Р Е Ц Е Н З И Я

относно кандидатурата на доц. д-р Кирил Симов за участие в конкурс
за заемане на академичната длъжност „професор“ по професионално направление
4.6. Информатика и компютърни науки, специалност Информатика
от проф. дмн Галя Ангелова, ИИКТ-БАН

Конкурсът е обявен в „Държавен вестник” бр. 49 (21.06.2019) за нуждите на ИИКТ-БАН, секция „Лингвистично моделиране и обработка на знания”. Единствен кандидат е доц. д-р Кирил Симов. Съгласно изискванията на *Правилника за специфичните условия за придобиване на научни степени и за заемане на академични длъжности в ИИКТ-БАН,* който задава завишени критерии за хабилитация в сравнение с минималните национални прагове и изискванията на БАН, кандидатите за заемане на академичната длъжност „професор” в област 4, професионално направление 4.6 Информатика и компютърни науки, трябва да имат над 100 точки в група показатели В, над 260 точки в група показатели Г, поне 140 точки в група показатели Д и поне 150 точки в група показатели Е. Доцент Симов представя справка за изпълнение на минималните изисквания на ЗРАСРБ със следния брой точки: 120 за група показатели В, 324 за група показатели Г, 228 за група показатели Д, и 648 за група показатели Е. Приемам пресметнатите точки по отделните показатели така, както ги е представил кандидатът. Доцент Симов е придобил научната и образователна степен доктор през 2006 г., има над 5 години трудов стаж като доцент и в научните му трудове не е установено плагиатство. С това формалните изисквания на ЗРАСРБ и Правилника на ИИКТ-БАН са изпълнени и дори значително надхвърлени особено при показателя Е.

Кратки биографични данни за кандидата

Доц. Симов получава магистърска степен по информатика от Факултета по математика и информатика – Софийски университет през 1986 г. През 2006 г., като докторант в Института по математика и информатика на БАН (секция Математическа лингвистика) защитава кандидатска (по сегашната терминология – докторска) дисертация на тема „Логически средства за обработка на лингвистични знания в опорната фразова граматика”. Научните му интереси се насочват не само към формализмите за обработка на информация в компютърната лингвистика, но и към систематичното създаване на езикови ресурси за българския език. Работил е в Централния институт за програмни продукти и системи, Института по математика и механика на БАН, Координационния център по информатика и изчислителна техника (КЦИИТ) към БАН, неговите наследници ЦЛПОИ и ИПОИ, където се хабилитира през 2007 година и в сегашния Институт по информационни и комуникационни технологии (ИИКТ) - БАН, където понастоящем е ръководител на секция „Лингвистично моделиране и обработка на знания”.

Общо описание на представените материали

За участие в конкурса са представени общо 27 публикации на английски език, видими в Scopus, като за покриване на критериите в група В са представени 8 публикации и за

покриване на критериите в група Г - 19 публикации. Те са издадени след годината на хабилитация (2007) и не са представяни при предишни процедури за получаване на степен или заемане на академична длъжност. Петнадесет от представените за конкурса статии са в списания или поредици с импакт ранг (SJR), а осем са индексирани от Web of Science. Макар че в Scopus кандидатът има 45 публикации след 2007 г., очевидно той се е ограничил да подаде за конкурса статиите, които представят най-съществените му резултати през последните 12 години. Би трябвало също да отбележим, че в предадената за конкурса Автобиография е включен пълен списък на публикациите на доц. Симов – общият брой е 180.

В справката за цитирания, представена от кандидата, са посочени 38 цитирания на 8 видими в Scopus публикации. Общият брой на цитиранията на доц. Симов в Scopus е 277 цитирания с h-индекс 10, а в Гугъл Сколар - 1919 цитирания с h-индекс 21. Макар че последните индикатори включват автоцитиранията, големият им брой и чуждестранните източници показват известността на най-добрите публикации на доц. Симов по целия свят.

Всички представени публикации са в съавторство, като кандидатът е първи автор в шест от тях. Приемам, че получените резултати, описани в публикациите по конкурса, са постигнати с равностойно участие на авторите на съответните трудове. Съавторството не намалява значението на постиженията на доц. Симов, а по-скоро подчертава неговата позиция на учен с международно-разпознаваема експертиза по компютърна лингвистика, който си сътрудничи с голям брой съавтори от цяла Европа в една активно-развиваща се интердисциплинарна област. Големият брой цитати от разнообразни международни източници е доказателство за значимостта и актуалността на научните постижения на доц. Симов.

Научни резултати и приноси

Научните резултатите на кандидата се отнасят към три под-области на компютърната лингвистика и изкуствения интелект: *(i)* създаване на езикови ресурси и езикови технологии за българския и други езици (главно английски); *(ii)* създаване на концептуални ресурси и семантични технологии и *(iii)* приложения на езикови и семантични ресурси и технологии. Представените в публикациите оригинални резултати в трите под-области са следните:

- *В областта създаване на езикови ресурси и езикови технологии*: Разработка на анотационни схеми и формати на езикови ресурси. Дизайн и реализация на процедури за извличане на информация с цел създаване на нови езикови ресурси. С този подход са създадени корпус с политическа реч, валентен речник на българския език и система за транскрибиране на DBpedia URIs от други езици на български.

  Предложено е общо представяне на лингвистичното знание, което позволява едновременно решаване на няколко задачи за автоматична обработка на текст. Представянето има вида на разширено депендентно дърво, което кодира информация необходима за синтактичен и семантичен анализ, както и разрешаване на кореферентни връзки. На базата на получените резултати е проектирана иновативна система за синтактичен анализ с използване на резултатите от няколко синтактични анализатора, чрез „гласуване“ и избор между локален и глобален анализ. Също така е проектиран алгоритъм, който построява представяне на семантиката на дадено изречение. Алгоритъмът работи върху депендентното синтактично дърво за изречението.

С цел създаване на езикови технологии за българския език са решавани задачи като граматично анотиране, разпознаване на наименовани същности, синтактичен анализ, както и техни комбинации. Резултатите при тестване са били най-добрите за български текст към времето на публикуване на статиите.

- *В областта създаване на концептуални ресурси и семантични технологии*: Създаден е модел на релацията онтология-текст чрез терминологичен лексикон и анотационна граматика. В речника се задават основните форми на термините, морфологични характеристики и граматична информация за синтактичната им реализация в текста. Предложени са стратегии за установяване на съответствия между елементи на лексикона и онтологията при липса на еквивалент на лексикален елемент или понятие. Предложена е методология за изграждане на онтологии на предметни области. Тъй като първоначалният етап при ръчно натрупване на онтологично знание е бавен и трудоемък, предлага се натрупването да започне от ядро от понятия, извлечени автоматично от текстови източници вкл. стандарти или готови терминологични речници. След определяне на ключови термини и фрази се прилага и синкатичен анализ, който позволява да се извлекат семантични релации между идентифицираните термини. Същността на формализацията е преход от ключовите думи към дефинирането на понятия и тяхното кодиране във формален език (например OWL).

  Измежду най-актуалните резултати на доц. Симов са предложените алгоритми за разширяване на граф със знание. Алгоритмите работят над релации, извлечени от понятия организирани като граф от знания построен върху семантичната мрежа на WordNet. В този граф понятията (съвкупност от синонимни думи) са възли, а релациите между тях – дъги. Така построените графи се използват за разрешаване на семантичната многозначност (Word Sense Disambiguation) чрез алгоритми за случайно блуждаене (Random Walk on Graphs), поради което свързаността на графа определя и качеството на обработката. Представени са няколко алгоритъма за извличане на нови релации или адаптиране на цели изречения за разширяване на графа, построен на базата на WordNet. Експериментално е показано, че използването на разширените графи подобрява резултата с около 10%.

- *В областта приложения на езикови и семантични ресурси и технологии*: Изградена е IT-онтология с около 2,500 понятия за целите на електронното обучение и друга онтология в областта на домашния текстил за целите на бранша за производство и реализация на такъв текстил. В проекта по 6-та рамкова програма LT4eL (Езикови технологии за електронно обучение) е разработена функционалност за разширяване на заявките с извод над онтологията по информационни технологии. Показано е, че това разширяване подобрява търсенето около три пъти в сравнение с пълнотекстовото търсене за всички седем езика, разглеждани в проекта. В друг проект е изграден подходящ речник и граматика чрез анализ на текстове, свързани с българската иконография; резултатът е използван за анотиране на описания към колекция на български икони. Създаден е онтологично базиран речник за българския език, по-късно разширен до сегашния български WordNet – BTB-WN, съдържащ в момента около 22,000 понятия.

  Разработена е комбинация от няколко свободно достъпни бази със свързани отворени данни (включващи DBpedia, Geonames, Freebase и някои други) като ресурс от знания,

който улеснява извършването на непротиворечиви изводи. Приносът на доц. Симов е дефинирането на обединяваща онтология, която позволява достъп до всички данни в интегрираната база.

Разработени са две системи за отговори на въпроси, които ползват предложените езикови технологии, описани по-горе. Приносът на доц. Симов е в дефинирането на архитектурата на системите и определянето на характеристиките за оценка на валидност на отговорите.

С използване на предложената от доц. Симов релация *онтология-текст* е създадена база (отворени данни) от около 650,000 юридически документа свързани чрез онтологиите EuroVoc и Geonames. Приносът на доц. Симов е създаването на структурната онтология за представяне на документите и извличането на RDF-фактите от анотираните документи.

В заключение на справката за академични приноси, кандидатът представя и плановете си за бъдеща работа – развиване на езикови технологии с използване на невронни мрежи и дълбоко самообучение, както и прилагане на езикови технологии за създаване на графи със знания. Тези въпроси са днес измежду най-актуалните теми на изкуствения интелект.

Участие в научни инициативи: проекти, организиране на научни форуми

Доц. Кирил Симов, заедно с проф. Петя Осенова от Факултета по Славянски филологии на СУ „Св. Кл. Охридски“, е изключително последователен в идеята за систематично създаване на езикови ресурси за българския език. От 2001 до 2007 г. в двуфазов проект между Университета в Тюбинген и ИПОИ-БАН, финансиран от Фондация „Фолксваген“ – Германия по програмата „Кооперииране с учени от естествените и инженерни науки в Централна и Източна Европа“, той ръководи разработката на Банка от 15,000 синтактични дървета за българския език с анотация по формализма HPSG. Наречен BulTreeBank, ресурсът е рядко явление за периода 2005-2007 година (а за славянски език – изключително), и привлича интереса на много компютърни лингвисти, занимаващи се със синтактичен анализ. Тази банка е съпроводена от морфологичен речник и анотиран корпус от текстове, които стават основа на т.нар. BLARK (Basic Language Resource Kit). Ако днес българският език е снабден с достатъчно езикови ресурси за базисна автоматична обработка, това се дължи на проектите BulTreeBank, настойчивостта на доц. Симов и предприемчивостта на ИПОИ-БАН да приеме разработката на лингвистични продукти като приоритетна задача. От 2008 до 2016 г. доц. Симов ръководи българския тим в осем други проекта, финансирани по научни програми на Европейската комисия, и се утвърждава като един от малкото компетентни компютърни лингвисти в Балканските страни. Доц. Симов е желан партньор в разработки, свързани с многоезикови задачи напр. машинен превод. Проектите му позволяват да формира група, която продължава да работи и до днес по нови и сложни предизвикателства като семантичния анализ на текста и е основното ядро за изпълнение на инфраструктурния проект КЛАДА-БГ.

След 2007 г. (т.е. след хабилитацията за доцент) Кирил Симов съвместно с Петя Осенова е организатор на 10 международни научни форума, между които престижното 30-то издание на Европейското лятно училище по логика, език и информация, проведено в рамките на две седмици в София през август 2018 г. с около 400 участници. Изключително високата посещаемост на лятното училище е още един указател за известността на организаторите и

доверието във високото качество, с което те изпълняват поетите задължения към международната общност по компютърна лигвистика.

Лични впечатления

Познавам Кирил Симов от 1984 г., когато той беше един от моите първи дипломанти в секция „Математическа лингвистика“ на Института по математика и механика на БАН. Голямо впечатление правеха неговата задълбоченост, упоритост, постоянство и мотивация. През 1988 г. той беше първият програмист, който се зае да работи сериозно по системата MorphoAssistant за морфологичен анализ и синтез на български думи. Днес, 30 години по-късно, виждаме че доц. Симов е запазил интереса си към създаване на речници и корпуси за българския език и продължава тази дейност с разработка на ресурси за синтаксиса и семантиката.

С годините доц. Симов разви разбиране и съпричастие към работата на колегите от хуманитарните специалности, особено тези свързани с изучаване на езика и културното наследство. Това рядко за информатик качество му помага извънредно много при ръководството на големи проекти като КЛАДА-БГ, в който повечето партньори идват от лингвистични звена и музеи. Широката професионална компетентност на доц. Симов привлича и много кандидати за задочна или платена докторантура, които го предпочитат за свой ръководител (например в областта на биомедицинските онтологии или базите от данни – такива са докторантите идващи от фирма Пенсофт или Йенс Колер от Университета по приложни науки в Манхайм, който защити докторска дисертация в ИИКТ през 2018 г. с доц. Симов като научен консултант).

Заключение

Доц. Кирил Симов е водещ европейски изследовател в областта на компютърната лингвистика, познат като създател на средата CLaRK за разработка на корпуси, на ресурса BulTreeBank, като координатор на CLARIN-BG и CLADA-BG и организатор на много научни форуми. Представените за конкурса материали доказват неговите задълбочени познания по езикови и семантични технологии, капацитета му за ръководене на проекти, способността за формиране на колективи, постоянството в привличане на докторанти - а всичко това са качества, които се предполагат като присъщи на "професор" в ИИКТ. **Подкрепям убедено избора на доц. Кирил Симов за професор в секция „Лингвистично моделиране и обработка на знания” на ИИКТ-БАН и предлагам на членовете на Научното жури единодушно да гласуват за приемане на такова решение.**

16 октомври 2019
София

Член на Научното жури за процедурата:

проф. дмн Галя Ангелова