ARTIQ - Centra Doskonałości AI

Zgłoszenie Instytucji Hostującej

Instytucja	Narodowe Centrum Badań i Rozwoju, Narodowe Centrum Nauki
Przedsięwzięcie	Wspólne Przedsięwzięcie krajowe: ARTIQ - Centra Doskonałości AI
Zakres czasowy zgłoszeń	8 kwietnia -11 maja 2021 r.

I. INFORMACJE O INSTYTUCJI HOSTUJĄCEJ

Dane identyfikacyjne Instytucji Hostującej

Nazwa (pełna)	**Instytut Podstaw Informatyki Polskiej Akademii Nauk**
Nazwa (skrócona)	**Instytut Podstaw Informatyki PAN**
Nazwa podstawowej jednostki organizacyjnej (jeśli dotyczy)	-
Adres siedziby	
Ulica	**Jana Kazimierza**
Nr budynku	**5**
Nr lokalu	

Kod pocztowy	**01-248**
Miejscowość/dzielnica	**Warszawa / Wola**
Poczta	**Warszawa**
Gmina	**m.st. Warszawa**
Powiat	**m.st. Warszawa**
Województwo	**mazowieckie**
Adres do korespondencji (jeśli inny niż adres siedziby)	
Ulica	
Nr budynku	
Nr lokalu	
Kod pocztowy	
Miejscowość/dzielnica	
Poczta	
Gmina	
Powiat	
Województwo	
Skrzynka EPUAP	**/IPIPAN/SkrytkaESP**
Forma prawna	**Państwowe jednostki organizacyjne**
Osoba wyznaczona do kontaktu z NCBR oraz z potencjalnym Liderem/kierownikiem projektu	
Imię	**Agnieszka**
Nazwisko	**Mykowiecka**
Stanowisko	**Z-ca Dyrektora ds. Naukowych**
Nr telefonu	**+48 22 380-05-48**
Adres e-mail	**Agnieszka.Mykowiecka@ipipan.waw.pl**

Osoba upoważniona do reprezentacji zgłaszającego	
Imię	**Wojciech**
Nazwisko	**Penczek**
Funkcja/Stanowisko	**Dyrektor**

II. ZDOLNOŚĆ INSTYTUCJI HOSTUJĄCEJ DO WYKONANIA PROJEKTU

1. Opis najważniejszych osiągnięć naukowych w zakresie realizacji projektów B+R jak również komercjalizacji ich wyników w tematyce sztucznej inteligencji z ostatnich 5 lat przed rokiem lub w roku zgłoszenia wraz z wykazem najważniejszych publikacji, patentów zgłaszającego (do 1 strony A4).

1.1 Opracowanie efektywnych pamięciowo modeli językowych: celem projektu była optymalizacja rozmiarów sieci neuronowych wykorzystywanych przy tworzeniu modeli językowych. Wyniki zostały skomercjalizowane przez Samsung Electronics Polska.

1.2 Rozpoznawanie fałszywych wiadomości: międzydomenowy system sprawdzania faktów przy użyciu zbiorów danych o wynikaniu, takich jak The Stanford Natural Language Inference Corpus lub XNLI (komercjalizacja - Samsung, 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 Ocena wiarygodności źródeł internetowych (Projekt HOMADOS): opracowanie metody wykrywania niewiarygodnych treści, takich jak fałszywe wiadomości, na podstawie cech językowych oraz złagodzenie ich wpływu poprzez automatyczne sugerowanie źródeł wiarygodnych odpowiednich dla lokalnego kontekstu.

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

1.4 Techniki weryfikacji systemów wieloagentowych (projekt VoteVerif): opracowanie i zaimplementowanie w dwóch narzędziach *open source* (STV i MsATL) algorytmów i wykorzystanie ich do analizy istniejących protokołów bezpiecznego i weryfikowalnego głosowania elektronicznego, takich jak Selene i 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 To*ol for SAT-Based ATL Satisfiability Checking*. AAMAS 2020: 2111-2113

1.5 MCFS - metoda wyboru cech dla regresji, którą można zastosować do danych o dużych wymiarach. Jest uznawana za jedną z podstawowych metod szeregowania cech pod kątem wpływu na wielkość wynikową, opartą na rodzinach drzew decyzyjnych.

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

1.6 Masowo równoległa wyszukiwarka do pracy z polskimi zasobami Internetu (https://nekst.pl/): program dokonuje podziału zasobów na poetykietowane grupy tematyczne. Z zasobów korzystał m.in. Jednolity System Antyplagiatowy (JSA) OPI.

1.7 CytoMeth (https://github.com/mdraminski/CytoMeth) - system do przetwarzania surowych danych Next Generation Sequencing, który zwraca poziom niciowo-specyficznej metylacji DNA w rozdzielczości pojedynczego nukleotydu. System użyto w pracy do Nature Communications https://www.biorxiv.org/content/10.1101/867861v2.

1.8 Metoda detekcji zwodniczych przykładów (ang. adversarial examples) dla głębokich sieci neuronowych oraz procedury treningu dające pewne gwarancje odporności.

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. Lista do 5 projektów badawczo-rozwojowych w ramach konkursów krajowych lub międzynarodowych z obszaru sztucznej inteligencji i realizowanych w ciągu ostatnich 5 lat przed rokiem lub w roku zgłoszenia przez zgłaszającego (tytuł, kierownik, źródło finansowania, wysokość dofinansowania) (do 1 strony A4).

1. „**Sztuczna Inteligencja na Sposób Wytłumaczalny i Społeczny**” (NCN; CHIST-ERA 2019; 821.790,00 zł) kierownik: prof. dr hab. Stanisław Matwin. Celem projektu jest zaproponowanie modeli maszynowego uczenia się, które będą (i) skoncentrowane na człowieku, (ii) możliwe do wyjaśnienia oraz (iii) bardziej rozproszone i zdecentralizowane (tj. nie kontrolowane centralnie).

2. “**Counteracting misinformation in digital media by deception detection and facilitating access to reliable sources using machine learning and natural language processing**” (NAWA; Polskie Powroty 1.355.000,00 zł) kierownik: dr Piotr Przybyła. Celem projektu jest próba rozwiązania problemu dezinformacji w publikacjach tekstowych w mediach cyfrowych. Pojawiające się tam nieprawdziwe informacje mogą spowodować, że ich czytelnicy uzyskają fałszywe przekonanie. Ważne jest, aby rozróżnić dwie sytuacje: autor takich treści może mieć świadomość, że wprowadzają one w błąd lub nie.

3. „**Budowa zintegrowanego systemu statystyki cen detalicznych (INSTATCENY)**” (NCBiR; GospoStrateg; 1.341.088,00 zł) kierownik IPIPAN: prof. Mieczysław Kłopotek. Prace badawcze w ramach projektu mają na celu opracowanie koncepcji modernizacji procesu pomiaru zmian cen detalicznych towarów i usług (pomiar inflacji) przez GUS. Uwzględnione zostaną nowe źródła danych i innowacyjne metody ich pozyskiwania. Efektem projektu będzie stworzenie systemu informatycznego, który pozwoli na zarządzanie danymi oraz integrację heterogenicznych i rozproszonych zbiorów.

4. „**Socio-Technical Verification of Information Security and Trust in Voting Systems (STV)**” (NCBiR; program PolLux; 1.971.500,00 zł) kierownik: dr hab. Wojciech Jamroga Głosowanie i wybory są niezwykle ważne dla społeczeństw demokratycznych. Aby demokracja była skuteczna, niezbędna jest ocena i łagodzenie zagrożeń związanych z oszustwami, manipulacjami i przymusem. W projekcie zostało zaproponowane wykorzystanie technik z teorii gier, systemów wieloagentowych i teorii systemów społeczno-technicznych, aby na nowo zdefiniować i przeanalizować różne wymagania w publicznych procedurach podejmowania decyzji, takich jak głosowanie i wybory.

5. „**Atlas obszarów regulatorowych specyficznych dla mózgu ludzkiego - nowe narzędzie odkrywania ścieżek powodujących wybrane choroby mózgu**” (NCN; program SYMFONIA; 2.072.314,00 zł) kierownik IPIPAN: dr Jan Komorowski, prof.. Celem tego projektu jest stworzenie wysokorozdzielczych map regulatorowego DNA w mózgu i odkrycie sieci regulacyjnych istotnych w patogenezie glejaka mózgu. Dodatkowo, atlas zostanie wykorzystany do określenia epigenetycznego środowiska obszarów regulatorowych w pobliżu mutacji punktowych (SNPs) opracowanych w GWAS dla chorób psychiatrycznych. Pozwoli to określić efekt mutacji oraz innych modyfikacji w obszarze regulatorowym na ekspresję genu podlegającemu regulacji przez określony region.

3. Dostępny sprzęt badawczy, aparatura/infrastruktura oraz własności niematerialne i Prawne WNiP pozostające w posiadaniu w kontekście realizacji projektu w tematyce sztucznej inteligencji (do 1 strony A4).

Od października 2011 IPI PAN dysponuje serwerownią, która umożliwia wykorzystywanie serwerów o łącznej wielkości do 570U. Obecnie Instytut wykorzystuje około 110 serwerów DELL i planuje rozwój infrastruktury sprzętowej. Dostęp Instytutu do sieci zewnętrznej jest realizowany przez dwa niezależne łącza światłowodowe.

W Instytucie działa biblioteka naukowa posiadająca duże zbiory własne oraz dostęp do publikacji zagranicznych, w szczególności baz Elsevier, IEEE, Oxford Journals, EBSCO, Scopus i JSTOR. Instytut posiada też bogate zasoby danych tekstowych i modeli językowych gotowe do wykorzystania w pracach nad rozwojem sztucznej inteligencji (m.in. Narodowy Korpus Języka Polskiego, Korpus Dyskursu Parlamentarnego, Słownik Gramatyczny Języka Polskiego, bank rozbiorów gramatycznych, korpusy językowe opisane lingwistycznie).

W IPI PAN opracowano i wdrożono kilka narzędzi opierających się na Sztucznej Inteligencji, m.in.

NEKST - masowo-równoległa wyszukiwarka internetowa NEKST (http://nekst.pl), która gromadzi zasoby całego polskiego Internetu. Naszą specjalnością jest systematyzowanie zasobów internetowych oraz ich udostępnianie użytkownikowi. Systematyzacja oznacza automatyczny podział zasobów internetowych na grupy tematyczne, wyróżnianie kanałów tematycznych w serwisach internetowych, oraz etykietowanie i kategoryzowanie dokumentów i ich grup. Z punktu widzenia użytkownika, przekłada się to nie tylko na bardziej precyzyjne identyfikowanie dokumentów wartościowych dla użytkownika. Systematyzacja daje również możliwość wyszukiwania kontekstowego zarówno pojedynczych dokumentów, jak i ich grup, np. kanałów tematycznych czy serwisów, oraz dywersyfikację odpowiedzi wyszukiwarki. Ze zgromadzonych zasobów korzysta m.in. Jednolity System Antyplagiatowy (JSA) OPI.

MCFS (Monte Carlo Feature Selection) - to metoda wyboru cech, którą można zastosować do danych o dużych wymiarach (tysiące/miliony cech). Algorytm jest zaimplementowany w Javie, ale istnieje przyjazny dla użytkownika pakiet R (rmcfs). Pierwsza wersja MCFS została opublikowana w latach 2004/2005, a w 2008 roku ostateczna wersja MCFS została opublikowana w czasopiśmie Bioinformatics:

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. Ułatwienia lub inne zachęty do utworzenia Centrum Doskonałości AI w tym podmiocie (do 1 strony A4).

Zespoły IPI mają znaczną wiedzę i know-how w wielu dziedzinach. Wielu pracowników jest członkami redakcji i komitetów naukowych międzynarodowych czasopism (IEEE Trans. on Knowledge and Data Engineering, Journal of Intelligent Information Systems, Data Mining and Knowledge Discovery) i konferencji poświęconych sztucznej inteligencji (AAAI, IJCAI, KDD, WWW, ECML-PKDD, ISMIS). W szczególności zespół Teorii Systemów Rozproszonych i Obliczeniowych zbudował już swoją reputację w UE i na świecie w dziedzinie formalnej weryfikacji interakcji między autonomicznymi inteligentnymi agentami. Powołanie Centrum otworzy nowe możliwości jeszcze szerszej współpracy z naukowcami w Polsce i za granicą, pracującymi nad formalnymi metodami sztucznej inteligencji i wspólne podejmowanie takich trudnych tematów jak pokonanie bariery złożoności obliczeniowej.

Zespół Inżynierii Lingwistycznej IPI PAN jest znaczącym w Polsce zespołem NLP, zatrudniającym ponad 20 etatowych pracowników naukowych i kilkudziesięciu współ-pracowników projektowych). Zespół jest uczestnikiem infrastruktur badawczych CLARIN i DARIAH, realizuje liczne projekty krajowe i międzynarodowe (finansowane przez NCN, NCBiR, NPRH, ze środków MNiSW i unijnych – CEF, H2020) oraz uczestniczy w wielu projektach komercyjnych dot. sztucznej inteligencji (mających na celu m.in. generowanie opisów produktów w języku naturalnym czy budowę modeli językowych gotowych do uruchamiania na urządzeniach mobilnych). ZIL wydaje czasopismo Journal of Language Modelling (https://jlm.ipipan.waw.pl/), organizuje konkurs dla narzędzi NLP dla języka polskiego (PolEval – http://poleval.pl/ oraz konferencję AI & NLP conference – http://nlpday.pl/). Zespół prowadzi ogólnodostępne seminarium (http://zil.ipipan.waw.pl/seminarium), utrzymuje Polską Listę Językoznawczą, stronę CLIP – Computational Linguistic in Poland (https://clip.ipipan.waw.pl/) oraz stronę Lingwistyka Komputerowa na Facebooku (https://www.facebook.com/lingwistyka.komputerowa/).

Zespół Biologii Obliczeniowej koncentruje się na analizach funkcji niekodujących regionów DNA, a tym samym na wykrywaniu zaburzeń regulacyjnych, które mogą skutkować nieprawidłowościami w szlakach biologicznych. Aby lepiej zrozumieć rozwój różnych chorób, akumulujemy różne typy informacji dotyczące regulacji ekspresji genów w tym zmienności genomicznej, epigenomicznej, proteomicznej i innych -omicznych elementów. Grupa jest silnie multidyscyplinarna i obejmuje ekspercką wiedzę z zakresu statystyki, modelowania matematycznego, maszynowego uczenia się, programowania, analiz Big Data, biochemii, ekologii, ewolucji i biologii molekularnej. Pozwala to na implementację nowatorskich algorytmów w celu weryfikacji hipotez obejmujących szerokie spektrum zagadnień biologicznych.

5. Inne informacje o umiędzynarodowieniu podmiotu, zagranicznych naukowcach zatrudnionych w tej instytucji, dostępności seminariów w języku angielskim, itp. (do 1 strony A4).

W Instytucie odbywają się seminaria zespołowe oraz ogólnoinstytutowe. Większość wykładów wygłaszana jest w języku polskim, ale gościliśmy też wielu naukowców z zagranicy, którzy seminaria prowadzą w języku angielskim. Zespół Biologii Obliczeniowej organizuje seminaria w języku angielskim na temat badań związanych z genomiką, transkryptomiką, epigenetyką (http://zbo.ipipan.waw.pl/seminars.html), a Zespół Inżynierii Lingwistycznej – na tematy związane z przetwarzaniem języka naturalnego (http://zil.ipipan.waw.pl/seminarium). Część seminariów jest nagrywana i dostępna na YouTube (https://www.youtube.com/ipipan). Przykładowi prelegenci spoza Polski w ostatnich paru latach to:

- Robert Moskovitch (University of the Nagev)
- Yan Kim (University of Luxembourg)
- Laure Petrucci (Université Paris 13)
- Benjamin Bordais (ENS Rennes)
- Jörg Keller (Fern Universität Hagen)
- Andrzej Mizera (Uniwersytet w Luksemburgu)
- Alexander Rosen (Uniwersytet Karola w Pradze)
- Igor Boguslavsky (Rosyjska Akademia Nauk / Politechnika w Madrycie)
- Agata Savary (Université François Rabelais Tours)
- Ekaterina Lapshinova-Koltunski (Uniwersytet Kraju Saary)
- Daniel Zeman (Uniwersytet Karola w Pradze)
- Jakub Waszczuk (Uniwersytet Heinricha Heinego w Düsseldorfie)

6. Inne istotne informacje potwierdzające doświadczenie oraz zasoby instytucji (do 1 strony A4).

Instytut jest współorganizatorem angielskojęzycznych studiów doktorskich: Szkoła doktorska Technologii Informacyjnych i Biomedycznych Instytutów PAN (TIB PAN). Obecnie pracownicy IPI PAN prowadzą tam kursy Bioinformatics i Natural Language Processing oraz seminarium Selected Topics in Machine Learning. Jeden ze studentów, którymi się bezpośrednio opiekujemy, pochodzi z Ukrainy.

IPI PAN uczestniczył w projektach europejskich typu COST. Obecnie bierze udział w akcji Nexus Linguarum (*European network for Web-centred linguistic data science*, https://www.cost.eu/actions/CA18209/), natomiast we wcześniejszych latach IPI PAN brał udział w akcji TextLink (Structuring Discourse in Multilingual Europe, https://www.cost.eu/actions/IS1312) oraz PARSEME (*PARSing and Multi-word Expressions. Towards linguistic precision and computational efficiency in natural language processing*, https://www.cost.eu/actions/IC1207)

IPI PAN współpracuje z infrastrukturą CLARIN ERIC oraz kilkunastoma instytucjami naukowymi w Europie w ramach prac nad zestawem korpusów danych parlamentarnych (projekt ParlaMint; https://www.clarin.eu/content/parlamint).

W ramach programu Mobilność Plus IPI PAN współpracuje z Uniwersytetem w Oxfordzie (kontakt: Mary Dalrymple). W ramach współpracy z międzynarodową grupą naukowców (Vincent Ng, Yulia Grishina, Sameer Pradhan, Massimo Poesio) opracowywana jest wspólna dla wielu języków metoda opisu zjawisk anaforycznych (Universal Anaphora). Bierzemy udział w licznych projektach międzynarodowych w ramach różnych schematów finansowania: CEF (ELRC – European Language Resource Coordination, MARCELL – Multilingual Resources for CEF.AT in the legal domain i 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). Pracownicy IPI PAN byli współorganizatorami międzynarodowych konferencji: LFG (Lexical-Functional Grammar Conference), International Symposium on Parallel and Distributed Computing 2020, i warsztatów CORBON/CRAC (Coreference Resolution Beyond OntoNotes/Computational Models of Reference, Anaphora, and Coreference), przy najważniejszych konferencjach NLP na świecie (2016–21) (https://www.ptbi.org.pl/website/conferences/aut20/).

Zespół Teorii Systemów Rozproszonych i Obliczeniowych współpracuje z wieloma grupami badawczymi w Europie, w szczególności z Uniwersytetem Luksemburskim (Luksemburg), Uniwersytetem Paris-Est Créteil, Uniwersytetem Telecom Paris oraz Université Sorbonne Paris Nord (Francja), Uniwersytetem Neapolitańskim (Włochy) i KTH Stockholm (Szwecja). W efekcie realizacji projektu Symfonia zespół biologii obliczeniowej nawiązał współpracę z profesorem Jackiem Majewskim z McGill University.

Od 2020 r. Zespół Kryptografii współpracuje z australijskim zespołem z ośrodka CISIRO, Data61 kierowanym przez Seyita Camtepe. Obszar współpracy to styk zagadnień związanych z bezpieczeństwem informacji i sztuczną inteligencją.