ARTIQ

ARTIQ - AI Centres of Excellence

Application for a Host Institution

Institution	National Centre for Research and Development, National Science Centre
Project Joint National Project:	ARTIQ – AI Centres of Excellence
Deadline for the submission of applications	8th of April-11th of May 2021

I. HOST INSTITUTION DATA

Identification data of the Host Institution

Name (full)	**Instytut Podstaw Informatyki Polskiej Akademii Nauk (Institute of Computer Science Polish Academy of Sciences)**
Name (short)	**Instytut Podstaw Informatyki PAN** **(Institute of Computer Science PAS)**
Name of the main organisational unit (where applicable)	
Address of the registered office	
Street	**Jana Kazimierza**
Building No.	**5**
Office No.	
Postal code	**PL01-248**
City/district	**Warszawa / Wola**
Post office	**Warszawa**

Municipality	**m.st. Warszawa**
County	**m.st. Warszawa**
Province	**Mazowieckie / Mazovian**
Correspondence address (if different than the address of the registered office)	
Street	
Building No.	
Office No.	
Postal code	
City/district	
Post office	
Municipality	
County	
Province	
EPUAP [Electronic Platform for Public Administration Services] mailbox	**/IPIPAN/SkrytkaESP**
Legal form	**State organizational units**
The person appointed for contact with NCBR and with the potential Leader/Project Manager	
First name	**Agnieszka**
Last name	**Mykowiecka**
Position	**Deputy Director for Scientific Affairs**
Phone number	**+48 22 380-05-48**
E-mail address	**Agnieszka.Mykowiecka@ipipan.waw.pl**
The person authorised to represent the applicant	
First name	**Wojciech**
Last name	**Penczek**
Function/Position	**Director**

II. CAPACITY OF THE HOST INSTITUTION TO PERFORM THE PROJECT

1. Description of major research achievements in the scope of implementation of R&D projects, as well as the commercialisation of deliverables of such projects regarding artificial intelligence for the last 5 years prior to or in the year of the application along with a list of the most important publications and patents of the applicant (max. 1 A4 page).

1.1 Creation of memory-efficient language models: The project was aimed at size-optimization of neural networks applied in the creation of language models. The deliverables were commercialized by Samsung Electronics Poland in 2018.

1.2 Fake news recognition: The project was aimed at development of the cross-domain fact-checking system using entailment data sets such as The Stanford Natural Language Inference Corpus or a crowd-sourced Cross-lingual NLI (XNLI) corpus. The results were commercialized by Samsung in 2019.

Wawer A., Wojdyga G., Sarzyńska-Wawer J. (2019). *Fact Checking or Psycholinguistics: How to Distinguish Fake and True Claims?* Proceedings of the 2nd Workshop on Fact Extraction and VERification (FEVER): 7–12. http://doi.org/10.18653/v1/D19-6602

1.3 Assessing credibility of online sources: The HOMADOS project is aimed to automatically detect unreliable fabricated content, such as fake news, based on its linguistic features; mitigate influence of misleading claims by designing techniques for automatically suggesting reliable sources relevant to a local.

Przybyła P. (2020). *Capturing the Style of Fake News*. Proceedings of the 34th AAAI Conference on Artificial Intelligence: 490–497. https://doi.org/10.1609/aaai.v34i01.5386

1.4 Verification of multi-agent systems: our designed algorithms have been implemented in two open-source tools, STV & MsATL. They were already used in case studies to analyse existing protocols for secure and verifiable e-voting, such as Selene and Pret-a-Voter.

D. Kurpiewski, W. Jamroga, M. Knapik: *STV: Model Checking for Strategies under Imperfect Information*. AAMAS 2019: 2372-2374

A. Niewiadomski, M. Kacprzak, D. Kurpiewski, M. Knapik, W. Penczek, W Jamroga: *MsATL: A Tool for SAT-Based ATL Satisfiability Checking*. AAMAS 2020: 2111-2113

1.5 MCFS: a feature selection method in regression problems that can be applied to high dimensional data. It is currently considered to be one of the basic methods of ranking features in terms of their impact on the resulting size, based on decision tree families.

M. Dramiński, J. Koronacki, *rmcfs: An R Package for Monte Carlo Feature Selection and Inter-dependency Discovery*, Journal of Statistical Software, vol 85, 2018,

1.6 The massively parallel search engine to work with the Polish Internet resources (https://nekst.pl/) performing their automatic distribution into thematic labelled groups. The resources are used by Uniform Anti-Plagiarism System (JSA) OPI.

1.7 The CytoMeth system: the system processes raw data after next-generation bisulfitation and sequencing data (Next Generation Sequencing, NGS) and returns the level of strand-specific DNA methylation at single nucleotide resolution. Project is available on GitHub and used to compile some of the results accepted for Nature Communications (www.biorxiv.org/content/10.1101/867861v2)

1.8 The training procedure of neural nets, which leads to the classification models with some guarantees against adversarial examples.

Paweł Morawiecki, Przemysław Spurek, Marek Smieja, Jacek *Tabor*: *Fast and Stable Interval Bounds Propagation for Training Verifiably Robust Models*. European Symposium on Artificial Neural Networks (ESANN), 2020

2. A list of 5 research and development projects within national and international competitions in the area of artificial intelligence and implemented within the last 5 years prior to or in the year of the application (title, manager, source of financing, amount of financing) (max. 1 A4 page).

1. "**Social Explainable Artificial Intelligence**" (National Science Centre; CHIST-ERA 2019; 821.790,00 PLN) PI: Stanisław Matwin, Ph.D. Prof.. The research community has already realised that the current centralised approach to AI is not an acceptable and sustainable model in the long run. We posit that the "next wave" of ML-driven AI will be (i) human-centric, (ii) explainable, and (iii) more distributed and decentralised (i.e., not centrally controlled).

2. "**Counteracting mis-informa-tion in digital media by deception detection and facilitating access to reliable sources using machine learning and natu-ral language processing**" (Polish National Agency for Academic Exchange NAWA; Program Polish Returns; 1.355.000,00 PLN) PI: Piotr Przybyła, Ph.D. The proposed research programme is aimed at dealing with the problem of misinformation, which herein refers to textual publications in digital media which may cause their readers to obtain a false belief. It is important to distinguish between two situations: an author of such content is aware of it being misleading, or not.

3. "**Building an integrated retail price statistics system (INSTATCENY)**" (National Centre for Research and Development; GospoStrateg; 1.341.088,00 PLN) PI: Mieczysław Kłopotek, Ph.D. Prof. The research work within the project is aimed at developing the concept of modernisation of the process of measuring changes in retail prices of goods and services (inflation measurement) by the Statistics Poland. New data sources and innovative data obtaining methods will be taken into account. The project will result in the creation of an IT system which will allow the management of acquired data and the integration of heterogeneous and distributed data sets.

4. „**Socio-Technical Verification of Information Security and Trust in Voting Systems (STV)**" (National Centre for Research and Development; PolLux; 1.971.500,00 PLN) PI: Wojciech Jamroga, Ph.D., D.Sc. Voting and elections are extremely important for democratic societies. If democracy is to be effective, it is essential to assess and mitigate the threats of fraud, manipulation, and coercion. In this project, it was proposed to use techniques from game theory, multi-agent systems, and theory of socio-technical systems to redefine and analyse various requirements in public decision-making procedures, such as voting and elections.

5. „**Atlas of brain regulatory regions and regulatory networks - a novel systems biology approach to pathogenesis of selected neurological disorders**" (National Science Centre; program SYMFONIA; 2.072.314,00 PLN) PI in IPIPAN: Jan Komorowski, Ph.D., Prof. The aim of this interdisciplinary project is to create high-resolution DNA regulatory maps in the brain and to discover regulatory networks important in the pathogenesis of glioma brain tumor. In addition, the Atlas will be used to define the epigenetic environment of regulatory regions in the vicinity of point mutations (SNPs) developed in GWAS for psychiatric diseases. This will determine the effect of mutations and other modifications in the regulatory region on the expression of the target gene. The results of the research will contribute to the discovery of the pathogenesis of gliomas and mental diseases, and elucidate the dysfunction of the networks controlling the regulation of gene expression and epigenetic relationships in these processes.

3. Available research equipment, apparatus/infrastructure and intangible assets held in the context of implementation of a project regarding artificial intelligence (max. 1 A4 page).

IPI PAN has a server room that supports servers with a capacity of up to 570U. Currently, the Institute uses about 110 DELL servers. Access to the Internet is provided by two independent fiber-optic connections to Wide Area Networks

The Institute has a scientific library with a large book collection and access to many on-line resources, in particular the Elsevier, IEEE, Oxford Journals, EBSCO, Scopus and JSTOR databases.

IPI PAN maintains large resources of textual data and linguistic models ready to be used in the development of AI-based solutions (The National Corpus of Polish, The Polish Parliamentary Corpus, The Grammatical Dictionary of Polish, linguistic treebanks and manually annotated corpora).

Several tools based on Artificial Intelligence have been developed and implemented at the Institute:

NEKST – Institute has developed and implemented the massively parallel search engine to work with the Polish Internet resources in a novel way (https://nekst.pl/). Our specialty is systematizing online resources, and making their systematics perceivable to the user. Systematization is understood as automatic distribution of online resources into thematic groups, highlighting thematic channels in websites, labelling and categorizing documents and their groups. From the user's point of view, this translates into not only a more precise document identification - systematization enables also contextual search of both individual documents and their groups, such as channels or services, and diversification of the search engine response.

MCFS (Monte Carlo Feature Selection) is a feature selection method that can be applied to high dimensional data (thousands/millions of features). Algorithm is implemented in Java but there is a user friendly R package (rmcfs). First version of MCFS was published in 2004/2005 and in 2008 the final version of MCFS was published in Bioinformatics journal:

M.Dramiński, A.Rada-Iglesias, S.Enroth, C.Wadelius, J. Koronacki, J.Komorowski "Monte Carlo feature selection for supervised classification", BIOINFORMATICS 24(1): 110-117 (2008).

M.Dramiński, J. Koronacki, J.Ćwik, J.Komorowski "Monte Carlo Gene Screening for Supervises Classificattion", Proceedings of the EUROFUSE 2004 Workshop on Data and Knowledge Engineering, B.De Beats, R. De Caluwe, G. de Tre, J. Fodor, J. Kacprzyk, S. Zadrozny (eds):Current Issues in Data and Knowledge Engineering, Akademicka Oficyna Wydawnicza EXIT Warszawa 2004.

4. Facilities or incentives to establish an AI Centre of Excellence in the entity (max. 1 A4 page).

The IPI teams have substantial expertise and know-how in many domains. Many of its researchers serve as editorial board members and members of scientific committees of international AI journals (IEEE Trans. on Knowledge and Data Engineering, Journal of Intelligent Information Systems, Data Mining and Knowledge Discovery) and conferences (AAAI, IJCAI, KDD, WWW, ECML-PKDD, ISMIS). In particular, the STV team has already built its reputation in EU and worldwide in formal verification of interaction between autonomous intelligent agents. Establishing the Centre will open up new possibilities to collaborate with researchers in Poland and abroad, working on formal methods for AI. This is essential for further progress, as the subject is extremely challenging. In particular, overcoming the computational complexity barrier requires completely novel approaches, that can be developed only in collaboration.

The Linguistic Engineering Group at IPI PAN is the largest NLP team in Poland, employing over 20 full-time researchers and several dozen project associates. The team participates in CLARIN and DARIAH research infrastructures, carries out numerous national and international projects (financed by National Science Centre, National Centre for Research and Development, National Programme for the Development of Humanities, from the funds of the Ministry of Science and Higher Education and the EU – Connecting Europe Facility and Horizon 2020) and participates in many commercial projects on AI (aimed at generating descriptions of products in natural language or building language models ready to run on mobile devices). The team publishes the Journal of Language Modelling (https://jlm.ipipan.waw.pl/), organizes a competition for NLP tools for Polish language (PolEval – http://poleval.pl and AI & NLP conference – http://nlpday.pl/). The team runs a public seminar (http://zil.ipipan.waw.pl/seminar), maintains the Polish Linguistics mailing list (in Polish), a CLIP website – Computational Linguistic in Poland (https://clip.ipipan.waw.pl/) and a Facebook page on Computational Linguistics (https://www.facebook.com/lingwistyka.komputerowa/, in Polish).

The focus of the Computational Biology Groupis on learning functions of non-coding DNA regions and thus detect regulatory disorders that may result in abnormalities in biological pathways. In order to better understand development of various diseases, we seek to rely on thorough studies of multiple informative gene expression regulatory layers, including the genomic, epigenomic, proteomic and other -omics variability in the course of evolution. This group incorporates multidisciplinary knowledge including statistics, mathematical modelling, machine learning, programming, Big Data analysis, parallel computing, biochemistry, ecology, evolution and molecular biology to unveil, with the help of our proprietary algorithms, the mechanistic structure of a wide spectrum of biological issues.

5. Other information concerning internationalisation of the entity, foreign scientists employed in this institution, availability of seminars in English, etc. (max. 1 A4 page).

Team seminars and institution-wide seminars are held at the Institute. Most are delivered in Polish, but we also hosted many foreign scientists who conduct seminars in English. Computational Biology Lab organizes seminars in English on research related to genomics, transcriptomics and epigenetics (http://zbo.ipipan.waw.pl/seminars.html and the Linguistic Engineering Group – the seminar on NLP: http://zil.ipipan.waw.pl/seminar). Some of the seminars are recorded and available on YouTube (https://www.youtube.com/ipipan). Examples of speakers from outside Poland in the last few years include:

- Robert Moskovitch (University of the Nagev)
- Yan Kim (University of Luxembourg)
- Laure Petrucci (Université Paris 13)
- Benjamin Bordais (ENS Rennes)
- Jörg Keller (FernUniversität Hagen
- Andrzej Mizera (University of Luxembourg)
- Alexander Rosen (Charles University in Prague)
- Igor Boguslavsky (Institute for Information Transmission Problems, Russian Academy of Sciences / Universidad Politécnica de Madrid)
- Agata Savary (Université François Rabelais Tours)
- Ekaterina Lapshinova-Koltunski (Saarland University)
- Daniel Zeman (Institute of Formal and Applied Linguistics, Charles University in Prague)
- Jakub Waszczuk (Heinrich-Heine-Universität Düsseldorf)

6. Other significant information confirming the experience and resources of the institution (max. 1 A4 page).

The Institute is a co-organizer of English-language doctoral studies: Doctoral School - Information and Biomedical Technologies Polish Academy of Sciences. Currently, IPI PAN employees run two courses Bioinformatics and Natural Language Processing, and a seminar Selected Topics in Machine Learning. One of our students comes from Ukraine.

IPI PAN participated in European COST projects. At present, we are taking part in the Nexus Linguarum (European network for Web-centred linguistic data science, https://www.cost.eu/ actions/CA18209/) campaign, while in the previous years, IPI PAN took part in the TextLink (Structuring Discourse in Multilingual) campaign and PARSEME (PARSing and Multi-word Expressions. Towards linguistic precision and computational efficiency in natural language processing, https://www.cost.eu/actions/IC1207).

We work with the CLARIN ERIC infrastructure and several scientific institutions in Europe on the parliamentary data corpora (ParlaMint project; https://www.clarin.eu/content/parlamint). As part of the Mobility Plus program, we cooperate with the University of Oxford (contact: Mary Dalrymple). In cooperation with an international group of scientists (Vincent Ng, Yulia Grishina, Sameer Pradhan, Massimo Poesio), a common method for describing anaphoric phenomena (Universal Anaphora) is being developed. We participate in numerous international projects under various financing schemes: CEF (ELRC - European Language Resource Coordination, MARCELL - Multilingual Resources for CEF.AT in the legal domain and CURLICAT - Curated Multilingual Language Resources for CEF AT), Preparatory Action / Coordination and Support Action (ELE - European Language Equality), H2020 (ELG - European Language Grid, PARTHENOS - Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies). IPI PAN employees were co-organizers of international conferences: LFG (Lexical-Functional Grammar Conference), International Symposium on Parallel and Distributed Computing 2020, and CORBON / CRAC (Coreference Resolution Beyond OntoNotes / Computational Models of Reference, Anaphora, and Coreference) workshops, NLP conferences around the world (2016–2021) and "PTBI Autumn Workshop'' (https://www.ptbi.org.pl/website/conferences/aut20/) in 2020.

The Theory Of Distributed And Computing Systems Team collaborates with a number of research groups in Europe, in particular at the University of Luxembourg (Luxembourg), University of Paris-Est Creteil, Telecom University Paris, and Université Sorbonne Paris Nord (France), University of Naples (Italy), and KTH Stockholm (Sweden). As a result of the Symphony project, the Computational Biology group established foreign cooperation with Professor Jacek Majewski from McGill University.

From 2020, the Cryptography Team closely collaborates with the Australian CSIRO, Data61 research group led by Seyit Camtepe. Its collaboration is at the intersection of information security and artificial intelligence.