REVIEW

regarding the application of Assoc. Prof. Stoyan Milkov Mihov for participation in a competition to hold the academic position of "professor" in the professional field 4.6. Informatics and computer sciences, specialty 01.01.12. Informatics

by Prof. Galia Angelova, IICT-BAS

The competition was announced in the State Gazette no. 45 (from May 28, 2021) for the needs of IICT-BAS, department "Artificial Intelligence and Language Technologies". The only candidate for the competition is Assoc. Prof. Stoyan Milkov Mihov. According to the Regulations for holding academic positions at IICT-BAS and fulfillment of the minimum requirements for professorships, the candidates for the academic position “professor” must have at least 50 points in indicator “A”, 100 points in indicator “B”, 260 points in indicator “Г”, 140 points in indicator “Д” and 150 points in indicator “E”. Assoc. Prof. Mihov presents a completed Registration Form for NACID, which contains 50 points in indicator “A”, 100 points in indicator “B” (from a monograph published in Cambridge University Press in 2019) , 470 points in indicator “Г”, 1704 points in indicator “Д” for citations and 318 points in indicator “E”. Assoc. Prof. Mihov has over 30 years of work experience in computer science - as a researcher in IICT-BAS and lecturer in informatics in Sofia University, as well as a PhD diploma issued by the Higher Attestation Commission in 2000 and diploma from IICT-BAS for awarding the scientific degree "Doctor of Sciences" in 2020. Thus the provisions of the formal requirements in the Regulations are not only fulfilled, but also significantly exceeded, especially for indicator “Д” for citations.

Brief biographical data about the candidate

Assoc. Prof. Mihov graduated with a master's degree in the Faculty of Mathematics and Informatics at Sofia University "Kl. Ohridski ”in 1993, defending a thesis on “Unification of coregular sets” with supervisor Assoc. Prof. Anatoly Buda. In 2000 he defended at IICT-BAS a doctoral (in the current terminology - PhD) dissertation on "Minimal acyclic automata: constructions, algorithms, applications" with supervisor Prof. Dimitar Skordev. He became an Associate Professor of informatics at IICT-BAS in 2006 in the Department "Linguistic Modeling" (now "Artificial Intelligence and Language Technologies"). In 2020 he defended at IICT-BAS a dissertation on "Finite automata, transducers and bimachines: algorithmic constructions and applications" for the degree of "Doctor of Science". During most of the years of professional activity he combines the scientific work in IICT-BAS, related to participation in various research activities and many projects, with regular teaching at Sofia University “St. Kl. Ohridski” and industrial research for Bulgarian companies. In recent years, his interests in deep theoretical investigations and creation of practical applications have allowed IICT to create a stable group of young scientists for research in finite state automata theory and development of approximate search systems, as well as speech synthesis and recognition software prototypes.

Description of materials submitted in the application

17 scientific articles (16 of them indexed by Scopus and / or Web of Science, as well as one in the archive) and one patent registered in the USA were submitted for the competition. All materials are co-authored and characterise the joint work of the candidate with young scientists, collaborators from Sofia University "St. Kl. Ohridski", foreign colleagues (mainly from Ludwig Maximilian University in Munich) and industrial partners. The co-authorship in the publications does not diminish the importance of the Assoc. Prof. Mihov's achievements, but rather emphasizes the importance of his status as a valuable and sought-after collaborator, partner, supervisor and project leader.

The citation list contains 213 citations to six of the candidate's most popular articles:

- 93 citations are presented for "Fast string correction with Levenshtein automata",
- 46 citations are presented for "Fast approximate search in large dictionaries",
- 30 citations are presented for "Incremental construction of minimal acyclic finite-state automata",
- 23 citations are presented for "Orthographic errors in Web pages: Toward cleaner Web corpora",
- 12 citations are presented for "Lexical postcorrection of OCR-results: The web as a dynamic secondary dictionary?", and
- 9 citations are presented for "Adaptive Text Correction with Web-Crawled Domain-Dependent Dictionaries".

All these articles are devoted to efficient processing of strings (words) and dictionaries by application of finite automata.

There are two PhD students who completed their theses under the supervision of Assoc. Prof. Mihov - Stefan Gerdjikov (who defended a dissertation on "Effective algorithms for approximate search in regular sets" in 2014, and now he is already an Associate Professor at the Faculty of Informatics and Computer Science at Sofia University "St. Kl. Ohridski") and Peter Mitankin (who defended a dissertation on "Universal automata for effective determination of proximity between strings" in 2010, currently a senior expert in "Ontotext" Ltd). At present Assoc. Prof. Mihov is supervisor of Georgi Shopov, a full-time doctoral student at IICT.

In the Registration Form prepared for NACID, indicator E mentions the participation of Assoc. Prof. Mihov in the AComIn project (2012-2016), management of the Bulgarian team in the IMPACT project (Improving Access to Text), as well as points corresponding to significant funds raised for industrial projects with companies H-TECH EOOD and STATSOFT EOOD. The applicant's activity in these projects shows his desire for systematic accumulation of resources and program code for automatic processing of text, speech and audio:

- in the AComIn project a phonetic corpus of the Bulgarian language was created, using the modern Speech Processing laboratory build by the project (this installation was suggested by Assoc. Prof. Mihov);
- in the IMPACT project the IICT team created a historical dictionary of the Bulgarian language at the end of the 19th century, a language corpus of the Bulgarian language at the end of the 19th century and a specialised OCR system for recognising old Cyrillic (together with the company ABBYY that created FineReader)
- in the projects with H-TECH, industrial software for approximate search of similar audio

recordings has been created, which is used for identification of broadcasts of TV commercials in real time;

- The company STATSOFT EOOD was provided with a Bulgarian phonetic corpus, containing speech recordings annotated at the phonetic level. The corpus contains 21891 recordings made by 140 speakers on 319 phonetically rich sentences in Bulgarian.

Research results and achievements in the application

The topic of candidate's achievements, described in the 17 papers and the patent submitted for the competition, can be grouped into 5 directions:

- **Theoretical** (paper 2 – introduces an alternative construction principle for bimachines called the equalizer accumulation principle. The space complexity of the suggested construction is close to optimal in terms of the number of states, and it can be applied to a broad class of rational functions; paper 3 – it introduces efficient algorithms for composition of conditional probabilistic subsequential transducers with probabilistic subsequential failure transducers and weight pushing (canonisation) of probabilistic subsequential failure transducers. The article shows useful applications of the suggested algorithms in e.g. speech analysis).
- **Text correction using finite automata** (papers 1, 4, 9, 10, 11, 12, 16 and 17 – an efficient method for text rewriting is presented by constructing a specific subsequential transducer in linear time; an algorithm for translating a given contextual rewrite dictionary into an f-transducer /a deterministic transducer that uses failure transitions in order to reduce the size/ and al algorithm for composing f-transducers is proposed; methods have been developed for effective detection of spelling errors of various types on the Internet and their marking using databases of language knowledge, automatically extracted from the web; an effective method for selection of candidates for lexical correction of misspelled words is suggested, by using an universal Levenshtein automaton with refinements of the basic Levenshtein distance; an approach for automatic calculation of the profile of errors that need spelling correction and adaptive selection of candidates for correction is shown; a technique to retrieve dictionaries and language models from the Internet that can be used to correct input text is proposed; a new approach for extracting spelling variations from a list of examples is described, together with a correction algorithm that offers and ranks candidates for correcting a certain wrong word; a novel general and language-independent approach for text correction, developed on the basis of the functional automata in the CULTURA project, is also presented);
- **Approximate search with finite automata** (papers 13, 14 and 15 – effective methods and algorithms are presented for searching similarity of an input pattern in a large database; a new efficient procedure for approximate search is proposed which is organised by subsearches that always start with an exact partial match where a substring of the input pattern is aligned with a substring of a lexicon word, and afterwards this partial match is extended stepwise to larger substrings; in addition the WallBreaker system for approximate search of strings is presented. It won a worldwide competition for efficient similarity search in 2013);
- **Speech processing** (papers 5, 6, 7 and 8 – successive versions of the system for recognition of the continuous Bulgarian speech are presented, starting with the first version made in 2009, with improved performance by partially compiling the word lattice as a deterministic

finite state machine in 2016, and further by designing and constructing of the BulPhonC speech database; the development of the speech corpus BG-PARLAMA using public records of plenary session of Bulgarian parliament in 2019 is described as well).

- **A method for automatic analysis of the interactions among influencers in social networks** – patent 18.

The research topics listed above characterise the vast area of Assoc. Prof. Mihov's scientific interests, the variety of tasks and applications he considers, as well as the numerous collaborators in various projects and undertakings.

A more comprehensive view to the candidate's achievements

Assoc. Prof. Stoyan Mihov is a world-famous specialist in the application of finite state machines in computational linguistics (i.e. as a tool for automatic processing of language and speech) and approximate search (both in strings and in other resources such as audio recordings). Often his applied results are based on original contributions to the theory of finite automata. Assoc. Prof. Mihov has over 60 scientific papers, most of them indexed in Scopus and / or WoS. In the CV presented in the application documents he mentions 426 citations of his works in Scopus, but in Google scholar they are 1148 and most of them are not self-citations.

As early as 2000, in his doctoral dissertation, Assoc. Prof. Mihov proposed an algorithm for direct construction of a minimal acyclic finite automaton given a dictionary of lexicographically ordered words, and published this result in the Computational Linguistics journal as an article cited to this day. His experience in investigation of different types of automata and the constant pursuit of combining theoretical justifications with practical applications allowed him to offer a comprehensive perspective to finite automata from an abstract algebraic point of view, which was built to design computationally efficient structures. These results are presented in the monograph published in Cambridge University Press in 2019.

Based on this monograph, in 2020 Assoc. Prof. Mihov defended a dissertation for the award of the scientific degree "Doctor of Science" with the following theoretical contributions: development of a procedure for deciding the bounded variation property of finite-state transducers, which can be integrated into the sequentialization construction; proposed a new algorithm with polynomial complexity for canonization of a subsequence converter; a new efficient construction with polynomial complexity for canonization of a subsequential transducer is presented; a new construction has been developed for direct composition of bimachines from finite-state transducers; a construction together with proof of correctness was obtained for direct composition of bimachines. As an applied result of the dissertation the programming language C (M) is presented together with a software library of 45 programs of C (M) for the construction of automatic structures and applications in real computational problems.

Assoc. Prof. Mihov leads the team that created the Wallbreaker system which won a world competition for approximate search in 2013 (https://www2.informatik.hu-berlin.de/~leser/searchjoincompetition2013/Results.html).

In recent years under his leadership, with a lot of work and patience to overcome difficulties with funding, a prototype of a dictaphone for Bulgarian speech (speech-to-text system) was created. The development started using a corpus of parliamentary speech as the first training data set and continues with training corpora of medical speech, which are prepared within the eHealth National Research Program. During the last few months the Bulgarian Union of the Blind has been investing in the development of a Bulgarian text-to-speech synthesis system, to be carried out by the IICT group. Thanks to the persistence and efforts of Assoc. Prof. Mihov, IICT can currently plan to create voice interfaces in Bulgarian for different systems.

Personal Impressions

I have known the candidate for many years since he joined IICT as a full-time doctoral student. Apart from the talent, his independence and dedication make a great impression. Through his own efforts, he was able to attract young people and build a group of scientists developing in an original manner the theory of finite state machines and its applications to natural language processing. His active and continuous teaching activity attracts students who are interested in in-depth research. I am also impressed by the patience of Assoc. Prof. Mihov to plan and execute contracts for industrial projects.

Conclusion

I believe that Assoc. Prof. Stoyan Mihov is a rare example of a talented mathematician who is strongly interested in creating real systems and is ready to work as a professional programmer for their construction. The materials presented for the competition prove his attitudes: the presence of deep knowledge, leading role in formulating ambitious research goals, ability to work in a team, as well as perseverance, precision and striving to reach the world level. The quantity and quality of his earlier articles submitted to the competition and their citations show that he has been recognised by the international scientific community as a leading researcher with original ideas for about 15 years. The projects funded by HTech Ltd. prove its ability to create industrial software (which is rare in academic environment). **I strongly support the election of Assoc. Prof. Stoyan Mihov as a professor in the Department "Artificial Intelligence and Language Technologies" of IICT-BAS and I invite the esteemed members of the Scientific Jury to vote unanimously in support of such a decision.**

27 September 2021
Sofia

Member of the Scientific Jury for the procedure:

Prof. Galia Angelova, IICT-BAS