията и произтичащите от нея конкретни задачи. На кратко е описана и структурата на дисертацията.

В **Глава 2** се дават формални определения на поставените задачи като въвеждат основните понятия и математически означения.

В **Глава 3** се прави обзор на основните резултати и съвременните изследвани в областта.

В **Глава 4** е представен разработения подход основан на рекурентни невронни мрежи за определяне частите на речта. В тази глава са представени и планираните и проведени експерименти за проверка и сравняване на предложения подход.

В **Глава 5** е представена подход за обогатяване на семантичния граф на WordNet чрез допълнителни връзки, които по-добре да отчитат контекста. Предложено е явно да се представят връзките към хиперним-хипоним в семантичната мрежата на WordNet, в случая на българския език, както и да се добавят синтактични връзки от допълнителни източници. В края на тази глава са представени и резултатите от опитна проверка на разработените методи. Проведени са експерименти, в които са използвани данни от корпуса с части на речта и лексикални значения BulTreeBank за българския език и корпуса SemCor за английския език.

В **Глава 6** е представено едно по-нататъшно развитие на подходите от предишната глава с цел „научаване“ на модели на разпределено представяне на контекста на думи. Основната хипотеза е, че обогатената семантичната мрежа ще доведе до по-добри резултати при редица задачи в и частност задачата за разрешаване на многозначност на думи. Опитно е проверена тази хипотеза, като е изследвано как обогатените семантични графи подпомагат създаването на по-добри векторни пространства. Комбинират се подходи основани на лексикални знания и такива научени от алгоритмичен анализ на големи корпуси от текстове.

В **Глава 7** се представят два подхода за разрешаване на многозначност на думи основани на рекурентни невронни мрежи, които учат модели на леми, синоними и контекст. Отново са представени и съответната опитна проверка и оценка на предложените подходи с реални данни.

В **Глава 8** изследва подхода на многозадното обучение с използване на рекурентни невронни мрежи. Съчетават се класически методи за разрешаване на многозначност на думи и научените контекстни вектори. И в тази глава се представят резултати от съответната опитна проверка и сравняване на предложените подходи.

Всяка от главите 2-8 завършва с подглава, в която се обсъждат постигнатите резултати, и се предлагат идеи за по-нататъшно развитие на предложените подходи.



В **Глава 9 (заключение)** се обобщават постигнатите резултати, представят се конкретните приноси на докторанта и се предлагат идеи за бъдещи развитие на предложените разработки.

Списъкът на **използваната литература** („Библиография“) е с обем 13 страници и съдържа над 140 заглавия: научни статии, книги, онлайн публикации и документации. Не забелязах такива, към които да няма препратки в текста на дисертацията.

Има **две приложения**, с общ обем от 6 страници, като в първото е даден списък на таблиците, а във второто – списък на фигурите.

Докторантът **демонстрира задълбочени знания** в научните области на дисертацията. Направените обзори са задълбочени и покриват широк спектър от методи, технологии и софтуерни инструменти. Тези обзори адекватно представят текущото състояние в съответните области. Списъкът с използвана литература е представителен за областта и е достатъчен за докторска дисертация. Разработените подходи и модели са добре обяснени, дадени са подходящи илюстриращи примери. Приложените алгоритми са проверени в редица експерименти, които са добре планирани и проведени, а резултатите са добре представени и дискутирани. За лесното разбиране на представените методи и резултати от експерименти спомагат не малкото фигури и таблици в дисертацията (8 фигури, 23 таблици).

Избраната **методологията на изследване е подходяща** за поставените цели в дисертацията. Тя включва: аналитичен обзор на областта, мотивиран избор на подходящи методи и тяхното творческо доразвиване и прилагане за решаването на поставените задачи. Разработените методи са реализирани в софтуерни инструменти и е проверен в реални проекти. Видно е, че докторантът е натрупал опит по целия път на научната работа: от създаването на абстрактни модели до приложението им за решаването на конкретни задачи по реални проекти, например проекта QTLeap.

Дисертацията е много **добре форматирана** и естетически оформена. Издържан е и от гледна точка на правопис, граматика и стилистика. Важните единици в текста са обозначени и почертани с подходящо форматиране. Фигурите и таблиците са надлежно номерирани, имат обяснителни надписи и са коментирани в текста. В края са приложени индекс на фигурите, индекс таблиците и списък със съкращенията.

Научните и научно-приложните **приноси** на дисертацията могат да бъдат обобщени, както следва:

1. Обогаляване на WordNet с множества от нови семантични и лексикални връзки. Опитно е установено, че тези нови връзки значително подобряват резултатите в някои задачи за разрешаване на лексикална многозначност, при които се използват бази от знания.

2. Предложени и опитно проверени са няколко архитектури на невронни мрежи с памет от вида „последователност-последователност“ в рамките на три лексикални задачи: определяне на части на речта, разрешаване на лексикалната многозначност и представяне на контекста.

3. Изследвани са различни векторнопространствени модели (с използване за думи, леми, синоними и наставки) за представяне на теста за алгоритмите за машинно обучение с учител. Някои от тези модели се основават на структурата на семантична мрежа WordNet и на разработените нейни обогатени варианти. Опитно е установено, че тези модели показват много добри резултати при решаване на задачите за намиране на сходство и свързаност между думи. Създадено е ново разпределено представяне на граматическите роли, което е използвано по нов начин за обогаляване на връзките в графа на WordNet.

4. Опитно са сравнени два подхода, основани на многозадачно машинно обучение, за моделиране на лексикални връзки. Резултатите показват, че моделите, използващи лексикални зависимости, се представят по-добре. По-точно казано, тези модели показват подобрения в задачите за „разрешаване на лексикална многозначност с отчитане на контекст“ и за „разрешаване на лексикална многозначност с отчитане на частите на речта“.

Представен е списък от 14 статии, в които са публикувани части от резултатите в дисертацията. Две от тях са самостоятелни, а на 3 от тях докторанта е водещ автор. Всичките публикации са в рецензирани международни издания. Осем от тях са индексирани в SCOPOS, включително и двете самостоятелни, а пет са индексирани в Web of Science, включително и двете самостоятелни. Докторантът е изнесъл 4 доклада по темата на дисертацията на реномирани международни научни конференции, 2 на студентки работни срещи и един на вътрешен семинар. За значимостта на получените резултати говорят и 8-те позовавания на статии, в които са публикувани приноси от дисертацията, като 3 от цитиранията са на публикация с водещ автор докторант. Представените публикации, като бройка и качество, не само изпълняват формалните изисквания на закона и правилниците за дисертационен труд за образователна и научна степен доктор, но и значително ги надхвърлят.



Някой от резултати, представени в дисертацията, са използвани в няколко изследователски проекта, което е свидетелство за тяхната **приложимост**, например обработващи последователности за български език и английски език са използвани в европейския проект QTLeap и европейския проект EUCases.

**Авторефератът е добре направен** и точно отразява съдържанието на дисертационния труд. Термините старателно са преведени на български език, но според мен, в тази насока има място за подобрене. Съкращенията са правилно въведени, като е приложен речник с тях, но според мен са прекалено много, което затруднява разбирането на съдържанието. Като цяло текстът на автореферата може да се подобри с една редакция с цел изчистване на терминологията и описанието.

Към представената дисертация **нямат съществени забележки**.

Не познавам лично дисертанта, но представената професионална биография и другите документи свидетелстват, че той е високо квалифициран специалист с богат опит и изследователски дух.

**Заключение:** **Оценката ми** за дисертационния труд, автореферата, научните публикации и научните приноси на Александър Попов **е положителна**.

Представеният **дисертационен труд отговаря напълно на всички изисквания**, условия и критерии по Закона за развитието на академичния състав в Република България (ЗРАСРБ), Правилника за прилагане на ЗРАСРБ, Правилник за условията и реда за придобиване на научни степени и за заемане на академични длъжности в БАН, и Правилник за специфичните условия за придобиване на научни степени и за заемане на академични длъжности в ИИКТ-БАН. На тази основа предлагам на уважаемото жури да **присъди образователната и научна степен „доктор“ на Александър Попов** в област на висше образование 4. Природни науки, математика и информатика; професионално направление 4.6. Информатика и компютърни науки.

09.10.2018 г.  
гр. София

