REPORT

on the competition with a single applicant
Assoc. Prof. DSc Stoyan Milkov Mihov
Scientific field: **4. Natural sciences, mathematics and informatics**
Professional field: **4.6. Informatics and computer science**
Announced in "Durzhaven vestnik" no. 45/28.05.2021.

The competition for the position of a full professor is announced in "Durzhaven vestnik" no. 45/28.05.2021. There is one application for the position by:

Assoc. Prof. DSc Stoyan Milkov Mihov.

I am going to evaluate the scientific and teaching activities of the candidate as described in the documents presented for this competition. The scientific research of the applicant corresponds thematically to the description of the position.

1. Personal Data

Stoyan Mihov is born on the 9th of April 1968 in Sofia. In 1993 he graduates from the Faculty of Mathematics and Informatics of Sofia University "St. Kliment Ohridski" with a master thesis entitled "Unification of coregular sets". In 2000 he obtains the Doctor degree from the Institute of Information and Communication Technologies of the Bulgarian Academy of Sciences with a doctor thesis "Minimal acyclic automata: constructions, algorithms and applications", scientific advisor Prof. Dimitar Skordev.

In the period from 1995 until 2006 he works as a programmer and an assistant at the Institute for Information and Communication Technologies. In 2006 Stoyan Mihov takes the position of an aasociate professor again in the IICT at BAS, where he is imployed up to the present moment. In 2020 Stoyan Mihov obtains the scientific degree Doctor of Sciences with a thesis "Finite automata, transducers and bimachines: algoritmic constructions and implementations".

Stoyan Mihov has been teaching at the Faculty of Mathematics at Sofia University for about twenty years. He helds also a parttime position of a researcher at Rila Solutions and Commec.

2. Research Activity

The candidate presents a habilitation thesis, 17 scientific papers, a chapter in a book and a patent. The scientific investigations of the candidate lie in the field of theoretical informatics and correspond fully to the subject of the competition. These include theory of the finite automata, processing of natural languages, speech recognition, text correction and text normalization.

As a habilitation text the candidate presents the book "Finite State Techniques: Automata, Transducers and Bimachines" by himslef and prof. Klaus Schulz from LMU-München published by the prestigeous publishing house Cambridge University Press. The book is a complete coverage of the field, starting from a conceptual introduction and building to advanced topics and applications. The central finite-state technologies are introduced with mathematical rigour, ranging from simple finite-state automata to transducers and bimachines as input-output devices. Special attention is given to the rich possibilities of simplifying, transforming and combining finite-state devices. All algorithms presented are accompanied by full correctness proofs and executable source code in a new programming language, C(M), which focuses on the transparency of steps and the simplicity of the code. Thus, by enabling readers to obtain a deep formal understanding of the subject and to put finite-state methods to real use, this book closes the gap between theory and practice. My opinion is that this book has all the qualities of a habilitation thesis.

Five of the papers are published in prestigeous scientific journals; four of them have an impact factor; five papers have an SJR. Eleven papers are published in the proceedings of international conferences and symposia. Four of the papers are published in Lecture Notes in Computer Science. One of the papers is a preprint published in arXiv. The papers are published in the following journals:

- Natural Language Engineering - 1; (IF 1.065)
- Theoretical Computer Science - 1; (IF 1.231)
- Studies in Computational Intelligence - 1 (IF 1.052)
- Computational Linguistics - 1 (IF 1.800)
- Journal of Automata Languages and Combinatorics - 1 (SJR 0.325)
- Lecture Notes in Computer Science - 4; (SJR 0.293, 0.338, 0.427, 0.249)
- Proc. of the Int. Conf. Recent Advances in Natural Language Processing - 1
- Proc. of the 10th Int. Conf. on Language Resources and Evaluation - 1
- Proc. of the Int. Conf. Document Analysis and Recognition - 2
- Proc. of IJCAI 2007 Workshop on Analytics for Noisy Unstructured Text Data - 1
- ACM International Conference Proceeding Series - 1
- Proceedings - 11th IAPR International Workshop on Document Analysis Systems, DAS 2014 - 1
- preprint in arXiv - 1

Five of the above papers have one co-author; seven of them have two co-authors; three have three co-authors; one has five, and one has six co-authors. All the papers are published after the habilitation and have not been used in previous procedures. The candidate presents a letter of recommendation from Prof. Klaus Shulz (Ludwig Maximilian Universität – München), where the latter declares that the contribution of the candidate in the ten joint papers , as well as in the joint monograph, is equally weighted. In my view, there is no doubt that the contribution of the candidate in the joint paper is significant and at least equal to that of the other authors.

The author classifies the presented papers in the following three subjects: theory of the finite automata; processing of natural languages and speach recognition; fast approximate search, text correction and text normalization.

(1) Theory of the finite automata

The first group contains papers [1–4]. In paper [1] an effecive method is presented, which for given dictionary of rewriting rules constructs a subsequential transducer which takes as an input a text and outputs the intended rewriting result under the so-called "leftmost-longest match" replacement with skips. Some drawbacks of the construction from [1] are improved in [4]. In [3] the authors present a new methodology, which allows the composition of probabilistic subsequential transducers with probabilistic subsequentional failure transducers and weight pushing of probabilistic subsequential failure transducers. The algorithms from this paper are applicable to many tasks for representing probabilistic models with subsequential failure transducers. In paper [2] the authors describe a principle for the construction of bimachines, in which the steps of the bimachine reflect the alternative parallel paths of the transducer. A class of functional transducers working in real time is presented having $n+2$ states, for which the above principle gives 2^n+n+3 states. It is demonstrated that the memory complexity achieved is close to the optimal complexity.

(2) Processing of natural languages and speach recognition

This group contains papers [5–8]. They are devoted to the processing of texts in natural languages and the authomatic language recognition. In paper [5] the candidate describes a system for language recognition in Bulgarian in the presence of a large dictionary. The ultimate goal is to create a system for language recognition in legal documents written in Bulgarian. One of the important contributions here is the creation of a phonetic system and a specific model based on bigrams using a large dictionary. In [7] the authors describe the main principles in the creation of a special corpus of speach, named BulPhonC, which is used in the training of the system from [5]. This reduces the error in legal documents below the one reported in [5].

Paper [6] is devoted to the development of a method for the generation of the first n best hypotheses in language recognition. It is reduced to the construction of deterministic finite automaton, which using the same amount of time presents a larger number of hypotheses. In [8] a methodology is presented, which is used for the compilation of a new corpus of Bulgarian texts suitable for the training and evaluation of modern systems for language recognition. It is derived from the recordings of the plenar sessions of the Bulgarian Parliament.

(3) Approximate search, text correction and text normalization

In this group fall in papers [9–17]. Many problems of huge practical importance can be formulated as problems for approximate search. In [9] a method for correction of orthographic errors in Internet texts. In particular, the methodology creates authomatically dictionaries with wrong entries, as well as statistics of the errors of certain type. Paper [10] presents a method which can be used to create a measure for closeness that is more relevant than the classical Levenshtein distance. This measure is used further for the construction of an abstract Levenshtein automaton. The performed experiments suggest that the developed methods are highly effective. Paper [11] is devoted to a methodology which is similar to that in [10], but which works without a corpus of documents with orthographic errors and their corrected copies. In [12] a language model based on bigrams is described which creates a ranking of the candidates for correction from the previous paper. The model is based on Markov chains. In [13] the problem of normalization of historical texts is considered. This problem is viewed as an approximate search problem. The solution is based on a computation of special kind used in the computational linguistics. In paper [14] the candidate develops an effecive algorithm for an approximate search. A concrete implementation in ANSI C is contained in [15]. In [16] a new methodology for extracting variations in the orthography in historical texts is presented, and, finally, in [17] a complete system for authomatic normalization of historical texts is described.

The patent is a registered methodology which is used to analyze the impact of separate objects in a media covering of certain events. It is based on the generation of a graph with vertices associated with names of journalists, experts,organizations etc, and edges – certain relations“cites”, “covers”, “speaks about” etc. The graph defined in this way is then analyzed by which one gets an objective picture of the impact of the separate objects on the events in question. The product is an instrument which might help organizations to measure, ananlyze and plan the media policiy.

3. Main results

In my opinion the most important scientific results of the candidate are the following:

(1) A subsequential transducer for text rewriting along with various improvements are constructed.

(2) Effecive fast approximate search algorithms are created.

(3) Probabilistic models are presented using subsequential failure transducers.

(4) Systems for Bulgarian languge recognition are created.

(5) A methodology for automatic generation of corpora of Bulgarian texts is created that can be used for the training of systems for natural language recognition.

(6) A system for reconstruction of historical texts is created.

4. Teaching Activity

Though teaching is not a mainstream activity in the Bulgarian Academy of Sciences, I would like to point out that the candidate has an impressive teaching experience which amounts to twenty years of teaching (lectures and tutorails) at the Faculty of Mathematics and Informatics of Sofia University. He has been the supervisor of two dostoral (PhD) students, one in IICT at BAS and one in FMI at SU. He has been also the supervisor of over ten diploma theses.

5. Projects, Conferences etc.

In the period after 2000 the candidate has acted as a coordiantor and member in many national and international scientific projects. It has to be noted that he has taken part in several projects from the Seventh Framework Programme, as well as in the National scientific programme “Electronic health system in Bulgaria”(e-health).

6. Numerical Data

According to the application documents the scientific papers of Assoc Prof. Stoyan Mihov can be classified as follows:

- scientific journals with impact factor:	4
- proceedings with conference talks having SJR :	5
- proceedings with conference talks without SJR :	7
- patents	1
- preprints	1

The total impact factor of the presented papers is 5.148, and the total SJR is 1.632. In his application the candidate presents 213 citations of his scientific papers. They are related to just six of his publications. One of his publications is sited 93 times. For me, this is a compelling evidence that the scientific work of the candidate is significant and well-accepted within the professional community.

I accept the NACID self-assesment presented by the candidate. It becomes clear that he not just fulfills, but exceeds by a wide margin all the national requirements, as well as the specific reuirements of the Institute for Information and Communication Technologies and the Bulgarian Academy of Scences for the position of full professor.

7. **Critical Remarks.**

   I have no notable critical remarks.

8. **Personal Remarks.**

   I have known the cadidate personally for about five years. I have attended many of his talks which are always well-prepared at a high professional level. My impression is that he is a serious researcher with deep knowledge in his field. It is beyond any doubt that he has all the merits of a full professor at the Bulgarian Academy of Sciences.

9. **General Assesment of the Applicant**

   In my opinion Assoc. Prof. Stoyan Mihov has obtained important scientific result that are original and match the level of contemporary informatics. His teaching and project activities are equally impressive. Based on this, I assess **positively** the application of Stoyan Milkov Mihov for the position of a full professor in IICT- BAS in the professional field 4.6. Informatics and computer science.

Conclusion

I am deeply convinced that **Assoc. Prof. Stoyan Milkov Mihov** has all the merits and professional qualifications required for the position of a full professor of the Institute for Infromation and Communication Technologies at the Bulgarian Academy of Sciences. He fulfills all the legal requirements plus the specific ones of IICT at BAS for the scientific field 4. Natural sciences, mathematics and informatics, professional field 4.6 Informatics and computer science. I strongly recomend his application for the position of a full professor at IICT - BAS.

Sofia, 07.09.2021

Member of the Scientific Panel:

(Prof. DSc Ivan Landjev)