

РЕЦЕНЗИЯ

по конкурс за заемане на академична длъжност „професор“
в професионално направление 4.6 Информатика и компютърни науки,
специалност Информатика, за нуждите на секция „Изкуствен интелект и езикови
технологии“, ИИКТ-БАН, обявен в ДВ бр. 45 от 28.05.2021 г.
и на интернет страницата на ИИКТ-БАН

Рецензията е изготвена от проф. д-р Тинко Величков Тинчев от Софийски университет „Св. Климент Охридски“, ФМИ, професионално направление 4.5 Математика, в качеството му на член на научното жури по конкурса съгласно Заповед № 185 от 27.07.2021 г. на Директора на ИИКТ-БАН и в съответствие с решение на научното жури.

За участие в обявения конкурс е подал документи единствен кандидат:
доц. дн Стоян Милков Михов от секция „Изкуствен интелект и езикови
технологии“ на ИИКТ-БАН.

I. ОБЩО ОПИСАНИЕ НА ПРЕДСТАВЕНИТЕ МАТЕРИАЛИ

1. Кратки биографични данни

Доц. дн Стоян Милков Михов от 1988 г. до 1993 г. е студент във Факултет по математика и информатика (ФМИ) на Софийски университет „Св. Климент Охридски“ като завършва с квалификация „магистър по математика“ с дипломна работа по математическа логика на тема „Унификация на корегулярни множества“. От 2000 г. е „доктор“ по информатика с дисертация „Минимални ациклични автомати: конструкции, алгоритми, приложения“. От 2020 г. е „доктор на науките“ в професионално направление 4.6 Информатика и компютърни науки с дисертация „Крайни автомати, преобразуватели и бимашини: алгоритмични конструкции и имплементации“.

От 1995 г. и до сега доц. Стоян Михов работи в ИИКТ-БАН (в т.ч. КЦИИТ, ЦЛПОИ и ИПОИ, на които ИИКТ е правопреемник) последователно като програмист (), асистент (), главен асистент, а от 2006 г. е доцент в секция „Изкуствен интелект и езикови технологии“. Той е утвърден учен в международната научна общност с 1 монография и над 60 публикации в реномирани специализирани научни списания и трудове на високо ценени научни конференции, 33 от които са в Scopus и/или WoS и които са с над 369 цитирания (без самоцитирания) Scopus и h-индекс 7 (без самоцитирания). Доц. Михов има солиден опит в научно-приложни изследвания с индустриални приложения: национален координатор, ръководител и участник в редица европейски, международни и национални проекти (IMPACT, OCoRrect, AComIn, CULTURA, Електронно здравеопазване в България (е-здраве), SpeechLab и др.),

изследователска дейност в Коммек ЕООД и Рила Солюшънс ЕАД, активен от 2011 г. патент в USPTO.

Доц. Стоян Михов има активна преподавателска дейност от 2003 г. в бакалавърските програми на ФМИ и особено ценна в магистърските програми Логика и алгоритми (спец. математика и спец. информатика) и Компютърна лингвистика (спец. информатика) на катедрата по математическа логика. Той е ръководител на 10 успешно защитени магистърски работи, 8 във ФМИ и 2 във Факултет Славянски филологии. Под негово ръководство са защитени успешно 2 докторски дисертации (1 във ФМИ и 1 в ИИКТ).

2. Общо описание на представените документи

Представените от доц. Стоян Михов документи за участие в конкурса са в пълно съответствие със ЗРАСРБ, ППЗРАСРБ и със специфичните изисквания в правилниците на БАН и на ИИКТ-БАН. Те включват:

- молба до директора на ИИКТ за допускане до участие в конкурса;
- автобиография по европейски образец;
- копие на диплома за образователна и научна степен „доктор“;
- копие на диплома за научна степен „доктор на науките“;
- удостоверение от ИИКТ с изходящ номер 395 от 24.06.2021 год. за стаж по специалността (28 год., от които 15 год. и 1 месец като „доцент“);
- списък на публикации за участие в конкурса (18 броя, неизползвани в предходни процедури);
- списък на цитирания от Scopus (213 броя, неизползвани в предходни процедури);
- резюмета на публикациите за участие в конкурса (на английски и на български език);
- копия на публикациите за конкурса (18 броя, неизползвани в предходни процедури);
- справка за изпълнение на минималните национални изисквания по чл. 2б, ал. 2 и 3, и на изискванията на ИИКТ по чл. 2б, ал. 5;
- справка за оригиналните научни и научно-приложни приноси на представените по конкурса публикации;
- декларация, че няма доказано по законоустановения ред плагиатство в публикациите;

- препоръка, съдържаща декларация за равностойно съавторство, от проф. Клаус Шулц от LMU Мюнхен;
- имейл от техническия редактор Bianca Truthe на Journal of Automata, Languages, and Combinatorics, удостоверяващо приемането за печат на статията с номер 4 от списъка на публикации за участие в конкурса;
- наукометрични показатели за регистриране в НАЦИД на доц. Стоян Михов като „професор“;
- копие на CD на всички представени документи за участие в конкурса.

3. Обща характеристика на публикациите за участие в конкурса

За участие в конкурса доц. Стоян Михов представя 18 тематично много добре фокусирани в 3 направления публикации – 16 статии, един препринт от Arxiv и един активен от 2011 год. патент в Американската патентна агенция.

Всичките 16 статии се съдържат в Scopus и/или WoS. Те носят следните точки за показател Г7 (съгласно Протокол от 48-то заседание на Седмото ОС на БАН коефициентът за Г7 е 2):

- статиите [1] и [2] са съответно в списанията Natural Language Engineering (2007) и Theoretical Computer Science (2019), които за тези години са в квартал Q4 за WoS, поради което всяка от тях носи 24 т.;
- статиите [3], [4], [5], [6] и [8] са в издания, реферирани от Scopus и са с SJR, поради което всяка от тях носи 20 т.;
- статията [9] е в списанието Computational Linguistics (2006), което е в квартал Q1 за 2006 год., значи носи 50 т.;
- статиите [10], [11], [13] и [15] са в издания, реферирани от Scopus и са с SJR, поради което всяка от тях носи 20 т.

Следователно точките по показател Г7 са $2 \times 24 + 5 \times 20 + 1 \times 50 + 4 \times 20 = 278$. Тъй като патентът [18] носи 25 т. по показател Г9, Точките по показател Г от националните минимални критерии, пресметнати съгласно решението от „Протокол от 48-то заседание на Седмото ОС на БАН“ са **303**, и значи изискването на специфичните правила на ИИКТ за поне 260 т. по показател Г е изпълнено.

Приложеният списък от цитирания на статии, налични в Scopus, на доц. Стоян Михов от статии, които са също в Scopus, съдържа 213 единици. Този списък не съдържа самоцитирания и вече използвани в предходни процедури. Съгласно решението от „Протокол от 48-то заседание на Седмото ОС на БАН“ точките по показател Д са **619**, значително повече от изискуемите от специфичните изисквания на ИИКТ 140 т.

Всички представени публикации са съвместни. Приемам равния принос на авторите, предвид декларираното от доц. Стоян Михов и декларираното от проф. Клаус Шулц. Всъщност, декларацията на проф. Шулц е в контекста на обстойно мотивирана препоръка за избора на доц. Стоян Михов за професор.

4. Съдържателен анализ на научните и научно-приложните постижения на кандидата, съдържащи се в публикациите за участие в конкурса

Научните изследвания на доц. Стоян Михов са до голяма степен стимулирани от приложимост към нетривиални практически проблеми за обработка на естествен език, корекция и нормализация на текстове, разпознаване и генериране на реч. Държа да отбележа, че визираните конкретни приложения, независимо от тяхната самостоятелна ценност, са свидетелства за смислено богатство на абстрактните структури и операции над тях, изучавани от доц. Михов. Разбира се, връзката между абстрактните модели и практическите езикови приложения обикновено не е линейна – необходими са различни специфични големи колекции от данни, корпуси. Разработването на методология за създаването, създаването на корпуси и оценката им изисква не само ресурси, но и особен вид творчество, присъщо на научно-приложната дейност. В представените за конкурса 18 публикации има редица съществени научни и научно-приложни приноси, които са правилно отразени в авторската справка, следвайки подходящо групиране на публикациите в три по необходимост пресичащи се направления на приносите.

А. Приноси в теория на крайните автомати. В тази група се отнасят статиите [1]-[4] от списъка с представените публикации. Статиите [1] и [4] се разглеждат два варианта на задачата за презапис на текст: при даден краен речник за презапис (списък от двойки думи) се построява се построява подпоследователен преобразувател, осъществяващ презаписа. При първия вариант заместването е безконтекстно, докато при втория се допускат ляв и десен контекст, като левият контекст се задава с произволен регулярен израз, а десният контекст е краен списък от думи. Ограничението за десния контекст е съществено за съществуването на подпоследователен преобразувател. И в двата варианта конфликтите при заместване се разрешават според най-разпространената стратегия: заменя се първото от ляво най-дълго срещане на дума за заместване. Представената ефективна конструкция на подпоследователен преобразувател, използващ преходи при неуспех и алгоритъм за композиция на такива преобразуватели, са съществени научни приноси.

В статията [3] са създадени алгоритми за композиране на условно вероятностен подпоследователен преобразувател с вероятностен подпоследователен преобразувател с преходи при неуспех. Това е съществен научен принос, който намира приложения в задачата за разпознаване на реч.

В статията [2] се развива нов принцип за строене на бимашини. В частност, дефинирана е конструкция на почти оптимална бимашина по даден функционален реалновременен преобразувател, което е съществен научен принос.

Б. Приноси в обработка на естествен език и разпознаване на реч. В тази група попадат публикациите [5]-[8] и [18].

В статията [5] е представена първата система за разпознаване на непрекъснатата българска реч при голям речник, която е базирана на скрити Марковски вериги с използване на Гаусови смеси. Съществено подобрене на системата е представено в статията [6], дължащо се на построяването на подходящ детерминиран краен автомат, икономично представящ по-широк кръг от кандидати от използваните в [5]. Показана е статистическата значимост на подобрието. Този алгоритъм за преценка на решетката с кандидати е съществен научен принос. Реализацията на тези системи е безспорен научно-приложен принос.

В статията [7] е представен създаден корпус от българска реч от записи на 147 диктора, съдържащ 21891 изказа с обща продължителност около 32 часа. Този корпус, използван за тренирането на системата от [5] значително подобрява разпознаването на юридически текстове след адаптиране към диктор. Оказва се, че този корпус е с недостатъчна големина за технологиите, базирани на дълбоки невронни мрежи. В статията [8] е представен корпус, съдържащ записи на 572 диктори с обща продължителност около 250 часа. За създаването на този корпус са използвани разнообразни техники за обработка на естествен език и технологии за разпознаване на реч. Създаването на тези два корпуса, базирано на различни технологии, а така също и методологията, са научно-приложни приноси.

Публикацията [18] е активен патент в американското патентно ведомство. Той е методология за анализ на връзките между различни субекти, базиран на извличане на специфична информация от големи масиви от текстове на естествен език с последващ анализ на подходящо построен граф. Инструментът, построен по тази методология, е полезен за измерване, анализиране и планиране на медийната комуникация и е несъмнен научно-приложен принос.

В. Приноси в областта на приближено търсене, корекция и нормализация на текстове. В тази група са публикациите [9]-[19]. В авторската справка на доц. Михов те са подходящо разделени на 3 подгрупи.

В1. Алгоритми за приближено търсене, [13]-[15]. Статията [13] е лекция по покана в специалната сесия Computational Linguistics на 7th Conference on Computability in Europe, CiE 2011, Models of Computation in Context. Представя добре мотивирано ~~представя~~ приближеното търсене като специален вид изчисление, което е несъмнен научен принос с методологичен характер. Един оригинален подход към търсенето е алгоритъмът, представен в [14], целящ избягването на така наречения ефект на стената. Алгоритъмът показва особена ефективност при огромни речници от много дълги думи (записи) и големи разстояния. Една конкретна негова реализация е представена в [15]. С тази реализация авторският колектив спечели състезанието String Similarity Search, организирано в рамките на съвместната конференция EDBT/ICDT 2013. Този алгоритъм, наречен WallBreaker, е съществен научен принос.

В2. Корекция на текстове, [9]-[12]. В статията [9] е представена методология за автоматично и полуавтоматично съставяне на речници с грешни изписвания на думи и методология за автоматично създаване на речници с правилно изписани думи в

интернет страници. Създадената е и методология за орфографическа корекция, позволяваща значително подобряване на качеството на корпусите, автоматично извлечени от интернет страници. Тези методологии са използвани за създаване на универсален автомат на Левенщайн за извличане на списъците с кандидати за корекции в статиите [10] и [11]. Алгоритъмът в [10] използва ръчно коригиран корпус на грешните думи, докато алгоритъмът от [11] дава сходни резултати, но без използването на такъв речник. Статията [12] е последващо тяхно развитие, при което се ранкира списъкът от кандидатите за корекция, като се използва биграмен езиков модел.

В3. Нормализация и модернизация на исторически текстове, [16] и [17]. В статията [16] с помощта на инфиксни структури, базирани на крайни автомати, се постига много добра корекция на исторически вариации на думи, като се използва ръчно създаден корпус на наблюдавани вариации. В [17] идеите от [16] са продължени с използване на така наречените функционални автомати, като задачата за нормализация на текста се интерпретира като специален случай на задачата за машинен превод. Резултатите от сравнителния анализ с известните специализирани системи VARD2 и Moses показват значително по-високо качество на разработения алгоритъм.

5. Характеристика и оценка на преподавателската дейност на кандидата

Доц. Стоян Михов има значителен опит и високи резултати в преподавателската дейност в рамките на Софийски университет с лекциите си във Факултет по славянски филологии (ФСлФ) и във ФМИ. Във ФСлФ той водеше курса Формални езици и граматика в магистърската програма „Компютърна хуманитаристика“ в периода 2003-2006 год. От 2015 год. води курса „Търсене и извличане на информация. Дълбоко машинно обучение“, който е основен за магистърската програма „Компютърна лингвистика“, но е отворен за останалите магистърски и бакалавърски програми. Това е един задълбочен курс по съвременните подходи към тематиката. Особено място в преподавателската дейност на доц. Михов заема курсът „Приложения на крайните автомати“ за магистърските програми към катедрата по математическа логика, отворен и за бакалавърските програми във ФМИ, който той чете от 2003 год. Този курс дава и специфична теоретично-приложна окраска на магистърските програми „Логика и алгоритми“. Той, наред с класически резултати за крайни преобразуватели, съдържа постоянно обновяема част с най-нови резултати в областта и практическа реализация на алгоритми. Предизвиква голям интерес у силните студенти и е източник на теми за дипломни работи за магистри. Защитени са успешно 8 дипломни работи, две докторски дисертации (една във ФМИ и една в ИИКТ) и има един докторант в ИИКТ, който е в процес на обучение.

6. Критични бележки и препоръки

Нямам съществени критични бележки.

7. Лични впечатления за кандидата

Познавам лично доц. Стоян Михов от 1992 год., когато беше изявен силен студент в магистърската програма към катедрата по математическа логика. От семинара, който доц. Михов води от двадесетина години в ИИКТ, и неговото участие в семинари и конференции, свързани с катедрата по математическа логика във ФМИ, съм свидетел на неговото професионално развитие. Високо ценя преподавателската му дейност във ФМИ. През годините сме имали и ползотворна съвместна работа по 3 проекта. Наред със завидната му ерудираност в редица области на познанието и творческа активност държа да отбележа и неговите способности да организира и ръководи екипна работа. Той има рядкото умение да види абстрактна форма на практически проблем, която да реши и да имплементира решението в завършен технологичен индустриален продукт.

8. Заключение за кандидатурата

След като се запознах с представените в конкурса материали и публикации, въз основа на направения анализ на тяхната значимост и съдържащи се в тях значителни научни и научно-приложни приноси, **потвърждавам**, че научните и научно-приложните постижения отговарят на изискванията на ЗРАСРБ, Правилника за приложението му и съответните Правилник на БАН и ИИКТ-БАН за заемане от кандидата на академичната длъжност „професор“ в научната област и професионално направление на конкурса. В частност, кандидатът доц. дн Стоян Милков Михов удовлетворява изискванията на ИИКТ-БАН (а следователно и минималните национални изисквания) в професионалното направление и не е установено плагиатство в представените по конкурса публикации.

Давам своята **положителна** оценка на кандидатурата.

II. ОБЩО ЗАКЛЮЧЕНИЕ

Въз основа на гореизложеното, **убедено препоръчвам** на научното жури да предложи на Научния съвет на ИИКТ-БАН да избере доц. дн Стоян Милков Михов да заеме академичната длъжност „професор“ в професионално направление 4.6. Информатика и компютърни науки, специалност Информатика.

23.09.2021 г.

NOT FOR PUBLIC RELEASE

(проф. д-р Тинко Тинчев)