University of Debrecen
Faculty of Informatics
Debrecen, Hungary

The 2024 IEEE 3rd Conference on Information Technology and Data Science

General Information

COMMITTEE MEMBERS

General Chair

András Hajdu	University of Debrecen, Hungary

Honorary Chair

Levente Kovács	IEEE Hungary Section Chair

Program Committee Chair

István Fazekas	University of Debrecen, Hungary

Session Chairs

Sándor Baran, István Fazekas, János Sztrik	Stochastic models in Data Science
Zoltán Gál, Alex Doboli	Internet of Things, Operational Technology and Networking
András Hajdu	Artificial Intelligence and its Applications
Márton Ispány	Temporal Data Science with Applications
István Oniga, Do Van Tien	Autonomous vehicles and embedded systems
Anett Rácz	Optimization and Decision Support
László Szathmáry, Attila Adamkó, Tamás Bérczes	Software Technology
Ágnes Vathy-Fogarassy	Predictive Analytics

Program Committee

Attila Adamkó	University of Debrecen, Hungary
Jaume Baixeries	Universitat Politècnica de Catalunya (Barcelona), Spain
Sándor Baran	University of Debrecen, Hungary
Tamás Bérczes	University of Debrecen, Hungary
Pascal Bondon	Université Paris-Saclay, CentraleSupélec, France
Henning Bordihn	University of Potsdam, Germany
Sándor Bozóki	Corvinus University of Budapest, Hungary
Sérgio Cardoso	Politécnico De Santarém, Portugal
Lucie Ciencialová	Silesian University in Opava, Czech Republic
Erzsébet Csuhaj-Varjú	Eötvös Loránd University, Hungary

Alex Doboli	Stony Brook University, USA
Csaba Domokos	Robert Bosch GmbH, Germany
Dmitry Efrosinin	Johannes Kepler University, Linz, Austria
István Fazekas	University of Debrecen, Hungary
Szilárd Zsolt Fazekas	Graduate School of Engineering Science, Akita University, Akita, Japan
Zoltán Gál	University of Debrecen, Hungary
Attila Gilányi	University of Debrecen, Hungary
Teresa Gonçalves	University of Évora, Portugal
Stathis Hadjidemetriou	University of Limassol, Cyprus
András Hajdu	University of Debrecen, Hungary
Piroska Haller	University of Medicine, Pharmacy, Sciences and Technology of Targu Mures, Romania
Miklós Hoffmann	Eszterházy Károly University, Hungary
Peter Horak	University of Washington, USA
Márton Ispány	University of Debrecen, Hungary
Jeganathan Kathirvel	University of Madras, India
Gergely Kovásznai	Eszterházy Károly University, Hungary
Roland Kunkli	University of Debrecen, Hungary
Sebastian Lerch	Karlsruhe Institute of Technology, Germany
Sven Lončarić	University of Zagreb, Croatia
Filipe Madeira	Politécnico De Santarém, Portugal
Agassi Melikov	Baku Engineering University, Azerbaijan
Khawaja Moyeezullah Ghori	National University of Modern Languages (NUML), Pakistan
István Oniga	University of Debrecen, Hungary
Éva Orbán-Mihálykó	University of Pannonia, Hungary
Andrea Pintér-Huszti	University of Debrecen, Hungary
Anett Rácz	University of Debrecen, Hungary
Valdério Reisen	Universidade Federal da Bahia, Salvador, Brazil
Wolfgang Schreiner	Johannes Kepler University, Linz, Austria
Shinnosuke Seki	University of Electro-Communications, Tokyo, Japan
Désiré Sidibe	Université Evry Val Essonne, France
Nataša Sladoje	Uppsala University, Sweden
László Szathmáry	University of Debrecen, Hungary
Gábor Szederkényi	Pázmány Péter Catholic University, Hungary
János Sztrik	University of Debrecen, Hungary
György Terdik	University of Debrecen, Hungary
Do Van Tien	Budapest University of Technology and Economics, Hungary
Arban Uka	Epoka University, Albania

Pál Varga	Budapest University of Technology and Economics, Hungary
György Vaszil	University of Debrecen, Hungary
Ágnes Vathy-Fogarassy	University of Pannonia, Hungary

Organizing Committee Chair

Imre Varga	University of Debrecen, Hungary

Organizing Committee

Csaba Benedek	John von Neumann Computer Society, Hungary
Piroska Biró	University of Debrecen, Hungary
István Fazekas	University of Debrecen, Hungary
Bence Hegedűs	University of Debrecen, Hungary
Roland Kunkli	University of Debrecen, Hungary
Sándor Pecsora	University of Debrecen, Hungary
Anett Rácz	University of Debrecen, Hungary
Ákos Szabó	University of Debrecen, Hungary
Anikó Szakál	IEEE Hungary Section
Erzsébet Tóth	University of Debrecen, Hungary

Treasurer

Arnold Pintér	University of Debrecen, Hungary

Editors

Piroska Biró	University of Debrecen, Hungary
Erzsébet Tóth	University of Debrecen, Hungary

General contact e-mail: citds2024@inf.unideb.hu

Contents

PLENARY AND KEYNOTE SPEAKERS

Novel Challenges and Opportunities for Improving the Collaborative Efficiency of Teams Using Speech Data-based Machine Learning Models for Understanding Human Behavior

Alex Doboli (Department of Electrical and Computing Engineering, Stony Brook University, Stony Brok, NY)

Internet-of-Things Edge (IoT-E) is a novel architecture expected to offer important and unique opportunities in applications, like automotive, smart traffic, smart homes, healthcare, power supply management, environmental protection, and many more. This architecture supports distributed and decentralized computing, in which data sensed through a myriad of sensors is used to make decisions and produce actuation responses. The tackled performance requirements include real-time constraints, robustness, low power/low energy consumption, and more recently decarbonization for sustainability. Moreover, a special class of applications, called "Humans-in-the-Loop", must consider the specifics of human behavior over time as part of the goals to be optimized, where humans are either acting individually, as teams, or even as communities. To this end, Machine Learning is increasingly combined with traditional, algorithmic methods to improve the smartness and flexibility of the implementations.

The "Humans-in-the-Loop" feature of applications require new approaches to manage in a comprehensive, non-invasive yet private way the integration of human behavior with the resources of the (IoT-E) architecture. This presentation focuses on applications centered around teams that collaboratively solve problems, and for which behavior can be traced using the verbal discussions between team members. In addition to tracking speakers over time using diarization algorithms, the talk also discusses novel opportunities of utilizing theoretical models on team behavior from social sciences to support the computational modeling of this behavior. For example, models must incorporate information about subjective aspects, like emotions, intentions, and beliefs. However, the understanding of such aspects often requires the interpretation of the semantics of verbal discussions, which is still a challenging task. The requirements specific to human behavior create opportunities to devise new ML algorithms. The application of the methods presented in this talk has been to understand and improve team problem solving during algorithm design but other new applications in healthcare and education are envisioned too.

AI Ethics: What Is the Real Question?

Viktor Dörfler (University of Strathclyde Business School, Glasgow, UK)

Ethics is one of our oldest disciplines, one of the main branches of philosophy, it is well over two millennia old. AI is also an intellectual quest, one of the newest ones, although not nearly as new as some think? It started in the mid-1950s. What happens when we bring the two together? I give a brief historical overview of the development of thinking in ethics, putting on the map the traditional schools of ethics, some less known further attempts, as well as moral psychology, which I believe to be a paramount importance. Then I ask how can any of this be done with AI, and review the various approaches that may be relevant, such as computer ethics, digital ethics, and robot ethics. The problem becomes very wide at this point, as we have to realize that it is not only about AI being ethical but also developing AI in an ethical way, using AI in an ethical way, behaving ethically towards AI, and so forth. I also consider the historical context, which is unique in the course of ethics, to a large extent thanks to technological developments. Until now, over the whole course of history, if mankind faced a big ethical challenge, we could take decades to figure out how to deal with this. Just think about the atomic bomb. Today, if we mess up something around AI, we may have only hours at our disposal. Then I reflect on what AI can and cannot do compared to the human mind, arguing that the human mind is better at some things, such as intuiting, storytelling, engaging in social relations, and AI is better at some other things, such as processing enormous amounts of data and identifying potential patterns. However, judging the significance of a potential pattern remains a uniquely human capability, at least for now. On this basis I suggest that we are asking the wrong questions about AI Ethics, and explore some new directions that I find more promising. What emerges from this discussion is an understanding that we are in a serious trouble in ethics, be it in relation to AI or just between us humans. It seems that despite of the numerous models on ethics, humankind did not manage to agree about much in terms of moral judgements. Over the past two and a half millennia, we learned to ask increasingly difficult questions, but answers are scarce and perhaps there is none that everyone accepts. This is a problem, as with AI in our everyday life, we need to answer fundamental ethics questions right now. Delay is not an option.

A Case Study from Algebra through Logic to Computer Science

Sándor Jenei (Hungary Institute of Mathematics and Informatics University of Pécs, Eszterházy Károly Catholic University, Eger, Hungary)

In this presentation, we will discuss the relationship between the interpolation (logical) property and (algorithmic) decidability, the connection between interpolation and (algebraic) amalgamation, and demonstrate the presence or absence of the amalgamation property in some algebraic structures that are significant in the field of substructural logics.

The First Instruct-Following Large Language Models for Hungarian

Zijian Győző Yang (HUN-REN Hungarian Research Centre for Linguistics, Budapest, Hungary)

In recent months, large language models have gained significant attention, with companies striving to develop models capable of solving various natural language processing tasks through extensive data training. The release of ChatGPT by OpenAI demonstrated unprecedented capabilities via a multi-step fine-tuning process. For Hungarian, pre-trained large language models include PULI GPT-3SX, PULI GPTrio and in the recent months SambaLingo. In our research, we pre-trained a new large language model based on Llama-2 and inspired by ChatGPT, focuses on fine-tuning with instruction-based prompts. We created a Hungarian prompt dataset and fine-tuned the PULI large language models into instruction-following models.

In our research, we discovered that transfer learning allows the model to gain insights from other languages. We found that further pre-training of the language model could leverage valuable knowledge from the originally pre-trained model. Additionally, we can adapt a LLaMA model to another language, such as Hungarian. Our PULI LlumiX models in three Hungarian benchmark could achieve significant better performance. Our instruction model in both HuSST and HuRTE zero-shot competitions could achieve more than 10 accuracy scores. Our further pre-trained Llama-2 model, the PULI LlumiX 32K and the fine-tuned PULI LlumiX 32K Instruct, became state-of-the-art models capable of solving various language technology problems.

ABSTRACTS OF CONTRIBUTED TALKS

Modelling and visualization of spatial data: A case study of COVID-19 data in Maine, USA

Ismail H. Abdi (Doctoral School of Informatics, University of Debrecen)

The coronavirus pandemic has rapidly spread worldwide, and has significantly disrupted the daily lives of people globally. This paper presents the results of a detailed spatial analysis of the COVID-19 cases reported between March 2020 and December 2022 in Maine, USA. The aim of this study was to estimate and test the spatial autocorrelation of disease cases among the counties of the state, determine relative risks, and assess excess risk in each county. Using the adjacency matrix, Moran's I statistic was computed, and a significant positive spatial autocorrelation was found. The study employed Bayesian Conditional Autoregressive (CAR) distribution with Integrated Nested Laplace Approximation (INLA) to map the relative risk of COVID-19 cases, and identified that 56.25% of the counties in the state had a relative risk greater than one. The study further found clustering of disease cases in the state, which may indicate areas that need intervention measures in the future.

Availability modeling in Software Architecture

Tamás Bartók (EPAM Systems, Institute of Informatics, University of Szeged), Ferenc Héjja (EPAM Systems, Doctoral School of Informatics, University of Debrecen) and Gergely Kocsis (Department of Informatics Systems and Networks, Faculty of Informatics, University of Debrecen)

Most engineers and Solution Architects are familiar in general with Availability as a Quality Attribute. Availability on its own can mean different system quality to different stakeholders and could have a different meaning in different stages of a product. In this paper we are going through different definitions of availability, how it links to the basic concepts in probability theory and will illustrate with examples how they can be calculated. Finally we will explain which steps should be followed to model availability with fault tree analysis.

A Framework and a Theoretical Model to Integrate Internet-of-Things with Human Behavior in Teams

Alex Doboli (Department of Electrical and Computing Engineering, Stony Brook University)

The effective integration of complex human behavior with the automated responses of engineered systems and networks of systems remains challenging because existing theoretical models on human behavior devised by social sciences are hard to combine with traditional computing approaches. This paper proposes a novel framework to study and devise a new theoretical model of individual and team behavior, so that the model remains grounded in social science theories but can be used to optimize the operation and performance of engineered systems, including Internet-of-Things (IoT). Our ongoing work employs the framework to devise new algorithms that optimize IoTs that are more sensitive to individual and team behavior.

Further Keyword Generation Experiment in Hungarian with Fine-tuning PULI LlumiX 32K Model

Réka Dodé (HUN-REN Hungarian Research Centre for Linguistics) and Zijian Győző Yang (HUN-REN Hungarian Research Centre for Linguistics)

Our research represents the next step in an ongoing investigation where we employed neural models for generating and extracting keywords from lengthy texts, utilizing texts gathered from the REAL repository along with manually provided keywords by authors.

As a precursor to this research, we conducted experiments using three models. The first model utilized was fastText, where keyword extraction was treated as a multi-label classification task, serving as our baseline. The second method involved the pre-trained Hungarian language model PULI (GPT-3), which we fine-tuned for keyword generation. The third model was a fine-tuned Llama 2 model, which underwent further training.

In our current study, we fine-tuned a new model, the PULI LlumiX 32K model, with the same training data. This model combines Hungarian language knowledge with the capability of the Llama 2 model's lengthy 32,000-token input.

The added value of language models could lie in generating new, relevant keywords. Therefore, we examined how many new, relevant keywords the language models generated compared to

the author-provided keywords, as well as how many relevant keywords they generated that were not present in the text.

Upon comparing the results, we observed that the PULI LlumiX 32K model outperformed both the PULI language model and the Llama 2 model.

As for the generated keywords not present in the text, in the cases of PULI LlumiX 32K and Llama-2-7B-32K, the ratio was approximately 20%, similar to the author's keywords. For PULI GPT-3SX, the ratio was higher, around 30%. Within these, we examined how many keywords neither appeared in the text nor among the author's keywords. The ratios were similar here as well: around 20% for the Llama-2-7B-32K and PULI LlumiX 32K models, and around 30% for PULI GPT-3SX. Some of the new keywords that the text did not contain were relevant, at other times the search proved inadequate due to erroneous phrases i.e. with truncated endings.

Analysis of different Indoor Environment for the Terahertz propagation

Nagma Elburki (INRS, Universite du Qubec) and Sofiene Affes (INRS, Universite du Qubec)

In this paper, we have used different frequencies in THz bands and different points in the given plan. The transmitter position differs from the packed space to open space which makes an impact on the RF Signatures which can be seen in the paper. The similar RF Signature gives the similar kind of accuracy and prediction when done with the machine learning models like Artificial neural Networks (ANN), Random Forest (RF) etc. This paper gives the comparative view of RF Signatures and their prediction by using the different prediction models for the compartive analysis of RF Signatures.

Transforming Erlang Server Applications

Zsófia Erdei (Eötvös Loránd University Budapest), István Bozó (Eötvös Loránd University Budapest) and Melinda Tóth (Eötvös Loránd University Budapest)

In Erlang, behaviours serve as specialized design patterns, offering numerous advantages. For example, they allow us to abstract common components when addressing similar problems. Additionally, recognizing design patterns enhances our understanding of software source code by providing structured insights into specific parts and the underlying design decisions.

In contrast to object-oriented languages, which have various tools for identifying design patterns, functional programming languages like Erlang have fewer readily available solutions. Our focus is on legacy Erlang systems and their transformation into more readable modern code. Previously we developed a possible method for identifying source code fragments in such systems based on static analysis. In this paper, we propose a method for transforming recognized server candidates into client-server Erlang design patterns. To achieve this, we use the RefactorErl framework to verify the feasibility of the transformation and then transform the code into a generic server process. Our approach is demonstrated through an illustrative example.

Impact of Random Number Generation Methods Usage on Swarm Intelligence Algorithms for Energy Optimization in Wireless Sensor Networks

Levente Filep (Faculty of Informatics, University of Debrecen) and Zoltán Gál (Faculty of Informatics, University of Debrecen)

Swarm Intelligence (SI) is a complex, adaptive, and intelligent collective behavior observed in decentralized, self-organized systems. These behaviors arise from the collective, yet simple actions of individual agents forming the group. SI algorithms gained significant attention in various fields of science due to their optimization and problem-solving applications. A variety of algorithms have been proposed in the literature and applied to different optimization problems. As the actions of individuals are partly governed by seemingly random behavior, as well as the system usually being initialized at random, these algorithms rely on random number generators. These generators are pseudo-random in nature, an important aspect of experiment result repeatability. We often don’t think about these generators affecting the algorithms' performance, especially after a large number of generated random numbers. However, as we will see in this paper, this is not always the case. This paper focuses not on the mathematical background of the RNG algorithms but on the effects of them on the SI algorithms’ behavior conducted in MATLAB. We further focus on the performance of SI algorithms in WSN antenna placement problems, as well as classical benchmark landscapes, such as Rastrigin, and Rosenbrock.

Application of digital twins for connected, cooperative and automated mobility

Alexandru Forrai (Siemens Industry Software, The Netherlands), Amit Gali (Eindhoven University of Technology) and Ion Barosan (Eindhoven University of Technology)

This paper investigates the use of digital twins (DT) in connected, cooperative automated mobility (CCAM). At first, it provides a basic introduction to digital twins and their significance in expanding automated driving systems, which are crucial to Connected, Cooperative, and Automated Mobility (CCAM). The paper outlines three distinct scenarios in which digital twins play a vital role: monitoring vehicles from a remote location, conducting mixed reality testing, and ensuring the safe control of autonomous vehicles during emergency maneuvers and stops. Furthermore, it describes the architecture of an experimental test environment specifically created to implement and validate these use-cases. The study finishes by discussing the software implementation and presenting preliminary test results that provide evidence for the proposed concepts and the built architecture.

Model-based neural networks for thermographic image reconstruction

Gergő Galiger (Eötvös Loránd University Budapest), Nazila Azadi (Silicon Austria Labs Linz), Bernhard Lehner (Silicon Austria Labs Linz), Mario Huemer (Johannes Kepler University Linz) and Péter Kovács (Eötvös Loránd University Budapest)

Thermographic imaging is a specific non-destructive evaluation (NDE) approach in which the investigated object is exposed to an initial thermal excitation. The utilization of infrared cameras to detect the induced surface temperature change enables the deduction of the internal structure of the inspected material. This requires addressing a large-scale ill-posed linear inverse problem, involving regularization and forward modeling of the observed thermal diffusion process.

In this paper, we introduce a hybrid approach for thermographic imaging, which combines a deep unfolded Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) algorithm with a U-Net. Consequently, the resulting network architecture incorporates established numerical regularization heuristics, such as sparsity and smoothness. Additionally, to mitigate the data

demand of DL approaches, we establish a simulated training database containing both point-like defects and line-like cracks. As a case study, we begin by considering applications in nondestructive material testing, starting with 2D problems and progressing towards thermal tomography in 3D.

Applying Hybrid Forecasting Model SARIMA–SVR for Daily Energy Consumption Data

Herry Kartika Gandhi (Department of Information Technology, University of Debrecen)

This paper brings a new option for developing a hybrid model to predict daily energy consumption data uncertainties. High forecasting errors can lead to poor management planning and impact load shedding, which disrupts the pace of a region's economy. We propose a hybrid linear and nonlinear SARIMA-SVR by placing the residual SARIMA (linear) as the input SVR (non-linear). We find that the SVR formula can reform the SARIMA residual and minimize the total error value by providing more accurate results on the residual model. In the comparison test, we use four benchmark models: SARIMA, SARIMA-ARCH, and SVR. The Diebold-Mariano test proves that the proposed model is significantly better than the others. Whereas in n-step horizon forecasting compared to other models, SARIMA-SVR shows better error measurement values for (2, 3)-steps, but for (5, 10, 15)-steps, SARIMA-ARCH gives superior results on eastern US daily energy consumption datasets.

Application of different option numbers in Thurstone motivated models

László Gyarmati (Department of Mathematics, University of Pannonia), Csaba Mihálykó (Department of Mathematics, University of Pannonia) and Éva Orbán-Mihálykó (Department of Mathematics, University of Pannonia)

In this paper paired comparison models with stochastic background are investigated and compared. The possibility of the choice of the allowed options in decisions arises the question which choice is better. We compare two models allowing different option numbers from different aspects. We investigate the 3 and 5-option models both on theoretical and simulation planes. 3-option models do not distinguish the degree of winning, but 5-option models can

discriminate small wins from big wins and small defeat from big defeat, therefore they may use more information. It is an interesting question what are the requirements for their applications and whether the evaluation will be more reliable or not in the case of more information. We present how these two models with different option numbers relate to each other and we analyze a real-life sport example. We can conclude that for few comparisons, the 3-option model is worth using, while for a large amount of data, the model allowing 5 options is preferred.

A stacked ensemble model to identify Bangla religious hate comments

Md. Nahid Hasan (Bangladesh University of Engineering and Technology), Raiyan Azim (Bangladesh University of Engineering and Technology), Mahmudul Hasan (Khulna University) and Md. Monarul Islam (Khulna University of Engineering and Technology)

Today's world is changing rapidly. There are lots of pros and cons on each side in every sector. Social media is no different. People are using social media for various purposes. Having different aspects has enriched the social media platform but problems like hate speech are increasing alarmingly. Bangladesh and India consist of most of the people who speak Bengali. The majority of people who live in Bangladesh are Muslim and the majority of the people who live in India are Hindu. There has always been a soft keyboard fight between these two communities about their religion. This escalates mostly in times of religious celebrations. A dataset was created with 3003 comments that represent both hate comments and not hate comments in a heated conversion. This dataset is used to create a stacked ensemble model to predict whether the comment is a religious hate comment or not. 4 machine learning models namely Naive Bayes, Random Forest, Support Vector Machine, and XGBoost have been used as the 1st classifier and a neural network of (4,3,1) architecture has been used as the 2nd classifier to get the final output. While individual machine learning could only attain 80.7% accuracy, the stacked ensemble model architecture yielded an aggregate accuracy of 96.6%.

Training Embedding Models for Hungarian

Péter Hatvani (PPKE Doctoral School of Linguistics, HUN-REN Hungarian Research Center for Linguistics) and Zijian Győző Yang (HUN-REN Hungarian Research Centre for Linguistics)

Building Retrieval-Augmented Generation (RAG) systems for underrepresented languages, such as Hungarian, presents significant challenges due to the lack of high-quality embedding models. In this study, we address this gap by developing three state-of-the-art encoder-only language models specifically designed to enhance semantic similarity understanding for Hungarian. Utilizing a combination of public and internal datasets, including a 226-item corpus of news article titles and leads and a Hungarian version of the Semantic Textual Similarity (STS) dataset, we rigorously evaluate these models' performance. Our models — xml_roberta_sentence_hu, hubert_sentence_hu, and minilm_sentence_hu — demonstrate substantial improvements in semantic similarity tasks, with the hubert_sentence_hu model achieving the highest accuracy and F1-Score on the test corpus. These results underscore the potential of our models to significantly advance NLP capabilities for Hungarian, paving the way for their integration into more comprehensive RAG systems. Future work will focus on further refinement and application of these models in diverse contexts to enhance their performance and robustness.

A comparative study of pre-trained models in breast ultrasound image segmentation

Dhafer G. Honi (Doctoral School of Informatics, University of Debrecen), Mohammed Nsaif (Doctoral School of Informatics, University of Debrecen), Szathmáry László (Faculty of Informatics, University of Debrecen) and Szeghalmy Szilvia (Faculty of Informatics, University of Debrecen)

In recent years, several deep learning architectures have emerged achieving impressive results in breast ultrasound image segmentation, despite the fact that the problem itself is challenging, because of the variation in lesion size and unequal distribution of intensity in the lesion area. Many of these methods were trained and evaluated on a specific dataset, the Breast Ultrasound Images (BUSI), as it was one of the first publicly available datasets in the field with expert annotations. However, recently, problems with the dataset have been discovered. We conducted our research to estimate, through a few selected methods, the extent to which problems with the

dataset make the performance values reported in recent years unreliable. To achieve this, the selected procedures were trained and evaluated along the same methodology on the original and the cleaned datasets. Our results indicate that results related to the BUSI collection should be treated with serious caution.

Machine Learning Classification Approach for Refractive Index Prediction of D-Shape Plasmonic Biosensor

Nazrul Islam (Institute of Information Technology, Jahangirnagar University), Md. Nahid Hasan (CSE Discipline, Khulna University), Mia Mohammad Shoaib Hasan (Department of EEE, RUET), Imam Hossain Shibly (Department of EEE, RUET), Mohammad Abu Yousuf (Institute of Information Technology, Jahangirnagar University) and Md. Zia Uddin (Sustainable Communication Technologies, SINTEF Digital)

Machine learning (ML) has become an essential factor in improving the capabilities of sensing technologies across a broad range of applications, including industrial automation, environmental monitoring, biosensing, agriculture, and beyond. This paper introduces a D-shaped plasmonic biosensor with an ML classification scheme. The findings demonstrate that the model achieved a maximum resolution of 5.26 ×10−6 RIU and a wavelength sensitivity of 19,000 nm/RIU. These results showcase its acute sensitivity to subtle refractive index (RI) variations. Additionally, the study aimed to evaluate the sensor model's proficiency in predicting RI values between 1.34 and 1.41. It utilized a confusion matrix for analytical assessment.

The confusion matrix analysis highlighted the model's precision, as most classifications achieved a near-optimal area under the curve (AUC) values. These values come from receiver operating characteristic (ROC) curve assessments. The synergy of the D-shaped plasmonic biosensor with an ML-based classification paradigm marks a significant leap forward. This advancement in optical biosensing technologies is poised to enhance the accuracy of biomolecular detection.

Yule-Walker estimation of periodic INAR signals

Márton Ispány (Faculty of Informatics, University of Debrecen)

A periodically correlated non-negative integer-valued autoregressive process is considered. We present the main properties of the model through its implicit vector representation. The paper aims to introduce a new estimation method for the model's parameters. The novel algorithm is a version of the classical Yule-Walker algorithm that utilizes the autocovariance function. We propose two versions of the estimation method depending on whether immigration is uncorrelated or correlated.

Monte Carlo simulations verify the effectiveness of the proposed methodology.

Next Generation of Assessment Evaluation System for IT Configuration Challenges

Tamás Kádek (Faculty of Informatics, University of Debrecen) and Piroska Biró (Faculty of Informatics, University of Debrecen)

Automated evaluation systems involve a number of challenges, especially when the purpose of the evaluation is to verify an IT task. This environment encourages us to perform as many subtasks as possible with the help of a computer.

However, setting up a system of exercises and exams is far from simple.

The automated nature of the measurement must have a minimal impact on the measured results. In addition to ensuring that the system cannot be manipulated, care must be taken to ensure that the way in which it is used does not make the students' work more difficult or divert the focus from the task to administrative problems. It is natural that in the IT faculty, there are a number of tasks that have to be carried out in a computer environment, and there is then an opportunity to build into the environment some form of automation for checking.

Given the sheer volume of exams, managing them objectively would be an insurmountable task without IT support. Yet, the rapid evolution of information technology poses a perpetual challenge for evaluation software developers, who are constantly confronted with new tools and challenges. It is imperative to periodically review and renew existing mechanisms and processes to stay abreast of these changes.

In 2020, in response to the challenges of the pandemic, a framework was developed in a rush to facilitate the assessment of IT Security lab exercises, but later on, it was also considered for

similar subjects (such as operating systems). The strength of the system was its ability to automate the testing of computer configurations. In the meantime, the system's shortcomings have become apparent, and the need has arisen for a comprehensive rethinking of the system based on the same basic concept but with an entirely new set of features. In this article, we present the design work results based on the experience of using the system in 2020, which led to the second revised version to be implemented in 2024.

Prediction Uncertainty and Node Usefulness in Shallow Unattributed Graph Neural Networks for Explainability

Richárd Kiss (Department of Telecommunications and Artificial Intelligence, Budapest University of Technology and Economics) and Gábor Szűcs (Department of Telecommunications and Artificial Intelligence, Budapest University of Technology and Economics)

Graph Neural Networks (GNNs) excel in graph-based tasks but understanding their predictions remains challenging. We propose a method to decompose GNN predictions into contributions from neighboring nodes, focusing solely on graph topology. This approach provides insights into prediction uncertainty and node usefulness. A single-layer GNN is utilized with learnable node features and Multiclass Logistic regression for classification. Training involves L2 regularization and the Node2Vec loss function to capture the structure of the graph while avoiding overfitting. Entropy is used to measure prediction certainty and node usefulness. A case study on the Cora dataset demonstrates effectiveness of our method, revealing the relationship between prediction certainty and node usefulness. Precision-recall curves validate entropy as a measure of uncertainty. We also introduce an entropy-based filtering strategy to enhance classification accuracy by ignoring noisy neighbors. Our framework enhances the interpretability and reliability of GNN predictions, offering valuable tools for graph machine learning applications.

Forecasting Roll Rate for Sensor Plausibility Check in Off-Road Vehicle Conditions

János Kontos (Dept. of Computer Science and Systems Technology University of Pannonia and Continental Automotive Hungary Ltd.), László Bódis (Continental Automotive Hungary Ltd.) and Ágnes Vathy-Fogarassy (Dept. of Computer Science and Systems Technology University of Pannonia)

The rising popularity of Sport Utility Vehicles (SUVs) and other vehicles with high centers of mass has heightened the importance of effective anti-rollover protection systems. Traditional methods for checking roll rates can be cost-prohibitive due to the need for redundant sensors. This paper presents a novel approach using a Long Short-Term Memory (LSTM) neural network to predict the roll rate of vehicles under off-road conditions, forecasting up to 0.04 seconds into the future. The model inputs include lateral and longitudinal accelerations, roll and yaw rates, the steering angle of the first axle, and the speeds of all four wheels, with data spanning the current and up to 0.04 seconds in the past. The neural network was trained and validated with five hours of preprocessed data from the Bremly, Michigan off-road test track. A comparison with results from existing literature, along with a thorough evaluation, robustly demonstrates the effectiveness and reliability of the proposed model in predicting vehicle roll rates.

Designing prompts and creating cleaned scientific text for Retrieval Augmented Generation for more precise responses from generative large language models

Róbert Lakatos (Faculty of Informatics, University of Debrecen), Eszter Klára Urbán (Faculty of Informatics, University of Debrecen), Zoltán János Szabó (Faculty of Informatics, University of Debrecen), János Pozsga (Faculty of Informatics, University of Debrecen), András Hajdu (Faculty of Informatics, University of Debrecen) and Eszter Csernai (Health Services Management Training Centre, GE HealthCare)

This paper presents a comprehensive methodology for extracting and processing data from the scientific literature to improve the performance of generative language models in the case of the application of Retrieval Augmented Generation. We show how a knowledge-based system can be created to extract information from scientific literature using generative language

models. The methodology involves a two-phase approach, utilizing the GROBID PDF processing system for initial data extraction, followed by refinement through a custom text-cleaning module. The processed data is formatted into JSON for integration into a semantic search engine, facilitating efficient searching and retrieval. Additionally, the setup of prompts and management of generative language models are meticulously detailed to optimize response quality. Evaluation of response performance using metrics such as BLEU, ROUGE, METEOR, and cosine similarity demonstrates the efficiency of the proposed methodology. As a result, the efficiency of generative language models using data embedded and cleaned in our knowledge-based system improves by 9% in terms of cosine similarity and by 27%, 6%, and 2% in the case of BLEU, ROUGE, METEOR scores, in contrast to the direct usage of the data extracted by GROBID. Overall, this work showcases the methodology's effectiveness in improving the quality of responses generated by language models and lays the groundwork for further advancements in natural language processing and semantic search systems within the scientific literature domain.

Improving Machine Translation Capabilities by Fine-Tuning Large Language Models and Prompt Engineering with Domain-Specific Data

László János Laki (Globalese GmbH.) and Zijian Győző Yang (HUN-REN Hungarian Research Centre for Linguistics)

This study examines the applicability and performance of large language models (LLMs) in the field of machine translation for in-domain texts, with particular focus on fine-tuning LLMs, few-shot prompting and word vector-based example sentence search methods. The aim of the study is to determine the extent to which the few-shot technique can improve translation quality for domain-specific texts. Our results indicate that the few-shot learning approach consistently improved translation quality across all examined LLM systems. Surprisingly, however, the word vector-based method, which uses the vectorial representation of words to select translation examples, did not perform as well as the character similarity-based fuzzy matching technique. The study discusses the performance of various systems.

Machine Learning-Based Fine-Tuning for Accented Speech Recognition in Trinidad and Tobago

Shaquille Jared De Lancy (Department of Computing and Information Technology, The University of the West Indies) and Koffka Khan (Department of Computing and Information Technology, The University of the West Indies)

This paper explores the increasing influence of conversational AI toolkits and voice recognition software across diverse industries while shedding light on the challenges faced by individuals whose accents are underrepresented in the speech corpora used for training automatic speech recognition systems. The Trinidad and Tobago accent serves as a focal point in this study. By engaging volunteers from Trinidad and Tobago and incorporating recorded speech samples sourced from the internet, a compact speech corpus was assembled. Leveraging this corpus alongside the Nvidia NeMo open-source generative AI toolkit, the paper evaluates the performance gap experienced by individuals with the Trinidad and Tobago accent and proposes a potential remedy through fine-tuned speech recognition models based on pre-existing Nvidia models. Among these, the STT En Citrinet 1024 Gamma 0.25 pretrained model emerged as the top performer, achieving a 96.86% character-level transcription accuracy on the test dataset with a Trinidad and Tobago accent. Building upon this foundation, multiple fine-tuned models were developed, with the best achieving a 99.06% character-level transcription accuracy on the same dataset. This study contributes to the limited body of research focusing on speech corpora within the broader Caribbean context, with future endeavors aiming to extend these findings to other Caribbean nations and incorporate accent classification solutions for a more comprehensive automatic speech recognition system.

Consistency and inconsistency in the case of a stochastic paired comparison model

Csaba Mihálykó (Department of Mathematics, University of Pannonia), Éva Orbán-Mihálykó (Department of Mathematics, University of Pannonia) and László Gyarmati (Department of Mathematics, University of Pannonia)

The concept of consistency is a central question if pairwise comparison matrices based methods are applied but it is less investigated in such models which have stochastic background. Nowadays incomplete comparisons are in the focus of the research, namely how the traditional

measurements can be generalized for such cases when some pairs are not compared. Data consistency in case of stochastic models has been previously investigated if two options (better and worse) are allowed in decision. In this paper we focus on the concept of consistency in the case of three options in a model defined by Davidson satisfying Luce's axiom. We present a definition for consistency and a measurement of inconsistency. From practical point of view, it can characterize the comparison results and can provide possibility for the decision makers to decide whether data are suitable or not to make reliable decision based on them.

SAWE: Signature-based Aggregated Window Embedding for Unsupervised Time Series Segment Clustering

Marcell Németh (Department of Telecommunications and Artificial Intelligence, Budapest University of Technology and Economics) and Gábor Szűcs (Department of Telecommunications and Artificial Intelligence, Budapest University of Technology and Economics)

This paper introduces SAWE (Signature Aggregated Window Embedding), an advanced methodology for unsupervised time series segment clustering. Central to our approach is the use of path signature embeddings, which excel in capturing the intricate nonlinear dynamics inherent in time series data. Path signatures transform sequences of data points into a feature set that succinctly encodes both the order and magnitude of data interactions, facilitating a robust representation that is invariant to common distortions and noise. By combining these embeddings with change point detection, and aggregating them within intervals defined by detected change points, SAWE significantly reduces sensitivity to noise and enhances feature stability. The adoption of a sliding window technique further optimizes the extraction and aggregation process, resulting in superior classification accuracy and clustering metrics compared to non-aggregated, single segment embeddings (SSE), particularly effective in handling multi-class time series data.

Evaluating RNN Models for Multi-Step Traffic Matrix Prediction

Mohammed Nsaif (Faculty of Informatics, University of Debrecen), Gergely Kovásznai (Faculty of Informatics, Eszterházy Károly Catholic University), Hasanein D. Rjeib (Faculty of Engineering, University of Kufa; Institute of Information Technology, University of Miskolc), Ali Malik (School of Electrical and Electronic Engineering, Technological University Dublin) and Ruairí de Fréin (School of Electrical and Electronic Engineering, Technological University Dublin)

Network traffic matrix prediction is used to estimate the patterns of future network flows before they are initiated. Traffic matrix prediction is needed by a wide range of network management functions such as network monitoring and it relies on historical data. In this paper, we address the task of multi-time step traffic matrix prediction using Recurrent Neural Networks (RNN). Our objective is to conduct a comparative study of different types of RNNs and to evaluate their ability to predict multi-time step Origin-Destination traffic matrices. Experiments show that RNNs are capable of predicting multiple steps of traffic matrices, however, the RMSE of the predictions increases as we move further away from the last true value. Our primary finding is that the RNN-GRU show has the best prediction ability in the very beginning steps with an RMSE of 0.048, while RNN-LSTM demonstrated higher capability with the last steps, having an RMSE value of 0.058.

Investigating the influence of hyperparameters on the optimal time-series prediction ability of generative large language models

Árpád Pándy (Department of Data Science and Visualization, Faculty of Informatics, University of Debrecen), Róbert Lakatos (Department of Data Science and Visualization, Faculty of Informatics, University of Debrecen), Péter Pollner (Semmelweis University), András Hajdu (Department of Data Science and Visualization, Faculty of Informatics, University of Debrecen) and Eszter Csernai (Health Services Management Training Centre, GE HealthCare)

This study investigates the optimal hyperparameter settings for Large Language Models (LLMs) in predicting time-series data. We conducted extensive experiments using both synthetic datasets and real agricultural data to evaluate the performance of several LLMs. Our

findings reveal that the predictive accuracy of these models is highly dependent on the configuration of the temperature and Top-p hyperparameters.

Specifically, models demonstrated optimal performance when the ratio between these two parameters was maintained below a threshold, with the most effective ratios being under 2.0 for Mistral-small and around 4.0 for Gemma-7B. Furthermore, Mistral-small showed lower variability in performance across different settings, indicating higher determinism. By aggregating data from multiple experimental conditions, we identified that maintaining a temperature to Top-p ratio below 1.36 covers 50% of top-performing pairs for Mistral-small and below 2.75 for Gemma-7B. If we consider only hyperparameters, that has such ratios, we can decrease the MSE value by 32% on average.

These insights provide a foundation for enhancing the accuracy of LLMs in time-series forecasting, offering practical guidelines for hyperparameter optimization. Our results underscore the potential of LLMs to outperform traditional methods, particularly in datasets with periodic patterns, thereby contributing to the growing literature on the application of generative models in time-series analysis.

Design of the Simulation Analysis Tool for Wireless Sensor Network

Péter Polgár (Faculty of Informatics, University of Debrecen), Levente Filep (Faculty of Informatics, University of Debrecen) and Zoltán Gál (Faculty of Informatics, University of Debrecen)

Wireless Sensor Networks (WSNs) have significantly expanded recently, demonstrating their potential in various fields including healthcare, environmental monitoring, and military applications. This research has introduced diverse programming paradigms and model-based approaches, simplifying development and enhancing productivity. Critical reviews classify current methodologies, proposing future research to address identified gaps. Robust simulation tools are essential for evaluating and optimizing WSNs before deployment; therefore, we created such tools for testing and validating WSN node mobility algorithms and for simulating real WSN scenarios in an easily configurable, modular, realistic, and parallelizable way.

Smart Ransac: A Robust Approach

Vinayak Raju (Department of Information and Communication Technology, Manipal Institute of Technology) and Chetana Pujari (Department of Information and Communication Technology, Manipal Institute of Technology)

This paper introduces Smart RANSAC, an enhancement to the traditional Random Sample Consensus (RANSAC) algorithm, aimed at increasing robustness in outlier detection by incorporating a weighted selection mechanism for point consideration. Unlike the conventional RANSAC that selects points randomly, Smart RANSAC calculates a
weighted average for the second point based on the proximity to previously fitted lines, promoting a bias towards denser point regions. Our experiments demonstrate faster convergence in line fitting tasks.

Valid Decoding in Gaussian Mixture Models

Branislav Rudic (Linz Center of Mechatronics GmbH), Markus Pichler-Scheder (Linz Center of Mechatronics GmbH) and Dmitry Efrosinin (Institute of Stochastics, Johannes Kepler University)

A novel recursive Bayesian inference method for state observation models with Gaussian mixture assumptions is presented. The proposed approach is located between marginal and maximum a posteriori (MAP) inference, both of which have been extensively explored over the last decades. A tight coupling is revealed between inferring the predicted and filtered marginal distributions and recursively decoding MAP predecessors. Based on these findings, an algorithm is presented to decode state sequences that are valid, i.e., consistent with underlying model assumptions. Since Gaussian mixtures can be used as universal approximators for density functions, an appropriate decoder holds considerable potential for various applications. Preliminary simulation results from ongoing research on object tracking, where observations are affected by multimodal noise, suggest that the proposed decoder may exhibit superior characteristics over traditional inference methods.

Harnessing Large Language Models: Fine-tuned BERT for Detecting Charismatic Leadership Tactics in Natural Language

Yasser Saeid (South Westphalia University of Applied Sciences), Felix Neubürger (South Westphalia University of Applied Sciences), Stefanie Krügl (University of Applied Sciences South Westphalia), Helena Hüster (University of Applied Sciences South Westphalia), Thomas Kopinski (University of Applied Sciences South Westphalia) and Ralf Lanwehr (University of Applied Sciences South Westphalia)

This work investigates the identification of Charismatic Leadership Tactics (CLTs) in natural language using a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model. Based on an own extensive corpus of CLTs generated and curated for this task, our methodology entails training a machine learning model that is capable of accurately identifying the presence of these tactics in natural language. A performance evaluation is conducted to assess the effectiveness of our model in detecting CLTs. We find that the total accuracy over the detection of all CLTs is 98.96% The results of this study have significant implications for research in psychology and management, offering potential methods to simplify the currently elaborate assessment of charisma in texts.

A ConvNeXt V2 Approach to Document Image Analysis: Enhancing High-Accuracy Classification

Md Saiful Islam Sajol (Dept. of CSE, Louisiana State University), A S M Jahid Hasan (Dept. of ECE, North South University), Md Shazid Islam (Dept. of ECE, University of California Riverside) and Md Saydur Rahman (Dept. of ECE, University of California Riverside)

Proper classification of documents is of tremendous importance for an organization. As digital copies of documents are now available due to technological advances, it has become convenient to classify them automatically using machine learning or deep learning algorithms. Deep CNNs have been widely applied for document image classification, whereas newly introduced transformer-based models have also presented favorable outcomes. In this paper, one of the advanced and very high-performing deep CNN models, ConvNext V2, developed very recently, has been adapted for document image classification task, that leverages the imitation of the self-

attention mechanism of transformers through the use of masked autoencoders. Prior studies have suggested that models pre-trained on ImageNet may not perform optimally when directly applied to document classification tasks. Nevertheless, our findings reveal substantial enhancements in performance using ConvNext V2, indicating that further domain-specific pre-training (for instance, on the RVL-CDIP dataset) might not be essential for attaining high accuracy.

Our results demonstrate that effective, direct application of ImageNet pre-trained models can yield significant benefits. The model has been applied to one of the most standard document image classification datasets Tobacco-3482. The results show a high overall accuracy of 92.25% with fast convergence, outperforming several other state-of-the-art methods.

Emotion recognition based on facial gestures and Convolutional Neural Networks

Francisco Emiliano Sanchez-Callejas (Faculty of Engineering, Autonomous University of Queretaro), Irving A. Cruz-Albarran (Faculty of Engineering, Autonomous University of Queretaro) and Luis A. Morales-Hernandez (Faculty of Engineering, Autonomous University of Queretaro)

Humans express emotions verbally and non-verbally through their voice, facial expressions, and body language. Facial expression recognition systems can identify the emotional state of any person by using different intelligent algorithms, such as Support Vector Machines, Hidden Markov Models, and Convolutional Neural Networks, among others. This study focuses on facial expression recognition using eye and mouth regions of images from the FER-2013 dataset by training convolutional neural network (CNN) models. Seven emotional states - happy, sad, fear, anger, disgust, surprise and neutral - were identified. The methodology included segmenting and concatenating the images to form three CNN models. The best-performing model, a four-layer CNN with 8, 16, 32, and 64 filters, achieved remarkable results: 99.05% accuracy, 100.00% precision, 93.75% recall, 96.77% F1-score, 95.95% validation accuracy, and a 0.15 validation loss with a processing time of 3.03 minutes. It was possible to develop a CNN model capable of identifying seven emotional states from only the data of the eye and mouth region using concatenated images.

ECG Data Classification with Privacy Preservation in the IoT Context

Codruta Maria Serban (Department of Computer Science, Technical University of Cluj-Napoca), Gheorghe Sebestyen (Department of Computer Science, Technical University of Cluj-Napoca) and Anca Hangan (Department of Computer Science, Technical University of Cluj-Napoca)

Myocardial infarction (MI) can go undetected and leads to permanent heart damage or death. Its early detection with the proper medication can save lives. These days, continuous monitoring of the heart is possible due to IoT devices. Various cardiovascular diseases can be identified without the expertise of a cardiologist, only by analyzing the ECG signals. Most of the time, the inputs provided to an AI model are sensitive data. In order to avoid sending patients private data through the network and be exposed to multiple attacks, Federated Learning (FL) is implemented in the myocardial infarction detection system. The ECG signals are classified by a one-dimensional convolutional neural network (1D-CNN) which efficiently differentiates the subjects with MI from those who are healthy.

Automatic Analysis of Thermal Images to Detect Hot Spot Regions

Norbert Serbán (Faculty of Informatics, University of Debrecen) and Balázs Harangi (Faculty of Informatics, University of Debrecen)

This paper presents an automated system that uses an RGB image and thermal data from a thermal camera. We aim to detect hot spot areas and fuse the two images into one to highlight the hot spot regions on the RGB image. With this solution, it is possible to identify any inflamed areas and examine the healing process after surgery. The paper discusses an image registration method, a hot spot detection algorithm that can detect abnormal hot spot regions, and a fusion algorithm that can highlight the interesting areas on the fused image.

A speed detection method for BLDC motor

Fehmi Sevilmiş (Dept. of Electrical and Electronics Engineering, Selçuk University), Ali Sait Özer (Dept. of Control and Automation Technology, Konya Technical University) and Hulusi Karaca (Dept. of Electrical and Electronics Engineering, Selçuk University)

It is very important to accurately determine the position and speed of BLDC motor. The use of an encoder to precisely control the motor is a suitable solution, but the placement of the encoder on the motor shaft causes the volume of the system to increase. In addition, the high cost of the encoder is another disadvantage. To overcome these problems, sensorless control algorithms can be used for position and speed detection. However, the complex structure of sensorless control methods and their inadequate performance at low speeds cause them not to be preferred in practical applications. For these reasons, hall-effect sensor control algorithms come to the forefront due to their simple structure and low cost. Since the resolution of hall sensor signals is low, these signals must be processed properly in speed detection. For this purpose, MAF-PLL-based speed detection method is proposed in this study. The proposed method is tested under different conditions. The results obtained show the accuracy of the proposed speed detection algorithm.

Predicting Engineering Students' Employability using Data Mining Classification Techniques

Maria Elisa Linda Taeza-Cruz (University of Nizwa), Mohammed Muneerali (University of Nizwa), Badria Hamed Al Ruqishi (University of Nizwa) and Bernard Guzman Cruz (University of Technology and Applied Sciences)

Higher education institutions in Oman emphasize the employability of their students as one of their top priorities. Hence, the primary objective of this study is to construct a model to predict engineering students' employability using data mining classification techniques. It also investigates relevant factors to improve the prediction model. The study applied a technique known as Knowledge Discovery in Databases (KDD), which utilized 250 engineering students' records stored in the University of Nizwa database from the Spring Semester of the Academic Year 2020-2021. WEKA—Waikato Environment for Knowledge Analysis— was used to create the research classifiers. The InfoGainAttributeEval and the Ranker search methods were used to identify the most significant attributes of student employability. After the testing phase,

10-fold cross-validation was carried out. The prediction models J48 and RepTree outperformed Random Tree in accuracy metrics and tree size. J48 showed the highest Kappa value, lower error rates, and compactness, making it the top choice. Attributes like internship mark, credit hours, department and specialization heavily influenced model performance, with J48 exhibiting superior accuracy and interpretability over RepTree. Using the most accurate data mining classification algorithm, a prediction model was developed as a subsequent stage of the data mining technique. The study suggested recommendations to improve the student's employability by emphasizing skills and competencies related to employment.

Analysis of the B5G/6G Communication Power Entropy Patterns Based on Generative AI Methods

Djamila Talbi (Faculty of Informatics, University of Debrecen) and Zoltán Gál (Faculty of Informatics, University of Debrecen)

Moving toward new and higher frequencies would bring the 6G communication network into practice. Using a new MAC mechanism will enhance and overcome the THz challenges. In our paper, we focused on analyzing the entropy interdependence between two subsystems, normal and lucky mobile terminals. The two introduced entropy metrics play a crucial role in interpreting the interdependence of the two subsystems. However, the lack of data obliged us to use GAN and its methods to generate similar data. Therefore, we used four different GAN methods plus the standard GAN to obtain the most significant data, and we determined the similarity using six different similarity metrics. The results showed that cosine and correlation similarities are not appropriate to capture the similarity, meanwhile, the rest; dynamic time warping, Frechet inception, root mean square error, and peak signal-to-noise ratio agreed that DCGAN was the one who generated the most accurate data series compared to the rest.

Using Speech Data to Automatically Characterize Team Effectiveness to Optimize Power Distribution in Internet-of-Things Applications

Gnaneswar Villuri (Department of Electrical and Computer Engineering, Stony Brook University) and Alex Doboli (Department of Electrical and Computer Engineering, Stony Brook University)

This paper focuses on a fresh paradigm where human actions and Machine Learning meet to maximize system performance of Internet-of-Things Edge with Humans-in-the-Loop applications. To optimally allocate resources, like the available energy, this paper explores the challenges of bridging the semantic gap between team dynamics and team efficiency, so that the more effective teams are given higher priority in resource allocation during operation. The paper proposes methods to interpret team activities using transformer models, like DistilBERT, to process team interactions conducted through speech, and then utilize the extracted insight to characterize team dynamics. Based on these characteristics, a dynamic power distribution scheme was designed to allocate the available power to teams with higher effectiveness. The results show that the proposed method can improve power allocation in IoT-E applications.

Towards machine learning based text categorization in the financial domain

Frederic Voigt (University of the West of Scotland; Department of Computer Science, Hamburg University of Applied Sciences), Jose Alcarez Calero (School of Computing, Engineering & Physical Sciences, University of the West of Scotland), Keshav Dahal (School of Computing, Engineering & Physical Sciences, University of the West of Scotland), Qi Wang (School of Computing, Engineering & Physical Sciences, University of the West of Scotland), Kai von Luck (Department of Computer Science, Hamburg University of Applied Sciences) and Peer Stelldinger (Department of Computer Science, Hamburg University of Applied Sciences)

Despite the widespread research on text categorization in various Natural Language Processing (NLP) domains, there exists a noticeable void concerning its application to financial data. This study addresses this gap by employing pre-trained Bidirectional Encoder Representations from Transformers (BERT) models, fine-tuned specifically for the financial domain, to categorize newspaper articles focusing on financial topics. We evaluate the efficacy of established models

in sentiment prediction using these rather long texts. Furthermore, we delve into the intricacies of company-specific sentiment and relevance prediction within these articles, acknowledging the prevalence of multiple companies being mentioned, thus contributing to a more nuanced understanding of text analysis in the financial sector.

Winner prediction model for basketball matches based on statistical data of teams, players, and line-ups

Chanin Witheesawas (Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University) and Suphakant Phimoltares (Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University)

Winning the National Basketball Association (hereinafter referred to as "NBA") championship is the ultimate goal for all NBA teams, as it represents the highest achievement in professional basketball player. This accomplishment increases the team's fan base, leading to a boost in revenue. Additionally, winning the championship can enhance team loyalty, inspire team players to improve their skills, and attract high-potential players to the team. This research aims to present an precise model for predicting the winning team of NBA games utilizing statistical data of teams, players, and line-ups. Seven models used in this research consist of Multilayer Perceptron (MLP), Support Vector Machine (SVM), Gradient Boosting (GB), Random Forest (RF), Naive Bayes (NB) classifier, Decision Tree (DT), and K-Nearest Neighbor (KNN) method. The dataset utilized in this research was obtained from https://www.nba.com, covering the period from 2012 to 2023 for predicting winner's team. For the test outcomes, the Multilayer Perceptron Model exhibited the highest accuracy of 85.30% by using the combination of statistical data of teams, players, and line-ups.

The First Instruct-Following Large Language Models for Hungarian

Zijian Győző Yang (Hungarian Research Centre for Linguistics), Réka Dodé (HUN-REN Hungarian Research Centre for Linguistics, Doctoral School of Linguistics, Faculty of Humanities, Eötvös Loránd University), Gergő Ferenczi (HUN-REN Hungarian Research Centre for Linguistics), Péter Hatvani (HUN-REN Hungarian Research Centre for Linguistics), Enikő Héja (HUN-REN Hungarian Research Centre for Linguistics), Gábor Madarász (HUN-REN Hungarian Research Centre for Linguistics), Noémi Ligeti-Nagy (Hungarian Research Centre for Linguistics), Bence Sárossy (HUN-REN Hungarian Research Centre for Linguistics), Zsófia Szaniszló (HUN-REN Hungarian Research Centre for Linguistics), Tamás Váradi (HUN-REN Hungarian Research Centre for Linguistics), Tamás Verebélyi (HUN-REN Hungarian Research Centre for Linguistics) and Gábor Prószéky (HUN-REN Hungarian Research Centre for Linguistics)

In recent months, large language models have gained significant attention, with companies striving to develop models capable of solving various natural language processing tasks through extensive data training. The release of ChatGPT by OpenAI demonstrated unprecedented capabilities via a multi-step fine-tuning process. For Hungarian, pre-trained large language models include PULI GPT-3SX, PULI GPTrio and in the recent months SambaLingo. In our research, we pre-trained a new large language model based on Llama-2 and inspired by ChatGPT, focuses on fine-tuning with instruction-based prompts. We created a Hungarian prompt dataset and fine-tuned the PULI large language models into instruction-following models.

In our research, we discovered that transfer learning allows the model to gain insights from other languages. We found that further pre-training of the language model could leverage valuable knowledge from the originally pre-trained model. Additionally, we can adapt a LLaMA model to another language, such as Hungarian. Our PULI LlumiX models in three Hungarian benchmark could achieve significant better performance. Our instruction model in both HuSST and HuRTE zero-shot competitions could achieve more than 10 accuracy scores. Our further pre-trained Llama-2 model, the PULI LlumiX 32K and the fine-tuned PULI LlumiX 32K Instruct, became state-of-the-art models capable of solving various language technology problems.

Purpose of mobility data usage and risk assessment: POSMO ethical data market case

Olena Yatsenko (School of Engineering and Computer Science, Bern University of Applied Sciences)

This paper scrutinizes the complex risks associated with the utilization of mobility data, highlighting the ethical considerations and potential harm to individuals whose information is collected. It stresses the imperative of adhering to ethical standards in data usage, particularly the importance of alerting to and averting the risk of re-identification and other misuse forms. Further, the paper explores the ethical challenges and responsibilities tied to data usage, centring on POSMO, a data market platform committed to ethical practices. It underscores the decision-making process within the ethics committee as a vital, non-automatable element of ethical data usage, highlighting the significance of human empathy, an understanding of sociocultural contexts, and the evolving nature of moral standards. The paper presents a case against automating ethical decision-making, given the complexity, high dimensionality, and context-sensitive nature of mobility data that necessitates meticulous integration with current systems and infrastructure. Moreover, the paper details the objectives behind POSMO's use of mobility data, from enhancing transportation systems to ensuring public safety, and stresses the importance of limiting purposes to protect the privacy and autonomy of data subjects. This conscientious data usage approach is crucial for upholding accountability, reducing data collection, and avoiding misuse, thereby leading to more transparent, reliable, and ethically oriented data management practices.

EXTENDED ABSTRACTS

On properties of feasibility in non-standard Heyting arithmetic

Péter Battyányi
Departement of Computer Science
Faculty of Informatics, University of Debrecen
Debrecen, Hungary
battyanyi.peter@inf.unideb.hu

***Abstract*—We examine nonstandard Heyting arithmetic extended with a feasibility predicate. Feasibility is defined as a downward closed property containing all numerals and closed under applications with primitive recursive functions. Making use of Kleene's arithmetical realizability, we demonstrate for a theory slightly weaker than Heyting arithmetic that provably feasible terms coincide with the set of numerals. Moreover, we show that disjunction and existential properties are preserved in our arithmetical theory.**

***Index Terms*—Heyting arithmetic, realizability, non-standard arithmetic**

I. Introduction

We are concerned with non-standard Heyting arithmetic extended with a feasibility predicate. We consider intuitionistic arithmetic, called Heyting aritmetic, and we augment it with a non-standard element. Then we add to the theory axioms expressing a feasibility relation. In more detail, Heyting arithmetic with non-standard elements is obtained by adding the axioms $\underline{n} < c$ for all natural numbers n to our arithmetical theory. It was as early as in 1934, when Skolem proved with an argumentation based on the compactness theorem for first-order logic that first-order arithmetic is consistent with these axioms [7]. Since then, non-standard models of arithmetic were intensively studied, see, for example, the monographs [3] and [4]. Following the idea of Dragalin [2], we introduce a feasibility predicate which is a formalized counterpart of the notion of a proper cut in a nonstandard model. That is, we assume that feasibility is downward closed and no primitive recursive function can lead out of the set of feasible elements. Moreover, the infinite element c is not feasible.

The overspill principle, due to Robinson, for non-standard Peano arithmetic states that we are not able to define a proper cut of a model of Peano arithmetic with an arithmetic formula [4]. In our case, we prove that F, however, defines the proper cut of standard elements. Of course, F does not belong to the original language. A theory admits the disjunction property, if, whenever $A \vee B$ is derivable for the closed formula $A \vee B$, then either A or B is derivable. Similarly, we say that for a theory the existential property fulfills, if, whenever $\exists x A(x)$ is derivable for the closed formula $\exists x A(x)$, then $A(t)$ is derivable for some closed term t. Constructive, non-standard theories may refute both properties [1], [6]. One of the reasons is that they usually require the fulfillment of axioms other than the ones considered in this paper, for example transfer principles [1]. However, we have deliberately chosen the simplest axioms that could possibly be more defendable from a constructive standpoint. As a result of this, our theory preserves both the disjunction and the existential properties.

979-8-3503- 8788-9/24/$31.00 © 2024 IEEE

II. Preliminaries

We consider the constant c and we add to the existing theory the axioms $\underline{n} < c$ for all natural numbers n, $\underline{n}$ denoting $S(\dots S(0))$, where S occurs n times in the expression.

We further extend our theory with a predicate F expressing feasibility. Let F be a one-place predicate symbol. The following axioms are from the paper of Dragalin [2]:

1) $F(0)$,
2) $\forall x \forall y (F(x) \wedge y < x \supset F(y))$,
3) $\forall x (F(x) \supset x < c)$,
4) $\forall x_1 \dots \forall x_n (F(x_1) \wedge \dots \wedge F(x_n) \supset F(g(x_1, \dots, x_n)))$, for each symbol g standing for a primitive recursive function.

In addition, we accomplish slight modifications to the predicate calculus. In effect, the new axioms state that no fresh term can be introduced in a derivation unless they are feasible.

Finally, in place of the usual induction axiom in arithmetic, we introduce the axiom

$$A(0) \wedge \forall^f x (A(x) \supset A(Sx)) \supset \forall^f x A(x) \quad (Ind^f),$$

where $A(x)$ does not contain F and the notation $\forall^f x A(x)$ stands for $\forall x (F(x) \supset A(x))$. We denote the theory obtained in this way by HAF^*.

III. Results

We state the following results. In what follows, let $\vdash$ denote provability in HAF^*.

Theorem 3.1: Let $\vdash F(t)$ for a closed term t. Then there exists an n such that $\vdash t = \underline{n}$.

The proof of the theorem relies on a realizability technique due to Kleene [5].

As a consequence of the Theorem, we obtain the disjunction and existential properties.

Corollary 3.2: Let A be a closed formula such that $\vdash A$. Then the following assertions are valid.

1) If $A = (B \vee C)$ then either $\vdash B$ or $\vdash C$.
2) If $A = \exists x C(x)$ then $\vdash C(\underline{k})$ for some natural number k.

IV. Future Work

In his paper [2], Dragalin raises some questions concerning the theory HAF, which is Heyting arithmetic with the usual axioms for predicate logic together with a non-standard element and a feasibility relation. For example, if we adopt the formal thesis of Church, that is, $\forall x \exists y A(x, y) \supset \exists e \forall x \exists y(\{e\}(x) = y \wedge A(x, y))$, then it follows from Dragalin's result that $HAF + CT$ is conservative over HAF provided CT contains no occurrence of F. The question emerges what happens when we consider CT with the premise $\forall x \exists y(F(y) \wedge A(x, y))$. Does this mean that we are able to find a y for every x by means of a feasible algorithm? In [2], Dragalin establishes an algebraic model to study provability in arithmetical theories. As a future undertaking, it could be interesting to make use of the characteristics of this model in order to examine the metamathematical properties of these theories.

References

[1] J. Avigad, J. Helzner, "Transfer principles in nonstandard intuitionistic arithmetic," Archive for Mathematical Logic 41(6), 581–602, 2002.

[2] A. G. Dragalin, "Explicit algebraic models for constructive and classical theories with non-standard elements," Studia Logica 55(1), 33–61, 1995.

[3] P. Pudlák, P. Hájek, Metamathematics of First-Order Arithmetic. Springer-Verlag, Berlin Heidelberg, 1993.

[4] R. Kaye. Models of Peano Arithmetic. Clarendon Press, Oxford, 1991.

[5] S. C. Kleene. Introduction to Metamathematics. North Holland, Amsterdam, 1952.

[6] E. Palmgren, "A note on mathematics of infinity," The Journal of Symbolic Logic, 58(4), 1195—1200, 1993.

[7] Th. A. Skolem, "Über die nichtcharakterisierbarkeit der zahlenreihe mittels endlich oder abzählbar unendlich vielen aussagen mit ausschliesslich zahlvariablen. Fundamenta Mathematicae 23, 150–161, 1934.

A family of random graph evolution models with moderate density

István Fazekas
Faculty of Informatics
University of Debrecen
Debrecen, Hungary
fazekas.istvan@inf.unideb.hu

László Fórián
Faculty of Informatics
University of Debrecen
Debrecen, Hungary
forian.laszlo@inf.unideb.hu

I. Introduction

Nowadays, network theory has become one of the most popular and important research topics. The mathematical models of the networks are graphs. The vertices of the graph represent the nodes of the network and the edges stand for the connections. There are several models that describe connections of two nodes, e.g. the preferential attachment model introduced by Albert and Barabási or a moderately dense random graph model by Backhausz and Móri [1]. However, connections between more than two nodes can also be of interest in many real-world and theoretic models.

II. The model

In this work, we shall study a discrete time network evolution model, a generalization of [1]. Our network (i.e. the graph) can contain multiple edges. The evolution of the graph is based on constructions and deletions of k-cliques, where $k \geq 2$ is a fixed integer. A k-clique is a sub-graph containing k vertices and $\binom{k}{2}$ edges, i.e. any two different vertices are connected by 1 edge. When we form a k-clique, then we draw $\binom{k}{2}$ new edges among k vertices, and we add this new clique to the list of k-cliques. The $\binom{k}{2}$ new edges will be considered as the own edges of the clique, but the formerly existing edges are not considered as own edges of the clique at hand. When we delete a k-clique, then we delete it from the list of k-cliques, and we delete its $\binom{k}{2}$ own edges. But we do not delete its vertices and we do not delete those edges which were not own edges of the clique at hand.

The initial graph at time $n = 0$ contains k vertices and no one edge. In the first step i.e. when the time is $n = 1$, we connect the k cliques to obtain a single k-clique. The second step is the following. We choose two vertices uniformly at random, let us denote them by v_1 and v_2. Then we add a new vertex and construct two new k-cliques. The vertices of the first k-clique are the existing $k+1$ vertices but v_1, while the vertices of the second k-clique are the existing $k+1$ vertices but v_2. Then the original k-clique is deleted.

From now on, in each step, we choose k vertices uniformly at random from the existing vertices. If they do not form a k-clique, then construct a new k-clique on these vertices (i.e. we connect them using $\binom{k}{2}$ new edges). In the other case, when the sub-graph consisting of the k vertices chosen is a k-clique, then that k-clique is deleted but its vertices are used to construct two new k-cliques like in the second step. That is a new vertex is added to the graph and using this new vertex and the k vertices of the just deleted k-clique, two new k-cliques are created in the same way as in the second step.

III. Our results

We have analyzed the asymptotic behaviour of these random graph models. We also present our first theorem here. Let V_n denote the number of vertices in the model after n steps. In our first theorem we state that the magnitude of V_n is $n^{\frac{2}{k+1}}$ and V_n is asymptotically normal. Let us denote the normal distribution with mean m and variance σ^2 by $\mathcal{N}(m, \sigma^2)$.

Theorem 1. *As $n \to \infty$, the following almost sure convergence holds for the number of vertices in the graph after n steps:*

$$\frac{V_n}{\left(\left[\frac{(k+1)!}{2}\right]^{\frac{1}{k+1}} n^{\frac{2}{k+1}}\right)} \to 1. \tag{1}$$

Furthermore, we have

$$\frac{1}{n^{\frac{1}{k+1}}}\left(V_n - \left[\frac{(k+1)!}{2}\right]^{\frac{1}{k+1}} n^{\frac{2}{k+1}}\right) \Rightarrow \mathcal{N}\left(0, \frac{1}{2k+1}\left[\frac{(k+1)!}{2}\right]^{\frac{1}{k+1}}\right) \tag{2}$$

as $n \to \infty$, where $\Rightarrow$ denotes convergence in distribution.

We also obtained a functional limit theorem for the number of vertices in our model. The degree of a fixed vertex is also considered, but the asymptotic behaviour of the degrees is different for the $k \geq 3$ case (and we also have to distinguish between the $k = 3$ and $k > 3$ case). However, we have asymptotic normality for a transformed quantity obtained from the degree of a vertex. A multidimensional functional limit theorem was also obtained for the degree of multiple vertices. The main tools to prove these results were limit theorems for certain martingales. Martingales were obtained from V_n and the degree of the vertices.

REFERENCES

[1] Backhausz, Á., Móri, T. F. A random graph of moderate density. *Electronic Communications in Probability*, **27**, *2022*, 1–12.

[2] Hall, P., Heyde, C.C. Martingale Limit Theory and its Application. *Academic Press: New York*, 1980.

[3] Neveu, J., Discrete-Parameter Martingales. *North-Holland: Amsterdam*, 1975.

TimeTune: Personalized Study Plans Generation with Google Calendar Integration

Chevon Fernando
Computer Science and Engineering
University of Westminster
London, UK
chevonfernando10@gmail.com

Banuka Athuraliya
Department of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
banu.a@iit.ac.lk

Abstract— The purpose of this research is to provide a solution to the students' time management, which usually becomes an issue because students must study and manage their personal commitments. "TimeTune," an AI-based study planner that provides an opportunity to maneuver study timeframes by incorporating modern machine learning algorithms with calendar applications, is unveiled as the ideal solution. The research is focused on the development of LSTM models that connect to the Google Calendar API in the process of developing learning paths that would be fit for a unique student's daily life experience and study history.

A key finding of this research is the success in building the LSTM model to predict optimal study times, which, integrating with the real-time data of Google Calendar, will generate the timetables automatically in a personalized and customized manner. The methodology encompasses Agile development practices and Object-Oriented Analysis and Design (OOAD) principles, focusing on user-centric design and iterative development. By adopting this method, students can significantly reduce the tension associated with poor study habits and time management.

In conclusion, "TimeTune" displays an advanced step in personalized education technology. The fact that its application of ML algorithms and calendar integration is quite innovative is slowly and steadily revolutionizing the lives of students. The excellence of maintaining a balanced academic and personal life is stress reduction, which the applications promise to provide for students when it comes to managing their studies.

Keywords — Personalized Learning, Study Planner, Time Management, LSTM (Long Short-Term Memory) Model, Calendar Integration

Decidability, Interpolation, Amalgamation - A Case Study from Algebra through Logic to Computer Science

Sándor Jenei
Institute of Mathematics and Informatics
Eszterházy Károly Catholic University
Eger, Hungary
jenei.sandor@uni-eszterhazy.hu
and
Institute of Mathematics and Informatics
University of Pécs
Pécs, Hungary
jenei@ttk.pte.hu

Abstract

In this presentation, we will discuss the relationship between the interpolation (logical) property and (algorithmic) decidability, the connection between interpolation and (algebraic) amalgamation, and demonstrate the presence or absence of the amalgamation property in some algebraic structures that are significant in the field of substructural logics. For details, see [1]–[3].

Index Terms

decidability, interpolation, amalgamation

ACKNOWLEDGMENT

Supported by the NKFI-K-146701 grant.

REFERENCES

[1] G. Metcalfe, F. Montagna, C. Tsinakis, "Amalgamation and interpolation in ordered algebras," Journal of algebra, 402, (2014). pp. 21–82.
[2] G. Metcalfe, F. Paoli, C. Tsinakis, "Residuated Structures in Algebra and Logic," AMS, ISBN: 978-1-4704-6985-6
[3] S. Jenei, "Amalgamation in classes of involutive commutative residuated lattices," arXiv: 2012.14181

979-8-3503-8788-9/24/$31.00 © 2024 IEEE

A Two-Component Approach of Granules in Rough Sets Theory

Tamás Kádek
Department of Computer Science
University of Debrecen, Faculty of Informatics
Debrecen, Hungary
kadek.tamas@inf.unideb.hu

***Abstract*—Granuels play a crucial role when rough-set-based approximations are used. They also become useful when an information system is approximated using their attributes. A corresponding attribute system can help us to introduce properties, giving us the ability to create a language of granules. This leads to the idea of separating the property level and the object level of the approximations. First, in the presentation, I will summarize, as co-authors of the work, a recent publication in this field [1], in which we introduce an approach that takes advantage of this viewpoint that results in approximations on both intensional and extensional levels; then, I will highlight the possible benefits and further challenges this system can cause.**

***Index Terms*—Granular computing, rough set theory, intensions of granules, extensions of granules**

REFERENCES

[1] Tamás Mihálydeák, Tamás Kádek, Dávid Nagy, Mihir K. Chakraborty, Intensions and extensions of granules: A two-component treatment, International Journal of Approximate Reasoning, Volume 169, 2024, 109182, ISSN 0888-613X, https://doi.org/10.1016/j.ijar.2024.109182

Non-cooperative Polymorphic P systems

Anna Kuczik
University of Debrecen, Faculty of Informatics
Debrecen, Hungary
kuczik.anna@inf.unideb.hu

***Abstract*—This abstract contains a summary of my previous year's research work in the field of membrane systems. I dealt with a special case of membrane systems: polymorphic membrane systems. I compared non-cooperative polymorphic P systems with ET0L systems and PC ET0L systems. I summarize my results in this short summary.**

***Index Terms*—Membrane computing, P systems with dynamic rules, Polymorphic P systems, P systems with non-cooperative rules, P systems with limited depth, Parikh sets of ET0L languages**

I. Introduction

Membrane systems, or P systems were introduced in [9] as a computing model based on an abstract view of the processes taking place in living cells. They consist of a nested structure of membranes, and the regions delimited by these membranes contain different multisets of objects. These multisets are transformed step by step during the functioning of the system, until the result is produced when the system reaches a halting configuration. In the basic variant of the model, the objects can be rewritten in each computational step by parallel application of multiset rewriting rules, or they can be transferred (communicated) between neighboring regions. More details on the different variants and their computational properties can be found in the monograph [7] or in the handbook [8].

Polymorphic P systems were motivated by the idea that the program of a computing device could be viewed as data; therefore, it could also be changed during a computation. In these types of P systems, rules are not statically defined but are dynamically inferred from the contents of pairs of membranes: The contents of one membrane of the pair define the multiset representing the left-hand side of the rule, the contents of the other membrane define the right-hand side.

The initial results presented in [3] show the power of the model. With cooperative rules (rules with left-hand sides having more than one object) any recursively enumerable set of numbers can be generated, but non-cooperative systems (systems with rules having just one object on the left-hand side) can also generate several interesting languages, mainly based on the fact that exponential, even super-exponential growth of the number of objects inside the system can be produced.

The study of non-cooperative variants of the model was continued further in [5] by considering the case of "no ingredients", that is, when no special features (not even target indicators) are added to the system. As its main contribution, [5] presented a hierarchy of computational power based on the depth of the membrane structure, but in general, many questions concerning the computational capabilities of the non-cooperative variant remained open for further investigation.

979-8-3503- 8788-9/24/$31.00 © 2024 IEEE

II. Polymorphic P systems and Parikh sets of ET0L systems

With our work, we wanted to take further steps in this direction, in determining the strength of polymorphic P systems. Non-cooperative systems with so called finitely representable regions were investigated in [1]. A region is finitely representable (FIN-representable in short), if the set of possible multisets that can appear as the contents of the region in question during any computation is finite. In [1] we have shown that languages of non-cooperative polymorphic systems where all regions (besides the skin region) are FIN-representable coincide with the Parikh sets of languages of ET0L systems [6], [4]. We showed that (1) Parikh sets of ET0L languages can be generated using non-cooperative polymorphic P systems (with no other ingredients), and that (2) ET0L systems can generate string languages corresponding to the multiset languages of non-cooperative polymorphic P system where the set of the possible contents of regions corresponding to right-hand sides of rules is finite.

III. Polymorphic P systems and nonreturning PC ET0L systems

We have continued the investigations concerning the generative power of non-cooperative polymorphic P systems. In [2] article, we began to investigate the relationship between non-cooperative polymorphic P systems and parallel communicating ET0L systems. We have shown that they can be simulated by nonreturning PC ET0L systems and, thus, that the class of Parikh sets of nonreturning PC ET0L languages includes the class of languages generated by non-cooperative polymorphic P systems.

It seems to be clear that polymorphic P systems cannot simulate PC ET0L systems in a straightforward manner, because the communication structure of parallel communicating systems can be more complicated. The exact characterization of languages of non-cooperative polymorphic systems with PC ET0L systems having some kind of a simplified communication structure seems to be an interesting topic for further research, which we also plan to investigate in the future.

REFERENCES

[1] A. Kuczik, Gy. Vaszil (2024, May). "Simple variants of non-cooperative polymorphic P systems". *Journal of Membrane Computing*. Available: https://doi.org/10.1007/s41965-024-00145-0.

[2] A. Kuczik, Gy. Vaszil (2024) "Non-cooperative Polymorphic P Systems and Parallel Communicating ET0L Systems" unpublished.

[3] A. Alhazov, S. Ivanov, Y. Rogozhin, "Polymorphic P Systems" in *Membrane Computing*, vol. 6501, M. Gheorghe, T. Hinze, G. Păun, G. Rozenberg, A. Salomaa, Eds. Springer-Verlag: Berlin, Heidelberg, 2011, pp. 81–94

[4] A. Ehrenfeucht and G. Rozenberg, S. Skyum,"A relationship between ET0L and EDT0L languages", *Theoretical Computer Science*, vol. 1, issue 4., pp. 325–330, April 1976

[5] S. Ivanov, "Polymorphic P Systems with Non-cooperative Rules and No Ingredients" in *Membrane Computing*, vol. 8961, M. Gheorghe,G. Rozenberg, A. Salomaa, P. Sosík, C. Zandron, Eds. Springer International Publishing: Cham, 2014, pp. 258–273

[6] A. Lindenmayer, "Mathematical models for cellular interactions in development I. Filaments with one-sided inputs", *Journal of Theoretical Biology*, vol. 18, pp. 280-299, March 1968

[7] G. Păun, "Membrane Computing: An Introduction", Eds. Springer-Verlag: Berlin, Heidelberg, 2002

[8] G. Păun, G. Rozenberg, A. Salomaa, "The Oxford Handbook of Membrane Computing", Oxford University Press: Oxford, 2010

[9] G. Păun, "Computing with Membranes" *Journal of Computer and System Sciences*, vol. 61, pp. 108–143, August 2000

Requirement Specification of an Information Bank

Gábor Kusper
Faculty of Informatics
University of Debrecen
Debrecen, Hungary
kusper.gabor@gmail.com

Gyula Erdei
Informatics Department
Mayor's Office of Eger
Eger, Hungary
erdei.gyula@gmail.com

Attila Szekeres
Doctoral School of Management and Business Administration
John von Neumann University
Budapest, Hungary
attila.szekeres@gmail.com

***Abstract*—An information bank is a data storage solution where users can store data they own. Nowadays our data are scattered over the internet stored by large companies, government agencies, or simply lost. Sooner or later there will be information banks that offer free or at least cheap solutions to store our data, the data owned by ourselves. The main use case is the following: Mr. X buys a sofa and gets a PDF file which describes how to clean the sofa. Mr. X. uploads this PDF to his information bank account. Then, after 3 years of using the sofa, he split some coffee on it. He completely forgot how to clean his sofa, but he knows that most probably he uploaded the corresponding information to his information bank account. So, he logs into his account and asks a Large Language Model-based artificial intelligence trained on his data how to clean his sofa. He should not check the name of the sofa, since he owns only one, so the AI understands that he is looking for the information from the 'how to clean the XYZ sofa' PDF. We collect use cases and requirements relevant to designing such an information bank.**

***Index Terms*—information bank, Large Language Models, requirement specification**

I. Introduction

The problem statement of an information bank is given in the abstract. Due to space constraints, we do not repeat it here.

An information bank, defined by one of the authors, see [4], is fundamentally a user-centric data storage solution, where the data stored is unequivocally owned and controlled by the user. This paradigm shift is crucial in an era where data ownership and privacy concerns are increasingly in the spotlight [1]. By allowing users to consolidate their data in a single, secure location, information banks promise enhanced accessibility and organization, as well as empowering individuals over their own digital information.

One of the primary advantages of an information bank is its potential to integrate seamlessly with advanced technologies, such as Large Language Model (LLM)-based artificial intelligence (AI). LLM solutions are also available in Hungarian [5]. This integration can significantly enhance the utility of stored data, providing personalized and context-aware responses based on the user's data.

II. Draft of a Requirement Specification of an Information Bank

An information bank is not a

- data marketplace [3], neither a
- data monetization framework [2], nor a
- information broker.

Some information banks might offer the services listed above to exploit the big data stored by users, but that should not be their core function.

We collect the main use cases of a future information bank:

- Mr. X creates his information bank account. To do that, he has to identify himself to the information bank.
- Mr. X closes his information bank account. If he wishes, all his data are transferred to his new information bank account.
- Mr. X uploads a document which contains information which is owned by him.
- Mr. X links his browser to his information bank account. So, his browser history is collected at his information bank account.
- Mr. X links his email client to his information bank account. So, his emails are collected at his information bank account.
- Mr. X links his smartphone to his information bank account. So, his smartphone usage history is collected at his information bank account.
- Mr. X buys a bridge software which links his healthcare profile to his information bank account. He has to sign that his healthcare data can be stored in his information bank account.
- Mr X. has some IoT gadgets, namely a security camera. Mr X. links his security camera to his information bank account. He has to pay some extra fee to rent 5GB extra storage. The stream is stored in this extra storage, which is able to store 4-5 days of the video stream.
- Mr. X purchases a sofa and receives a PDF manual detailing the cleaning instructions. By uploading this document to his information bank, Mr. X ensures that the information is readily accessible. Years later, when he needs to clean a coffee spill, he can simply query his AI assistant, which, leveraging the stored data, can provide specific cleaning instructions without Mr. X having to recall the document's details or even the sofa's model.
- Mr. X opens his information bank, and opens his AI assistant. He can ask questions like: What was the article I read yesterday about sonar energy?

Thanks for GINOP-2.1.2-8-1-4-16-2017-00176

- Mr. X install his AI assistant on his smartphone. To do that, he has to agree that this app uses an internet connection to communicate with his information bank account. After that, he can use his AI assistant directly from his smart phone.
- Mr. X can ask questions to his AI assistant which needs to combine information from multiple sources which are uploaded or linked to his information bank account. For example: I have an electric car and some solar panels, as you know. Are they compatible? Can I charge my car directly from the solar panels?
- Mr. X synchronizes his information bank account across multiple devices. This allows him to access his data from his smartphone, tablet, and desktop computer, ensuring he can retrieve or upload information anytime, anywhere.
- Mr. X uploads multiple versions of a document to his information bank account. The system automatically tracks changes and maintains a version history, allowing Mr. X to revert to previous versions if needed.
- Mr. X shares certain documents in his information bank account with family members or colleagues. He can set permissions to control who can view, edit, or comment on these documents.

Requirements of a future information bank:

- An information bank has to be able to serve several users. Each user must have an information bank account. There is no active user without an account.
- An information bank must implement multi-factor authentication to enhance security. Users should verify their identity using two or more methods, such as a password and a mobile phone code.
- All data stored in the information bank should be encrypted both at rest and in transit. This ensures that data is protected from unauthorized access and breaches.
- The information bank should have an intuitive and user-friendly interface. Users should be able to easily upload, manage, and retrieve their data without technical difficulties.
- The information bank should undergo regular security audits by third-party experts to ensure the system remains secure against emerging threats and vulnerabilities.
- An information bank account can store documents in TXT, PDF, or in any other well-known formats.
- The AI assistant should be trained using the documents uploaded to the account.
- An information bank account may not store sensitive data (like healthcare data), unless the owner of the sensitive data is the user of this account, and the user gives his or her legal consent, which grants the right to store that sensitive data. The form of legal consent should comply with the laws of both the user's nationality and the country of the bank.
- An information bank account may not store sensitive data (like healthcare data), unless the owner of the sensitive data is the user of this account, and the user gives their legal consent, which grants the right to store that sensitive data. The form of legal consent should comply with the laws of both the user's nationality and the country of the bank.
- An information bank account should be able to store also images and video streams. The AI assistant should be trained using also that images and video streams.
- An information bank account should be able to store browser history.
- An information bank account should be able to store smartphone usage history.
- An information bank might rent extra storage capacity for additional fees to its users.
- An information bank can charge money according the usage of its services.
- An information bank is a legal entity.
- An information bank should provide guarantees that hackers cannot access data stored within it.
- The deep learning model used by the AI assistant of the information bank should be easy to update.
- The information bank must comply with international data protection regulations such as GDPR, HIPAA, and CCPA. This includes providing users with the right to access, modify, and delete their data.

III. Future Work

We plan to do questionnaires to understand which use cases are more important, and which ones are less important. We should also identify additional use cases that have not yet been collected. We should use OO analysis and design techniques to describe the details of the important use cases.

We do not plan to implement or operate any information banks. That should be done by software development companies, banks, or maybe telecommunication companies that think it is a good business to implement and operate an information bank.

And the final question: Why do we, researchers, design an information bank? If we, researchers, define the way an information bank should work, then large companies with strong capital cannot patent those features, and the competition remains more open.

References

[1] A. Acquisti, C. Taylor, L. Wagman, "The Economics of Privacy," *Journal of Economic Literature*, vol. 54, issue 2, pp. 442–492, 2016.

[2] A. Z. Faroukhi, et al., "Big data monetization throughout Big Data Value Chain: a comprehensive review," *Journal of Big Data*, vol. 7, article number 3, 2020.

[3] L. Huang, et al., "Toward a research framework to conceptualize data as a factor of production - The data marketplace perspective," *Fundamental Research*, vol. 1, issue 5, pp.586–594, 2021.

[4] G. Kusper, "Definition of Digital Ownership, Digital Data Treasure, Informationbank, some Use Cases and a Business Model," *Proceedings of FutureRFID 2014*, pp. 101–113, 2014.

[5] Z. Gy. Yang, L. J. Laki, T. Váradi, G. Prószéky, "Mono- and multilingual GPT-3 models for Hungarian," *Lecture Notes in Computer Science*, vol. 14102, pp. 94–104, 2023.

Parsing Boolean grammars with a three-valued semantics

Patrik Adrián
Department of Computer Science
University of Debrecen
Kassai út 26, 4028 Debrecen, Hungary
adrianpatrik@mailbox.unideb.hu

The formalism of context-free grammars (CFGs), introduced by Chomsky [1] in 1956, has been, without a question, the most popular formalism for the description of formal languages, thanks to its easy-to-understand semantics and admission of simple and effective parsing algorithms. Even though CFGs were originally defined as a rewrite system over a dichotomous alphabet of terminal and nonterminal symbols, one may just as well consider them as a set of constraints on the possible strings made up of the terminal symbols; in this interpretation, the parsing problem becomes a problem of logic where the a proof of the proposition that a word belongs to the language of the starting symbol is sought and the grammar rules are used for inference within the proof.

Context-free grammars, however, are insufficient to properly describe all languages; natural languages are famously non-context free, and neither are some programming languages used in everyday life (for example C++, even past the preprocessor stage, or the indentation mechanism of Python). The description of these languages requires a formalism stronger than CFGs.

The Boolean grammars of Okhotin [2] are a natural extension of the context-free grammars of Chomsky, where the choice between applicable rules, formalized as a disjunction is extended to allow the use of both a conjunction and negation. The ability of Boolean grammars to describe languages that are provably not context-free has already been demonstrated in literature, but in certain cases it is also possible to obtain a simpler specification of grammars that are context-free.

As for semantics, the classical approach to truth in a formal system is the closed-world assumption (CWA); everything that is not derivable as true through the rules of the system is considered to be false. Naive application of the CWA and its related rule of inference, negation as failure, breaks down in the presence of negation, leading to possible inconsistencies in the resulting valuation, and, when a procedural approach is taken, the deductive algorithm may even fail to terminate.

A possible resolution of this inconsistency involves the extension of the underlying logic to a three-valued one, where a third judgement, understood as being indeterminate, is allowed. Deduction in three-valued logic requires not only the derivation of what is true, but also what is false, as these concepts are still disjunct but no longer complementary. Three-valued rule-based inferences have long been the studied in the field of logic programming, where two well-known models have emerged: the fully open-world semantics of Fitting [3] that only considers entailment as a valid source of determinacy and the well-founded semantics of Gelder et al. [4] that relies on a restricted version of the closed-world assumption.

Application of a three-valued approach to parsing requires the redefinition of most concepts used in formal languages, including that of a formal language. The extended theory is mostly straightforward, but the handling of concatenation as a logic construct is not. Effective techniques that prove over all possible partitions of some string are required to keep the algorithm within low polynomial bounds.

We developed a novel constuction that strongly resembles the GLR family by Tomita [5] and others. The novelty of the approach comes from the fact that our algorithm performs the explicit deduction of negative knowledge, something that, to the best of our knowledge, has not been done previously in any variant of the GLR technique. Unlike the GLR of Okhotin [6], our algorithm is monotone in the sense that it does not remove new edges from the GSS, only creates them.

The algorithm performs recognition in time $O(|w|)^3$, where w is the input word, but its extension to a parser within the same bounds could be possible.

***Index Terms*—formal languages, Boolean grammars, parsing**

References

[1] N. Chomsky, "Three models for the description of language," *IEEE Transactions on Information Theory*, vol. 2, pp. 113–124, Sept. 1956.
[2] A. Okhotin, "Boolean grammars," *Information and Computation*, vol. 194, pp. 19–48, Oct. 2004.
[3] M. Fitting, "A Kripke-Kleene semantics for logic programs," *The Journal of Logic Programming*, vol. 2, pp. 295–312, Dec. 1985.
[4] A. V. Gelder, K. A. Ross, and J. S. Schlipf, "The well-founded semantics for general logic programs," *Journal of the ACM*, vol. 38, pp. 619–649, July 1991.
[5] M. Tomita, "An efficient context-free parsing algorithm for natural languages," in *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'85, (San Francisco, CA, USA), p. 756–764, Morgan Kaufmann Publishers Inc., 1985.
[6] A. Okhotin, "Generalized LR parsing algorithm for boolean grammars," *International Journal of Foundations of Computer Science*, vol. 17, pp. 629–664, June 2006.
[7] A. Patrik and V. György, "A GLR-like parsing algorithm for three-valued interpretations of boolean grammars with strong negation." Non-Classical Models of Automata and Applications, 2024, Göttingen, Germany, 2024.

Robust Pole Identification and Model Reduction for SISO-LTI Systems: A Frequency Domain Approach

Mátyás Márton Szabari
Systems and Control Laboratory
HUN-REN Institute for Computer Science and Control (SZTAKI)
Budapest, Hungary
szabari.matyas.marton@sztaki.hun-ren.hu
Department of Numerical Analysis, Faculty of Informatics
Eötvös Loránd University (ELTE)
Budapest, Hungary
wnmm2n@inf.elte.hu

We present a novel algorithm for identifying the poles of transfer functions that characterize discrete time, BIBO (Bounded-Input Bounded-Output) stable, SISO (Single-Input Single-Output) LTI (Linear Time-Invariant) systems. Our method operates in the frequency domain and is particularly effective with band-limited measurements, demonstrating the ability to extrapolate to other frequencies accurately, even in the presence of high noise levels.

For a sequence of parameters $\mathbf{a} = (a_0, a_1, \dots)$ (called the inverse poles) in the complex unit disk $\mathbb{D}$ the n-th Malmquist-Takenaka function is defined as

$$\Phi_n^{\mathbf{a}}(z) = \frac{\sqrt{1-|a_n|^2}}{1-\overline{a_n}z} \prod_{k=0}^{n-1} \frac{z-a_n}{1-\overline{a_k}z} \quad (z \in \mathbb{C} \setminus \{1/\overline{a_j}\}_{j\in\mathbb{N}})$$

[1]. The Malmquist-Takenaka functions form an orthonormal system in the Hardy-Hilbert space $H^2(\mathbb{D})$, and they are also complete with some mild conditions on the inverse poles (e.g. $|a_n| \leq \delta < 1$) [2]. Moreover, for any strictly proper rational function H in $H^2(\mathbb{D})$ (i.e. the transfer function of a realizable finite dimensional stable SISO LTI system) lies in the subspace generated by $\Phi_n^{\mathbf{a}}$ ($n \leq N = \deg H$), where a_n are the inverse poles of H. Thus, we can have a system representation by the equation

$$H = \sum_{n=0}^{N} c_n \Phi_n^{\mathbf{a}} = \sum_{n=0}^{N} \langle H, \Phi_n^{\mathbf{a}} \rangle \Phi_n^{\mathbf{a}}. \tag{1}$$

The algorithm involves solving (1), which is separable nonlinear optimization problem [3] involving c_n as linear and a_n as nonlinear parameters. For system identification, (1) can be reformulated as finding the minimum of the error function:

$$\min_{\mathbf{a}\in\mathbb{D}^N} r_2(\mathbf{a}) = \min_{\mathbf{a}\in\mathbb{D}^N} \|\mathbf{P}_{\mathbf{a}} - \hat{H}\|_2,$$

where $\mathbf{P}_{\mathbf{a}}$ is the projection operator to the subspace generated by the functions $\Phi_{\mathbf{a}_n}$ and $\hat{H}$ is the transfer function estimation coming from either frequency or time domain measurements. The optimization aims to minimize the least squared error of the parameterized model, which can be interpreted as finding the best subspace to which the orthogonal projection of $\hat{H}$ most accurately describes the measured system. The corresponding nonlinear optimization can be solved using gradient-based approaches by the so called Variable Projection method [4], where the gradient of the error functional of with respect to $\mathbf{a}$ is given by

$$\frac{1}{2}\nabla r_2 = \left[-(\mathbf{P}_{\mathbf{a}} D_{\mathbf{a}} \Phi_{\mathbf{a}}^{+} + (\mathbf{P}_{\Phi} D_{\mathbf{a}} \Phi^{+})^T)\hat{H})\right]^T \mathbf{P}_{\Phi}^{\perp} \hat{H}.$$

This method exhibits high noise tolerance and outperforms other popular methods such as the Output Error. We have rigorously tested the algorithm with simulated systems and real-life measurements on a permanent magnet DC motor, for which the frequency response and its poles have been accurately identified. The results indicate that the proposed method performs well even in scenarios where the transfer function has a high number of poles.

In addition to the aforementioned algorithm, we have also developed and tested a variant specifically designed for model reduction [2]. The goal is to simplify the system representation to a lower order system, while remaining close to the original system in H^2 norm. This variant maintains the core principles of the original algorithm but introduces modifications to incorporate the explicit knowledge about the higher order system. It has been subjected to similar rigorous testing as the original algorithm and demonstrated its effectiveness in accurately approximating the original system's dynamics. This makes the proposed algorithm suitable for applications where computational efficiency is important, without compromising the accuracy of system representation. In conclusion, our work presents not only a new method for pole identification of discrete time SISO LTI systems but also a tool for model reduction, offering a robust solution for both scenarios.

***Index Terms*—System Identification, SISO LTI, Model Reduction, Malmquist-Takenaka, Variable Projection**

Supported by the DKÖP-23 Doctoral Excellence Program and the ÚNKP-23-3 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

REFERENCES

[1] S. Takenaka, "On the orthogonal functions and a new formula of interpolations," Japanese Journal of Mathematics, vol. 2, pp. 129–145, 1925.

[2] P. S. C. Heuberger and P. M. J. Van den Hof and B. Wahlberg, "Modelling and Identification with Rational Orthogonal Basis Functions," Springer-Verlag London, 2005.

[3] Golub, G.H. and Pereyra, "V. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate," SIAM Journal on Numerical Analysis (SINUM), vol. 10, pp. 413–432, 1973.

[4] Golub, G.H. and Pereyra, V. "Separable nonlinear least squares: The variable projection method and its applications," Inverse problems, 19(2), R1–R26. 2003.