Annals of Computer Science and Information Systems
Volume 16

Position Papers of the 2018 Federated Conference on Computer Science and Information Systems

September 9–12, 2018. Poznań, Poland

Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki (eds.)

PTI

Annals of Computer Science and Information Systems, Volume 16

Series editors:

Maria Ganzha (Editor-in-Chief),
Systems Research Institute Polish Academy of Sciences and Warsaw University of Technology, Poland

Leszek Maciaszek,
Wrocław Universty of Economy, Poland and Macquarie University, Australia

Marcin Paprzycki,
Systems Research Institute Polish Academy of Sciences and Management Academy, Poland

Senior Editorial Board:

Wil van der Aalst,
Department of Mathematics & Computer Science, Technische Universiteit Eindhoven (TU/e), Eindhoven, Netherlands

Marco Aiello,
Faculty of Mathematics and Natural Sciences, Distributed Systems, University of Groningen, Groningen, Netherlands

Mohammed Atiquzzaman,
School of Computer Science, University of Oklahoma, Norman, USA

Barrett Bryant,
Department of Computer Science and Engineering, University of North Texas, Denton, USA

Ana Fred,
Department of Electrical and Computer Engineering, Instituto Superior Técnico (IST—Technical University of Lisbon), Lisbon, Portugal

Janusz Górski,
Department of Software Engineering, Gdańsk University of Technology, Gdańsk, Poland

Giancarlo Guizzardi,
Free University of Bolzano-Bozen, Italy, Senior Member of the Ontology and Conceptual Modeling Research Group (NEMO), Brazil

Mike Hinchey,
Lero—the Irish Software Engineering Research Centre, University of Limerick, Ireland

Janusz Kacprzyk,
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Irwin King,
The Chinese University of Hong Kong, Hong Kong

Juliusz L. Kulikowski,
Nałęcz Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland

Michael Luck,
Department of Informatics, King's College London, London, United Kingdom

Jan Madey,
Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw, Poland

Stan Matwin,
Dalhousie University, University of Ottawa, Canada and Institute of Computer Science, Polish Academy of Science, Poland

Marjan Mernik,
University of Maribor, Slovenia

Michael Segal,
Ben-Gurion University of the Negev, Israel

Andrzej Skowron,
Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw, Poland
John F. Sowa,
VivoMind Research, LLC, USA

Editorial Associates:
Katarzyna Wasielewska,
Systems Research Institute Polish Academy of Sciences, Poland
Paweł Sitek,
Kielce University of Technology, Kielce, Poland

TEXnical editor: Aleksander Denisiuk,
University of Warmia and Mazury in Olsztyn, Poland

Position Papers of the 2018 Federated Conference on Computer Science and Information Systems

Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki (eds.)

POLSKIE TOWARZYSTWO INFORMATYCZNE
POLISH INFORMATION PROCESSING SOCIETY

2018, Warszawa, Polskie Towarzystwo Informatyczne

Annals of Computer Science and Information Systems, Volume 16
Position Papers of the 2018 Federated Conference on Computer Science and Information Systems

USB: ISBN 978-83-949419-9-4
WEB: ISBN 978-83-949419-8-7

ISSN 2300-5963
DOI 10.15439/978-83-949419-8-7

© 2018, Polskie Towarzystwo Informatyczne
Ul. Solec 38/103
00-394 Warsaw
Poland

Contact: secretariat@fedcsis.org
http://annals-csis.org/

Cover photo:
Michał Skroboszewski,
Elbląg, Poland

Also in this series:

Volume 17: Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-952357-0-2, ISBN USB: 978-83-952357-1-9**
Volume 15: Proceedings Papers of the 2018 Federated Conference on Computer Science and Information Systems, **ISBN Web 978-83-949419-5-6, ISBN USB 978-83-949419-6-3, ISBN ART 978-83-949419-7-0**
Volume 14: Proceedings of the First International Conference on Information Technology and Knowledge Management, **ISBN WEB: 978-83-949419-2-5, ISBN USB: 978-83-949419-1-8, ISBN ART: 978-83-949419-0-1**
Volume 13: Communication Papers of the 2017 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-922646-2-0, ISBN USB: 978-83-922646-3-7**
Volume 12: Position Papers of the 2017 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-922646-0-6, ISBN USB: 978-83-922646-1-3**
Volume 11: Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-946253-7-5, ISBN USB: 978-83-946253-8-2, ISBN ART: 978-83-946253-9-9**
Volume 10: Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering, **ISBN WEB: 978-83-65750-05-1, ISBN USB: 978-83-65750-06-8**
Volume 9: Position Papers of the 2016 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-93-4, ISBN USB: 978-83-60810-94-1**
Volume 8: Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-90-3, ISBN USB: 978-83-60810-91-0, ISBN ART: 978-83-60910-92-7**
Volume 7: Proceedings of the LQMR Workshop, **ISBN WEB: 978-83-60810-78-1, ISBN USB: 978-83-60810-79-8**
Volume 6: Position Papers of the 2015 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-76-7, ISBN USB: 978-83-60810-77-4**

DEAR Reader, it is our pleasure to present to you Position Papers of the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), which took place in Poznań, Poland, on September 9-12, 2018.

We consider position papers to be important enough to recognize them as a separate publication. As you will see, they comprise two categories of contributions – ***challenge papers*** and ***emerging research papers***. *Challenge papers* propose and describe research challenges in theory or practice of computer science and information systems. Papers in this category are based on deep understanding of existing research or industrial problems. Based on such understanding and experience, they define new exciting research directions and show why these directions are crucial to the society at large. *Emerging research papers* present preliminary research results from work-in-progress based on sound scientific approach but presenting work not completely validated as yet. They describe precisely the research problem and its rationale. They also define precisely the intended future work including the expected benefits from solution to the tackled problem. Subsequently, they may be more conceptual than experimental.

FedCSIS 2018 was Chaired by prof. Krzysztof Jassem, while dr. Paweł Skórzewski acted as the Chair of the Organizing Committee. This year, FedCSIS was organized by the Polish Information Processing Society (Mazovia Chapter), IEEE Poland Section Computer Society Chapter, Systems Research Institute Polish Academy of Sciences, Warsaw University of Technology, Wrocław University of Economics, and Adam Mickiewicz University.

FedCSIS 2018 was technically co-sponsored by: IEEE Region 8, IEEE Poland Section, IEEE Computer Society Technical Committee on Intelligent Informatics, IEEE Czechoslovakia Section Computer Society Chapter, IEEE Poland Section Gdańsk Computer Society Chapter, SMC Technical Committee on Computational Collective Intelligence, IEEE Poland Section Systems, Man, and Cybernetics Society Chapter, IEEE Poland Section Control System Society Chapter, IEEE Poland Section Computational Intelligence Society Chapter, ACM Special Interest Group on Applied Computing, International Federation for Information Processing, Committee of Computer Science of the Polish Academy of Sciences, Polish Operational and Systems Research Society, Mazovia Cluster ICT Poland and Eastern Cluster ICT Poland. FedCSIS 2018 was sponsored by Intel, Gambit, Samsung, Silver Bullet Labs, eSensei and Data Center PPNT.

During FedCSIS 2018, keynote lectures have been delivered by:

- Aksit, Mehmet, University of Twente, "*The Role of Computer Science and Software Technology in Organizing Universities for Industry 4.0 and Beyond*"
- Bosch, Jan, Chalmers University Technology, "*Towards a Digital Business Operating System*"
- Duch, Włodzisław, Nicolaus Copernicus University, "*Neurocognitive informatics for understanding brain functions*"
- O'Connor, Rory, V., Dublin City University, "*Demystifying the World of ICT Standardisation: An Insiders Viewpoint*"

FedCSIS 2018 consisted of the following events (conferences, symposia, workshops, special sessions). These events were grouped into FedCSIS conference areas**,** of various degree of integration. Specifically, those listed without indication of the year 2018 signify "abstract areas" with no direct paper submissions to them (but with submissions to their enclosed events).

- **AAIA'18 – 13th International Symposium Advances in Artificial Intelligence and Applications**
  - AIMaViG'18 – 3rd International Workshop on Artificial Intelligence in Machine Vision and Graphics
  - AIMA'18 – 8th International Workshop on Artificial Intelligence in Medical Applications
  - AIRIM'18 – 3rd International Workshop on AI aspects of Reasoning, Information, and Memory
  - ASIR'18 – 8th International Workshop on Advances in Semantic Information Retrieval
  - DMGATE'18 – 1st International Workshop on AI Methods in Data Mining Challenges
  - SEN-MAS'18 – 6th International Workshop on Smart Energy Networks & Multi-Agent Systems
  - WCO'18 – 11th International Workshop on Computational Optimization
- *CSS – Computer Science & Systems*
  - BEDA'18 – 1st International Workshop on Biomedical & Health Engineering and Data Analysis
  - CANA'18 – 11th Workshop on Computer Aspects of Numerical Algorithms
  - C&SS'18 – 5th International Conference on Cryptography and Security Systems
  - CPORA'18 – 3rd Workshop on Constraint Programming and Operation Research Applications
  - LTA'18 – 3rd International Workshop on Language Technologies and Applications
  - MMAP'18 – 11th International Symposium on Multimedia Applications and Processing
- *iNetSApp – International Conference on Innovative Network Systems and Applications*
  - INSERT'18 – 2nd International Conference on Security, Privacy, and Trust
  - IoT-ECAW'18 – 2nd Workshop on Internet of Things - Enablers, Challenges and Applications
- *IT4MBS – Information Technology for Management, Business & Society*
  - AITM'18 – 15th Conference on Advanced Information Technologies for Management
  - ISM'18 – 13th Conference on Information Systems Management
  - KAM'18 – 24th onference on Knowledge Acquisition and Management
- *SSD&A – Software Systems Development & Applications*
  - MDASD'18 – 5th Workshop on Model Driven Approaches in System Development
  - MIDI'18– 6th Conference on Multimedia, Interaction, Design and Innovation
  - LASD'18 – 2nd International Conference on Lean and Agile Software Development
  - SEW-38 & IWCPS-5 – Joint 38th IEEE Software Engineering Workshop (SEW-38) and 5th International Workshop on Cyber-Physical Systems (IWCPS-5)
- **DS-RAIT'18 – 5th Doctoral Symposium on Recent Advances in Information Technology**

Each paper, found in this volume, was refereed by at least two referees.

The program of FedCSIS required a dedicated effort of many people. Each event constituting FedCSIS had its own

Organizing and Program Committee. We would like to express our warmest gratitude to all Committee members for their hard work in attracting and later refereeing 394 regular submissions.

We thank the authors of papers for their great contribution to research and practice in computing and information systems. We thank the invited speakers for sharing their knowledge and wisdom with the participants. Finally, we thank all those responsible for staging the conference in Poznań. Organizing a conference of this scope and level could only be achieved by the collaborative effort of a highly capable team taking charge of such matters as conference registration system, finances, the venue, social events, catering, handling all sorts of individual requests from the authors, preparing the conference rooms, etc.

We hope you had an inspiring conference and an unforgettable stay in the beautiful city of Poznań. We also hope to meet you again for FedCSIS 2019 in Leipzig, Germany.

Co-Chairs of the FedCSIS Conference Series

Maria Ganzha, *Warsaw University of Technology, Poland and Systems Research Institute Polish Academy of Sciences, Warsaw, Poland*

Leszek Maciaszek, *Wrocław University of Economics, Wrocław, Poland and Macquarie University, Sydney, Australia*

Marcin Paprzycki, *Systems Research Institute Polish Academy of Sciences, Warsaw Poland and Management Academy, Warsaw, Poland*

Annals of Computer Science and Information Systems, Volume 16

Position Papers of the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS)

September 9–12, 2018. Poznań, Poland

TABLE OF CONTENTS

13TH INTERNATIONAL SYMPOSIUM ADVANCES IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS

Call For Papers 1

Estimation of Student Understandings from Pulse Wave Changes Caused by Load in Preparatory 3
Atsushi Hagihara, Hiromitsu Shimakawa, Yusuke Kajiwara

Robotic Process Automation of Unstructured Data with Machine Learning 9
Anna Wróblewska, Tomasz Stanisławek, Bartłomiej Prus-Zajączkowski, Łukasz Garncarek

3RD INTERNATIONAL WORKSHOP ON ARTIFICIAL INTELLIGENCE IN MACHINE VISION AND GRAPHICS

Call For Papers 17

Dataset Enhancement in Hair Follicle Detection: ESENSEI Challenge 19
Jan Jakubik

An ensemble of Deep Convolutional Neural Networks for Marking Hair Follicles on Microscopic Images 23
Łukasz Podlodowski, Szymon Roziewski, Marek Nurzyński

COMPUTER SCIENCE & SYSTEMS

Call For Papers 29

3RD INTERNATIONAL WORKSHOP ON LANGUAGE TECHNOLOGIES AND APPLICATIONS

Call For Papers 31

From Building Corpora for Recognizing Faceted Entailment to Recognizing Relational Entailment 33
Martin Víta

INTERNATIONAL CONFERENCE ON INNOVATIVE NETWORK SYSTEMS AND APPLICATIONS

Call For Papers 39

2ND INTERNATIONAL CONFERENCE ON SECURITY, PRIVACY, AND TRUST

Call For Papers **41**

Proposal for simplified implementation of risk assessment method for measuring instruments **43**
Federico Grasso Toro, Martin Koval, Marko Esche

2ND WORKSHOP ON INTERNET OF THINGS - ENABLERS, CHALLENGES AND APPLICATIONS

Call For Papers **49**

Towards a Supportive City with Smart Urban Objects in the Internet of Things: The Case of Adaptive Park Bench and Adaptive Light **51**
Marvin Hubl, Philipp Skowron, Michael Aleithe

INFORMATION TECHNOLOGY FOR MANAGEMENT, BUSINESS & SOCIETY

Call For Papers **59**

24TH CONFERENCE ON KNOWLEDGE ACQUISITION AND MANAGEMENT

Call For Papers **61**

The Potential of the Internet of Things in Knowledge Management System **63**
Artur Rot, Małgorzata Sobińska

SOFTWARE SYSTEMS DEVELOPMENT & APPLICATIONS

Call For Papers **69**

5TH WORKSHOP ON MODEL DRIVEN APPROACHES IN SYSTEM DEVELOPMENT

Call For Papers **71**

A Model-Driven Approach to Microservice Software Architecture Establishment **73**
Branko Terzic, Vladimir Dimitrieski, Slavica Kordić, Ivan Luković

6TH CONFERENCE ON MULTIMEDIA, INTERACTION, DESIGN AND INNOVATION

Call For Papers **81**

Reading is Vital, but will it be Invisible? Screens vs. Paper on Our Way to Naturalized Technology of Reading **83**
Jarosław Kowalski, Piotr Toczyski, Cezary Biele, Aldona Zdrodowska

JOINT 38TH IEEE SOFTWARE ENGINEERING WORKSHOP (SEW-38) AND 5TH INTERNATIONAL WORKSHOP ON CYBER-PHYSICAL SYSTEMS (IWCPS-5)

Call For Papers **91**

Simulation Driven Development of Distributed Systems – Coupling of virtual and real system components **93**
Bernd Pfitzinger, Tommy Baumann

Challenges in Causal Inference from Personal Monitoring Devices **99**
Tomasz Wiktorski

5TH DOCTORAL SYMPOSIUM ON RECENT ADVANCES IN INFORMATION TECHNOLOGY

Call For Papers **103**

Static typing and dependency management for SOA **105**
Nikita Gerasimov

Acoustic Model Training, using Kaldi, for Automatic Whispery Speech Recognition **109**
Piotr Kozierski, Talar Sadalla, Szymon Drgas, Adam Dąbrowski, Joanna Ziętkiewicz, Wojciech Giernacki

Author Index **115**

13th International Symposium Advances in Artificial Intelligence and Applications

AAIA'18 brings together scientists and practitioners to discuss their latest results and ideas in all areas of Artificial Intelligence. We hope that successful applications presented at AAIA'18 will be of interest to researchers who want to know about both theoretical advances and latest applied developments in AI.

Topics

Papers related to theories, methodologies, and applications in science and technology in the field of AI are especially solicited. Topics covering industrial applications and academic research are included, but not limited to:

- Decision Support
- Machine Learning
- Fuzzy Sets and Soft Computing
- Rough Sets and Approximate Reasoning
- Data Mining and Knowledge Discovery
- Data Modeling and Feature Engineering
- Data Integration and Information Fusion
- Hybrid and Hierarchical Intelligent Systems
- Neural Networks and Deep Learning
- Bayesian Networks and Bayesian Reasoning
- Case-based Reasoning and Similarity
- Web Mining and Social Networks
- Business Intelligence and Online Analytics
- Robotics and Cyber-Physical Systems
- AI-centered Systems and Large-Scale Applications

We also encourage researchers interested in the following topics to submit papers directly to the corresponding workshops, which are integral parts of AAIA'18:

- AI in Medical Applications (AIMA'18 workshop)
- AI in Machine Vision and Graphics (AIMaVIG'18 workshop)
- AI in Reasoning Foundations (AIRIM'18 workshop)
- AI in Information Retrieval (ASIR'18 workshop)
- AI in Data Mining Challenges (DMGATE'18 workshop)
- AI in Smart Energy Networks (SEN-MAS'18 workshop)
- AI in Computational Optimization (WCO'18 workshop)

All submissions accepted to the main track of AAIA'18 and to the above workshops are treated equally in the conference programme and are equally considered for the paper awards.

Professor Zdzisław Pawlak Best Paper Awards

We are proud to continue the tradition started at the AAIA'06 and grant two "Professor Zdzisław Pawlak Best Paper Awards" for contributions which are outstanding in their scientific quality. The two award categories are:

- Best Student Paper. Papers qualifying for this award must be marked as "Student full paper" to be eligible.
- Best Paper Award.

Each award carries a prize of 300 EUR funded by the Mazowsze Chapter of the Polish Information Processing Society.

Event Chairs

- **Kwaśnicka, Halina,** Wrocław University of Science and Technology, Poland
- **Markowska-Kaczmar, Urszula,** Wrocław University of Science and Technology, Poland

Advisory Board

- **Kacprzyk, Janusz,** Polish Academy of Sciences, Poland
- **Marek, Victor,** University of Kentucky, United States
- **Matwin, Stan,** Dalhousie University, Canada
- **Michalewicz, Zbigniew,** University of Adelaide, Australia
- **Skowron, Andrzej,** University of Warsaw, Poland
- **Ślęzak, Dominik,** University of Warsaw, Poland

Area Supervisory Committee

- **Derksen, Christian,** SEN-MAS'18
- **Janusz, Andrzej,** DMGATE'18
- **Lasek, Piotr,** AIMA'18
- **Loukanova, Roussanka,** AIRIM'18
- **Markowska-Kaczmar, Urszula,** AAIA'18
- **Mozgovoy, Maxim,** ASIR'18
- **Śluzek, Andrzej,** AIMaViG'18
- **Zaharie, Daniela,** WCO'18

Program Committee

- **Baron, Grzegorz**
- **Bartkowiak, Anna,** Wrocław University, Poland
- **Bazan, Jan,** University of Rzeszów, Poland
- **Bembenik, Robert**
- **Betliński, Paweł,** Security On Demand, Poland
- **Błaszczyński, Jerzy,** Poznań University of Technology, Poland
- **Chakraverty, Shampa,** Netaji Subhas Institute of Technology, India
- **do Carmo Nicoletti, Maria,** UFSCar & FACCAMP, Brazil
- **Franova, Marta,** CNRS, LRI & INRIA, France
- **Froelich, Wojciech,** University of Silesia, Poland
- **Gawrysiak, Piotr**
- **Girardi, Rosario,** UNIRIO, Brazil

- **Jaromczyk, Jerzy,** University of Kentucky, United States
- **Jatowt, Adam,** Kyoto University, Japan
- **Jin, Xiaolong,** Institute of Computing Technology, Chinese Academy of Sciences, China
- **Kasprzak, Włodzimierz,** Warsaw University of Technology, Poland
- **Korbicz, Józef,** University of Zielona Góra, Poland
- **Kryszkiewicz, Marzena,** Warsaw University of Technology, Poland
- **Kulikowski, Juliusz,** Institute of Biocybernetics and Biomedical Engineering, Poland
- **Lopes, Lucelene,** PUCRS, Brazil
- **Matson, Eric T.,** Purdue University, United States
- **Menasalvas, Ernestina,** Universidad Politécnica de Madrid, Spain
- **Miyamoto, Sadaaki,** University of Tsukuba, Japan
- **Moshkov, Mikhail,** King Abdullah University of Science and Technology, Saudi Arabia
- **Myszkowski, Paweł B.,** Wroclaw University of Technology, Poland
- **Nowostawski, Mariusz,** Norwegian University of Technology and Science (NTNU), Norway
- **Ohsawa, Yukio,** University of Tokyo, Japan
- **Peters, Georg,** Munich University of Applied Sciences, Germany
- **Po, Laura,** Università di Modena e Reggio Emilia, Italy
- **Porta, Marco,** University of Pavia, Italy
- **Przybyła-Kasperek, Małgorzata,** University of Silesia, Poland
- **Raś, Zbigniew,** University of North Carolina at Charlotte, United States
- **Rauch, Jan,** University of Economics, Prague, Czech Republic
- **Reformat, Marek,** University of Alberta, Canada
- **Schaefer, Gerald,** Loughborough University, United Kingdom
- **Sikora, Marek,** Silesian University of Technology, Poland
- **Sikos, Leslie F.,** University of South Australia, Australia
- **Skonieczny, Lukasz**
- **Śluzek, Andrzej,** Khalifa University, United Arab Emirates
- **Stańczyk, Urszula,** Silesian University of Technology, Poland
- **Subbotin, Sergey,** Zaporizhzhya National Technical University, Ukraine
- **Sydow, Marcin,** Polish Academy of Sciences & Polish-Japanese Academy of Information Technology, Poland
- **Szczęch, Izabela,** Poznań University of Technology, Poland
- **Szczuka, Marcin,** University of Warsaw, Poland
- **Szpakowicz, Stan,** University of Ottawa, Canada
- **Szwed, Piotr,** AGH University of Science and Technology, Poland
- **Tomczyk, Arkadiusz,** Łódź University of Technology, Poland
- **Unland, Rainer,** Universität Duisburg-Essen, Germany
- **Unold, Olgierd,** Wrocław University of Technology, Poland
- **Zakrzewska, Danuta,** Łódź University of Technology, Poland
- **Zielosko, Beata,** University of Silesia, Poland
- **Ziółko, Bartosz,** AGH University of Science and Technology, Poland

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 3–8
DOI: 10.15439/2018F124
ISSN 2300-5963 ACSIS, Vol. 16

Estimation of Student Understandings from Pulse Wave Changes Caused by Load in Preparatory Course

Atsushi Hagihara
Ritsumeikan University
Graduate School of Information Science and Engineering
Kusatsu-city, 525-8577 Shiga, Japan
Email: is0273rs@ed.ritsumei.ac.jp

Hiromitsu Shimakawa
Ritsumeikan University
College of Information Science and Engineering
Email: shimakawa@cs.ritsumei.ac.jp

Yusuke Kajiwara
Komatsu University
Faculty of Production Systems Engineering and Sciences
Email: yusuke.kajiwara@komatsu-u.ac.jp

***Abstract*—We propose a method to estimate whether student hold knowledge on the contents and tasks of the class, using cognitive load and mental load. We estimate the cognitive load of the student using a pulse wave and the pulse change.The mental load can be estimated from the activity of the autonomic nerve. In order to clarify student knowledge on course contents, we estimate student schemata from the load on the student.We conducted an experiment, preparing 2 kinds of tasks; S1 task can be answered with student schemata, while S2 task cannot be answered with student schema. Through the experiment using the proposed method, we classified tasks given to students into S1 tasks and S2 tasks.It was possible to estimate the student schemata from the standard deviation of the heart rate.This method makes it possible to figure out student understanding earlier than traditional knowledge measurement method.**

I. Introduction

FLIPPED CLASSES are introduced to actual educational classes, as open teaching materials, e-learning, and MOOCs(Massive Open Online Courses) spread[1][2]. In Flipped classes, learners prepare before lessons, based on open teaching materials and assignments. Meanwhile, in the classes, the teachers check the learners' knowledge. If they find contents the students fail to understand, they explain the contents.

In face-to-face classes, the learners study at designated places such as classrooms and at fixed times. The learners review the lesson contents after the lesson. They engage in homework to develop their learning. However, in face-to-face classes, teachers put their main points in giving knowledge to learners. Learners also concentrate on acquiring knowledge. As a result, there are concerns that initiative learning of students would decline, because they cannot enhance interests in acquiring new knowledge and problem solving skills.

In the flipped classes, there is no constrains on place or time for learners to prepare the content of the class. They can learn at any place and any time suitable for each learner. The learners can determine contents to study and the learning time for themselves. In addition to that, since the learners study the prepared knowledge in classes, they can increase opportunities to use the knowledge. Through these aspects, flipped classes are expected to have good effects to increase the learning time and motivation of learners, to promote initiative learning. Some colleges in U.S. have introduced flipped classes, showing the effect of improving learning motivation and raising the completion rate of students [1][2].

However, there are some defects for flipped classes. For example, in flipped classes, it is not secure the learning time[3]. Since teachers cannot grasp the comprehension degree of individual learners, they cannot conduct classes suitable for each of learners. If assignments or class content given to students at the time of preparation is too easy, learners would acquire less knowledge, which makes the learning efficiency low. Therefore, it is necessary for teachers to grasp the knowledge acquired by learners at the time of preparation and the knowledge they have already acquired, in order to make the learning more efficient and the progress of the flipped classes smooth. The paper proposes a method to estimate whether or not learners hold knowledge about lesson contents. We focus on the cognitive load in learner during preparation and the mental load due to emotional changes such as impatience and tension. Existing studies have revealed that when cognitive load or emotion changes, the sympathetic and parasympathetic functions change, which causes the pulse wave to fluctuate. Based on the results, this method uses a pulse wave sensor to measure the cognitive load and the mental load of learners. We extract learners' features representing cognitive load and mental load from various data elements of pulse waves and pulse fluctuations. In the experiment using this method, we have found that the standard deviation of the heart rate of the learner at the time of problem solving is effective to discriminate whether the learner holds knowledge on the given task.

In this paper, we introduce related research in ChapterII. ChapterIII explains the method proposed in this research. ChapterIV gives experiments and evaluations. ChapterV discusses the experimental results. ChapterVI summarizes this research.

©2018, PTI

II. Existing Research

Learner's stumbling and troubles occur at the time of preparation in the flipped classes, because of lack of knowledge of learners on tasks and contents in classes. At this time, the learner is overloaded. The overload makes the leaners avoid their learning. On the contrary, learning effect is low when they are forced to answer the task they have already achieved. This task has low cognitive load for learners. More than one researches have been conducted to measure the load on learners at the time of learning, aiming at providing them with appropriate load according to the ability of the learners and reducing the burden on learners[5].

A. Existing research on cognitive load

The cognitive load theory[4] points out there are working memory and long term memory (LTM) in the human memory structure for learning. The working memory is a memory which has limited capacity and holds information temporarily. When humans perceive experiences and knowledge, burdens are imposed on the working memory. Schema is stored in LTM. A schema is a onein which human experiences and knowledge are accumulated and organized.

There are three types of cognitive loads which occur during learning[6]. The intrinsic cognitive load is the one caused by task itself such as difficulty level and complexity in task solving. It is defined by the number of factors considered in learning. Although it is not related to the task itself, the extraneous cognitive load occurs in recognition of design and format of the teaching material used for learning. The germane load occurs when learners are establishing schemas to fix the knowledge in problem solving and memorization. Since these three kinds of cognitive loads are additive relationships, they never appear in a completely separate way. Many studies try to measure these three types of cognitive loads individually, operating experiment environments[8].

B. Measurement of cognitive load

The cognitive load is measured using subjective evaluation, physiological index, performance index such as the learner ' s exam results. Mizuno et al[9]. measured cognitive load using reaction time in problem-solving learning as a performance index. In order to measure the learner's load on line at an early stage during learning, it is effective to acquire a physiological index using a sensor. Examples of physiological indices used for measuring cognitive function include skin conductance response, pupillary reaction, and heart rate[11]. Tsunoda et al[13] measured cognitive load at mental load work of brain workers, using heart rate variability as a physiological index.

Most of studies measure the cognitive load based on whether or not a load is applied. There are few studies which measure the cognitive load, focusing on schema construction. We cannot know whether learners hold knowledge on learning contents from the cognitive load of learners at their learning time.

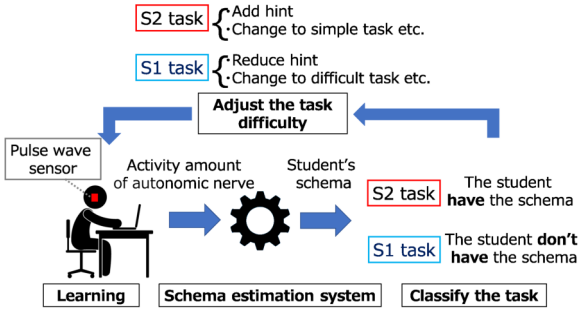

Fig. 1. The outline of the schema estimation system

III. Estimation of Learner's Schema Using Load on Learner

A. Improve flipped classes by estimating learner's schema

From ChapterII, There are many researches to estimate cognitive load from pulse fluctuation, and it can be estimated with high accuracy. However, few researches focus on the relationship between pulse fluctuation and the schema held by the learner. Therefore, in this research, we propose a system which classifies whether or not the learner holds the schema for the task from the pulse wave and the pulse fluctuation at the time of answer the task by the learner. Figure1 shows the outline of the system (schema estimation system) that estimates the schema held by the learner using this method.

The learner's learning content is improved by identifying from learner's pulse wave and pulse fluctuation whether it is a task that can be answered by the learner's schema. The procedure is shown below.

1) By using the pulse wave sensor, acquire pulse wave and pulse fluctuation during answer task of learner.
2) Analyze pulse wave fluctuation and pulse fluctuation of the acquired by the learner, and calculate 18 variations of learners' autonomic nerves.
3) From the calculated 18 variations of the learner ' s autonomic nervousness, it is identified whether the task given to the learner is a task or not that can be answered by the learner's schema.
4) When it becomes possible to grasp the schema held by the learner from heart rate variability, it is possible to identify whether learners are able to organize knowledge on the task. And it can be done earlier and in real time than traditional knowledge measurement methods such as hearing and testing.

Through these procedures, depending on the schema held by the learner, it is possible to support such as giving hints or changing the difficulty level of the task. And it is possible to improve the learning efficiency at the time of preparation in the flipped class.

B. Relationship between learner's schema and cognitive processing

In this research, we focus on differences in cognitive processing at the time of solution of tasks that can be answered

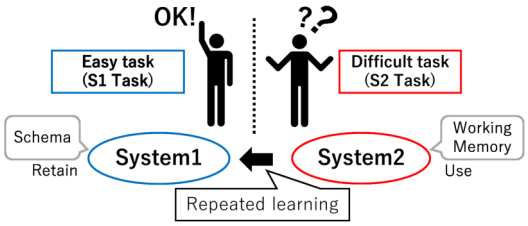

Fig. 2. Cognitive processing when the learner holds the schema and does not hold it

by the learner's schema and tasks that can not be answered. Figure2 shows how cognitive processing is performed when learners answer tasks. There are two ways of human cognitive processing, System 1 and System 2 proposed by Daniel Kahneman et al[14].

System 1: This Cognitive processing that can automatically understand without much perceptual activity. Because the experiences and the knowledge about the task are already organized and structured as a schema and it is kept in long-term memory. System 2: This Cognitive processing that consciously understands the objects. Because, it is unfamiliar to experiences and knowledge concerning the object to be recognized, construction of the schema is not done, and perceptual activity is necessary for understanding.And, by repeatedly performing cognitive processing of System 2, a schema related to the experiences and knowledge is constructed, and cognitive processing of System 1 can be perceived. Replacing the ideas of System 1 and System 2 with learning, tasks can be divided into the following two types.

- The tasks that the learner can construct the schemata has been able to organize the knowledge on the tasks, and the tasks that do not require much perceptual activities to answer.
- The tasks that learner have not been able to construct schemata have not been able to organize knowledge about the tasks, and the tasks that require much perceptual activities to answer.

Therefore, in this research, the tasks that can be answered by the schema held by the learner are set as S1 tasks, and the tasks difficult to solve by the schema held by the learner are set as S2 tasks.

C. Estimation of load using learner's pulse wave and pulse fluctuation at task solving time

When solving the S1 task, the cognitive load on the learner is small, and the usage of the working memory is small. In solving the S2 task, the cognitive load on the learner is large, and the working memory is pressed. Also, when the learner solves the S2 problem, it is difficult to solve the S2 task, and emotional changes such as impatience and tension appear[10]. Therefore, by using pulse fluctuation that can estimate cognitive load and emotion, S1 task and S2 task are identified. It is known that the cognitive load on learners at task solving time can be estimated from pulse wave and pulse fluctuation of learners [11].

In addition, the load due to mental factors such as impatience and tension of the learner can be estimated from autonomic nervous fluctuation that can be calculated from the pulse wave or pulse fluctuation of the learner. As a method of acquiring the pulse rate and pulse fluctuation of the learner, there is a method of attaching a pulse wave sensor to the learner. By using a pulse wave sensor, it is possible to measure pulse wave and pulse fluctuation easily from the fingertip and earlobe, so there is little extra load due to attaching a pulse wave sensor to the learner. In this research, we estimate the load at the time of task solving by the learner from the pulse wave and pulse fluctuation acquired by attaching the pulse wave sensor to the learner. By analyzing the acquired pulse wave and pulse fluctuation, we calculate the following nine variables which are indices of the autonomic nerve.

From the heart rate variability, the following components can be obtained; Heart Rate, RR Interval (RRI), TP (total of VLF, LF, HF), VLF (Very Low Frequency), LF (Low Frequency), HF (High Frequency), LF / HF(Indicator of sympathetic function), HFnorm(Expression (1)), LFnorm(Expression (2)). HFnorm, LFnorm are calculated by the following expressions.

$$HFnorm = (\frac{HF}{HF + LF}) \times 100 \quad (1)$$

$$LFnorm = (\frac{LF}{HF + LF}) \times 100 \quad (2)$$

We calculate the mean and standard deviation of the indices of these nine autonomic nerves and estimate the load on the learner when solving the tasks from the total of 18 variables.

D. Classification of tasks by machine learning

In this study, we classify tasks as S1 tasks or S2 tasks for learners, based on the difference between cognitive load and mental load on learners at the time of solving tasks of S1 task and S2 task. The S1 task and the S2 task are classified by machine learning. We use the variation of the autonomic nerves of the 18 learners written in SectionIII-C. As an explanatory variable and two kinds of tasks given to the learner as objective variables. There are two types of S1 task and S2 task. In this system, the S1 task and the S2 task are presented to the learner beforehand. At that time, acquire the variation amount of the learner's autonomic nerve, and use it as teacher data. Random Forest (RF)[15] with ensemble learning is used for machine learning algorithm. RF has a learning phase and an classification phase. In the learning phase, the RF constructs and learns an ensemble of decision trees using the obtained teacher data as training data. Thereafter, in the classification phase, each decision tree of RF classifies whether the task given to the learner from the load of the learner at the time of problem solving is the S1 task or the S2 task. And outputs the most frequent classification result.

E. Adjustment of tasks by estimating learner's schema

By using this method, it is possible to classify whether the task given at the preparation from the load at the time of task solving is the S1 task or the S2 task for the learner. If the task given to the learner is the S1 task, the learner can answer the task without much thinking. And the learner has little knowledge to acquire at the time of the task solving and poor learning efficiency. On the other hand, if the task given to the learner is the S2 task, the leaner needs perceptual activities for answers and it may be thought that the task is too difficult for the learner. Therefore, the learner can not answer the task, leading to problems such as learners' giving up and stumbling[18].

In order to improve the learning efficiency, it is required for the learner to have an appropriate difficulty level that the task is not too simple and not too difficult. Therefore, if the assignment given to the learner is the S1 task, the learner can construct the schema for the assignment and it is too easy, so it is necessary to adjust the task, such as reduction of hints and changes to applied task, in order to improve the learning efficiency. If the task given to the learner is the S2 task, the learner can not be able to construct the schema for that task and it may be too difficult. It is necessary to adjust the task, such as addition of hints and changes to more basic tasks, in order to set tasks suitable for learners[7]. Also, if the learner can correctly answer the S2 task, the learner repeatedly answers the same difficulty task, so that the learner can construct the schema and shift to the S1 task. By using this schema estimation system at the time of preparation in the flipped class, it is possible to adjust appropriate tasks and improve the learning efficiency of the learner.

IV. Purpose and Overview of Experiment

In this experiment, we examined whether the tasks presented to the learner can be identified as S1 task or S2 task, from the pulse rate and pulse fluctuation at the time of answer by the learner. The research participants were 12 university students aged 20 to 24 and answered six programming tasks as shown in Figure3. The programming task which was set by research participants is the selection problem of the filling format of programming code written in C language. And, we set questions of the following two patterns three questions at a time. And the task has the following two patterns. Each question was set three questions for each subject[16].

- S 1 Task(Tasks:1,2,3); Even if the learner does not understand the contents of the whole programming code, the question that can be solved if the leaner understands basic syntax
- S 2 Task (Tasks:4,5,6); The learners can not solve the question unless they understand the flow of the whole programming code or the functions in the code.

In order to avoid stress which is unrelated to the solution of the programming task, we did not set time limit of answer time. We allowed research participants to use the Web Browser and search when questions arise. To obtain the pulse

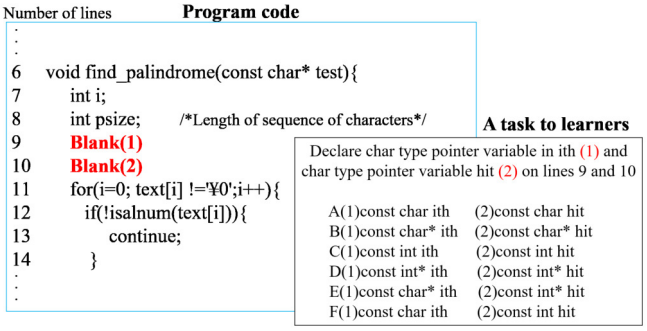

Fig. 3. Example of the programming tasks

rate and pulse fluctuation at the time of research participant programming solution, the research participant wears Vital Meter that a wireless earlobe pulse wave device manufactured by TAOS Laboratories. We get the following nine indicators from learners at the time of answering programming tasks. And calculate their mean and variance. The nine indicators are Heart Rate, RRI, LF, HF, LF / HF, VLF, TP, HFnorm, LFnorm. The sampling period in acquiring the pulse wave and pulse fluctuation was 1 kHz. Using these calculated values as explanatory variables, the task solved by the learner classifies either the S1 task or the S2 task based on the difference between the load amount when the learner solves the S1 task and the S2 task.

A. Confirmation of the validity of the tasks

In order to confirm the validity of the S1 task and the S2 task, the research participants were asked to answer the questionnaire. Questionnaire survey was conducted at the end of each task solution. We asked the questioner about the extent of perceptual activity such as calculation and memorization and answered in 6 steps of Linkert scale[17]. TableI shows the number of correct answers for each task. As a result of the questionnaire, the average of tasks 1, 2, and 3 were 3 points or less, and the average of 4, 5, and 6 were 4 points or more. Among the 12 research participants, the number of correct answers of tasks 1,2, and 3 were more than half, and the number of correct answers of tasks 4,5, and 6 were less than half. Based on the questionnaire result and the number of correct answers, it was confirmed that the tasks 1, 2, and 3 were S1 tasks because the learner needed less perceptual activity to the task solution and the correct answer rate were high. Also, tasks 4, 5, and 6 confirmed that the task were S2 because the learner needed perceptual activity to solve the problem and the correct answer rate were low.

B. Identification using Random Forest

In this experiment, using the analysis software attached to Vital Meter, the Heart Rate, RR Interval, LF, HF , LF / HF, VLF, TP, HFnorm, LFnorm, and a total of 18 variables of the mean and standard deviation of each of the nine variables. We used variance analysis to select effective variables from these 18 variables for classifying S1 task and S2 task. The

TABLE I
NUMBER OF CORRECT ANSWERS

Task Number	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
Number of correct answers	10	10	9	5	5	5

TABLE II
RESULTS OF VARIANCE ANALYSIS OF EACH 18 VARIABLES FOR CLASSIFYING S1 TASK AND S2 TASK

Explanatory variable		Heat Rate	RRI	HF	LF	LF/HF	VLF	TP	HF_norm	LF_norm
P value	Mean	0.6763395	0.7962887	0.3208657	0.2567065	0.9865234	0.5166716	0.3151521	0.0453415	0.3158083
	Standard deviation	0.0161003	0.2538482	0.1276715	0.443223	0.5046133	0.2063358	0.3018221	0.2666182	0.2666182

TABLE III
RESULT OF RANDOM FOREST USING STANDARD DEVIATION OF HEART RATE AND MEAN OF HFNORM

	S1 Task	S2 Task
S1 Task	1910	1701
S2 Task	1690	1899
Recall	0.531	0.528
Precision	0.529	0.529
F-measure	0.530	0.528

TABLE IV
RESULT OF RANDOMFOREST FOR EACH VARIABLE

Explanatory variable	Recall	Precision	F-measure
Standard deviation of Heart Rate	0.625	0.625	0.625
Mean of HFnorm	0.500	0.500	0.497

results of the analysis of variance are shown in TableII. This result shows a variable with a significant difference with the significance level set at 5%.

As a result of analysis of variance, it was shown that there is a significant difference between the standard deviation of heart rate and the mean of HFnorm. Explanatory variables were two variables, the standard deviation of heart rate and the mean of HFnorm. In order to compute the universal discrimination result, 12 leave-one-out cross verification were carried out by using Random Forest. Random Forest constructs a decision tree by randomly sampled training data. In order to reduce the influence due to bias in random sampling, 100 times identification was performed using Random Forest. TableIII shows the result of totaling the identification results by 100 times Random Forests. From the results of Random Forest, when using the standard deviation of heart rate and the mean of HFnorm as explanatory variables, the recall, precision, and F-measure were each about 0.53.

Also, 12 cross-validation was performed using Random Forest for each variable with the standard deviation of heart rate and the mean of HF norm. TableVII shows the results of identification using Random Forest for each variable. From the classification results using Random Forest, when the average of HFnorm is used as an explanatory variable, the recall, precision, and F-measure were each about 0.5. On the other hand, If the standard deviation of the heart rate is used as an explanatory variable, both the recall, precision, and F-measure were about 0.62, and the S1 task and the S2 task can be classified.

V. DISCUSSION

A. Deviation of heart rate

In ChapterIV, we conducted an experiment to distinguish the S1 task where learners hold the schema from the S2 task where learners do not hold the schema, using the pulse and pulse fluctuation. Based on the experimental results, we showed that the task given to the learner can be classified more accurately than random judges, using the standard deviation of the heart rate of the learner at the time of task engagement. The standard deviation of heart rate is high in tasks 4, 5, and 6, which falls into S2 task. It is conceived that the cognitive load imposed on the learners caused the high standard deviation of the heart rate. At the same time, we found the learners often sighs during the S2 task. It is conceived that the variation in heart rate was great because the heart rate is affected by the big breathing.

B. Dependence on individual learners

In this experiment, we did not take into consideration whether it is a high-load task or a low-task for individual learners. Here, we discuss this point. Using questionnaire, the tasks given to each learner were classified into high-load tasks or low-load tasks. We calculated the deviation value of the perceptual activity based on Likert scale 6. Tasks with deviation values of 50 or less were regarded as low load tasks, and the others high load tasks. TableV shows the task loads calculated from the questionnaire. For each of the 18 variables acquired at the time of the task engagement, we performed the variance analysis to examine significant differences between the high-load tasks and the low-load tasks. The results are shown in TableVI.

The results show that there was a significant difference in the standard deviation of the heart rate. We adopt the standard deviation of the heart rate as the explanatory variable to judge whether the tasks given to the learner were high-load ones. As a result of 12 cross-validation using Random Forest, the recall, the precision, and the F-measure were about 0.57, which means we cannot judge only with the standard deviation of the heart rate. Since the standard deviation of HF had the second lowest P value, we added the standard deviation of HF to explanatory variables. TableVII shows the results of 12 cross-

TABLE V
A QUESTIONNAIRE RESULT ON WHETHER EACH TASK IS A HIGH-LOAD TASK OR A LOW-LOAD TASK

Task Number	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
Number of learners who answered high-load tasks	2	2	1	11	11	7

TABLE VI
RESULTS OF ANALYSIS OF VARIANCE OF EACH 18 VARIABLES. IN ORDER TO FIND THE DIFFERENCE BETWEEN HIGH-LOAD TASK AND LOW-LOAD TASK.

Explanatory variable		Heart Rate	RRI	HF	LF	LF/HF	VLF	TP	HF_norm	LF_norm
P value	Mean	0.7330816	0.5980086	0.3791852	0.1775325	0.928875	0.6977778	0.2977588	0.227955	0.227955
	Standard deviation	0.0082179	0.4444583	0.0861439	0.5396019	0.7337266	0.6905658	0.7072911	0.5477441	0.5477441

TABLE VII
IDENTIFICATION OF A HIGH-LOAD TASK OR A LOW-ORDER TASK USING RANDOMFOREST

	Low-Load Task	High-Load Task
Low-Load Task	2807	1364
High-Load Task	993	2036
Recall	0.739	0.599
Precision	0.673	0.672
F-measure	0.704	0.633

validation using Random Forest. We identified it 100 times by Random Forest, to calculate the average.

The judgement using Random Forest taking the standard deviation of the heart rate and the standard deviation of HF as explanatory variables presents about 0.67 of the recall, the precision, and the f-measure. Based on the subjective difficulty level, it was possible to identify whether the tasks given to learners were difficult tasks or simple tasks.

VI. CONCLUSION

In this research, in order to give lesson contents and tasks suitable for learners at the time of preparation in the flipped class, we proposed a method to estimate whether the learner holds the schema related to the contents or not, using the cognitive load of the learner. We obtained cognitive load from autonomic nerve fluctuation calculated from learner's pulse wave and pulse fluctuation. By using machine learning, we identify whether the tasks given to the learner are hard ones from the cognitive load, to estimate whether the learner holds the schema for the tasks.

By experiment, we identified the difficulty level of tasks given to learners, using standard deviation of heart rate. As a result, the F-measure was about 0.62. By using the system proposed in this research, it is possible to adjust the contents of lessons and tasks for the learner, to improve flipped classes. In the future, in order to improve the accuracy, we will analyze data other than the pulse waves and the pulse fluctuation.

REFERENCES

[1] OnlineUniversities.com : "12 Inspiring Schools Using Khan Academy", http://www.onlineuniversities.com/blog/2012/04/12-inspiring-schools-using-khan-academy.

[2] The Chronicle of Higher Education : "California State U. System Will Expand MOOC Experiment", https://www.chronicle.com/blogs/wiredcampus/california-state-u-system-will-expand-mooc-experiment.

[3] Katsusuke Shigeta. : "Flipped Classroom : Educational reform utilizing information technology.", j-stage, Vol.56, No.10, pp.677-684, (2014).

[4] Sweller, John : "Cognitive load theory, learning difficulty, and instructional design.", Learning and instruction, Vol.4, No.4, pp.295-312, (1994).

[5] Schnotz, Wolfgang, and Christian KÃijrschner : "A reconsideration of cognitive load theory.", Educational psychology review, Vol.19, No.4, pp.469-508, (2007).

[6] Sweller, John, Jeroen JG Van Merrienboer, and Fred GWC Paas : "Cognitive architecture and instructional design." Educational psychology review , Vol.10, No.3 , pp.251-296, (1998).

[7] Paas, Fred GWC, and Jeroen JG Van MerriÃńnboer. "Instructional control of cognitive load in the training of complex cognitive tasks." Educational psychology review Vol.6, No.4, pp.351-37, (1994).

[8] DeLeeuw, K. E., Mayer, R. E. : A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. Journal of Educational Psychology, Vol.100, No.1, pp.223, (2008).

[9] Youichi Mizuno, et al. : " Experimental Investigation of Relationship between Cognitive Loads and Cognitive Processes in Learning through Problem Solving", SIG-ALST, Vol5, No.3, pp.7-11, (2016).

[10] Galy, Edith, Magali Cariou, and Claudine MÃl'lan. "What is the relationship between mental workload factors and cognitive load types?." International Journal of Psychophysiology, Vol.83,No.3, pp.269-275, (2012).

[11] Ikehara, Curtis S. Martha E. Crosby : "Assessing cognitive load with physiological sensors." , System Sciences, 2005, HICSS'05, Proceedings of the 38th Annual Hawaii International Conference on. IEEE, pp.295a-295a, (2005).

[12] Mulder, Gijsbertus, Lambertus JM Mulder : "Information processing and cardiovascular control." Psychophysiology, Vol.18, No.4, pp.392-402, (1981).

[13] Keisuke Tsunoda et al. : "Estimating Cognitive Performance Change using Heart Rate Variability". Journal of Information Processing Society of Japan, Vol.57, No.8, pp.1835-1844, (2016).

[14] Kahneman, Daniel. : " Thinking, fast and slow." Macmillan, (2011).

[15] Breiman, Leo : "Random forests.", Machine learning, Vol.45, No.1, pp.5-32, (2001).

[16] Gerjets, Peter, Katharina Scheiter, and Richard Catrambone. "Designing instructional examples to reduce intrinsic cognitive load: Molar versus modular presentation of solution procedures." Instructional Science Vol.32,No.1-2, pp.33-58, (2004).

[17] Pollock, Edwina, Paul Chandler, and John Sweller. "Assimilating complex information." Learning and instruction, Vol.12, No.1,pp.61-86, (2002).

[18] Likert, Rensis. "A technique for the measurement of attitudes." Archives of psychology, (1932).

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 9–16
DOI: 10.15439/2018F373
ISSN 2300-5963 ACSIS, Vol. 16

Robotic Process Automation of Unstructured Data with Machine Learning

Anna Wróblewska*,†, Tomasz Stanisławek*,†, Bartłomiej Prus-Zajączkowski*,†, Łukasz Garncarek†
*Faculty of Mathematics and Information Science, Warsaw University of Technology
ul. Koszykowa 75, Warszawa, Poland
†Applica.ai
ul. Wiślana 8, Warszawa, Poland
E-mail: {anna.wroblewska, tomasz.stanislawek, bartlomiej.prus, lukasz.garncarek}@applica.ai

***Abstract*—In this paper we present our work in progress on building an artificial intelligence system dedicated to tasks regarding the processing of formal documents used in various kinds of business procedures. The main challenge is to build machine learning (ML) models to improve the quality and efficiency of business processes involving image processing, optical character recognition (OCR), text mining and information extraction. In the paper we introduce the research and application field, some common techniques used in this area and our preliminary results and conclusions.**

I. Introduction

NOWADAYS, business processes are not enough time and cost effective, so much effort is made to automate them. Researchers together with business partners are expecting that a vast majority of repetitive human work can be eliminated with the aid of artificial intelligence. Besides human work automation, there are also other advantages to gain: economic leverage of large scale, reduction of a process duration, repetitiveness, analytics of a process, higher specialization in core business, lower maintenance costs [1]

In a typical Robotic Process Automation (RPA) tool human work is automated by observing the user perform a task in the graphical user interface (GUI), and consecutively with the help of artificial intelligence repeating the same task in the GUI. There are a lot of companies that help to automate parts of business processes that fit that model, like 'Blue Prism', 'UiPath', 'Verint', 'WorkFusion' [2][3][4][5].

In our case, we want to focus on dealing with unstructured data (e.g. letters, e-mails, invoices). This is one of the hardest parts to solve because one does not take exactly the same actions to process each document. Examples of such tasks are: finding relevant monetary amounts in invoices, automatically understanding customer complaints, chat bots for call centers, etc.

The main contribution of this paper is to state and organize challenges associated with RPA based on unstructured data. Moreover, we present our baseline machine learning models for processing formal documents and try to challenge them with innovative approach incorporating deep learning methods, e.g. transfer learning using language models, state-of-the-art deep learning architectures, etc.

In the following sections we introduce requirements for RPA, show a general architecture of such a system (section II), and sketch state-of-the-art ML methods that are used for particular parts of our system (section III). Then we show our approach to automation of processing formal documents, describing use cases and preliminary results (section IV). Finally, we conclude our work with future plans and important tips for RPA with ML methods (section V).

II. Robotic Process Automation of Unstructured Data

It is observed that in big companies a lot of communication with clients or business partners takes place through some kind of documents (e.g. letters, e-mails, images, pdfs, text files). Nowadays, the typical solution is to create Business Process (BP) to handle that kind of information channel.

Robotic Process Automation can be defined as a support system which automates human work that involves routine tasks [6]. In our study we solely deal with business processes that handle unstructured data (we call them documents). According to Fung [7] not all of BPs can be automated, so we define criteria and a general architecture of a business process that may adopt RPA mechanism. Processes which meet these requirements could be automated and bring improvements to operational efficienvy of any organization.

Criteria that must be fulfilled by a business process to apply a RPA solution:

1) *High volume* – thousands and thousands of documents are processed every month, involving much workforce in a business process.
2) *Standardization* – a business process must be standardized at some level, i.e. a process must have specifications that are fixed most of the time. For instance, we must know what kind of information should be extracted from a document.

On the other hand, we also define criteria which must be met by a RPA solution:

1) *Simple architecture* – ensures easy integration at various stages of a business process.
2) *High scalability* – ensures ability to cope with a high volume of documents. RPA service should be easy to scale and it should process documents in a matter of milliseconds or seconds rather than minutes.

©2018, PTI

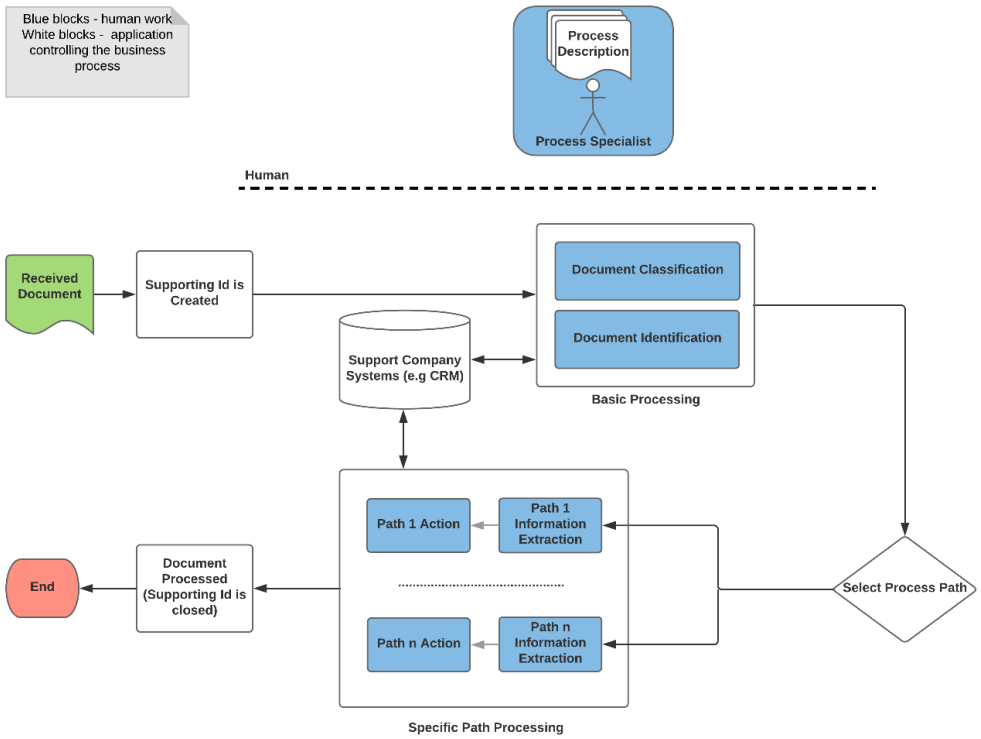

Fig. 1. General Business Process Architecture.

3) *Flexibility* – each model should be implemented as a separable module to make it easy to use.
4) *Quality* – machine learning models must be very accurate in order to ensure profitability from a business perspective.
5) *Security* – ensures appropriate access control to company's data and resources.

A. General Business Process Architecture

According to our criteria, we present a common architecture of a business process (Fig. 1), in which a RPA system can be applied. In the proposed architecture we can distinguish the following blocks:

1) *Document Received* – a document was sent to the process and assigned an ID as a marker. Commonly, companies use queues to store documents for further processing.
2) *Basic Processing* – at this stage we define two main modules: *Document Classification* and *Document Identification*. In the classification task a human specialist must assign a category/categories to the document, based on a taxonomy specified in the Process Description document. Each category has some specific meaning regarding the particular business process, e.g. e-mails with the 'Complaint' category will have a greater priority than e-mails with the 'Payment Confirmations' category. The main goal of the Document Identification module is to link the existing database records from the company's CRM system to each document, e.g. finding a client ID or a contract ID related to this document.
3) *Process Path Selection* – the business process owner defines a set of rules based on the information collected from the *Basic Processing* stage which determines a processing path for the document, e.g. the document category determines an object list that must be extracted in the next step. Those rules are implemented in an application to support the business process.
4) *Specific Path Processing* – when a path is selected, we can process the document based on Process Specification. Each path has two modules: *Information Extraction* and *Action*. It means that first we must extract specific objects from documents and then we should take a specific action related to the collected information. For instance, if we know that a client sent an e-mail with the 'Payment Confirmation' category, we can extract an amount of the payment and afterwards check if the amount is the same as the value retrieved from CRM. If the two amounts agree, we can unblock the user account, and if not, we can send an e-mail to the client. Additionally, we can store all extracted valuable information in the CRM system.
5) *Document Processed* – the final stage of the process. It means that the document is marked as processed and all relevant actions have been taken.

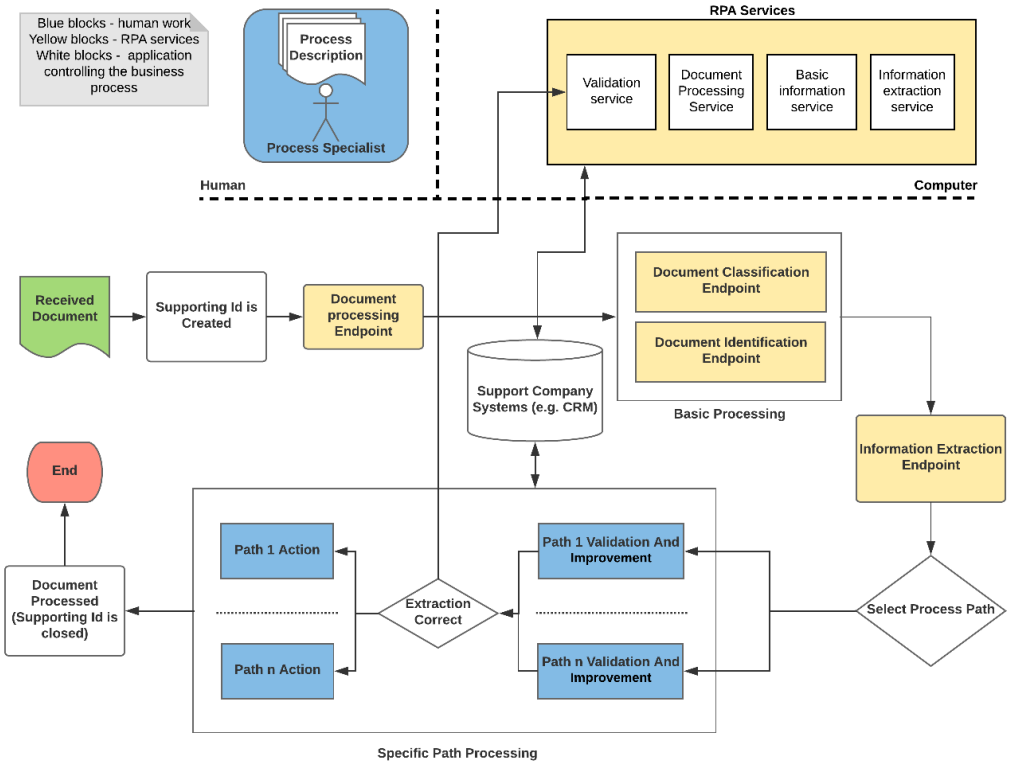

Fig. 2. Business Process Architecture with RPA Services.

B. Automation of Business Process

As we can see in Fig. 1 a lot of work is done manually by company employees in any business process. Most of such tasks can be replaced by machine learning methods such as:

1) *Image processing* – scanning letters, invoices, attachments from emails, etc.
2) *Classification* – labeling documents according to a taxonomy described in Process Specification documents, linking CRM information to documents.
3) *Information extraction* – extracting specific objects from image/text that are needed to process documents.

We offer our own system architecture (Fig. 2), in which most of modules are replaced by services using machine learning methods. That solution reduces the amount of work done by human workforce. To make reading easier blue blocks indicate tasks done by humans and yellow ones indicate tasks done by RPA services.

The workflow of the business process is mostly the same. From the company's perspective this is a big advantage because they do not need to change much of their business process environment. Additionally, our solution assumes that all communication between RPA services and a business process system is based on REST web-services, which is a standard solution these days. The difference is only in three extra blocks, which were added for integration purposes:

1) *Document Processing Service* – a service to pre-process documents for further use in all RPA services (perform OCR on images, process a text by NLP tools, etc.).
2) *Validation And Improvement Path* – at this stage humans check if the data have been correctly classified and extracted. If some models made a mistake it is sent to *Validation Service*.
3) *Validation Service* – a service to ensure high accuracy of models. This service is provided with information from model errors which are diagnosed and used to correct our predictions in the future.

III. Machine Learning Methods

In this chapter we describe some state-of-the-art and classic techniques that are used in tasks similar to these that we need in our RPA solution.

A. Image Processing

The most common and standard methods to improve the quality of images involve image filtering in the spatial or frequency domain. In a spatially filtered image, the value of each output pixel is the weighted sum of neighboring input pixels. The weights are provided by a matrix called the convolution kernel or filter. There are following kinds of filters: texture filtering to smooth an image or to sharpen it, filters that extract edges, morphological filters, etc. [8]. There are environments to do these operations in an easy way: Matlab environment [9] and OpenCV library with API for the most common programming languages [10].

To perform easier tasks there are still common classical methods that need preparing feature extraction from images and then building ML models using these features as an input. These models can be: random forest [11], linear regression, etc.

Currently the most popular ML methods for computer vision are deep learning (DL) neural networks. The latest research shows that DL can outperform even humans in tasks of object class recognition, etc. [12]. For example, to detect and classify particular (previously defined) classes of objects, the most efficient architecture is 'Faster-RCNN' [13].

These DL techniques take advantage of the knowledge gleaned from very big open datasets (e.g. ImageNet [14], COCO [15]) and models built to classify, detect and even segment objects in them (e.g. AlexNet, ResNet, VGG architectures [16]). This technique is called transfer learning.

B. Classification

The problem of classification can be defined as a prediction of a discrete class or classes y given a vector of features $\mathbf{x} = (x_1, \ldots, x_k)$. The domain of y is a set of possible classes depending on a problem. In machine learning area a whole range of methods may be used to tackle this problem.

In recent years, linear models (e.g. logistic regression or support vector machines) have seen an increase in usage, mainly when dealing with big data. This is thanks to their ability to scale very well when using appropriate optimization techniques (e.g. stochastic gradient descent) and possibility of on-line learning [17][18][19]. However, we can observe a great improvement in classification accuracy when using deep neural networks[20][21] or XGBoost method [22].

1) Logistic Regression: Logistic regression (LR) is in its basic form a linear binary model. It tries to estimate the probability of a positive class given the input observations. Formally, it can be expressed by the following equation

$$P(Y = 1|x) = \sigma(\mathbf{x^T w})$$

where $\mathbf{x}$ is input data, $\mathbf{w}$ is a vector of model weights (or coefficients) and σ is sigmoid function which transforms the value of $\mathbf{x^T w}$ from $(-\infty, +\infty)$ into $[0, 1]$.

To find parameters $\mathbf{w}$ LR uses *log loss* function during optimization, defined as

$$L(y, \mathbf{wx}) = \log(1 + e^{-y\mathbf{wx}}).$$

2) Support Vector Machines: Support Vector Machines (SVMs) were proposed by Cortes and Vapnik in 1995 [23] and later employed for text classification by Joachims in [24]. A SVM model attempts to find a hyperplane separating positive and negative training examples with additional condition of maximizing the margin (i.e. the distance of training examples to the hyperplane). In case of linearly separable classes, to this end one needs to use only a small fraction of training examples called *support vectors* – hence the name.

Assume we are given a sequence of data points $\mathbf{x}_i \in \mathbb{R}^d$ together with their class labels $y_i \in \{\pm 1\}$. The SVM assigns to each point a score by the formula

$$A(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b,$$

where $\mathbf{w} \in \mathbb{R}^d$, and $b \in \mathbb{R}$ are the trainable parameters of the model.

The optimization objective is defined with help of the *hinge loss* function which penalizes not only misclassification, but also, to a lesser extent, predictions with low confidence. The loss to be minimized is defined as

$$\frac{C}{N}\sum_{i=1}^{N}\max\{0, 1 - (y_i A(\mathbf{x}_i))\} + \frac{1}{2}\|\mathbf{w}\|^2,$$

where C is a regularization constant.

3) Convolutional Neural Networks (CNN) : Convolutional neural networks improve their efficiency not only in image processing techniques, as they were initially used, but also in text classification. Overall, CNNs are extremely effective in mining semantic clues in contextual windows [25]. The latest research underlines their great achievements in text field [26].

Assume we have an embedding matrix $\mathbf{W} \in \mathbf{R}^{nxd}$ which is a $\mathbf{d}$-dimensional representation of embeddings of $\mathbf{n}$ words (or letters). This input layer can be pre-trained or initialized randomly. Convolutions are performed on consecutive parts of this input embedding layer (on a consecutive row of input embedding matrix).

The convolutional layer involves a filter $\mathbf{k} \in \mathbf{R}_{hxd}$ which is applied to a window of $\mathbf{h}$ words to produce a new feature and a non-linear activation function, for example, ReLU. The filter $\mathbf{k}$ is applied to all possible windows using the same weights to create the feature map. A set of convolutional filters of different widths slides over the entire embedding matrix and they extract patterns of n-grams. Convolutions are usually followed by max-pooling, which subsamples the input to provide fixed-length output and keep the most salient n-gram features across the whole input [25].

C. Information extraction

In the information extraction task we try to find a set of objects (e.g. Person, Localization, Amount, etc.) in unstructured text data relevant to a specific domain [27]. The most common approach is to represent a text as a sequence of tokens (words or characters) denoted as vector $\mathbf{x} = (x_1, \ldots, x_n)$, and the corresponding vector $\mathbf{y} = (y_1, \ldots, y_n)$, representing output class labels (each value $\mathbf{x_i}$ has its own output class $\mathbf{y_i}$ with domain $\mathbf{D}$), where $\mathbf{i}$ is a position of a token in the text. As a baseline we have used Conditional Random Fields (CRF) but this is being replaced by deep learning architecture involving state-of-the-art recurrent neural networks.

1) Conditional Random Fields: Conditional Random Fields (CRF) was proposed in 2001 by Lafferty [28] and for a long time it was considered one of the most accurate methods of sequence labeling.

CRF is a probabilistic graphical model that defines a single log-linear distribution over label sequences $\mathbf{y} = (y_1, \ldots, y_n)$,

given a particular observation sequence $\mathbf{x} = (x_1, \ldots, x_n)$,. Lafferty [28] defines that probability as a normalized product of potential functions

$$\mathbf{P}(\mathbf{y}|\mathbf{x}, \lambda) = \frac{1}{Z(\mathbf{x})} exp \sum_{i=1}^{N} \sum_{j} \lambda_j f_j(y_{i-1}), y_i, \mathbf{x}, i)$$

where $Z(\mathbf{x})$ is a normalization factor, λ_j is a parameter to be estimated and $f_j(y_{i-1}), y_i, \mathbf{x}, i)$ is a feature function that expresses some characteristic of the empirical distribution that we wish to hold in the model distribution [29].

2) Recurrent Neural Networks: Although recently challenged by *temporal convolutional networks*, *recurrent neural networks* (RNNs) are the traditional neural architectures used for sequence labeling [30][31][26]. Given a sequence of vectors

$$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$$

of arbitrary length n, a RNN model produces an output sequence $(\mathbf{y}_i)$ of the same length, which in general can be described by the following recurrent formula

$$\mathbf{y}_i = \begin{cases} F_0(\mathbf{x}_0) & \text{if } i = 0, \\ F(\mathbf{x}_i, \mathbf{y}_{i-1}) & \text{otherwise,} \end{cases}$$

where F_0 and F are functions of a prescribed form with trainable parameters. The two most common choices for these functions lead to *long short-term memory* (LSTM) and *gated recurrent unit* (GRU) modules. We refrain from presenting the exact formulas, as they are a bit complex and out of scope of this paper.

What is important though is the fact that suitably defined recurrent modules, such as LSTMs or GRUs, exhibit a behavior that can be described as possessing *memory*. Namely, the model may learn to store information encountered in earlier steps in the output vectors $\mathbf{y}_i$, and use it later, when needed.

IV. Application

A. Use-case Description

In this section we describe a problem of automating BP from a debt collector company. The company receives more than 150 000 documents each month sent by courts and bailiffs as letters via the post office. Such an enormous information stream is hard to handle by human workers alone. It is also important to process each document immediately because some debt collection cases have limited time to react to. Below we present the results of our baseline solution, which is currently being improved by employing new innovative techniques.

First, each document is scanned and company workers assign the main category to it: either 'Bailiff document' or 'Court document'. Next, the process looks very similar to the one presented in Fig. 2:

1) Document Received – supporting ID is created.
2) Document Preprocessing – documents are preprocessed and OCR-ed.
3) Basic Processing – documents are classified and then Case and Client are identified and matched to the existing client's database.
4) Information Extraction – relevant information is extracted from the document, based on its category.
5) Extraction Validation – data is verified and sent to a validation service if not correct.
6) Action – a specific action is taken based on the extracted information.
7) Document Processed – supporting ID is closed.

One of the many challenges in the process is to interpret and understand Polish language texts that are retrieved with OCR. Thus we make an effort to process document scans and obtain texts in as good quality as possible to improve further processing. Having texts after OCR we try to recognize important entities (e.g. persons, addresses, ID numbers of documents, dates), extract monetary amounts, etc. These tasks are very challenging regarding that NLP tools for Polish are not of as high a quality as NLP for English, mostly because Polish is a highly inflected language with a complex grammar.

B. Approach and Results

1) Image Processing: In order to improve OCR results, we need to have high quality document images. Thus we need image processing methods to remove noisy background from images, detect important parts of images that should be treated differently from other parts of scans, e.g. tables and multiple columned texts.

To improve document image quality, we use morphological filtering and correcting text line skewness with Hough transform techniques [8]. Then to detect table headers, which are particularly difficult to OCR mechanism, we classify them as having particular grayish background or inverse colors of the background and text using image feature extraction and random forest models. Then we can distinguish those different backgrounds and take proper further steps.

At the same time we detect multicolumn texts, e.g. tables with monetary amounts calculations using DL methods. We adopt transfer learning – 'VGG-19' learned on ImageNet dataset and 'Faster-RCNN' architecture. Then we train them on our small dataset of about 400 documents.

With these technique we achieve preliminary results on a test set of about 300 documents to identify tables with about 91% of precision and 87% of recall, even without preprocessing steps.

The next process is to select important regions of document images to OCR them with adjusted parameters for these particular tasks, e.g. OCR of tables. Additionally we try to improve OCR results in order to help information extraction methods and extract columns and rows of tables, e.g. to extract proper monetary amounts and their classes. Here we use heuristics that have a table and column area in the images and OCR results (that is bounding boxes and recognized text tokens). Thus we have prepared text documents of as high a quality as possible.

2) Document classification: The first and the most important task during the whole process is to classify incoming documents into appropriate categories. There are two possible sources of documents: either bailiff or court documents. Bailiff documents are classified into one of 18 categories and court documents are classified into one of 25 categories. This is a single-label multi-class categorization problem. The task is really important because further processing depends on the classification results.

The training set for court documents contains 26639 documents and for bailiff documents – 25339. Additionally, we have independently prepared test sets and they have 2501 and 2691 documents for court and bailiff respectively.

The solution presented here uses SVM classifier to complete this task. Throughout years SMVs were shown to achieve good performance in text classification tasks and are a good starting point [32]. This is why we chose it as a baseline method to compare our future results with. Yet our current experiments show that SVMs can be outperformed using more sophisticated methods, even with less training data.

In the beginning we perform classic document preprocessing (lowercasing, removing diacritics and stopwords, tokenizing, etc.). We take unigrams and bigrams as features, weighted with TF-IDF algorithm. Features so prepared are passed to the SVM classifier. For SVM training we use popular LIBLINEAR library [18].

With such a baseline solution, we were able to achieve 92.08% accuracy for bailiff documents and 92.36% accuracy for court documents on the test sets. Moreover, we have verified our predictions on real data and reached about 91% on court and about 93% accuracy on bailiff documents.

As mentioned above, document classification is an essential part of the whole process, as it directly impacts further processing. That is why we put a lot of effort into improving its quality. Currently, we are experimenting with different deep learning methods for classification. We tried convolutional neural networks [21] and temporal convolutional networks [26]. We were able to improve upon SVM performance by about 0.5-1.0 pp.

Moreover, apart from deep learning efforts, we experiment with transfer learning [33]. To do this we have built a language model based on a very large data set and this model is then used as an input to a classifier. Preliminary results show that it is a very promising approach. Not only can we obtain better accuracy, but also do it with much less data.

3) Document identification: After classifying a document, we need to assign it to an existing case using the information from our client's database. This process has four steps.

The first one is information extraction. We extract such objects as court reference numbers, personal identification numbers, addresses, debt amounts etc. We typically either find only a subset of information that could be extracted from the document or we end up with noisy information that render extracted objects unhelpful in further steps. This step ends with an incomplete set of textual features, some of which may contain errors.

The second step is to search the client's database for cases which partially match the extracted information. We call these cases 'candidates'.

After finding the candidates, we once again compare each candidate's database entry with the extracted information, and use compatibility measures to produce numerical features, thus describing each candidate by a real-valued vector.

In the last step we pass the feature vectors through a machine learning model (our baseline solution uses logistic regression) that produces a probability that a given candidate is the correct match for the document in question, and return the candidate with the highest probability, provided that it exceeds some fixed threshold.

With our benchmark solution we are able to achieve 99.84% precision and 76.02% recall. Obtained recall results are still below our expectations but we are optimistic that using a pre-trained language model to improve quality of the text will result in a much better performance of document identification.

4) Information extraction: Apart from the information extracted in the process of document identification (section IV-B3), we are currently dealing with the problem of extracting monetary amounts of different kinds, which appear directly in the document. There are a few groups of amounts we are working on and each of them requires a different architecture.

In general, this problem can be modeled as follows. Given a document $\mathbf{x}$ represented as a sequence of characters

$$\mathbf{x} = (x_1, \ldots, x_n),$$

we may infer a sequence

$$\mathbf{p} = (p_1, \ldots, p_n),$$

where $p_i \in [0, 1]$ is the probability that the character x_i belongs to the amount to be extracted. Having produced such a sequence of probabilities, we extract maximal contiguous subsequences of the form $(x_k, x_{k+1}, \ldots, x_\ell)$ such that $p_i > T$ for all $k \leq i \leq \ell$ and a given threshold value T. These excerpts then undergo validation and normalization procedures, mainly due to the noise in OCR-ed texts, but also to eliminate some obvious errors such as partial extraction.

A single model can be also used to extract multiple related objects. In this case we just use $p_i \in [0, 1]^k$, where k is the number of extracted objects.

To produce the score sequence we use state-of-the-art character-level recurrent neural networks (section III-C2). Architectures that we employ consist of an embedding layer, possibly followed by some convolutional layers, a single-layer GRU module, and some densely connected layers producing the final scores. Intuitively, using convolutional layers means that we try to look not only at single characters, but also syllables or even words – depending on the parameters of the convolutional module. We are also experimenting with applying a pre-trained character based language model instead of a classical embedding layer.

TABLE I
PERFORMANCE OF MONETARY AMOUNT EXTRACTORS

Task	Precision	Recall	F1 Score	# of test documents	# of test objects	# of train documents	# of train objects
Bailiff costs	83.07%	86.40%	84.70%	3316	10729	12500	40637
Court costs	77.46%	97.77%	86.44%	20557	6131	126665	38058

We did not found a significant difference when using LSTM instead of GRU, and increasing the number of recurrent layers beyond 1 leads to unacceptable prediction times.

There is a technical issue here, related to the potentially unlimited length of the documents. Unlike some other applications, information extraction requires us to process the whole document. We solve the problem by cutting the text into overlapping fragments of the form

$$(x_{k(\ell-s)}, x_{k(\ell-s)+1}, \ldots, x_{k(\ell-s)+\ell-1})$$

of prescribed length ℓ and overlap size s. We pass these fragments to the model separately, and suitably interpolate the resulting probability sequences on overlapping fragments, once again obtaining a single sequence of probabilities.

Now, let us get back to the problem at hand. We are working on two groups of amounts: attorney representation costs, and costs related to various actions performed by the bailiff in search of debtor's assets. Table I presents the performance of extractors trained with 80%-20% train-test split. It is worth noting that our current data sets were obtained in a partially automated manner. Thus they need to be verified by linguists to achieve results of a higher quality.

V. CONCLUSIONS

In this paper we described general business processes for documents flow and current approaches to automating them; we also presented our baseline solutions for a specific business process. Our particular use-case is dedicated to a debt collecting process. We described possible applications of ML methods to improve the efficacy of these processes. This is a significant benchmark for all future work in this field.

We are devoted to an ongoing aim to improve our existing models. Each part of our RPA solution is constantly being upgraded. Much emphasis is given to further research into deep learning methods and into transfer learning, using previously trained language models.

We believe that the core value of our solutions in the near future will be the reduction of training data size required to achieve performance comparable to the currently used state-of-the-art methods. Corporate clients will need to provide significantly less data and the whole process of implementing automation will be a lot more efficient (lower cost, quicker integration, robust, etc.).

Besides, we plan to experiment with automation of *Action* (Fig. 1) stage. We want to be able to automatically take actions for more complicated use-cases, e.g. responding to a client via e-mails based on collected information from *Basic Processing* and *Information Extraction* modules.

ACKNOWLEDGMENT

This paper provides a description of a project currently conducted at Applica.ai, an AI business application firm. As such, the paper has a large number of contributors, including, but not limited to, Adam Dancewicz and Piotr Surma, our managing directors, and Paweł Dyda, Przemysław Lipka, Arakdiusz Trzepacz, Szymon Sikorski, Paweł Józiak and a team of linguists.

The authors would like to acknowledge the support the Applica.ai project has received as being co-financed by the European Regional Development Fund (POIR.01.01.01-00-0144/17-00).

REFERENCES

[1] M. Kukreja and A. singh Nervaiya, "Study of robotic process automation (rpa)," *International Journal on Recent and innovation trends in computing and communication*, vol. 6, pp. 434–437, 2016.

[2] Robotic process automation, "Robotic process automation — Wikipedia, the free encyclopedia," 2018, [Online; accessed 7-June-2018]. [Online]. Available: https://en.wikipedia.org/wiki/Robotic_process_automation

[3] Capgemini Consulting, "Robotic Process Automation – Robots conquer business processes in back offices," 2016, accessed: 2018-06-07. [Online]. Available: https://www.capgemini.com/consulting-de/wp-content/uploads/sites/32/2017/08/robotic-process-automation-study.pdf

[4] PWC, "Rethinking retail: Artificial Intelligence and Robotic Process Automation," 2017, accessed: 2018-06-07. [Online]. Available: https://www.pwc.be/en/documents/20171123-rethinking-retail-artificial-intelligence-and-robotic-process-automation.pdf

[5] David Schatsky and Craig Muraskin and Kaushik Iyengar, "Robotic process automation A path to the cognitive enterprise," 2016, accessed: 2018-06-07. [Online]. Available: https://www2.deloitte.com/content/dam/insights/us/articles/3451_Signals_Robotic-process-automation/DUP_Signals_Robotic-process-automation.pdf

[6] S. Aguirre and A. Rodriguez, "Automation of a business process using robotic process automation (rpa): A case study," in *Workshop on Engineering Applications*. Springer, 2017, pp. 65–71.

[7] H. P. Fung, "Criteria, use cases and effects of information technology process automation (itpa)," 07 2014.

[8] W. K. Pratt, *Digital Image Processing: PIKS Inside*, 3rd ed. New York, NY, USA: John Wiley & Sons, Inc., 2001.

[9] Matlab official website, "Matlab," 2018, [Online; accessed 7-June-2018]. [Online]. Available: https://www.mathworks.com/products/matlab.html

[10] OpenCV official website, "Opencv," 2018, [Online; accessed 7-June-2018]. [Online]. Available: https://opencv.org

[11] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[12] J. Howard and S. Ruder, "Fine-tuned language models for text classification," *CoRR*, vol. abs/1801.06146, 2018. [Online]. Available: http://arxiv.org/abs/1801.06146

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[14] ImageNet dataset official website, "Imagenet," 2018, [Online; accessed 7-June-2018]. [Online]. Available: http://www.image-net.org

[15] COCO dataset official website, "Coco," 2018, [Online; accessed 7-June-2018]. [Online]. Available: http://cocodataset.org

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[17] X. Li, "Classification with large sparse datasets: Convergence analysis and scalable algorithms," 2017.

[18] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1390681.1442794

[19] K.-c. Lee, B. Orten, A. Dasdan, and W. Li, "Estimating conversion rate in display advertising from past performance data," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 768–776. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339651

[20] H. T. Le, C. Cerisara, and A. Denis, "Do convolutional networks need to be deep for text classification ?" *CoRR*, vol. abs/1707.04108, 2017. [Online]. Available: http://arxiv.org/abs/1707.04108

[21] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014. [Online]. Available: http://arxiv.org/abs/1408.5882

[22] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep 1995. [Online]. Available: https://doi.org/10.1007/BF00994018

[24] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98*, C. Nédellec and C. Rouveirol, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 137–142.

[25] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *CoRR*, vol. abs/1708.02709, 2017. [Online]. Available: http://arxiv.org/abs/1708.02709

[26] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *CoRR*, vol. abs/1803.01271, 2018. [Online]. Available: http://arxiv.org/abs/1803.01271

[27] J. Piskorski and R. Yangarber, *Information Extraction: Past, Present and Future*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 23–49. [Online]. Available: https://doi.org/10.1007/978-3-642-28569-1_2

[28] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. [Online]. Available: http://dl.acm.org/citation.cfm?id=645530.655813

[29] H. M. Wallach, "Conditional random fields: An introduction," *Technical Reports (CIS)*, p. 22, 2004.

[30] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," *CoRR*, vol. abs/1511.08308, 2015. [Online]. Available: http://arxiv.org/abs/1511.08308

[31] Z. Yang, R. Salakhutdinov, and W. W. Cohen, "Multi-task cross-lingual sequence tagging from scratch," *CoRR*, vol. abs/1603.06270, 2016. [Online]. Available: http://arxiv.org/abs/1603.06270

[32] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, Mar. 2002. [Online]. Available: http://doi.acm.org/10.1145/505282.505283

[33] Z. Yang, R. Salakhutdinov, and W. W. Cohen, "Transfer learning for sequence tagging with hierarchical recurrent networks," *CoRR*, vol. abs/1703.06345, 2017. [Online]. Available: http://arxiv.org/abs/1703.06345

3rd International Workshop on Artificial Intelligence in Machine Vision and Graphics

THE main objective of the 3rd Workshop on Artificial Intelligence in Machine Vision and Graphics (AIMaViG'18) is to provide an interdisciplinary forum for researchers and developers to present and discuss the latest advances of artificial intelligence in the context of machine vision and computer graphics. The workshop covers the whole range of AI-based theories, algorithms, technologies and systems for diversified and heterogeneous areas of vision and graphics.

TOPICS

The topics and areas include but are not limited to:

- image processing
- scene analysis, modeling, and understanding
- machine vision
- pattern matching and pattern recognition
- image synthesis, including three-dimensional imaging and solid modeling
- computer-aided graphic arts and animation
- mathematical approaches to image processing, analysis, and synthesis
- computational geometry
- image models and transforms
- visualization and graphical data presentation
- diagrammatic knowledge representation and reasoning
- monocular and stereo vision
- modeling of human visual perception
- innovative uses of various graphic and vision devices and systems

EVENT CHAIRS

- **Kwaśnicka, Halina,** Wrocław University of Science and Technology, Poland
- **Śluzek, Andrzej,** Khalifa University, United Arab Emirates

PROGRAM COMMITTEE

- **Dias, Jorge,** Khalifa University, United Arab Emirates
- **Foresti, Gian Luca,** University of Udine, Italy
- **Janusz, Andrzej,** University of Warsaw, Poland
- **Karhang, Maylor Leung,** Universiti Tunku Abdul Rahman, Malaysia
- **Kasprzak, Włodzimierz,** Warsaw University of Technology, Poland
- **Kulikowski, Juliusz,** Institute of Biocybernetics and Biomedical Engineering, Poland
- **Sikos, Leslie F.,** University of South Australia, Australia
- **Subbotin, Sergey,** Zaporizhzhya National Technical University, Ukraine
- **Tomczyk, Arkadiusz,** Łódź University of Technology, Poland
- **Werghi, Naoufel,** Khalifa University of Science and Technology, United Arab Emirates

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 19–22
DOI: 10.15439/2018F388
ISSN 2300-5963 ACSIS, Vol. 16

Dataset Enhancement in Hair Follicle Detection: ESENSEI Challenge

Jan Jakubik
Wrocław University of Science and Technology
Faculty of Computer Science and Management, Department of Computational Intelligence
Wrocław, Poland
Email: jan.jakubik@pwr.edu.pl

***Abstract*—In this paper, a solution to ESENSEI data mining challenge concerning the analysis of microscopic hair images is described. The task of the challenge was to detect locations of hair follicles in closeup images of a human scalp. The proposed solution is based on a convolutional neural network architecture. To improve generalization performance, we enhance training and test datasets using image transformations applied to both input and output. The chosen transformations are two axis symmetries and switching axes, all of which are possible to apply regardless of resolution without producing interpolation artifacts. Since these can be combined, $2^3 = 8$ possible views of each image can be created to expand both training and test data. We demonstrate the effects of dataset enhancement in both training and classifying on results achievable on the competition dataset. The solution placed 2nd in the final challenge evaluation.**

I. Introduction

IN RECENT years, convolutional neural networks (CNN) have become the standard for machine learning tasks concerning image analysis. Following the introduction [1], breakthrough results in recognition of handwritten digits [2], and significant speed improvements achieved with GPU computing [3], over the past decade the deep learning approach has been widely recognized as the current state of the art in image classification tasks. The capability of convolutional networks to perform well on raw data has freed machine learning researchers of the need to design features with specific invariance properties that are expected in image analysis (invariance to scaling, rotations, transposition). Simultaneously, developments in areas of GPU processing and big data made it more feasible to deploy these complex architectures, requiring large amounts of training data, on real datasets for practical problems.

The weaknesses of convolutional networks related to the complexity of the models have been since an object of attention. In widely publicised result presented in [4], an observation was made that it is possible to completely "confuse" neural networks by altering images in ways imperceptible to humans. This has shown that the CNN's generalization ability still does not directly correspond to our perception of visual data. One possible approach to lessen this problem comes in form of data enhancement: creation of additional samples in the dataset that cover potential cases in which the network could not otherwise learn by generalizing training data. These can include noise disturbances, adversarial samples, and transformations such as rotation by few degrees or scaling.

In this paper, we describe a CNN-based solution developed during the ESENSEI data mining competition [5] to the task of detecting hair follicles in microscopy images. The crucial insight that allowed the proposed approach to place 2nd in the final evaluation is that expanding the dataset by rotation and symmetrical reflections can significantly improve results. However, due to the nature of the data (specifically, the output being a low-resolution binary image), enhancing the data using transformations that require interpolation does not produce good results. Therefore, the final submission limited transformations to combinations of three different ones that do not require interpolation.

The paper is organised as follows: in section II, challenge data is described and analysed. Section III outlines the implemented approach: preprocessing applied to the data, convolutional network architecture and data enhancement. In section IV we experimentally demonstrate the influence on data enhancement on results. Section V summarizes the conclusions.

II. Challenge Description

The task of ESENSEI data mining challenge was to predict the locations of hair follicles on microscopic images. The dataset consists of 1920x1080 images in color. Fig. 1 shows sample images from the dataset, demonstrating the significant variability in hue, lighting, hair color and the number of follicles on each image. The dataset has 4880 images for training and 1000 for testing. During the competition, leaderboard evaluation was based on 10% of the test set.

Hair follicles are marked on a 16x9 grid. Output variables are binary, i.e., either the 144x144 square corresponding to the position on the grid contains a follicle or does not. The submissions format was the same as annotations, i.e., binary values on a 16x9 grid. These were evaluated using F-score as a metric, with F-score calculated for each image separately and then averaged over the dataset.

III. Method Description

The proposed approach uses a fairly standard convolutional neural network architecture with downscaled images as an

©2018, PTI

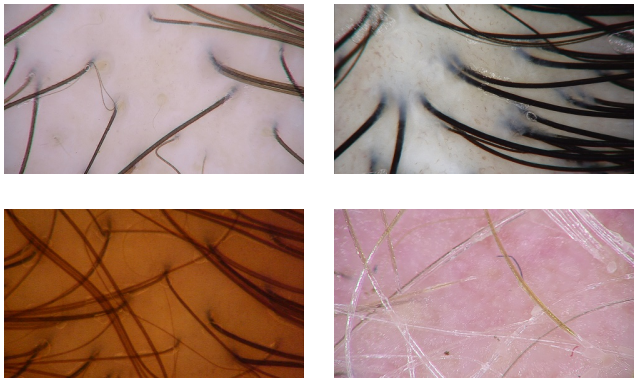

Fig. 1: Sample images from training data with significant differences in hue, lighting, hair color and hair density

input. As CNNs are known to function well with raw data, we strived to minimize the preprocessing steps.

A. *Data preprocessing*

All images are first resized to 1/8 of the original height and width. This partially removes significant noise present in the challenge images, while making processing of the images significantly less resource-heavy. Then, simple normalization is applied by first scaling all channels to (0,1) interval by dividing them by 255, and then transposing each channel to zero mean. This is done for each image separately to alleviate differences in hue and lighting between images.

B. *Network Architecture*

The neural network employed for follicle detection consists of convolutional layers and pooling only. Each convolutional layer utilizes ReLU activation and 3x3 filters. Layer size was set to 40 for all layers. Full network is shown in Fig. 2. A single residual connection [6] between layers is used, adding the output of the 1st layer to the output of layer 4 as the input of layer 5. This allows faster training given the model depth. We built this architecture though iterative deepening, i.e., we started with 6 layers of convolution with 2 pooling layers in between, and then added new layers as long as performance improvements were seen on both validation set during training, as well as leaderboard after uploading test set results. The residual connection was added when problems in propagating gradients first appeared at 10 layers depth.

With two pooling layers, the image is downscaled 15 times, to 16x9, which means the network output can be compared directly with the ground truth grid. As the problem is effectively mapping images onto images, we did not find it necessary to use any fully connected layers on top of the network. For optimization objective, we chose standard Mean Square Error loss. The network weights are optimised with Adadelta adaptive gradient descent method [7].

As the output is real-valued, thresholding must be applied to obtain locations marked as follicles. The threshold that has to be exceeded in order to mark a particular pixel in the 16x9

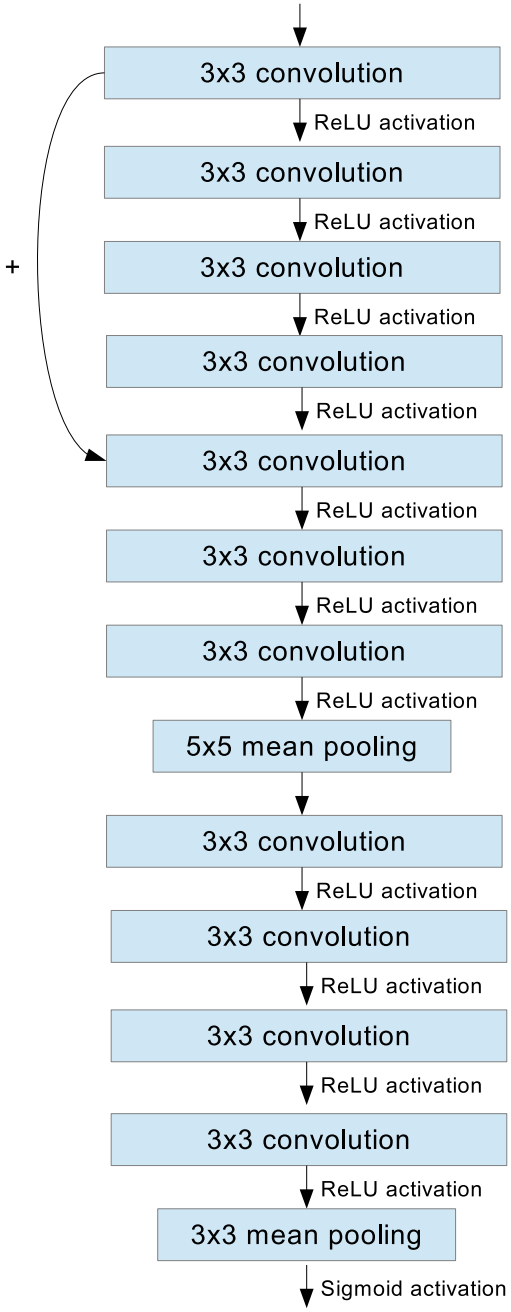

Fig. 2: CNN architecture for the final submission. Plus sign denotes a residual connection.

Fig. 3: 20x20 resolution picture and its rotation by 10 degrees. Interpolation results are clearly seen and will affect training of a CNN.

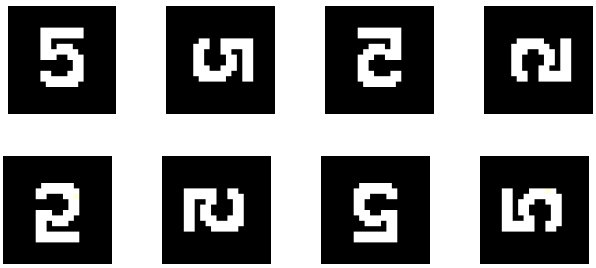

Fig. 4: Possible transformations resulting from different combinations of two axis symmetries and axis switching.

output as a follicle is chosen based on validation set results. We test treshold values from 0.01 to 0.5 (every multiple of 0.01) and choose the one that maximises F1 score on the validation set. The split of data into training and validation is done randomly, using 4400 images out of 4880 for training.

C. Dataset Enhancement

As convolutional networks do not possess invariance properties regarding scaling or rotation naturally and are expected to learn these from data, we utilize data enhancement to improve results.

For a model where (x, y) is a valid input-output training sample, we seek transformations T for which $(T(x), T(y))$ is also a valid input-output training sample that can be added to the data. Rotations and symmetries are among these transformations - the locations of follicles in a rotated image should be possible to obtain by rotating the follicle locations on the original image, and the same is true for symmetric reflections. However, due to the very small resolution of the output grid, interpolation artifacts produced by rotations that are not multiples of 90 degrees could influence results in a significant way. Fig. 3 presents results of rotation requiring interpolation on a 20x20 resolution picture. Due to concerns of interpolation affecting the output values, we limit the transformations to three that do not require interpolation:

- X axis symmetry
- Y axis symmetry
- switching X and Y axes

By combining these, we can obtain 8 possible transformations, including rotations by multiples of 90 degrees. These are shown in Fig. 4.

Additionally, some other enhancement transformations that could be considered are scaling and random noise distortion. However, we decided not to include these in the dataset based on our intuitions concerning the data. Enhancing the training dataset is applicable only when the enhancement improves generalization ability on the test data. For rotations and symmetries, it can be intuitively justified. As the hair on training and test images grow in different directions and at different angles, we would like training samples to cover most angles possible, and one way to do this is to expand the data with rotations and symmetries.

For random noise and scaling, such intuitions are not in place. Random noise distortions would be justifiable if there was a noise pattern present in test data that was missing from training data and can be replicated. We failed to identify such a pattern. Scaling, similarly to rotations, would require interpolation, with the possible exception of scaling by an integer number. However, an upscaled picture will show hair at a different zoom level than the original photos. Our knowledge of microscopic hair imaging is insufficient to tell whether widely varying zoom levels are expected between training and test data.

In practice, the dataset enhancement is implemented as follows: during training, for a particular batch, we apply each of the three listed transformations with 50% probability. During testing, all 8 possible views are considered by the same neural network, and a position on the grid is considered a follicle if in more than half of the views it was marked as a follicle.

IV. Results

Fig. 5 demonstrates the performance on validation set over the duration of training. For obtaining these results, the same training-validation split was used for each experiment.

The comparison includes proposed approach applied to the training data, test data or both, as well as a baseline with no dataset enhancement. We also test enhancement with random rotations with rotation angle uniformly sampled from $(-10, 10)$ interval, applied in addition to already mentioned symmetries and axis switching. This is to test whether the problem of interpolation outweighs the potential gains from expanding the data by a wider range of rotations.

It can be seen that the biggest boost in performance is achieved by enhancing the training set. Without it, the network starts overfitting very fast. This can be seen in the graph as decreasing performance on the validation set from epoch 10 onwards. Networks trained with training data enhancement do not overfit within the 100 epochs period, but it can be seen that adding an expanded range of rotations worsens performance instead of increasing it.

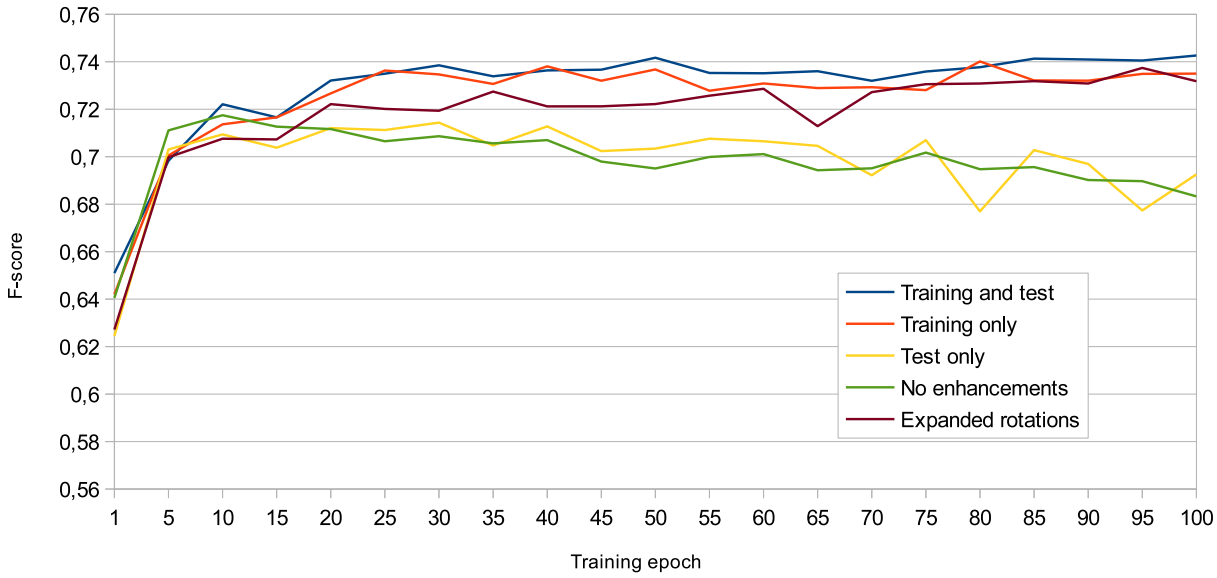

Fig. 5: Performance over the course of training, depending on where the data enhancement was applied. Expanded rotations include rotations between -10 and 10 degrees, possibly requiring interpolation.

TABLE I: Results of the competition - top submissions

Team	RMSE
rs	0.76299467
jj	**0.74934757**
podludek	0.71388834

Based on these results, the approach of enhancing training and test data with symmetries and axis switching was chosen for the final competition submission. Scores achieved by top 3 submissions in the competition are shown in Table 1.

V. Conclusions

We have described an approach to the task of hair follicle detection, based on a CNN trained with dataset enhancement. The proposed approach to the task has proven successfull in ESENSEI data mining challenge, achieving 2nd place out of 16 submissions. Experimental results show that the good results are largely reliant on the chosen approach to enhancing the training and test data. In general, the described approach can be applied to any kind of image data, regardless of resolution, due to the fact that all applied transformations require no interpolation.

Our results showcase the importance of dataset enhancement for training CNN. However, care should always be taken when choosing specific ways to enhance data. Tests with transformations that require interpolation on the competition dataset show that the produced artifacts have a negative impact on performance as expected, likely due to the small resolution of the desired output image. However, the observed improvement from utilizing rotations and symmetries is also an effect specific to the competition data. For hair imaging, the choice of symmetries and rotations by multiples of 90 degrees results from two basic insights. First is that we need to expand the range of angles at which hair grow in training pictures, and second is that the very low resolution of the output grid prevents us from using transformations with interpolation. In practice, any approach to enhancing a dataset should be reliant on reasonable intuitions concerning the particular type of data.

Acknowledgements

We would like to thank eSensei and Knowledge Pit for providing the data and a platform for the competition.

References

[1] LeCun, Yann; Leon Bottou; Yoshua Bengio; Patrick Haffner (1998). "Gradient-based learning applied to document recognition". Proceedings of the IEEE. 86 (11): 2278-2324
[2] Ciresan, Dan; Meier, Ueli; Schmidhuber, Jurgen (June 2012). "Multi-column deep neural networks for image classification". 2012 IEEE Conference on Computer Vision and Pattern Recognition.
[3] Dave Steinkraus; Patrice Simard; Ian Buck (2005). "Using GPUs for Machine Learning Algorithms". 12th International Conference on Document Analysis and Recognition (ICDAR 2005). pp. 1115-1119.
[4] Goodfellow, Ian, Shlens, Jonathon, Szegedy, Christian. (2014). Explaining and Harnessing Adversarial Examples. arXiv 1412.6572.
[5] https://knowledgepit.fedcsis.org
[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, Deep Residual Learning for Image Recognition, arXiv preprint arXiv:1512.03385, 2015
[7] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method", arXiv preprint arXiv:1212.5701, 2012.

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 23–28
DOI: 10.15439/2018F389
ISSN 2300-5963 ACSIS, Vol. 16

An ensemble of Deep Convolutional Neural Networks for Marking Hair Follicles on Microscopic Images

Łukasz Podlodowski, Szymon Roziewski, Marek Nurzyński
National Information Processing Institute
Natural Language Processing Laboratory
al. Niepodległości 188b 00-608 Warsaw, Poland
Email: lpodlodowski@opi.org.pl, sroziewski@opi.org.pl, mnurzynski@opi.org.pl

***Abstract*—This paper presents an application of a Convolutional Neural Network as a solution for a task associated with ESENSEI Challenge: Marking Hair Follicles on Microscopic Images. As we show in this paper quality of classification results could be improved not only by changing architecture but also by ensemble networks. In this paper, we present two solutions for the task, the first one based on benchmark convolutional neural network, and the second one, an ensemble of VGG-16 networks. Presented models took first and third places in the final competition leaderboard.**

I. Introduction

IN THIS paper, we present our solution to the ESENSEI Challenge: Marking Hair Follicles on Microscopic Images. We used Deep Convolutional Neural Networks (DCNNs) models [1] to solve the given problem. Our final result consisting of an ensemble of DCNNs models was the best performing solution of the challenge. DCNNs training was a massive task, which took a few days using GPU based computation. Before that, we had tested different architectures and did many experiments in order to find a better approach. Subsequently, we trained derived architectures with gentle modification applied. After all, we prepared an ensemble which achieved better score with F-measure. In this paper, we also present 3rd place solution of the challenge based on benchmark CNN model.

II. Related work

We found one study dealing with intelligent image analysis of hair follicles. In [2], advanced automatic target recognition (ATR) algorithms for identification of hair follicles were used. The approach was not related to neural networks, though.

Instead of learning convolutional filters, the authors used the ATR method which performed wavelet filtering of the targets for enhancing the contrast of hair features in the images. Subsequently, the collected features were sent to Adaboost classifier for training and recognition. The system itself was able to provide the intended hair follicle locations. The follicle locations were isolated by continuous wavelet transform in a grayscale image. The authors observed that plotting a grayscale picture of skin with follicles in three dimensions, lead to a print with local minima in the surface plot, in which hair follicles resided. The cross-section of a hair follicle resembled an inverted version of the second derivative of a Gaussian wavelet, referred as Mexican hat wavelet. The correlation between both, cross-section and a Mexican hat wavelet was used to locate a hair follicle. Finally, correlations were used to find regions of interest from the image. Subsequently, features for follicles identification were extracted e.g. entropy, skewness, mean, standard deviation, feature area, minimum value, correlation peak and Euler Number.

The first stage of the ATR system was the optical correlation operation, with the intentionally low set threshold. The output contained many false positives, which were reduced in the second stage, by application of adaptive boosting. The lower bound of prediction performance was 77 % with a maximum of 2 false positives per image. However, the images used in [2] were not containing any hair, but only hair follicles, which seemed to be a fairly easier task to solve.

III. Problem description

The dataset consisted of microscopic digital images of a human scalp. The picture was in full HD resolution, with size 1920x1080 pixels. Each image was divided into equal non-overlapping 144 sectors. Each sector forms a subimage square with size 120x120 pixels. It means that each original image was made of 144 bitmaps, framed in 16 columns and 9 rows. Those subimages represented an defined input of classification problem and contained the visual representation of possible the hair follicle occurrence. In another bitmaps file, the knowledge reflecting the presence of follicles was collected. If a subimage contained at least one follicle, then it was marked by 1, otherwise by 0. Finally, the task was to classify all sectors from an image.

The training set contained 4880 full HD pictures in jpeg format, with corresponding bitmaps containing encoded follicle presence. However, dividing this images into 144 sectors provided 702720 training samples. Approximately 14% of this samples had been labeled as positive.

The test set was made of 1000 images. The task was to predict the follicle positions in the test pictures by classification of each sector. Each prediction for an image had to contain

©2018, PTI

144 binary values, concerning 16x9 square areas making one image entity. Finally, 144000 test samples were classified.

IV. Data processing

A. Data preprocessing

We applied some preprocessing steps in order to improve quality of classification. Those steps were performed on both sets, training and test. We treated all pictures as one channel representing an image in grayscale. We experienced that using 3 color channels was not worth all the effort needed, greatly increasing input data size without significant result improvement. All image data were standardized with respect to the training and test set. Since the images were with size 120x120 pixels, we scaled them to lower resolutions for different models, i.e. 24x24, 36x36 and 48x48 pixels. We were concerned about prediction quality on the sector edges. We also recognized classified that a sector could strongly rely on information from neighbour sectors, i.e. recognizing hair geometry in neighbourhood could provide important information about hair facility occurrence in the analyzed sector. In order to make it more robust, we merged near neighbourhood areas with analyzed subimage, making one larger new input image, to better recognize follicles on the image boundaries. We used a few approaches to this problem, one of them used whole neighbouring sectors. As result, we used area of 9x9 sectors, where classified sector was always in the center of the area. In this way, we created new datasets, with different surrounding cell sizes. The first set was made of images in size 24x24 pixels, joined with their 8 neighbours (also 24x24 px). As a result, we got 72x72 pixels images. We obtained further two datasets, but in this case, we limited the surrounding area by lowering the size of the neighbouring squares by factor 2. Thus, we produced new input images with sizes 72x72 and 96x96 pixels. For the boundary images, when no neighbours were available, we applied images filled with zeros.

B. Data augmentation

Data augmentation is a very common technique in image classification, used for the training dataset enlargement. What we had to do, was to perform minor alterations of existing images. Those minor changes might be following: translations; rotations; both, horizontal and vertical flipping; zooming or more sophisticated transformation. The neural network would treat such generated picture as a distinct one, and thus benefit from it. ”Data augmentation has been shown to produce promising ways to increase the accuracy of classification tasks” [3]. This technique ”has played an active role in achieving state-of-the-art results on many vision tasks” [4].

V. Architecture of network

We used two different architectures for each solution. In this section, we describe both of them and final models based on a benchmark CNN and an ensemble of deeper CNNs based on VGG-16 architecture. We also want to discuss the problem of damping position information in CNN.

A. Position information in CNN

As we said, our input data included neighbour sectors which did not represent directly visual information about the classified sector. It was very important to keep information about that signal of hair facility occurrence raised from central sector of input data. CNN could damp this information by their architecture nature. Convolutional filters are applied to the whole image and allow to propagate information about pattern occurrence in the specific area. However, the output of filters is normally processed by pooling layer which makes ”mixture” of information from nearest sectors. In the most common approach, the max signal value is chosen from this area. A solution to this problem could be removing pooling layers from a network. However, pooling layers allow to greatly decrease the number of parameters in the network, reducing the size of an analyzed topology of outputs from the previous convolutional layer. This property of pooling is very desired in CNN and allow to greatly simplify model complexity. As shown in [5], in some cases CNN could not include pooling layers and perform parameter reduction by applying a bigger stride of a filter. Whereas, it can perform with similar results as classical CNN architecture with pooling layer. This approach allows keeping pattern occurrence position information through whole signal propagation. However, the bigger stride could also cause skipping some patterns in the analyzed area. Stride-based architecture could also have the problem with generalization made network rigidly associated with input data pattern positions, which in turn, could lead to generalization problem. Based on this information we decided to use average pooling, which does not damp position information from the specific area like max pooling does but make this information as a mixture from neighbour areas which finally provide noised information about pattern occurrence position.

The problem of pattern position also limited possible data augmentation techniques, i.e. shifting could noisy relation between the input and the desired output.

B. VGG-16

We made use of cross-validation to choose better models and to estimate learning rate parameter. Finally, we have experienced that a VGG-16 [6] neural network is outperforming other architectures. We applied many experiments with different dense layers, class weights and image sizes. Besides the architecture, the most important factor, in well-performing prediction, is image size. We have observed that F1-score value grows in general, with image resolution. Higher resolutions lead to better results. On the other hand, increasing image size demands more memory in a graphics card and slows down the training process. Likewise, training of a high-resolution model demands more data in general. Thus, we had to balance all that criteria to build a suited and well-performing model. All in all, we built on top of solution an ensemble model made from four models.

The VGG-16 architecture consisting of 23 layers is shown in the Figure 1.

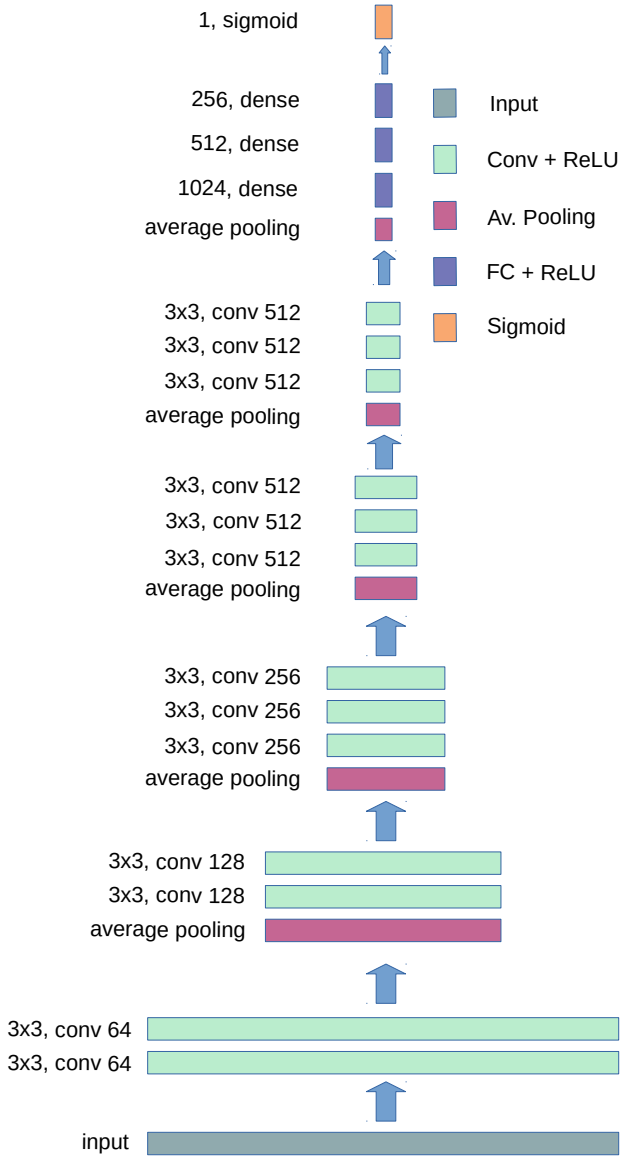

Fig. 1. The architecture of VGG-16 deep convolutional neural network. Different input sizes were used. Before each of convolution and average pooling layers, zero-padding and batch normalization were applied, respectively. After each VGG-16 section, we performed dropout layers to avoid over-fitting. Regarding dense layers, we added dropout and batch normalization after each dense layer.

Instead of max pooling, we used the average, which seemed to perform a bit better. This decision had also analytic background associated with information about pattern position in the image.

We have also modified the number of dense layers, we have four of them. We used dropout to prevent model from overfitting [7] and batch normalization technique [8] to scale and adjust activation functions outputs. Zero padding technique removes unwanted boundary effects. As a loss function, we adapted binary cross-entropy with Adam optimizer for gradient descent optimization.

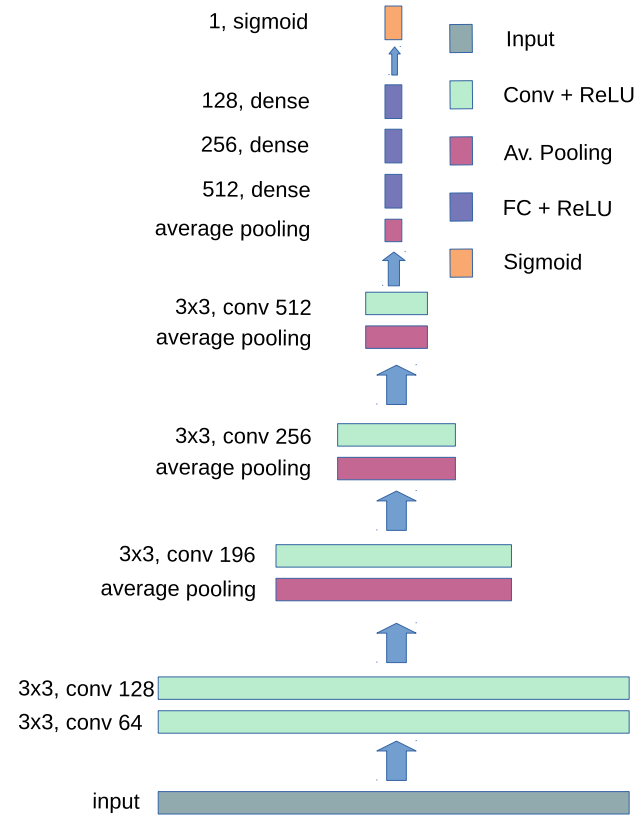

Fig. 2. The architecture of benchmark CNN used as one of the solutions for a defined problem.

VI. Models

Our solution are based on convolutional neural network which are the state-of-the-art models for many image classification problems[9][10][11]. For training networks, we used Adam optimizer[12] which "has been immensely successful in development of several state-of-the-art solutions for a wide range of problems" [13].

A. Training

We applied the learning rate of 0.005 for all models. We have used Adam optimizer for optimization, β_1 and β_2 parameters were set close to 1. This method allows to easily adapt to data and model architecture without a rich parametrization. Networks were trained on 90 percentages of data, rest of set was used for model validation. The early stopping technique was used based on a progress of evaluated on validation data F1-score metric value. Network training optimized binary cross-entropy which do not correspond directly to F1-score. However, we could decrease the gap between both optimization goals by manipulation class weights which could allow us to find the balance between optimizing precision and recall. Different strategies were chosen for different networks. Figure 3 presents metrics values during a training process. It is noticeable that shallow network train is more stable, with fewer

outliers points. The mean values of loss functions and f1 scores for models m_1, m_2, m_3, m_4 for both training and validation sets are given in tuples respectively: $(0.341 \pm 0.052, 0.715 \pm 0.035)$ and $(0.214 \pm 0.026, 0.715 \pm 0.023)$. One can observe the equality of F1-scores for training and validation sets.

B. Benchmark CNN

At first, we want to describe CNN architecture which took 3rd place in the competition. For convenience, we will call this model as benchmark CNN. Figure 2 presents the architecture of model. To prevent this network from an overfitting problem, between all convolutional layers we used a dropout with $p = 0.4$, for fully connected layers $p = 0.5$. Batch normalization allowed to damp too strong signals spikes. This network was processing images with 48x48 pixels representing an area of 9x9 sectors. The classified sector was always in the center of the area. Class weights were fixed as 3 per positive class and 1 per class negative class. This value was adjusted experimentally. In next sections, we will discuss and compare results of this model with an ensemble of VGG-16 networks. In our experiments, we treat the results of this model as a benchmark. The model consists of 2,3 mln trainable parameters.

C. DCNN ensemble

As we stated before, we trained 5 models with respect to the given DCNN architecture in Table 1. As shown in [14] ensemble of CNN could enhance results quality. For convenience, we will call members of ensemble m_1, m_2, m_3, m_4, m_5.

Our two models m_1, m_2 have 72x72 pixels images on input. The input image construction has been mentioned in Section IV. Basically, one image is taken from 9 original subimages. Those models differ only in class weights. Although there are almost 6 times more negative samples than positive, we found that the class weights 2 or 3 perform better. The first model has class weights 1 and 2, and the second one has 1 and 3, where the first weight is for class 0 and second for class 1. We monitor the training process to not overfit the model, by using early stopping method. We feed the model under training with a validation dataset, in order to stop training if the model performance measured in F1-score is getting lower, while we keep the best performing model at hand. Both of the models contained roughly 17.5 mln trainable parameters each. The validation set was a 10% subset of total training data.

Subsequently, we have a model m_3 with input images 72x72 pixels, but with higher resolution. The central cell is made from 36x36 pixels image. We kept the same setup and applied class weights 1 and 3. This model has the same number of parameters to estimate. The outcome was slightly better than in the previous approach.

In the end, we trained a model m_4 with the highest resolution. The training itself was a massive task. Since we applied the input images with 96x96 pixels, we had to compute values of more than 20 mln parameters. We trained this model like others with using early stopping. But this time we wanted to re-train a new model m_5 from scratch, with make use of

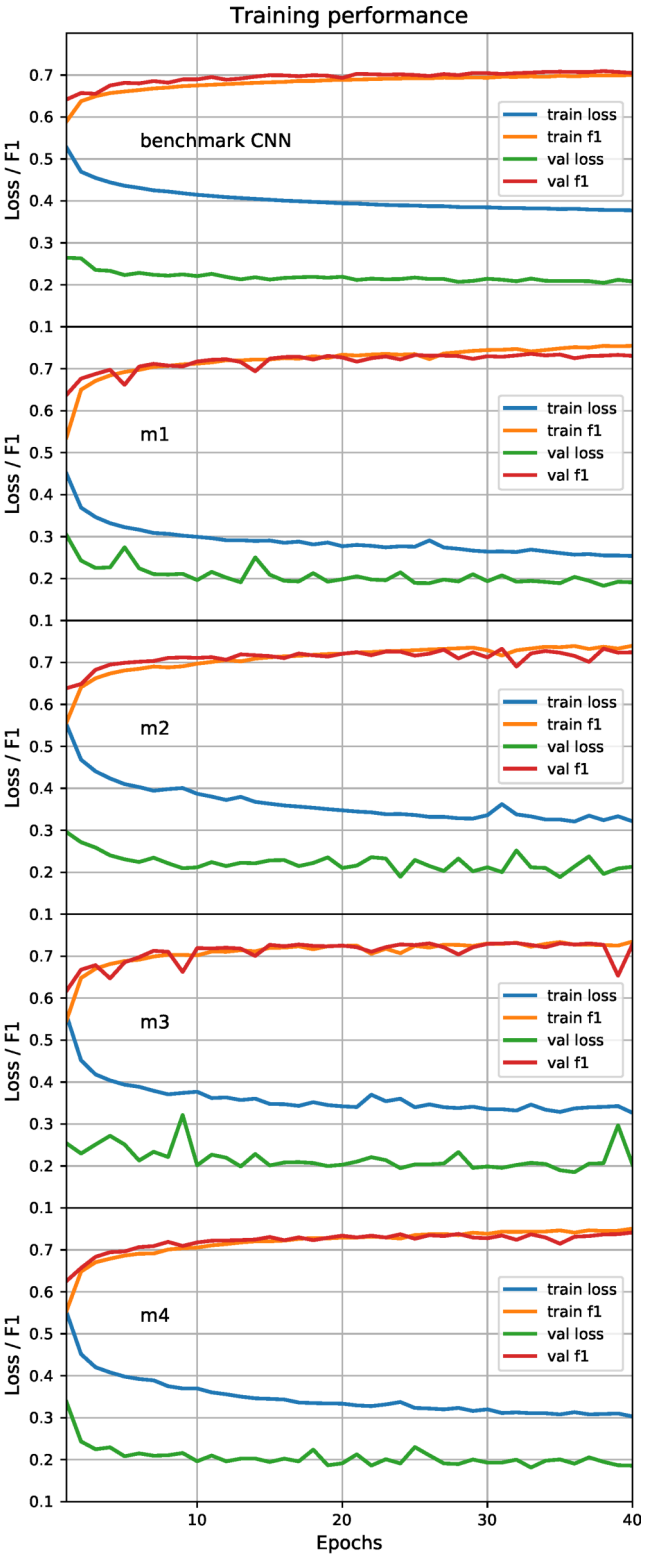

Fig. 3. Training performance charts, for each models we present both, loss function and F1 score performed on training and validation data sets.

all data, including validation data. During the first training, we estimated the number of epochs needed to establish the parameter. Then, in the next round, we just added 2 epochs more and started computing.

In the end, we made an ensemble of those described 5 models, which performed nicely on the leaderboard.

In the Figure 3 shown measurements from a training process for all models from the ensemble. F1 score values and binary cross-entropy loss function values seem to not always been correlated in outliers points. However, trends on both function are similar. We will discuss this problem in the next section.

VII. Results

Our models m_1, m_2 are trained for more than 30 epochs, each epoch lasts for about 33 minutes. F-measure calculated on the validation data set is about 0.736. The model m_3 has a bit higher F1 score – 0.738. The last model m_4 has been trained for 37 epochs with F-measure 0.741. The epoch time length was close to 54 minutes on GPU P40 machine. It means that training of the biggest network took more than 33 hours of excessive computations.

After all, we created an ensemble of 5 models: m_1, m_2, m_3, m_4, m_5 by computing arithmetic average based on predicted probabilities of the follicle presence.

TABLE I
Experimental results - classification quality based on crossvalidation

Model	F1 score	Binary cross-entropy loss
benchmark CNN	0.7132	0.2116
m_1	0.7357	0.1859
m_2	0.7329	0.1960
m_3	0.7384	0.1913
m_4	0.7415	0.1858

As shown in table I and fig. 5 m_1's loss function is almost equal to m_4's and m_3's loss function results. However F1 score difference between this models is noticeable (fig. 4). The main reason for this behaviour is associated with lack of direct relation between F1 score and optimized binary cross-entropy function. F1 score could be seen as second level training quality function. It is not directly implied by cross-entropy but strongly correlated. This correlation is depended on a prior distribution of classes in training set. However, we could influence on strength of this correlation by changing class weights. The training set was unbalanced and contained approximately 14% of positive classes samples. This state leads optimization process to focus on precision of positive class classification. Increasing positive class weight leads model to choose a positive class as an output more frequently which implies increasing a recall importance in an optimization process. F1 score is the harmonic average of the precision and recall, which means that desired optimization method should focus equally on precision and recall to reach the best results.

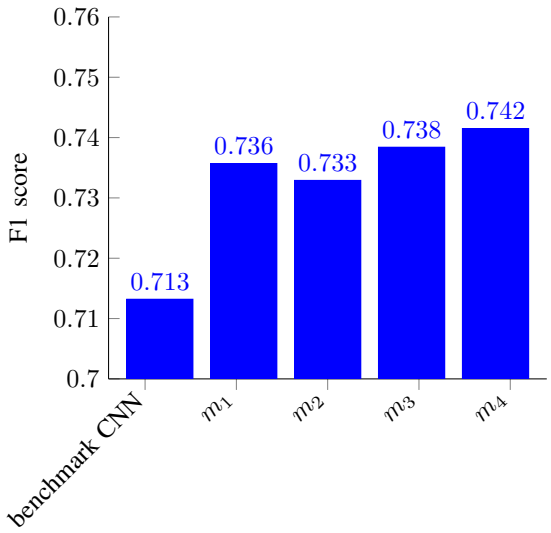

Fig. 4. F1 score results for specific models based on 3-fold cross-validation

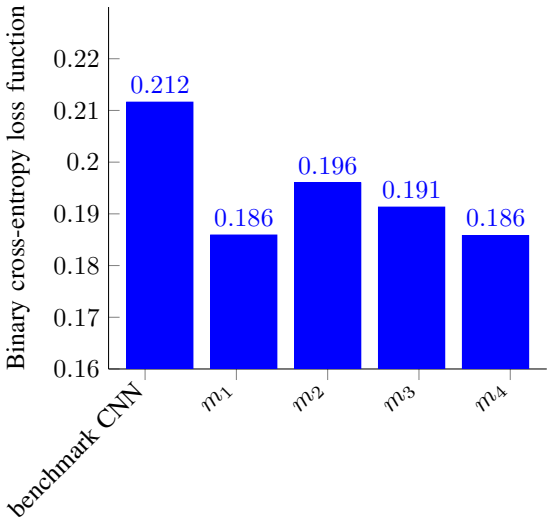

Fig. 5. Binary cross-entropy loss function results for specific models based on 3-fold cross-validation

The second reason could be damping output signal. A network could prefer to minimizing error penalty by taking less risk in making the decision and return probabilities near threshold value of 0.5. This behaviour could have many reasons. First of them could be wrong labeling in training set, which forces the network to generate different output from almost the same inputs. The second one could be a too limited capacity of the model to handle data and force network to bring outputs closer to a prior distribution. A possible solution for increasing a quality of the final class prediction could be changing a threshold or manipulating class weights. However, class weights manipulation had the influence on an entire training process which allows us to use directly optimization process as a tool to solve problems with unbalanced data. This

the reason we choose this option to handle this problem.

The results achieved by VGG-16 models outperform benchmark CNN model which shows that deeper network architecture allows to better solve this classification task. Finally, an ensemble of VGG-16 models improve classification quality and took first place in the competition with the final score of 0.763.

REFERENCES

[1] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85 – 117, 2015. doi: https://doi.org/10.1016/j.neunet.2014.09.003

[2] B. Walker, T. Lu, and T.-H. Chao, "Intelligent image analysis for image-guided hair removal and skin therapy," vol. 8207, p. 820707, 2012. doi: 10.1117/12.910741. [Online]. Available: http://adsabs.harvard.edu/abs/2012SPIE.8207E..07W

[3] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *CoRR*, vol. abs/1712.04621, 2017. [Online]. Available: http://arxiv.org/abs/1712.04621

[4] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard, "Adaptive data augmentation for image classification," 01 2016.

[5] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," *CoRR*, vol. abs/1412.6806, 2014. [Online]. Available: http://arxiv.org/abs/1412.6806

[6] W. D. Shuying Liu, *Very deep convolutional neural network based image classification using small training sample size*. IEEE, 2015.

[7] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2627435.2670313

[8] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[10] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, Aug 2003. doi: 10.1109/ICDAR.2003.1227801 pp. 958–963.

[11] D. C. Cireşan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, ser. IJCAI'11. AAAI Press, 2011. doi: 10.5591/978-1-57735-516-8/IJCAI11-210. ISBN 978-1-57735-514-4 pp. 1237–1242. [Online]. Available: http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-210

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[13] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=ryQu7f-RZ

[14] M. Izadyyazdanabadi, E. Belykh, M. Mooney, N. Martirosyan, J. Eschbacher, P. Nakaji, M. C. Preul, and Y. Yang, "Convolutional neural networks: Ensemble modeling, fine-tuning and unsupervised semantic localization," *CoRR*, vol. abs/1709.03028, 2017. [Online]. Available: http://arxiv.org/abs/1709.03028

Computer Science & Systems

CSS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to more technical aspects of computer science and related disciplines. The CSNS area spans themes ranging from hardware issues close to the discipline of computer engineering via software issues tackled by the theory and applications of computer science and to communications issues of interest to distributed and network systems. Events that constitute CSNS are:

- BEDA'18—1st International Workshop on Biomedical & Health Engineering and Data Analysis
- CANA'18—11th Computer Aspects of Numerical Algorithms
- C&SS'18 - 5th International Conference on Cryptography and Security Systems
- CPORA'18—3rd Workshop on Constraint Programming and Operation Research Applications
- LTA'18—3rd International Workshop on Language Technologies and Applications
- MMAP'18—11th International Symposium on Multimedia Applications and Processing

Area Supervisory Committee

- Burdescu, Dumitru Dan, MMAP'18
- Damasevicius, Robertas, LTA'18
- Ge, Mouzhi, DaSCA'18
- Janicki, Artur, BigDAISy'18
- Królak, Aleksandra, BEDA'18
- Ksiezopolski, Bogdan, C&SS'18
- Paprzycki, Marcin, 4A'18
- Ristov, Sashko, WSC'18
- Sitek, Pawel, CPORA'18
- Stpiczyński, Przemysław, CANA'18

3rd International Workshop on Language Technologies and Applications

DEVELOPMENT of new technologies and various intelligent systems creates new possibilities for information processing. Natural Language Processing (NLP) addresses problems of automated understanding, processing, evaluation and generation of natural human languages. LTA workshop provides a venue for discussion and presenting innovative research in NLP domain, but not restricted, to: computational and mathematical modeling, analysis and processing of any forms (spoken, handwritten or text) of human language, interactions via Virtual Reality and Augmented Reality, Computational Intelligence models and applications but also other various applications in decision support systems. We welcome papers covering innovative applications and practical usage of theoretical aspects. The LTA workshop will provide an opportunity for researchers and professionals to discuss present and future challenges as well as potential collaboration for future progress in the field.

TOPICS

The submitted papers shall cover research and developments in all NLP aspects, such as (however this list is not exhaustive):

- Computational Intelligence methods applied to language & text processing
- text analysis
- language networks
- text classification
- language networks, resources and corpora
- document clustering
- various forms of text recognition
- machine translation
- intelligent text-to-speech (TTS) and speech-to-text (STT) methods
- authorship identification and verification
- author profiling
- plagiarism detection
- sentiment analysis
- NLP applications in education
- knowledge extraction and retrieval from text and natural language structures
- multi-modal and natural language interfaces
- innovative language-oriented applications and tools
- interactions models and applications via Virtual Reality and Augmented Reality
- NLP for text analysis in forensic linguistics and cybersecurity

EVENT CHAIRS

- **Damasevicius, Robertas,** Kaunas University of Technology, Lithuania
- **Martinčić – Ipšić, Sanda,** University of Rijeka, Croatia
- **Napoli, Christian,** Department of Mathematics and Informatics, University of Catania, Italy
- **Woźniak, Marcin,** Institute of Mathematics, Silesian University of Technology, Poland

PROGRAM COMMITTEE

- **Artiemjew, Piotr,** University of Warmia and Mazury, Poland
- **Burdescu, Dumitru Dan,** University of Craiova, Romania
- **Calixto, Iacer,** University of Amsterdam, The Netherlands
- **Čukić, Bojan,** UNC Charlotte, United States
- **Cuzzocrea, Alfredo,** University of Trieste, Italy
- **Dobrišek, Simon,** University of Ljubljana, Slovenia
- **Ganchev, Ivan,** University of Limerick, Ireland / University of Plovdiv "Paisii Hilendarski", Bulgaria, Ireland
- **Gelbukh, Alexander,** Instituto Politécnico Nacional, Mexico
- **Grigonytė, Gintarė,** University of Stockholm, Sweden
- **Harbusch, Karin,** Universität Koblenz-Landau, Germany
- **Kapočiūtė-Dzikienė, Jurgita,** Vytautas Magnus University, Lithuania
- **Krilavičius, Tomas,** Vytautas Magnus University, Lithuania
- **Kurasova, Olga,** Vilnius University, Institute of Mathematics and Informatics, Lithuania
- **Lopata, Audrius,** Vilnius University, Lithuania
- **Marszałek, Zbigniew,** Silesian University of Technology, Poland
- **Maskeliūnas, Rytis,** Kaunas University of Technology, Lithuania
- **Matson, Eric T.,** Purdue University, United States
- **Meštrović, Ana,** University of Rijeka, Croatia
- **Mikelić-Preradović, Nives,** University of Zagreb, Croatia
- **Nowicki, Robert,** Czestochowa University of Technology, Poland
- **Połap, Dawid,** Institute of Mathematics, Silesian University of Technology, Poland
- **Pulvirenti, Alfredo,** University of Catania, Italy
- **Rosen, Alexandr,** Charles University, Czech Republic
- **Sanada, Haruko,** Rissho University, Japan

- **Skadina, Inguna,** University of Liepaja, Latvia
- **Śluzek, Andrzej,** Khalifa University, United Arab Emirates
- **Šnajder, Jan,** University of Zagreb, Croatia
- **Stanković, Ranka,** University of Belgrade, Serbia
- **Starczewski, Janusz,** Czestochowa University of Technology, Poland
- **Steinberger, Josef,** University of West Bohemia, Czech Republic
- **Szymański, Julian,** Gdansk University of Technology, Poland
- **Tahmasebi, Nina,** University of Gothenburg, Sweden
- **Tambouratzis, George,** Institute for Language and Speech Processing, Athena Research Centre, Greece
- **Tramontana, Emiliano,** University of Catania, Italy
- **Wang, Lipo,** Nanyang Technological University, Singapore
- **Wei, Wei,** School of Computer Science and Engineering, Xi'an University of Technology, China
- **Yanushkevich, Svetlana,** University of Calgary, Canada
- **Žabokrtský, Zdeněk,** Charles University

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 33–38
DOI: 10.15439/2018F381
ISSN 2300-5963 ACSIS, Vol. 16

From Building Corpora for Recognizing Faceted Entailment to Recognizing Relational Entailment

Martin Víta
NLP Centre, Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
Email: info@martinvita.eu

***Abstract*—Recognizing textual entailment (RTE) became a well established and widely studied task. Partial textual entailment—and faceted textual entailment in particular—belong to tasks that are derived from RTE. Although there exist many annotated corpora for the original RTE problem, faceted textual entailment is in the sense of easy-accessible corpora highly neglected. In this paper, we present a semi-automatic approach to deriving corpora for faceted entailment task from a general RTE corpus using open information extraction (open IE) tools.**

As a generalization of this approach and general principles of open IE, we introduce a notion of relational entailment and provide its basic properties and relations to other entailment-like issues. We would like to introduce the problem of relational entailment as an important task with potentially wide range of real-world applications.

I. Introduction

TEXTUAL entailment is a relation is between two texts or text snippets that expresses whether one text can be inferred from the other. Textual entailment has quickly become one of the most prominent tasks of NLP with a wide spectrum of (potential) applications – including multi-document summarization, plagiarism detection, machine translation evaluation etc. In the last years, a huge effort was made in the field of approaches and algorithms for recognizing textual entailment as well as in building annotated corpora for this purpose.

Nevertheless, "classical" recognizing textual entailment is not able to treat situations when the text is only "almost entailed". A notion of *partial entailment* was introduced by Nielsen et al. [1] to address this problem. The original motivation arises from automatic assessment of students' tests where it is necessary to cover situations when a certain answer is only partially correct. Facets, i. e., pairs of words accompanied by a semantic relationship connecting these two words together, were introduced in the same paper in order to present a model of decomposing the sentence into fragments and provide a basis for recognizing partial entailment. Later, in SemEval 2013 challenge, a pilot task of partial – faceted entailment in particular – was introduced [2].

Recognizing textual entailment is a classification task. Supervised methods are naturally a predominant approach in this setting. Especially deep learning approaches (often ensemble methods) currently highly outperform other methods [3], [4], [5]. Nevertheless, other methods than DNN are also investigated, see [6].

Deep learning methods generally require generally a great amount of training data. For textual entailment relevant sources, i. e., annotated corpora are already available, but, in contrast, in partial/faceted entailment, researchers are facing a problem of lack of the training data. One of the main goals of this paper is to suggest a method for deriving a new corpus for faceted entailment task using open information extraction tools from an existing RTE corpus.

Based on the idea of using open information extractors, we propose a novel task of relational entailment that can be viewed as a form of a partial entailment and as a modified generalization of faceted entailment.

II. Textual Entailment, Faceted Entailment and Corresponding Corpora

For completeness, we briefly recall the notion of textual entailment and provide a selection of notable corpora for this task.

A. Recognizing Textual Entailment – Task Definition

Textual entailment is defined as a directional relationship between pairs of text expressions, denoted by T – the entailing "Text", and H – the entailed "Hypothesis". We say that T entails H if humans reading T would typically infer that H is most likely true, see [7]. Deep insight into the nature of textual entailment is provided by a current work of Korman et al. [8].

Recognizing textual entailment (RTE) is fundamentally a binary classification task to decide whether a given text T entails a given hypothesis H – with a "boolean" answer. Later, a modification of this, so called two-way task, appeared: no entailment category was split into two classes: *unknown*, where the truth of H cannot be determined on the basis of T, and *contradiction*, where T contradicts H, [9].

Example. If the text T has a form of a sentence *Apple says the new Intel dual-core chips improve performance by up to 39 percent over the previous single-core variety.* and the hypothesis H is a sentence *Dual-core chips make the performance better.*, then a correct RTE system should label this item as ENTAILMENT.

©2018, PTI

As already mentioned, current deep ensemble architectures are beating other architectures in RTE task. [1]

B. Recognizing Textual Entailment – Notable Corpora

During the last decade, several annotated corpora for RTE task were created. Among most important corpora belong:

- *The Boeing-Princeton-ISI (BPI) Textual Entailment Test Suite* – one of the smallest and oldest corpora that contains only 250 pairs of pairs. The value of this corpus is that it provides items "syntactically simpler but semantically challenging, with the intension of focusing more on the knowledge rather than just linguistic requirements."[2]
- *RTE-1–RTE-8* – a set of corpora used in textual challenges organized under several frameworks, starting in 2004 as PASCAL Recognizing Textual Entailment (RTE) Challenges, later as the tracks of the Text Analysis Conference (TAC) and finally as a part of SemEval challenges. First RTE-1–RTE-5 were designated as collections of standalone (T, H) pairs, whereas later RTE-6–RTE-8 datasets were created in order to tackle more realistic scenarios, including knowledge base population, slot filling task etc. The last one was motivated by educational purposes: to assess the correctness of a student answers. Some of RTE-x corpora were annotated in a three-way manner. A comprehensive description of these corpora is available in [9].
- *The Stanford Natural Language Inference (SNLI) Corpus* – a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels entailment, contradiction, and neutral. One of the most important and the biggest corpora for RTE purpose [10].

Annotated corpora are available mostly for English, nevertheless there exist corpora for other languages such as Portuguese [11]. General issues of RTE corpora preparation, i. e., generating entailment pairs as well as connection between knowledge acquisition and RTE are discussed in [12].

C. Partial and Faceted Entailment

An ordered pair $(T; H)$ forms a partial textual entailment if *a fragment* of the hypothesis H is entailed by T. According to this definition, the fragment of the hypothesis is no more defined. Thus, we need to answer a question how to decompose the hypothesis into fragments. The model of facets tackles this problem.

A *facet* is an ordered pair of key terms (w_1, w_2) (that are contained in the considered hypothesis). These terms can be single words or multi-word expressions like "carbon dioxide" etc. These terms are connected together with a semantic relation. A simplified version of this approach – used in RTE-8 – deals only with the pair of these terms *without* the semantic relation mentioned explicitly.

[1] A comprehensive list of DNN successful architectures is available at: https://nlp.stanford.edu/projects/snli/

[2] See: http://www.cs.utexas.edu/users/pclark/bpi-test-suite/

For example, if the hypothesis has the form of a sentence *The water was evaporated, leaving the salt.*, one of corresponding facets is *(evaporated, water)*, a different one (leaving, salt), etc.

1) Recognizing Faceted Entailment – Task definition: Recognizing faceted entailment is again a binary classification task having three inputs: text T, hypothesis H and a facet (w_1, w_2) of words that occur in the hypothesis H. The task consists of determining whether T expresses, either directly or by implication, the same relationship between the facet words as in H, see [2]. Positive case is labeled by "expressed", other one as "unaddressed" (rather then entailment–no entailment).

2) SemEval 2013 Task 7 RTE-8 Corpus for Faceted Entailment: Nowadays, there exists only one widely accepted corpus for faceted entailment – a corpus used in SemEval 2013 Task 7, RTE-8: it is based on SciEntsBank corpus of student answers (together with corresponding questions and reference answers). The reference answers are broken down into the facets and annotated in order to support educational systems in recognizing when specific parts of a reference answer are expressed in the student answer, even if the reference answer is not entailed as a whole [1]. As mentioned in [2], "the correspondence between educational labels and entailment judgments was not perfect due to the difference in educational and textual entailment perspectives. Nevertheless, the two classes of assessment appeared to be sufficiently well correlated."

An item taken from SciEnts Bank / SemEval 2013 Task 7, RTE-8

QUESTION: *You used several methods to separate and identify the substances in mock rocks. How did you separate the salt from the water?*

STUDENT ANSWER: *Let the water evaporate and the salt is left behind.*

REFERENCE ANSWER: *The water was evaporated, leaving the salt.*

FACET: *(evaporated, water)*

In this case, the result is "Expressed" (thus the student's answer can be regarded as partially correct).

In contrast, when student answers "I don't know." the facet *(evaporated, water)* is obviously not expressed.

III. Open Information Extraction

Open Information Extraction (open IE or OIE for short), introduced by Banko et al. [13], is a task of extracting (ordered) textual n-tuples containing *relation phrase* and *its arguments* from a given sentence. For example, from a following simple sentence "FedCSIS took place in Prague.", an open IE system (open information extractor) should extract a textual triple (FedCSIS, took place in, Prague). Unless traditional information extraction concept, open IE systems *do not require* a predefined list of relations (given by relation vocabularies or ontologies).

Following [14], an open information extractor can be formally defined as a function from a document, d, to a set of triples, `{(arg1, rel, arg2)}`, where the `args` are noun

phrases and `rel` is a textual fragment indicating an implicit, semantic relation between the two noun phrases. The extractor should produce one triple for every relation stated *explicitly* in the text, but is not required to infer implicit facts. (For example, from the sentence "US president Donald Trump, visited Singapore.", it is not required to obtain a triple (Donald Trump, be president of, US)), nevertheless, some of current open information extractors provide a limited functionality in this field.

New generation of open IE systems as OpenIE4 and higher are also able to extract even n-tuples corresponding with semantic n-ary relations, $n > 3$, as well as nested relations (used for dealing with modal sentences like "Early scientists believed that earth is the center of the universe.", where extracting (Earth, is the center of, universe) is inappropriate – all described in detail in [15].

A. Open IE: Goal and Applications

Applications of open IE can be generally divided into two main groups [15]: end user tasks and NLP end tasks. End user tasks include primarily building Open knowledge bases that allow simple querying, fact finding, as well as browsing through information extracted from a collection of textual documents and/or summarizing information about given entities. Open IE approaches also provide a particular kind of data compression (a source free text $\rightarrow$ a collection of textual tuples), that preserves important information (and is still in a human-readable form).

NLP end tasks include text comprehension, computing sentence similarity and (multi-)document summarization. On more general level, open IE results, i. e. textual tuples, can be effectively used for training word embeddings. The results of "open IE embeddings" at lexical similarity and analogy tasks outperforms embeddings obtained by using other types of contexts (like probably most widely used bag-of-words) [15].

B. Open IE Tools

Since 2007, when the concept of open IE was introduced – together with the first open information extractor called TextRunner in [13], further systems were created. The following paragraphs summarize some notable systems for open IE task.

The following list contains a selection of leading systems for Open IE task:

- ReVerb, [16]
- ArgLearner and R2A2, [15]
- Ollie, [15]
- OpenIE4, [15] and its successor OpenIE5,
- ClauseIE, [17]
- Stanford openIE, [18].

IV. Deriving Annotated Corpus for Recognizing Faceted Entailment Task

In this section we outline a method for generating faceted entailment items, i. e., triples in form (T, H, a facet) accompanied by a label (expressed/unaddressed). We assume we already have an annotated corpus for RTE from an external source containing items in form (T, H), label (entailment/no entailment).

A. Description of the Method

1) From each pair (T, H) such that $T \rightarrow H$, generate positive instances $T \rightarrow h_i$,
2) If a pair (T, H) is labeled in RTE corpus as neutral/unknown (if the corpus is annotated in a three-way manner), then generate a set of negative candidates $T \rightarrow h_i$ to be manually checked,
3) If a pair (T, H) is not labeled not as entailed nor unrelated, then (manually) check for which h_i, $T \not\rightarrow h_i$ and add this pair T, h_i to negative instances
4) Perform postprocessing issues over h_i.

h_i is an output of open information extraction on H, such that h_i is a triple and each element of the triple does not contain more than two words/tokens (open information extractors are able to extract longer phrases as arguments as well as longer predicates, but they are irrelevant from our "faceted" point of view).

The postprocessing consists of removing stopwords and taking an arbitrary pair of each triple h_i.

Note that we assume that if the entailment $T \rightarrow H$ does not hold, then there is at least one facet such that faceted entailment according to T, H does not hold. Although the proposed method includes a certain part of manual work, in case of preparing balanced corpus, half of the work (positive instances) is done automatically and, moreover, the negative instances can be recommended from the list of potential candidates (obtained in the third step) by some simple algorithm like [19].

B. Example

Let us consider an entailment pair *T: The beetle can rub its elytra making noises in order to communicate.* and *H: The elytra are used to make sounds.* – in this case, the entailment holds. (This example was taken from SciEntsBank, [2], the original question was *One function of the bess beetle's elytra (the hard, black wing set) is protection. What is another function of the elytra?*).

From the sentence *The elytra are used to make sounds.* the open information extractor produced among others a triple: *(The elytra; to make; sounds)*. After removing the stopwords *the* and *to*, and taking for example a pair (elytra, make), we have – together with T and H – a positive faceted entailment item.

C. Challenge of Evaluation

The quality of obtained corpus depends mainly on the the quality of the open IE system involved, as well the quality of the initial RTE corpus. Evaluation of open IE systems is generally difficult, often expressed only in term of precision, not recall. First steps in repeatable and large benchmarking of open IE systems were done by Stanovsky et al. in [20]. The issue of measuring quality of the corpus obtained by described method is a challenge for further investigations.

V. Novel Notion: Relational Entailment

As mentioned earlier, faceted entailment deals with ordered pairs of words or multi-word expressions, i. e., with binary relations (or ternary relations, when the semantic relation is explicitly mentioned). In this position paper we suggest relaxing the condition that "the considered object is a binary relation" and use a general textual n-tuples instead.

A. Task definition and Examples

Given (i) a text T (ii) a hypothesis H, both in a form of a sentence and (iii) an open IE text n-tuple t_H extracted from the sentence H (where $n \geq 3$), the first proposed task consists of determining whether T expresses, either directly or by implication, the same relationship between exhibited by t_H in the context of H (i. e. we have again a binary classification problem).

In this case, the meaning of t_H is "emphasized" by the context given by H. This definition corresponds with the original definition of faceted entailment in [21].

For practical purposes arising from potential applications we can go further and omit the role of the hypothesis H and state the definition of a relational entailment.

Definition. Given a text T and a textual n-tuple t, the task is to decide whether the meaning of t, i. e., the semantic relationship expressed by t can be inferred from T. Similarly as in the previous case, we have a binary decision task of recognizing relational entailment – RRE for short. (Positive answer can be labeled as entailed – with notation $T \rightarrow t$, negative as non-entailed – with notation: $T \not\rightarrow t$.)

Example. Given a sentence $T =$ "Patrick flew from Boston to Los Angeles with Delta Airlines with one stopover." and a textual quadruple $t_1 =$ (Patrick, flew, from East Coast, with Delta Airlines) and $t_2 =$ (Patrick, flew, from East Coast, to Los Angeles, via Atlanta).

Following the previous definition, we can (obviously) say that $T \rightarrow t_1$, and, in contrast, $T \not\rightarrow t_2$.

B. Potential Applications of Relational Entailment

From the viewpoint of faceted entailment motivation, recognizing relational entailment has a potential application as a component for automatic assessment of students' tests: having open IE tuples extracted from the reference answer and taking student's answer as an input text, we can use a RRE system for judging whether student's answer is at least partially correct.

Next, probably more important application of recognizing relational entailment, arise from checking/proving facts in open knowledge bases, i. e., in knowledge bases that consist of open IE tuples. Open knowledge bases can be easily interpreted by humans and form an intermediate representation of information in the text(s), [15]. The RRE can be used for checking whether a given text provides an evidence for a given piece of information contained in a considered open knowledge base and subsequently in case of searching relevant text supporting certain fact of an open KB.

Other possible applications may cover various aspects of multi-document summarization, computing textual similarity as well as some other common NLP tasks.

C. Recognizing Relational Entailment, Recognizing Textual Entailment and Paraphrase Identification

In [21], Levy et al. used recognizing faceted entailment system for recognizing ("complete") textual entailment: in the paper, they propose a following, three-stage architecture:

1) Decompose the hypothesis into facets.
2) Determine whether each facet is entailed.
3) Aggregate the individual facet results and decide on complete entailment accordingly.

This approach provided promising results, nevertheless it relies on the assumption that the result about complete entailment can be done (only) from the individual facet results. Generally, there exist textual tuples not entailed by the given text T, although the facets (as pairs of word expressions) covering the tuple are expressed/entailed from T.

This approach presented in [21] also assumes the facet decomposition is already performed.

Obviously, this approach can be modified in order to exploit the benefits of "open IE" perspective:

1) Instead of decomposing the hypothesis into the facets, we generate a set of open IE tuples (this set of tuples can be viewed as a "small" open knowledge base.
2) Instead of determining whether each facet is entailed, we decide, whether each open IE tuple is entailed with respect of the definition above.
3) Aggregating of individual "tuple" results.

This proposed approach brings two main advantages comparing to the faceted-like approach:

1) We leave the decomposition of the hypothesis to an open information extractor.
2) We also decide about entailment of facts that cannot be decomposed into a set of ordered pairs, i. e. facets (that could cause false positives).

By this architecture, we have already linked recognizing relational entailment with recognizing "complete" textual entailment. Now, we are able to do one more natural step towards widely investigated NLP problem: *paraphrase identification*. Since paraphrase identification can be viewed as a bidirectional entailment, the modification for our setting is straightforward. Let us consider we have a couple of sentences P and S and corresponding sets of open IE tuples $\{p_1, p_2, \ldots p_n\}$ and $\{s_1, s_2, \ldots s_m\}$, respectively. The final judgment about paraphrasing can be viewed as an aggregation of results about individual relational entailments $P \rightarrow s_i$ and $S \rightarrow p_j$, where $1 \leq i \leq n$ and $1 \leq j \leq m$.

D. Creating Annotated Corpora for Recognizing Relational Entailment

Similarly as by "classical" recognizing textual entailment problems, we can expect that in recognizing relational entailment challenges supervised approaches will play a key role.

Hence there arises a need of annotated corpora for recognizing relational entailment.

Analogously as in the "facet" case, we can use corpora for recognizing textual entailment that were already built. We assume an open IE extractor is available. In contrast to the faceted case, the situation is simpler since there is no need to modify extracted textual tuples.

We can summarize the semi-automatic method of creating RRE corpus in the following steps:

1) From each pair (T, H) where entailment $T \rightarrow H$ holds, generate positive instances $T \rightarrow h_i$,
2) If a pair (T, H) is labeled in RTE corpus as neutral/unknown (if the corpus is annotated in a three-way manner), then generate a set of negative candidates $T \rightarrow h_i$ to be manually checked,
3) If a pair (T, H) is not labeled not as entailed nor unrelated, then (manually) check for which tuple t_h extracted from H, $T \not\rightarrow t_h$ and add this pair (T, t_h) as a negative instance.

(h_i is an output of open information extraction on H).

Similarly as in the "facet case", we assume that if the entailment $T \rightarrow H$ does not hold, then there must be at least one textual open IE tuple such that relational entailment from T does not hold. Again, half of the work in building balanced annotated corpora (positive instances) is done automatically and, moreover, the negative instances can be recommended from the list of potential candidates by some simple algorithm.

In some cases, it may be reasonable to remove pairs (T, h_i) where the ratio "number of tokens in h_i/ number of tokens in H" is higher than a given threshold in order to avoid cases where obtained tuples are mostly, in fact, permutations of tokens of the hypothesis, such as H: *Slovakia is a member of EU.* and h_i: (Slovakia; is a member of; EU). In these cases, the task of RRE would be highly similar to "classical RTE".

Example. Let us consider a text T: *A Filipino hostage in Iraq was released.* and a hypothesis H_1: *The hostage is a citizen of the Philippines.* labeled as entailment. An open information extractor provides a triple *(The hostage; is a citizen of; the Philippines)*, hence we have a positive instance of relational entailment: *A Filipino hostage in Iraq was released.* $\rightarrow$ *(The hostage; is a citizen of; the Philippines)*. In contrast, if the hypothesis is *The hostage was a citizen of Iraq.* and related extraction is *(The hostage; was a citizen of; Iraq)*, than we got a non-entailment item for RRE corpus.

VI. Conclusion

In this paper we have proposed a method for semi-automatic building of a corpus for recognizing faceted entailment when annotated corpus for RTE task is given and open information extractor is available. Current absence of mid- or large scale annotated corpora for faceted entailment is a serious restriction for using modern deep learning approaches.

Subsequently, we have established a connection between "the world of entailment" and "the world of open IE" by introducing the notion of *relational entailment*. We have demonstrated its strengths, possible applications and we have presented it as a natural generalization of faceted entailment.

A. Discussion

Obviously, the success of most of presented issues hardly relies on the quality of used open information extractor(s). Nowadays, evaluation and benchmarking of open IE tools become emerging and important topic [20]. In current investigations, Groth et al. note [22] that OIE systems perform significantly worse on scientific text than encyclopedic text. Thus various aspects – the domain of the text for instance – should be taken into the account, and the results should be interpreted in such context.

Regarding the newly proposed notion of relational entailment, there arises a need for a sound methodology for creating annotated corpora in order to ensure a suitable quality of training data. There will be also a demand for benchmarks of potential RRE engines and employment of modern metalearning/algorithm selection issues [23].

It also should be take into the account, that proposed notions and architectures are "high level", hence the overall process of training recognizing relational entailment will be prone to error propagation from the initial levels: the quality of data used for training/tuning/development of open information extractor affects the coverage and correctness of extracted textual tuples. The quality of corpus for RRE is also influenced by the quality of RTE annotated corpus we are extracting the tuples from. The overall amount of training instances of RRE corpus then affects the quality of RRE system, that is subsequently propagated in downstream application. Thus each component of this process requires careful error analysis.

B. Further Work

Since we have proposed a new decision problem, the first part of further work will include building and improving corpora for relational entailment, create first architecture for RRE system and provide its evaluations, not only for English but also for languages where RTE corpora and open IE language models are available.

A general issue will be investigations of open IE tuples embeddings, i. e. their vector representations.

Other direction of research is building suitable corpora for faceted entailment and assessment of their quality.

References

[1] R. D. Nielsen, W. Ward, and J. H. Martin, "Recognizing entailment in intelligent tutoring systems," *Natural Language Engineering*, vol. 15, no. 4, pp. 479–501, 2009. doi: 10.1017/S135132490999012X

[2] M. Dzikovska, R. D. Nielsen, C. Brew, C. Leacock, D. Giampiccolo, L. Bentivogli, P. Clark, I. Dagan, and H. T. Dang, "Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge," NORTH TEXAS STATE UNIV DENTON, Tech. Rep., 2013.

[3] R. Ghaeini, S. A. Hasan, V. Datla, J. Liu, K. Lee, A. Qadir, Y. Ling, A. Prakash, X. Z. Fern, and O. Farri, "Dr-bilstm: Dependent reading bidirectional lstm for natural language inference," *arXiv preprint arXiv:1802.05577*, 2018.

[4] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[5] Y. Tay, L. A. Tuan, and S. C. Hui, "A compare-propagate architecture with alignment factorization for natural language inference," *arXiv preprint arXiv:1801.00102*, 2017.

[6] R. Zanoli and S. Colombo, "A transformation-driven approach for recognizing textual entailment," *Natural Language Engineering*, vol. 23, no. 4, pp. 507–534, 2017. doi: 10.1017/S1351324916000176

[7] I. Dagan, O. Glickman, and B. Magnini, "The pascal recognising textual entailment challenge," in *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*. Springer, 2006, pp. 177–190.

[8] D. Z. Korman, E. Mack, J. Jett, and A. H. Renear, "Defining textual entailment," *Journal of the Association for Information Science and Technology*, vol. 69, no. 6, pp. 763–772, 2018. doi: 10.1002/asi.24007

[9] L. Bentivogli, I. Dagan, and B. Magnini, "The recognizing textual entailment challenges: Datasets and methodologies," in *Handbook of Linguistic Annotation*. Springer, 2017, pp. 1119–1147.

[10] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015.

[11] G. Rocha and H. Lopes Cardoso, "Recognizing textual entailment: Challenges in the portuguese language," *Information*, vol. 9, no. 4, p. 76, 2018. doi: 10.3390/info9040076

[12] M.-S. Guo, Y. Zhang, and T. Liu, "Research advances and prospect of recognizing textual entailment and knowledge acquisition," *Jisuanji Xuebao/Chinese Journal of Computers*, vol. 40, pp. 889–910, 04 2017. doi: 10.11897/SP.J.1016.2017.00889

[13] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the web." in *IJCAI*, vol. 7, 2007, pp. 2670–2676.

[14] F. Wu and D. S. Weld, "Open information extraction using wikipedia," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 118–127.

[15] M. Mausam, "Open information extraction systems and downstream applications," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 4074–4077.

[16] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 1535–1545.

[17] L. Del Corro and R. Gemulla, "Clausie: clause-based open information extraction," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 355–366.

[18] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. 1, 2015, pp. 344–354.

[19] M. Víta and V. Kríž, "Word2vec based system for recognizing partial textual entailment," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016, pp. 513–516.

[20] G. Stanovsky and I. Dagan, "Creating a large benchmark for open information extraction," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2300–2305.

[21] O. Levy, T. Zesch, I. Dagan, and I. Gurevych, "Recognizing partial textual entailment," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2013, pp. 451–455.

[22] P. Groth, M. Lauruhn, A. Scerri, and R. Daniel Jr, "Open information extraction on scientific text: An evaluation," *arXiv preprint arXiv:1802.05574*, 2018.

[23] P. Brazdil and C. Giraud-Carrier, "Metalearning and algorithm selection: progress, state of the art and introduction to the 2018 special issue," 2018.

International Conference on Innovative Network Systems and Applications

MODERN network systems encompass a wide range of solutions and technologies, including wireless and wired networks, network systems, services and applications. This results in numerous active research areas oriented towards various technical, scientific and social aspects of network systems and applications. The primary objective of Innovative Network Systems and Applications (iNetSApp) conference is to group network-related events and promote synergy between different fields of network-related research. To stimulate the cooperation between commercial research community and academia, the conference is co-organised by Research and Development Centre Orange Labs Poland and leading universities from Poland, Slovak Republic and United Arab Emirates.

The conference continues the experience of Frontiers in Network Applications and Network Systems (FINANS), International Conference on Wireless Sensor Networks (WSN), and International Symposium on Web Services (WSS). As in the previous years, not only research papers, but also papers summarising the development of innovative network systems and applications are welcome.

- CAP-NGNCS'18—1st International Workshop on Communications Architectures and Protocols for the New Generation of Networks and Computing Systems
- INSERT'18 - 2nd International Conference on Security, Privacy, and Trust
- IoT-ECAW'18—2nd Workshop on Internet of Things—Enablers, Challenges and Applications
- WSN'18 - 7th International Conference on Wireless Sensor Networks

AREA SUPERVISORY COMMITTEE

- Awad, Ali Ismail, INSERT'18
- Furtak, Janusz, IoT-ECAW'18
- Hamrioui, Sofiane, CAP-NGNCS'18
- Ševčík, Peter, WSN'18

2nd International Conference on Security, Privacy, and Trust

ADMITTEDLY, information security works as a backbone for protecting both user data and electronic transactions. Protecting communications and data infrastructures of an increasingly inter-connected world have become vital nowadays. Security has emerged as an important scientific discipline whose many multifaceted complexities deserve the attention and synergy of the computer science, engineering, and information systems communities. Information security has some well-founded technical research directions which encompass access level (user authentication and authorization), protocol security, software security, and data cryptography. Moreover, some other emerging topics related to organizational security aspects have appeared beyond the long-standing research directions.

The 2nd International Conference on Security, Privacy, and Trust (INSERT'18) focuses on the diversity of the information security developments and deployments in order to highlight the most recent challenges and report the most recent researches. The conference is an umbrella for all information security technical aspects, user privacy techniques, and trust. In addition, it goes beyond the technicalities and covers some emerging topics like social and organizational security research directions. INSERT'18 is intended to attract researchers and practitioners from academia and industry, and provides an international discussion forum in order to share their experiences and their ideas concerning emerging aspects in information security met in different application domains. This opens doors for highlighting unknown research directions and tackling modern research challenges. The objectives of the INSERT'18 can be summarized as follows:

- To review and conclude researches in information security and other security domains, focused on the protection of different kinds of assets and processes, and to identify approaches that may be useful in the application domains of information security
- To find synergy between different approaches, allowing elaborating integrated security solutions, e.g. integrate different risk-based management system.
- To exchange security-related knowledge and experience between experts to improve existing methods and tools and adopt them to new application areas

TOPICS

Topics of interest include but are not limited to:

- Biometric technologies
- Human factor in security
- Cryptography and cryptanalysis
- Critical infrastructure protection
- Hardware-oriented information security
- Social theories in information security
- Organization- related information security
- Pedagogical approaches for information security
- Social engineering and human aspects in security
- Individuals identification and privacy protection methods
- Information security and business continuity management
- Decision support systems for information security
- Digital right management and data protection
- Cyber and physical security infrastructures
- Risk assessment and risk management
- Tools supporting security management and development
- Trust in emerging technologies and applications
- Ethical trends in user privacy and trust
- Digital forensics and crime science
- Security knowledge management
- Privacy Enhancing Technologies
- Misuse and intrusion detection
- Data hide and watermarking
- Cloud and big data security
- Computer network security
- Security and safety
- Assurance methods
- Security statistics

EVENT CHAIRS

- **Awad, Ali Ismail,** Luleå University of Technology, Sweden
- **Bialas, Andrzej,** Institute of Innovative Technologies EMAG, Poland

PROGRAM COMMITTEE

- **Banach, Richard,** University of Manchester, United Kingdom
- **Bun, Rostyslav,** Lviv Polytechnic National University, Ukraine
- **Clarke, Nathan,** Plymouth University, United Kingdom
- **Cyra, Lukasz,** DM/OICT/RMS (UN)
- **Daszczuk, Wiktor Bohdan,** Warsaw University of Technology, Poland
- **Felkner, Anna,** Research and Academic Computer Network NASK
- **Furnell, Steven,** Plymouth University, United Kingdom
- **Furtak, Janusz,** Military University of Technology, Poland

- **Gawkowski, Piotr,** Institute of Computer Science, Warsaw University of Technology, Poland
- **Geiger, Gebhard,** Technical University of Munich, Faculty of Economics
- **Grzenda, Maciej,** Orange Labs Poland and Warsaw University of Technology, Poland
- **Hämmerli, Bernhard M.,** Hochschule für Technik+Architektur (HTA), Switzerland
- **Hasssaballah, M.,** South Valley University, Egypt
- **Kapczynski, Adrian,** Silesian University of Technology, Poland
- **Krendelev, Sergey,** Novosibirsk State University,JetBrains research, Russia
- **MD Faisal, Mohammad,** Integral University, India
- **MD Rafiqul, Islam,** School of Computing and Mathematics, Charles Sturt University
- **Misztal, Michal,** Military University of Technology, Poland
- **Pańkowska, Małgorzata,** University of Economics in Katowice, Poland
- **Rot, Artur,** Wroclaw University of Economics, Poland
- **Stokłosa, Janusz,** WSB University in Poznan, Poland
- **Suski, Zbigniew,** Military University of Technology, Poland
- **Szmit, Maciej,** University of Lodz; IBM GSDC, Poland
- **Wahid, Khan Ferdous,** Airbus, Germany
- **Yahya, Eslam,** Ohio State University, Columbus
- **Zamojski, Wojciech,** Wrocław University of Technology
- **Zieliński, Zbigniew,** Military University of Technology, Poland

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 43–47
DOI: 10.15439/2018F377
ISSN 2300-5963 ACSIS, Vol. 16

Proposal for simplified implementation of risk assessment method for measuring instruments

Martin Koval
Czech Metrology Institute, Brno, Czech Republic
Email: mkoval@cmi.cz

Federico Grasso Toro, Marko Esche
Physikalisch-Technische Bundesanstalt, Berlin, Germany
Email: {federico.grassotoro, marko.esche}@ptb.de

***Abstract* - Legal Metrology is the economic sector where measuring instruments subject to legal control (taximeters, electricity meters, etc.) are used. In this field, constant growth of Measuring Instruments using ICT technology is evident. For this reason, higher security requirements need to be imposed as stated by the relevant EU directives. Risk assessment is an additional security assurance requirement for software, based on current regulations Directive 2014/31/EU and Directive 2014/32/EU (MID) that state: "The documentation shall make it possible to assess the instrument's conformity to the relevant requirements and shall include an adequate analysis and assessment of the risk(s)." [1]. Several methods for risk assessment of software exist, but based on this statement above, it is necessary to find appropriate solutions for the realization of risk assessment for metrological software, on the base of its technical documentation. The WELMEC Working Group 7 has developed a Risk Assessment method, based on international standards. In this article a simpler method is proposed, aiming for advantages such as universality, simplicity and transparency, in contrast with already existing methods. The combination of these advantages in the proposed method will allow its simple understanding and implementation for all active stakeholders (both the Notified Bodies and the manufacturers).**

I. Introduction

THE term *risk* can be defined in many ways for different purposes. One of these definitions is "Combination of the consequences that would follow from the occurrence of an unwanted event and the likelihood of the occurrence of the event." [2]. It should be emphasized that, in this case, the risk is the likelihood of occurrence combined with the impact of the event, and it can be transformed into the event that has potentially happened. This potential risk can also be a factor for system hazard or weak security provisions against dangerous/unauthorized influences.

II. The Proposal of Risk Assessment Method

The proposal of this method is on the base of the selection of technical parameters, present in measuring instruments in the real world. These parameters have an influence on the legally relevant software (LRS) and are assigned a point rating. The WELMEC Working Group 7 has developed a Risk Assessment method, based on international standards [4] to define these ratings. There exists a wide range of measuring instruments to be regulated. Each measuring instrument has its own specific parameters, which depend on the purposes of its usage. For the present risk assessment method, a model of an abstract measuring instruments (MAMI) is proposed. Real-world examples were used for the development of the risk method (Chapter VI). The diagram shown in Fig.1 describes the proposed method. The first step is checking the completeness of the technical documentation of the measuring instrument. The second step is the assessment of important parameters of the metrological software functionality. The assessment of parameters is based on 25 combinations of parameters (assets, threats and impacts: specification of technical parameters). The last step is an evaluation which summarizes the points and results in a potential risk. According to the MID [1] the assessment of the risk is a part of the documentation, which is necessary for the process of validation of metrological software. Notified Bodies (NB) must check the correctness of the documentation, meaning that they must also check if the evaluations and the result of potential risk are correctly realized.

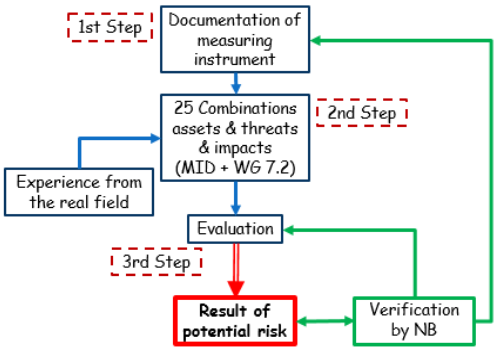

Figure 1. The basic structure of proposal of risk assessment method for metrological software.

III. Technical Parameters of Measuring Instrument Influencing Metrological Software

The following technical parameters that have an influence on metrological software are inspired by WELMEC Guide 7.2 [3], the directive MID [1] and practical knowledge.

©2018, PTI

The resulting list of parameters is proposed based on developments and regulations, according to the current situation of ICT in the measurement area. The parameters are chosen on the base of combinations of three main factors:

1) Most used,

2) Most risky technologies, according to practical knowledge and

3) Parameters with different levels or the configuration possibilities.

Table I contains the proposed list of parameters, divided into 5 sections.

TABLE I: The basic division the parameters into 5 sections.

A	PC, Download LRS, Separation
B	User interface
C	Communication interface
D	Storage
E	Security

IV. ASSIGNING POINTS TO TECHNICAL PARAMETERS

Table I contains the list of parameters which has possibility for configuration. For risk assessment, it is necessary to assign points to all parameters. These points represent important characters for the analysis of the measuring instrument and its risk assessment:

- When the potential risk is zero: Absence of technical parameters which can increase the potential risk.
- When the potential risk is assigned with 1 point: The technical parameters are integrated part of the measuring instrument with metrological software.
- When the potential risk is assigned with 2 points: The technical parameters are a common part of the measuring instrument with metrological software.
- When the potential risk is assigned with 5 points: The technical parameters have weak security against the possibilities which are available. And, for this rea-son, the potential risk increases or the used technical parameters increase the overall potential risk.
- When the potential risk is assigned with 10 points: The technical parameters are sophisticated and their characteristics are considered as carriers of higher potential risk. These parameters can contain hidden functions or be considered as weakly secured elements.
- When the potential risk is assigned with 15 points: The technical parameters are sophisticated and their characteristics are considered as carriers of higher potential risk, including the parameters at the previous point. These parameters are considered riskier than with 10 points from the point of view of practical knowledge.

Once the potential risks are assigned, the risk assessment can be performed.

V. SPECIFICATION OF TECHNICAL PARAMETERS

Some of the technical parameters existing in the designed concept of the risk assessment method for metrological software have possibility for configuration. In Tables II-VI the configuration possibilities are represented and defined. These tables contain points rating for each specific technical parameter.

TABLE II. Specification of technical parameters: PC, Download LRS and Separation.

PC, Download LRS, Separation			Points
A	1	Using PC as like primary	10
	2	Download LRS	10
	3	Separation SW	15

Based on the possible configurations specific rating points can be assigned to specific technical parameters of the measuring instrument. For example, if the measuring instrument has a configuration based on a universal computer, then there is a higher likelihood for other functions that can have negative influences on the legally relevant software.

The download of the LRS without breaking the physical seal of measuring instrument means that there is a method for modification of the LRS, and it can also cause negative influences on the LRS even though conditions for security are fulfilled.

Separation of SW, dividing metrological software into legally relevant SW and legally non-relevant SW (LnRS) can have different realizations, e.g. both LRS and LnRS are in one source code or divided into two microprocessors. If the LnRS needs data from the LRS, then there must be a connection between the LRS and the LnRS. And since the LnRS is not under control, that means that potentially dangerous applications can exist, which can have negative influences on the LRS.

TABLE III. Specification of technical parameters: User interface.

User interface			Points
B	4	Button/s	2
	5	Keyboards (include numbers field)	10
	6	Without user interface	0

If measuring instruments have any button/s, there is potential risk that there can be a hidden combination for negative influences on the LRS. In the case that the measuring instrument has a keyboard, then the possibilities for hidden functions is even higher. If the measuring instrument has no user interface that has possibilities of entry to the metrological SW, then, in this case, the potential risk can be considered zero.

TABLE IV. Specification of technical parameters: Communication interface

Communication interface			Points
C	7	Physical	2
	8	Wireless	5
	9	Internet	15
	10	Without communication interface	0

Like in previous cases with the user interface, also, if the measuring instrument has no communication interfaces, then the potential risk can be considered zero. The communication interfaces can have different realizations. If the measuring instrument has a communication interface only by physical connections (USB, RS-232, etc.) there is a lower likelihood for the potential risk as there is for wireless connections. In case of wireless connection, there is higher likelihood for not allowed or hidden connections to the measuring instrument, where there is not enough security by means of physical seals. The cases where the measuring instrument has the possibility of connecting to the Internet, the potential risk is higher, since the measuring instrument may be exposed to cyber-attacks.

TABLE V Specification of technical parameters: Storage.

Storage			Points
D	11	Part of microprocessor	1
	12	Internal storage	2
	13	Removable storage	10
	14	Without storage (data on display only)	0

The memory for metrological software is often divided into storage for software and other storage for the measurement data – The Harvard architecture. There are cases when the storage is part of the microprocessor, which is physically sealed. But since the space for software must exist, it cannot be potential risk zero. Very often the micro-processor has a small memory, and for this reason the measuring instrument contains other internal storage. These storages offer the possibility for realizing separation of software (in LRS and LnRS). But the capacity of the internal storage can also be as big as to offer hidden space for non-allowed applications [5]. In cases where the storage is removable, the potential risk is higher, because there are many ways for negative exploitation.

The security aspect from the proposed solution is the most complicated to define, because each technical parameter has an additional configuration.

Event recording in metrological SW occurs in two forms: 1) Event Counter and 2) Event Logger. Event Counter records each change by binary value (1/0) or counting of changes. The Event Logger can record each change with date, time, and additional description of each change. In cases where there is no Event Record, the potential risk is the highest, because it is possible to realize any changes without being recorded. The cases where passwords are used have potential risk, since every password can be broken. The passwords can be realized by number or combinations characters (alphanumerical). Most often a numerical combination is used, and the most used are 4 digits passwords. If a 4 digits password is used without a block protection, then there are only 10.000 possibilities, easy to break with IT technology. There are still measuring instruments without user or communication interfaces with LRS. In these cases, there is no need for passwords, since no access can happen to the LRS without breaking a physical seal. In the opposite example, if the measuring instrument has user and/or communication interfaces, there are many possibilities of changing or influencing the LRS, due to the lack of secure elements. Then, the conditions for security are not fulfilled. These cases are not acceptable.

TABLE VI. Specification of technical parameters: Security

Security: Event counter/logger			Points
E	15	Event counter	2
	16	Event logger	0
	17	None	15
	Security: Password		
	18	Numerical 4digits	10
	19	Numerical more than 4 digits	5
	20	Alphanumerical + block systems	2
	21	None	0
	Security: Seal cover		
	22	After break: device is not functional	0
	23	After break: device is still functional	15
	Security: Checksum		
	24	CRC 32 and weak	10
	25	better then CRC-32	1

The next part of the security aspect of the method are physical seals. The sealing of the measuring instrument is realized in different ways (it can be physical seal-lead/plastic, stamp). Physical seals that can be broken or removed are not applicable in legally metrology area. In the cases when the LRS of the measuring instrument and the physical seal are interconnected, where after the seal has been broken, the measuring instrument is not functional, the potential risk is considered zero. Given that there are cases where after seal breaking, the measuring instrument still works, there is a probability that the seal of measuring instrument is not enough for the security of LRS.

The final category for security are checksums. Currently, there are a lot of variations of checksums on the market.

The manufacturers mostly use the type CRC-16, for economic reasons. This type of CRC belongs to the weakest. Current WELMEC Guide 7.2 pushes for an acceptable solution CRC-32 [3]. This is the reason for this type of checksums in Table VI.

VI. EXAMPLES OF EVALUATION

On the base of the present technical parameters it is possible to propose different MAMI (Model Abstract Measuring Instrument).

The configurations MAMI A-D are considered as border cases, where it may occur:

- MAMI A: Proposed lowest potential risk.
- MAMI B: Proposed highest potential risk.
- MAMI C: Proposed basic technical parameters.
- MAMI D: Proposed advanced technical parameters.

TABLE VII. Evaluation of Risk assessment for metrological SW of MAMI (A-D).

Sect.	Num.	Technical parameters	Pts.	A	B	C	D
PC, Download LRS, Separation							
A	1	Using universal computer	10	0	1	0	1
	2	Download LRS	10	0	1	0	1
	3	Separation SW	15	0	1	0	1
User Interface							
B	4	Button/s	2	0	1	1	1
	5	Keyboards (include NF)	10	0	1	0	1
	6	Without user interface	0	1	0	0	0
Communication interface							
C	7	Physical	2	0	1	1	1
	8	Wireless	5	0	1	0	1
	9	Internet	15	0	1	0	1
	10	Without Comm. Inter.	0	1	0	0	0
Storage							
D	11	Part of microprocessor	1	0	1	1	1
	12	Internal storage	2	0	1	0	1
	13	Removable storage	10	0	1	0	1
	14	Without storage	0	1	0	0	0
Security							
E	15	Event counter	2	0	0	0	0
	16	Event logger	0	1	0	1	1
	17	Without Event C/L	15	0	1	0	0
Passwords							
	18	Numerical 4 digits	10	0	1	1	0
	19	Numerical more than 4 digits	5	0	0	0	0
	20	Alphanumerical	2	0	0	0	1
	21	Without password/s	0	1	0	0	0
Seal Cover							
	22	After breaking: is not functional	0	1	0	0	0
	23	After breaking: is still functional	15	0	1	1	1
Checksum							
	24	CRC-32 and weak	10	0	1	1	0
	25	Better than CRC-32	1	1	0	0	1
Summary of points				2	132	42	90

Table VII shows some examples of the evaluation of potential risks for metrological software for different MAMI (from A to D). Each technical parameter has been assigned rating points, where the columns for MAMI contains 1 or 0, depending on the technical parameters option. The last row of Table VII is the sum of points, where it is possible to realize further evaluations or adjustments to get the result of the potential risk by statistic or other methods. One of the possible solutions can be considered the MAMI B, for the maximum potential risk, assigning 132 points to the 100% of potential risk. Then, it is possible to create comparative graphs for the conclusions, see Figure 2.

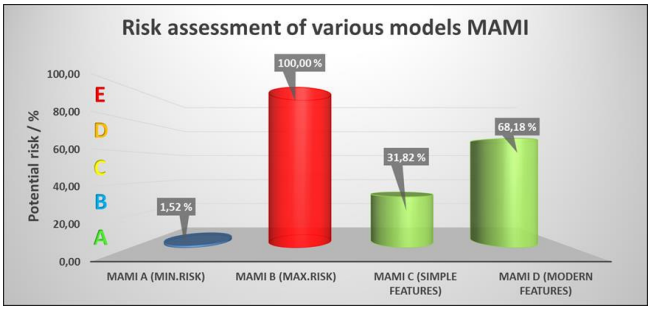

Fig. 2. Examples of Risk assessment for MAMIs (A-D).

The WELMEC Working Group uses a Risk Assessment method [4] that combines elements from the international ISO/IEC standards 27005 and 15408, to support the theoretical comparability of the risk assessment results. On the other hand, the present method focuses on existing measuring instruments, mostly type P according to WELMEC Software Guide 7.2[3].

The present method defines within the MAMIs four proposed model levels, allowing a quick approximation of the risk assessment for the modeled measuring instrument, supporting quick repeatability of analysis between very similar and well-known measuring instruments.

The method focused on existing measuring instruments with many previously detailed and collected reports regarding risk assessment, while the WELMEC Working Group 7 focuses on any kind of measuring instrument, even with previously unknown designs and the new developments.

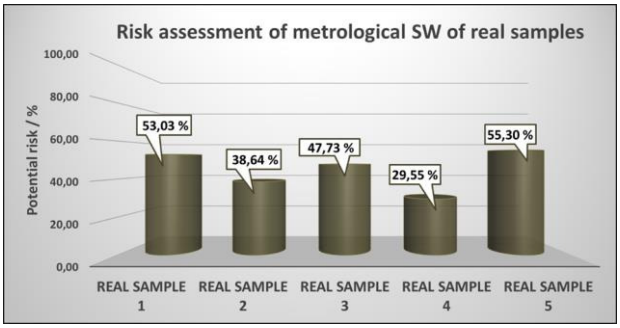

Fig. 3. Risk assessment for metrological SW of real samples.

VII. Conclusion

The present method for risk assessment for metrological software is not only focused on potential risk results in graphs, but it also analyses the resulting table with rating points. Table VII indicates which MAMI shows weak software protection (B and D). The proposal of the method is based on the potential risk that cannot be zero, but after the determination of technical parameters with assign rating point is possible the set within a maximum of potential risk of the metrological software. The cases MAMI A-D are specific cases, but after applying the method on real world examples (Fig.3), the potential risk results in the interval between 20% and 60%. This test on 5 real world samples is only for orientation purposes. For better results, it is necessary to apply this method on more real-world examples to check the quality of these results and the potential application of this simple method in the future.

Acknowledgements

This project is funded by Institutional Subsidy for Long-Term Conceptual Development of a Research Organization granted to the Czech Metrology Institute by the Ministry of Industry and Trade.

References

[1] Directive 2014/32/EU of the European Parliament and of the Council of 26 February 2014 on the harmonization of the laws of the Member States relating to the making available on the market of measuring instruments, / European Union, Council of the European Union // European Parliament, Directive, February 2014.

[2] ISO/IEC 27005:2011(e) Information technology - Security techniques - Information security risk management, International Organization for Standardization // Geneva, CH, Standard, June 2011.

[3] WELMEC 7.2 Software Guide, / European cooperation in legal metrology, // WELMEC Secretariat, Delft, Standard, March 2012.

[4] M. Esche and F. Thiel, "Software risk assessment for measuring instruments in legal metrology," 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), Lodz, 2015, pp. 1113-1123. DOI: 10.15439/2015F127.

[5] Ch.-B. do Prado, D.-R. Boccardo, R.-C.-S Machado, L.-F.-R. da Costa Carmo, T. -M. do Nascimento, L.-M.-S. Bento, R.-O. Costa, C.-G de Castro, S.-M. Camara, L. Pirmez and R. Oliveira / Software Analysis and Protection for Smart Metering in // NCSLI Measure: The Journal of Measurement Science. - 2014.-vol.9.-No.3 – p. 22-29. DOI: 10.1080/19315775.2014.11721691.

2nd Workshop on Internet of Things—Enablers, Challenges and Applications

THE Internet of Things is a technology which is rapidly emerging the world. IoT applications include: smart city initiatives, wearable devices aimed to real-time health monitoring, smart homes and buildings, smart vehicles, environment monitoring, intelligent border protection, logistics support. The Internet of Things is a paradigm that assumes a pervasive presence in the environment of many smart things, including sensors, actuators, embedded systems and other similar devices. Widespread connectivity, getting cheaper smart devices and a great demand for data, testify to that the IoT will continue to grow by leaps and bounds. The business models of various industries are being redesigned on basis of the IoT paradigm. But the successful deployment of the IoT is conditioned by the progress in solving many problems. These issues are as the following:

- The integration of heterogeneous sensors and systems with different technologies taking account environmental constraints, and data confidentiality levels;
- Big challenges on information management for the applications of IoT in different fields (trustworthiness, provenance, privacy);
- Security challenges related to co-existence and interconnection of many IoT networks;
- Challenges related to reliability and dependability, especially when the IoT becomes the mission critical component;
- Zero-configuration or other convenient approaches to simplify the deployment and configuration of IoT and self-healing of IoT networks;
- Knowledge discovery, especially semantic and syntactical discovering of the information from data provided by IoT;

The IoT conference is seeking original, high quality research papers related to such topics. The conference will also solicit papers about current implementation efforts, research results, as well as position statements from industry and academia regarding applications of IoT. The focus areas will be, but not limited to, the challenges on networking and information management, security and ensuring privacy, logistics, situation awareness, and medical care.

TOPICS

The IoT conference is seeking original, high quality research papers related to following topics:

- Future communication technologies (Future Internet; Wireless Sensor Networks; Web-services, 5G, 4G, LTE, LTE-Advanced; WLAN, WPAN; Small cell Networks. . .) for IoT,
- Intelligent Internet Communication,
- IoT Standards,
- Networking Technologies for IoT,
- Protocols and Algorithms for IoT,
- Self-Organization and Self-Healing of IoT Networks,
- Trust, Identity Management and Object Recognition,
- Object Naming, Security and Privacy in the IoT Environment,
- Security Issues of IoT,
- Integration of Heterogeneous Networks, Sensors and Systems,
- Context Modeling, Reasoning and Context-aware Computing,
- Fault-Tolerant Networking for Content Dissemination,
- Architecture Design, Interoperability and Technologies,
- Data or Power Management for IoT,
- Fog—Cloud Interactions and Enabling Protocols,
- Reliability and Dependability of mission critical IoT,
- Unmanned-Aerial-Vehicles (UAV) Platforms, Swarms and Networking,
- Data Analytics for IoT,
- Artificial Intelligence and IoT,
- Applications of IoT (Healthcare, Military, Logistics, Supply Chains, Agriculture, ...),
- E-commerce and IoT.

The conference will also solicit papers about current implementation efforts, research results, as well as position statements from industry and academia regarding applications of IoT. Focus areas will be, but not limited to above mentioned topics.

EVENT CHAIRS

- **Cao, Ning,** College of Information Engineering, Qingdao Binhai University
- **Furtak, Janusz,** Military University of Technology, Poland
- **Zieliński, Zbigniew,** Military University of Technology, Poland

PROGRAM COMMITTEE

- **Amanowicz, Marek,** Military University of Technology
- **Antkiewicz, Ryszard,** Military University of Technology, Poland
- **Chudzikiewicz, Jan,** Military University of Technology in Warsaw, Poland
- **Cui, Huanqing,** Shandong University of Science and Technology, China

- **Ding, Jianrui,** Harbin Institute of Technology, China
- **Fouchal, Hacene,** University of Reims Champagne-Ardenne, France
- **Fuchs, Christoph,** Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Germany
- **Gluhak, Alexander,** Intel Labs Europe
- **Hodoň, Michal,** University of Žilina, Slovakia
- **Johnsen, Frank Trethan,** Norwegian Defence Research Establishment (FFI), Norway
- **Krco, Srdjan,** DunavNET
- **Lenk, Peter,** NATO Communications and Information Agency, Other
- **Li, Guofu,** University of Shanghai for Science and Technology, China
- **Marks, Michał,** NASK - Research and Academic Computer Network, Poland
- **MURAWSKI, Krzysztof,** Military University of Technology, Poland
- **Niewiadomska-Szynkiewicz, Ewa,** Research and Academic Computer Network (NASK), Institute of Control and Computation Engineering, Warsaw University of Technology
- **Paprzycki, Marcin,** Systems Research Institute Polish Academy of Sciences, Poland
- **Sikora, Andrzej,** Research and Academic Computer Network (NASK)
- **Skarmeta, Antonio,** University of Murcia
- **Suri, Niranjan,** Institute of Human and Machine Cognition
- **Te Paske, Bert Jan,** TNO Netherlands
- **Wrona, Konrad,** NATO Communications and Information Agency
- **Xu, Jian,** Northeastern University, China
- **Zhang, Tengfei,** Nanjing University of Post and Telecommunication, China
- **Zhao, Yongbin,** Shijiazhuang Tiedao University, China

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 51–58
DOI: 10.15439/2018F118
ISSN 2300-5963 ACSIS, Vol. 16

Towards a Supportive City with Smart Urban Objects in the Internet of Things: The Case of Adaptive Park Bench and Adaptive Light

Marvin Hubl
Chair of Information Systems II,
University of Hohenheim,
Stuttgart, Germany
Email: marvin.hubl@uni-hohenheim.de

Philipp Skowron, Michael Aleithe
Chair of Information Management
Leipzig University,
Leipzig, Germany
Email: {skowron, aleithe}@wifa.uni-leipzig.de

***Abstract*—IoT technology is a key driver for building smart city infrastructure. The potentials for urban management problems requiring process control and allocation mechanisms has long been acknowledged. However, up to now the potentials of equipping urban objects with sensors, information processing capability, and actuators to enable adaptation to pedestrians' individual needs have not yet been exploited. The objective of our research is to design smart urban objects that enhance usability and safety of the urban space for pedestrians. We report on our conceptual design for an *IoT platform* that connects the novel smart urban object *adaptive park bench* with an *adaptive light system* to actively support pedestrians in the urban environment, in particular senior citizens with handicaps.**

I. Introduction

Internet of things (IoT) technology is increasingly used to make the urban environment "smart". With IoT-enabled process control and allocation mechanisms, to date many innovations aim at new possibilities for more effective traffic load balancing [1], [2], energy supply [3], [4], or waste management [5], [6]. While these innovations address important ecological problems the potentials to support pedestrians in the urban environment with IoT enabled physical urban objects are still in its infancy.

IoT technology connects physical objects via digital information processing and communication technology. By further equipping urban objects with sensors and actuators the urban environment is transformed into a cyber physical system [7] providing advanced potentials for pedestrian support. To make use of these potentials we design *smart urban objects (SUO's)* that adapt individually to pedestrians' specific requirements for the usability of the urban environment. Connecting SUOs with IoT allows for useful interactions between both, the objects with pedestrians, and the objects with each other.

While SUOs are supposed to improve usability for every pedestrian, a special focus is on pedestrians with handicaps. For handicapped pedestrians, the adaptivity of the SUOs shall not only enhance usability but effectively compensate (some of) their handicaps and thus enable safe usage of the urban space. A growing group of stakeholders who are in need for an adaptive and supportive urban environment are older adults [8]. Hence "smartification" of the urban environment to meet pedestrians' needs is not only a technical gadget but plays a crucial societal role.

This research contributes to the knowledge base of design-oriented research on the application of IoT technology in supportive urban environments to shape the smart city of the future. To that end, we provide a conceptual design of an IoT platform that connects the novel SUO adaptive adaptive park bench with an innovative adaptive light system. The leading question is: *How to design the SUOs and their connection via IoT to enable supportive interactions between the objects and pedestrians for enhancing safe use of the environment?*

The paper proceeds as follows: In the next section we briefly present recent developments within the smart city paradigm, IoT technology, smart street lighting and smart park benches. Then we report on our design of an IoT platform and the two SUOs adaptive park bench and adaptive light. Thereafter, we discuss a scenario for an interplay of the two SUOs, the IoT platform and pedestrians. We conclude and provide an outlook on future research.

II. State of the Art

A. The Smart City Paradigm

For the characterization of a smart city, several perspectives play a role: Technology innovations, organizational innovations and policy innovations [9]. While technology innovations seem to be only one part among three, they are essential for all innovations within the smart city paradigm: Organizational innovations are directly related to an effective use of technology and policies shall actively promote the use of effective technology innovations [9].

An overall objective of smart city innovations is sustainability, economic growth and quality of life [10]. Current trends in smart city developments involve the domains of natural resources and energy, transport and mobility, buildings, living, government, as well as economy and people [11].

From a design-oriented perspective, we contend that information technological innovations provide most effective means to enhance smartness of cities in particular with respect

©2018, PTI

to quality of life. Smart city initiatives use high technology innovations, such as IoT, to create benefits for well-being, inclusion and participation [12]. In conclusion, smart city relates to socio-technological developments for a harmony of material and virtual world [9]. Information technology plays an important role in contributing to solutions for emerging problems of urban living [11].

The information technological foundation to bring smart city innovations into built infrastructure are smart objects [13]. Smart objects in the urban area provide sensor data that can be used for urban environment monitoring, social networking, intelligent transport systems [14], or other information-intensive urban management problems. The aim of our research, however, is to go beyond exploiting smart objects as sensor networks. Our proposed SUOs have actuators and create benefits by actively adapting to citizens in interconnected scenarios.

B. Internet of Things Platforms

In the field of IoT architectures, many frameworks, models and platforms have already been developed seeking to solve both, specific concerns and general problems [15]. Nevertheless, there is still a significant lack of IoT middleware, which can be adapted to specific domains as well as to general applications [15], [16]. While one of our aims is to enhance safety of the users who indirectly use the IoT platform in the urban environment, we identify that existing solutions focus only pure technical aspects of safety and security [17]. A bridge to safety of IoT system users is rarely provided and is an essential point of research. A proposal for a solution for an IoT architecture that focuses on the problems describend in [15], [16] and thus provides a suitable basis for further developments is provided in [18]. Although in [18] security aspects of the solution play a role at the technical level, they are not specifically considered on the user level.

C. Smart Urban Objects

1) Smart Park Benches: Although seldom recognized as these, park benches are one of the most ubiquitous elements in public space [19]. By park bench we denote seats in the public area dedicated for more than one person. In urban life those seats play an important role for well-being. Yet, the idea of enabling park benches for adaptivity has not been widespread. From the perspective of user requirements adaptivity of park benches would provide clear gains for its usability. In practice, there are so called smart park benches which are supplied by solar energy and provide USB- and charging slots for smartphones, WiFi, and environmental sensor equipment [20], [21]. However, these park benches do not have moving actuators but at best screens to display information. There are further ideas to augment public seats with information technology but these are rather for entertainment purposes than for usability purposes [19].

In designing park benches, knowledge about usability has long been incorporated, even though in a static manner. In a comprehensive study on design principles that shall make park benches particularly usable for senior citizens, it has been surveyed based on focus-groups and expert interviews, that seniors require elevated sitting positions, slightly tilted forward seats, and armrests but footrests should be avoided due to tripping hazard [22]. While there are several vendors of park benches with these special characteristics, as a matter of fact static alignment to these requirements can only account for the average but not for the individual. The most adaptive type of park benches, which are available, are those with pneumatic springs [23]. The opportunities of making park benches adaptive is up to now exploited only hardly.

2) Smart Street Lighting: In the past, many different concepts have been presented that include an intelligent lighting system. Several objectives have been addressed and several milestones have been achieved.

Motivated by using natural light more efficiently in buildings and consequently to reduce the energy consumption of the electrically generated light, a lighting system has been presented in order to meet this challenge [24]. The intelligence consists of a sensor-based detection of the available daylight inside the facility, whereupon the artificially generated light is regulated.

In addition to the lighting systematics within a building [24], there exist achievements presented in [25], [26], [27] that augment street lighting with intelligence. The light intensity is adapted depending on the traffic situation. The overriding objective of these efforts is to minimize the required energy consumption. Networking of the individual lamps already exists in [25]. The user's view of smart lighting systems is examined in [28], [29]. The focus here is on the psychological effect of different light modes on the user. Particularly in [28], the mood of older people is positively influenced by different lighting configurations. However, the core of the investigations here lies more on psychological mood analyses than on the concept of a smart lighting system.

III. Design of Interconnected Smart Urban Objects

A. Internet of Things Platform

To coordinate all SUOs in the overall system, we use an IoT platform. The platform is based on IoT technology and is also the central unit for processing joint information. The platform is designed against the following requirements: It should

1) be expandable with respect to SUOs and other IoT devices,
2) facilitate customization and connection of processes,
3) support and use standards,
4) exhibit little latency for a tactile internet,
5) be secure and guard privacy and user safety,
6) apply intelligent algorithms on interaction log data.

(1.) Expandability is required to add potentially new SUOs or IoT devices. Therefore, open interfaces, protocols and data formats are relevant.

(2.) As different users of SUOs may have different demands for the administration and behavior of the SUOs, the underlying processes must be determined flexibly. Also, the platform

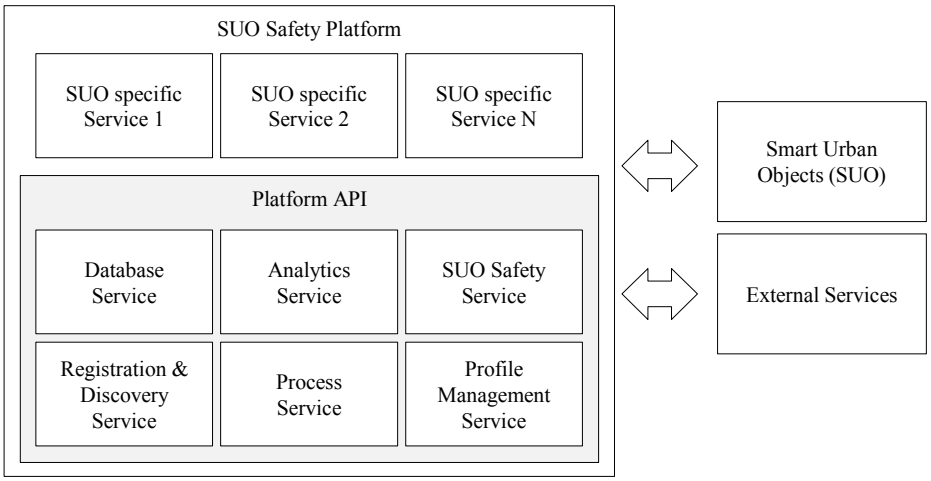

Fig. 1. Overview of the architectural design for the IoT platform

shall facilitate to combine any SUO, independently from the specifics of the SUO.

(3.) To meet the aforementioned requirements standardized interfaces and communication technology are prerequisite.

(4.) The platform is also a middleware for SUO-communication. High latencies of control messages within the SUO-network may be disadvantageous or even harmful because the SUOs are situated in public spaces.

(5.) Communicated data must be guarded against deletion, manipulation, and unauthorized reading. The functionality must be reliable and save.

(6.) Generated and collected data shall be analyzed and utilized for improving the service provision for the end users.

Based on these requirements and the challenges identified in [18] we design the architecture of our IoT platform as depicted in figure 1.

Basically, the platform contains two parts. The first part is the API part. There, those services run which must be available to the other, more specific services. The second part is for the specifics of the SUOs. There, those services run which are specific for single SUO or particular groups of SUOs. Arbitrary external services can be connected to the platform, too, as long as they use the supported protocol standards.

In the API part different basic services run to realize the core functionality of the IoT platform. This entails a central service for data storage. So all services can use a separate data storage according to the microservice pattern but can access these with a joint interface.

To connect and customize processes with SUOs the platform provides a processing service. Processes can be modeled with BPMN (Business Process Modeling and Notation) and are executed with a process engine. A process is a sequence of services that link several API and SUO-services. An analytics service can be integrated to enable descriptive, predictive, and prescriptive analyzes.

To ensure reliability and safety for end users in the public space, the SUO safety service provides required information based on user profiles and interaction models. To allow SUO-specific services to access SUO-data and user data the registration and profile management service provide a central interface to obtain user information and SUO meta-data.

Focusing the requirement for low latency, expandability and security several well-known protocols and technologies are used. Communication between platform and SUOs as well as external services is standardized with REST respectively MQTT interfaces. Both are encrypted. For platform internal communication either the HTTP/2 protocol (REST) can be used or for particularly critical services with respect to latency a gRPC interface with protocol buffers can be used. Data are stored in a datalake on the platform. The services run with container based orchestration to account for high flexibility.

B. Adaptive Park Bench

The perception of the usability of park benches varies among people. Thereby, a crucial factor is the extent to which the design and functionality of a park bench fits the individual anthropometric measures. The designs of conventional park benches constitute, due to their static characteristics, necessarily always a compromise between the divergent requirements between the variety of users. Although in everyday life citizens rarely notice imperfectness of park benches, enhanced individual alignment in design and functionality is expected to make a significant difference. When focusing people with infringed constitution, the present imperfectness may even be an effective impediment for usage.

Anthropometric studies provide, for instance, evidence that differences exist in relevant measures between senior citizens, aged 65 or more, and younger citizens [30], [31]. Aligning seat height to individual measures is one way to increase ergonomics, i.e. usability. However, this is only effective *while sitting*. For people with weakened musculature, and limitations to hold the balance the *process of sitting down and standing up* is a serious source of difficulties. And here, a contradiction arises which cannot be resolved by static design: For standing

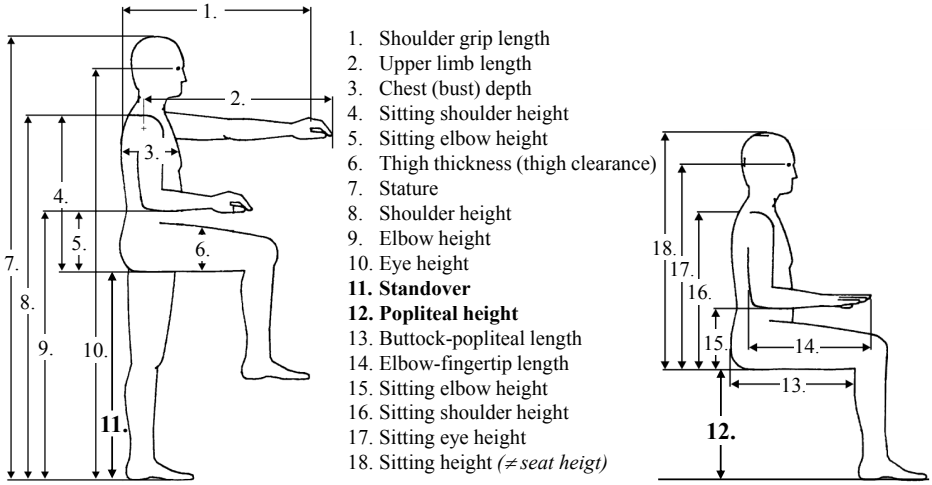

Fig. 2. Anthropometric measures, adopted from [30]; relevant for the adaptive park bench are 11. standover and 12. popliteal height.

	ParkbenchUser	
	Attribute	**Datatype**
core data	id	String
	age	Integer
	authorizationGroup	String
	authorized	Boolean
	standover [mm]	Integer
	poplitealHeight [mm]	Integer
	sittingElbowHeight [mm]	Integer
	maxDistanceUntilRestIsAdvised [m]	Integer
process data	distanceSinceLastRest [m]	Integer

Fig. 3. Datamodel for a park bench user.

up and sitting down, park benches with elevated seat height are easier to use but while sitting they are unergonomic. The same type of contradiction applies to the tilt of the seat. A seat which is tilted forwards facilitates gliding out of the seating position and eases standing up. While sitting, however, a slight tilt backwards is ergonomic. The chance with making park benches adaptive is not only that they are enabled to align to various but static measures. It also allows for a novel type of support for citizens. Active support by the urban environment does not only enhance its usability but makes it also a safer space for pedestrians with handicaps. Among the anthropometric measures depicted in figure 2 the relevant ones, to which the bench shall adapt, are 11. standover and 12. popliteal height.

To reduce difficulties with standing up and sitting down the park bench adjusts it seat height and seat tilt according to the respective phase: At the beginning of sitting down the seat is elevated and tilted forwards. Then it smoothly descends and tilts backwards for ergonomic sitting. For standing up the procedure is just reverse. When a person who needs support with sitting down to a seat approaches the adaptive park bench, the bench, first aligns its seat height to the standover of this person and tilts forwards. After the person has sat or leaned on the elevated seating surface the bench smoothly adjusts the surface to the person's popliteal height and tilts backwards. When the person wishes to stand up the bench slowly tilts forwards and elevates up to the standover again. For persons who do not need support in sitting down and standing up the first and last phase may be simply omitted or reduced in its amplitude (that is: the bench does not elevate up until the standover but only a little). To detect a person, who approaches the bench, short distance radio technology can be used, such as Bluetooth. People carry a respective portable device (which is also required for the adaptive light system). To clearly detect that an approaching person actually wishes to sit down a "manual" interaction is required. The simplest way is to attach a button to the bench or to send a corresponding signal via the required portable device. If no manual interaction is used to indicate the wish to sit down, but a weight sensor recognizing persons sitting on the seat

	AdaptiveParkbench	
	Attribute	**Datatype**
core data	id	Integer
	geoLocation	Array of Real [3]
	widthOfSeat [mm]	Integer
	depthOfSeat [mm]	Integer
	amountOfSeparateSections	Integer
	amountOfArmrests	Integer
	isSeatHeightAdaptive	Boolean
	isSeatTiltAdaptive	Boolean
	isArmrestHeightAdaptive	Boolean
	surfaceMaterial	Enum [timber, plastic metal, stone]
	backrestMaterial	Enum [na, timber, plastic, metal, stone]
process data	currentSeatHeight [mm]	Integer
	currentArmrestHeight [mm]	Integer
	currentSeatTilt [degree]	Integer
	numberOfSittingPersons	Integer

Fig. 4. Datamodel for an adaptive park bench.

surface then the bench can automatically adapt to the person's popliteal height (always assumed that the person carries the device with the required data). The "smartest" way, however, would be to infer that a person must be wishing to sit down. Figure 3 outlines the data structure of a park bench user required for executing the supportive function of the adaptive park bench and figure 4 outlines the data structure for the adaptive park bench.

C. Adaptive Light

As pointed out previously, there is up to now no street lighting system that implements an adaptive lighting configuration depending on individual positions of pedestrians. Therefore we are designing a concept for realizing a smart lighting system and present basic requirements. In section IV we discuss how these basic requirements relate to the overall system.

The adaptive light system shall be aligned to increase well-being of citizens. A particular goal is to enhance participation in urban life for older adults. The adaptive lighting system is therefore used as a means to induce positive impact on well-being in the urban area and to stimulate participation by varying light intensity or color.

As a use case we employ routing to specific points of interest by means of varying light colors and intensity and derive the following basic requirements:

1) Street lamps must be interconnected and be able to communicate with each other via the interconnection technology.
2) The person to whom the lights adapt must be identified within this interconnection structure (same technology as for the identification of park bench users).
3) Information exchange between the person an the light system must be available.
4) The lights must be able to change color and intensity.

IV. Scenario for Interconnected Smart Urban Objects

Figure 5 sketches a scenario to demonstrate the usefulness of interconnecting the SUOs. Figures 6 – 8 show sequence diagrams for interactions of an SUO with the IoT platform and a user by the example of an adaptive park bench. The sequences `sit down` and `stand up` (figures 7 + 8) are specific interactions with the park bench and replace the `interaction` placeholder in figure 6 for the example case that the SUO is an adaptive park bench. For other SUOs the respective interaction is inserted there. For adaptive light, for instance, the interaction could be that the light changes intensity or color for a pedestrian. Note, that a `ParkbenchUser` is meant to be extending the general SUO user class `Pedestrian` by containing additionally the specific attributes required for the park bench interactions.

Figure 6 outlines the intertwined logic between SUOs and the platform. SUOs interact locally with their users but the platform permanently monitors the SUOs and periodically stores the monitored states as the respective SUO data objects. These enable data analytics on interactions. Beyond that, data of several SUOs can be jointly analyzed, the platform facilitates data exchange between SUOs and can actively start control sequences for the SUOs. We now describe a scenario where the two example SUOs interact with each other and generate a benefit for a pedestrian.

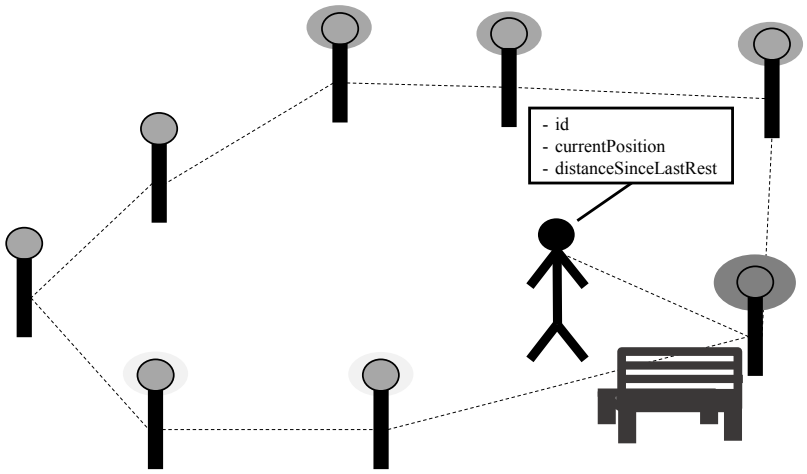

Fig. 5. Interconnected system of adaptive lighting and adaptive park bench

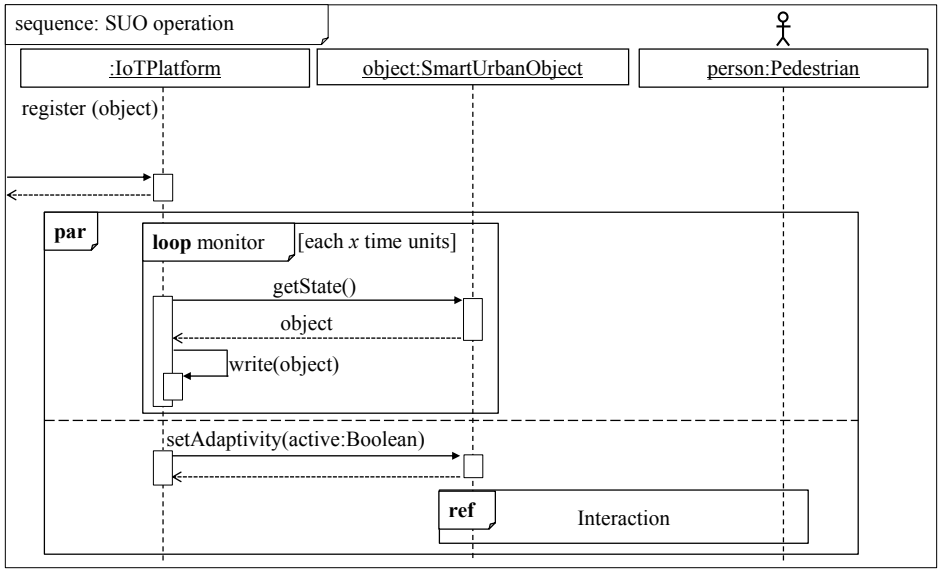

Fig. 6. Sequence diagram for operating SUOs.

The SUOs of this scenario are the adaptive park bench and the adaptive light system. The user is an older adult who needs to rest from time to time.[1] The environment is a pedestrian area which is equipped with adaptive park benches and adaptive lights. The benches are installed every 100 meters, the light posts have a spacing of 25 meters to each other, and they are arranged in such a way that at each park bench, there is a light. As a matter of fact, park benches are sometimes occupied, when needed. This may be indeed a severe problem for older adults needing a rest. In the presented scenario, however, the adaptive lights and park benches may help reducing this problem.

The IoT platform records the interactions of the older adult with the adaptive park bench. So the system knows when the user had his last rest. The adaptive light posts recognize when the person passes so the IoT platform can automatically infer the distance since the person's last rest, too. Having information about the distances after which the older adult is recommended to take a rest, the smart city infrastructure can take actions to maintain a high well-being of this citizen. Indirectly, the system cares for the safety of the older adult.

If an older adult needs a rest in the next time, then the smart city system changes the colors of the lights at reachable and free benches. If the next free park bench is farther away then the lights increase their intensity so the older adult can see the hint to a free park bench from far. While approaching the park bench, which is traceable via the lighting system, the bench adjusts its seat height already to the respective anthropometric measure of this person so that the person can directly sit down comfortably. Position tracking with the light posts promises a higher precision of position determination than using standard

[1]Depending on the constitution this may be even the case every 300 meters.

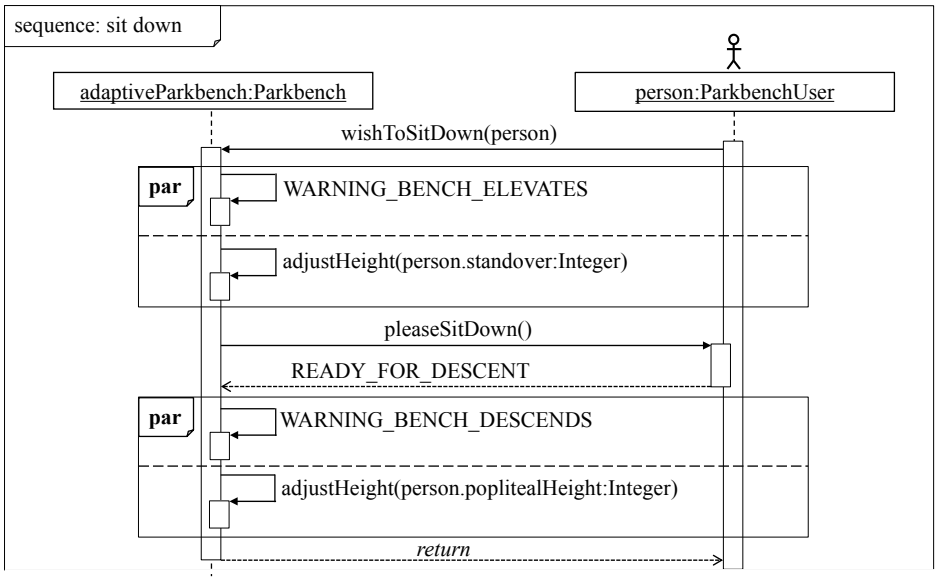

Fig. 7. Sequence diagram for sitting down with support of the adaptive parkbench.

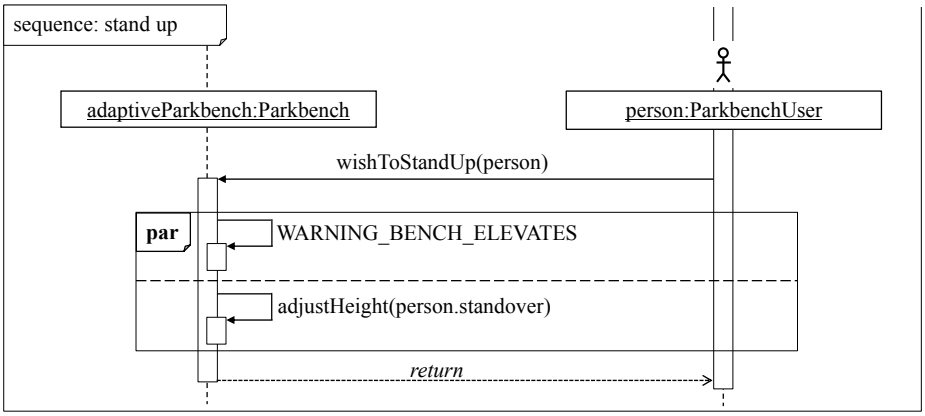

Fig. 8. Sequence diagram for standing up with support of the adaptive parkbench.

GPS alone.

More differentiated support by the smart infrastructure if multiple users need a rest at the same time in the same area is subject to our ongoing research (for the two-user case an adaption rule for simultaneous use of adaptive park bench has been recently studied in [32]).

V. Conclusions and Outlook

We presented designs for SUOs, particularly to facilitate urban life for people with handicaps. As SUOs adapt to personal needs, they are likely to enhance urban life for all citizens. We discussed a scenario where SUOs provide additional benefits for a smart city by interconnecting them with an IoT platform. The IoT platform serves as middleware and provides services to maintain various SUOs. At the same time the platform gathers data from the SUO for the purpose of analyzes. The inferred information are expected to be an essential factor to make modern cities and urban life smarter.

We are aware of privacy issues and societal risks that may occur in environments that digitize individuals. However, research suggests that IT-based services for the urban area can support in counteracting age-related decline of outdoor activities [33], [34]. Our concern is focused on employing SUOs to provide active IT-based support. One of the benefits for which we aim is enhanced safety for individuals in urban environments. To this end, we will further explore which hazards are relevant in citizens everyday activities and how we can mitigate them with interaction and interconnection between smart urban objects and citizens.

Acknowledgement

This work is has been part of the research project URBAN-LIFE+, supported by the Federal Ministry of Education and Research, Germany, under grants 16SV7438K and 16SV7442.

REFERENCES

[1] Cheng, Z. (A.), Pang, M.-S., Pavlou, P. A.: On Intelligent Transportation Systems and Road Congestion. In: 37th International Conference on Information Systems (ICIS 2016), Dublin, Ireland.

[2] Li, Z., Hong, Y., Zhang, Z.: An empirical analysis of on-demand ride sharing and traffic congestion. In: 37th International Conference on Information Systems (ICIS 2016), Dublin, Ireland.

[3] Fridgen, G., Häfner, L., König, C. Sachs, T.: The Value of IS-Enabled Flexibility in Electricity Demand – a Real Options Approach. In: 12th International Conference on Wirtschaftsinformatik (WI 2015), Osnabrück, Germany, 993-1007.

[4] Wanger, S., Brandt, T., Neumann, D.: IS-Centric Business Models for a Sustainable Economy – The Case of Electric Vehicles as Energy Storage. In: 12th International Conference on Wirtschaftsinformatik (WI 2015), Osnabrück, Germany, pp. 1055-1070.

[5] Chowdhury, B., Chowdhury, M. U.: RFID-based Real-time Smart Waste Management System. In: Australasian Telecommunication Networks and Applications Conference 2007 (ATNAC 2007), pp. 175–180. Institute of Electrical and Electronics Engineers, Inc (IEEE), Piscataway, New Jersey.

[6] Glouche, Y., Couderc, P.: A Smart Waste Management with Self-Describing objects. In: 2nd International Conference on Smart Systems, Devices and Technologies (SMART 2013), pp. 63–70.

[7] Wu, F.-J., Kao, Y.-F., Tseng, Y.-C.: From wireless sensor networks towards cyber physical systems. Pervasive and Mobile Computing 7, 397–413 (2011).

[8] Haustein, T., Mischke, J., Schönfeld, F., Willand, I. (eds.): Older people in Germany and the EU. Federal Statistical Office, Wiesbaden, Germany (publ.), 2016.

[9] Nam, T., Pardo, T. A.: Smart City as Urban Innovation: Focusing on Management, Policy, and Context. In: Proceedings of the 5th International Conference on Theory and Practice of Electronic Governance (ICEGOV 2011), 185–194. ACM (2011).

[10] Caragliu, A., Del Bo., C., Nijkamp, P.: Smart Cities in Europe. Journal of Urban Technology 18:2, 65–82 (2011).

[11] Neirotti, P., De Marco, A., Cagliono, A. C., Mangano, G., Scorrano, F.: Current trends in Smart City initiatives: Some stylised facts. Cities 38, 25-36 (2014).

[12] Dameri, R. P.: Searching for Smart City definition: a comprehensive proposal. International Journal of Computers & Technology 11:5, 2544–2551 (2013).

[13] Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J. R., Mellouli, S., Nahon, K., Pardo T. A., Scholl, H. J.: Understanding Smart Cities: An Integrative Framework. In: Proceedings of the 45th Hawaii International Conference on System Sciences (HICSS 2012), 2289–2297, IEEE.

[14] Vasseur, J.-P., Dunkels, A.: Smart cities and urban networks. In: Interconnecting Smart Objects with IP: The Next Internet, ch. 22, pp. 335–352, Morgan Kaufmann, Burlington (2010).

[15] Mineraud, J., Mazhelis, O., Su, X., Tarkoma, S.: A gap analysis of Internet-of-Things platforms. Computer Communications, 89, 5–16 (2016).

[16] Razzaque, M. A., Milojevic-Jevric, M., Palade, A., Clarke, S.: Middleware for internet of things: a survey. IEEE Internet of Things Journal 3(1), 70-95 (2016).

[17] Vasilomanolakis, E., Daubert, J., Luthra, M., Gazis, V., Wiesmaier, A., Kikiras, P.: On the Security and Privacy of Internet of Things Architectures and Systems. In: Proceedings of the International Workshop on Secure Internet of Things (SIoT 2015), 49–57, IEEE.

[18] Wehlitz, R., Häberlein, D., Zschörnig, T., Franczyk, B.: A Smart Energy Platform for the Internet of Things – Motivation, Challenges, and Solution Proposal. In: Abramowicz W. (eds.) Business Information Systems (BIS). Lecture Notes in Business Information Processing 288. Springer, Cham (2017).

[19] Dekel, A., Simon, Y., Dar, H., Tarazi, E., Rabinowitz, O., Sterman, Y.: Adding Playful Interaction to Public Spaces. In: Intelligent Technologies for Interactive Entertainment. Springer, Berlin, Heidelberg, 225229 (2005).

[20] *https://www.3fficient.com/shop/solar-smart-bench/* , last accessed: 2018-07-18.

[21] *http://www.soofa.co/getsoofa/* , last accessed: 2018-07-18.

[22] Beucker, N., Zurnatzis, M.: Stadtmobiliar für Senioren. Ausstattungskriterien für eine altengerechte Stadt [Urban furniture for senior citizens. Configuration criteria for an elderly-friendly city]. Technical Report, University of Applied Sciences Niederrhein, Germany (2011).

[23] *https://abes-online.com/en/product/bench-1-110-1/* last accessed: 2018-07-18.

[24] Nippuni Kumaar, A. A., Kiran, G., Sudarshan, T. S. B.: Intelligent Lighting System Using Wireless Sensor Networks. Int. Journ. of Ad hoc., Sensor & Ubiquitous Computing (IJASUC) 1:4, 17–27, (2010). doi:10.5121/ijasuc.2010.1402

[25] Crowther, J., Herzig, C., Feller, G.: The Time Is Right for Connected Public Lighting Within Smart Cities. Cisco Internet Business Solutions Group (IBSG) (2012)

[26] Escolar, S., Carretero, J., Marinescu, M., Chessa, S.: Estimating Energy Savings in Smart Street Lighting by Using an Adaptive Control System. International Journal of Distributed Sensor Networks, 117, (2014). doi:10.1155/2014/971587

[27] Jackson, H., Jackson, S., Jackson, C., Siminovitch, M.: Saving Energy in Buildings with Adaptive Lighting Solutions. California Energy Commission. California Lighting Technology Center, UC Davis (2015).

[28] Huldtgren, A., Katsimerou, C., Kuijsters, A., Redi, J. A., Heynderickx, I. E. J.: Design Considerations for Adaptive Lighting to Improve Seniors Mood. Inclusive Smart Cities and e-Health, 9102. In: Lecture Notes in Computer Science, 1526, Springer International Publishing (2015).

[29] Kuijsters, A., Redi, J., Ruyter, B. d., Heynderickx, I.: Lighting to Make You Feel Better: Improving the Mood of Elderly People with Affective Ambiences. PloS one 10:7 (2015). doi:10.1371/journal.pone.0132732

[30] Kothiyal, K., Tettey, S.: Antropometry for Design for the Elderly. Int. Journ. of Occupational Safety and Ergonomics 7:1, 15–34 (2001).

[31] Adler, M., Herrmann, H.-J., Koldehoff, M., Meuser, V., Scheuer, S., Müller-Arnecke, H., Windel, A., Bleyer, T.: Ergonomiekompendium: Anwendung ergonomischer Regeln und Prüfung der Gebrauchstauglichkeit von Produkten [Ergonomics compendium: Application of ergonomic rules and assessment of usability of products]. Federal Institute for Occupational Safety and Health (eds.), Germany (2010).

[32] Hubl, M.: Adaption Rule for Simultaneous Use of Smart Urban Objects from a Fairness Perspective. In: Proceedings of the 20th International Conference on Business Informatics (CBI 2018), 89–98, IEEE. doi:10.1109/CBI.2018.00019

[33] Fietkau, J.: The case for including senior citizens in the playable city. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, Leipzig, Germany, August 2017 (WI'2017), 1072–1075. doi:10.1145/3106426.3109042

[34] Leukel, M., Schehl, B., Wallrafen, S., Hubl, M.: Impact of IT Use by Older Adults on Their Outdoor Activities. In: Proceedings of the 38th International Conference on Information Systems (ICIS 2017).

Information Technology for Management, Business & Society

IT4MBS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the disciplines of information technology and information systems. The IT4BMS area emphasizes the issues relevant to information technology and necessary for practical, everyday needs of business, other organizations and society at large. This area takes a sociotechnical view on information systems and relates also to ethical, social and political issues raised by information systems. Events that constitute IT4BMS are:

- AITM'18—16th Conference on Advanced Information Technologies for Management
- ISM'18—13th Conference on Information Systems Management
- KAM'18—24th Conference on Knowledge Acquisition and Management

AREA SUPERVISORY COMMITTEE

- Carnero Moya, Maria del Carmen, AITSD'18
- Chmielarz, Witold, ISM'18
- Gontar, Beata, IT4L'18
- Komenda, Martin, TEMHE'18
- Korczak, Jerzy, AITM'18
- Pondel, Maciej, KAM'18

24th Conference on Knowledge Acquisition and Management

KNOWLEDGE management is a large multidisciplinary field having its roots in Management and Artificial Intelligence. Activity of an extended organization should be supported by an organized and optimized flow of knowledge to effectively help all participants in their work.

We have the pleasure to invite you to contribute to and to participate in the conference “Knowledge Acquisition and Management". The predecessor of the KAM conference has been organized for the first time in 1992, as a venue for scientists and practitioners to address different aspects of usage of advanced information technologies in management, with focus on intelligent techniques and knowledge management. In 2003 the conference changed somewhat its focus and was organized for the first under its current name. Furthermore, the KAM conference became an international event, with participants from around the world. In 2012 we've joined to Federated Conference on Computer Science and Systems becoming one of the oldest event.

The aim of this event is to create possibility of presenting and discussing approaches, techniques and tools in the knowledge acquisition and other knowledge management areas with focus on contribution of artificial intelligence for improvement of human-machine intelligence and face the challenges of this century. We expect that the conference&workshop will enable exchange of information and experiences, and delve into current trends of methodological, technological and implementation aspects of knowledge management processes.

TOPICS

- Knowledge discovery from databases and data warehouses
- Methods and tools for knowledge acquisition
- New emerging technologies for management
- Organizing the knowledge centers and knowledge distribution
- Knowledge creation and validation
- Knowledge dynamics and machine learning
- Distance learning and knowledge sharing
- Knowledge representation models
- Management of enterprise knowledge versus personal knowledge
- Knowledge managers and workers
- Knowledge coaching and diffusion
- Knowledge engineering and software engineering
- Managerial knowledge evolution with focus on managing of best practice and cooperative activities
- Knowledge grid and social networks
- Knowledge management for design, innovation and eco-innovation process
- Business Intelligence environment for supporting knowledge management
- Knowledge management in virtual advisors and training
- Management of the innovation and eco-innovation process
- Human-machine interfaces and knowledge visualization

EVENT CHAIRS

- **Hauke, Krzysztof,** Wroclaw University of Economics, Poland
- **Nycz, Malgorzata,** Wroclaw University of Economics, Poland
- **Owoc, Mieczyslaw,** Wroclaw University of Economics, Poland
- **Pondel, Maciej,** Wroclaw University of Economics, Poland

PROGRAM COMMITTEE

- **Abramowicz, Witold,** Poznan University of Economics, Poland
- **Andres, Frederic,** National Institute of Informatics, Tokyo, Japan
- **Bodyanskiy, Yevgeniy,** Kharkiv National University of Radio Electronics, Ukraine
- **Chmielarz, Witold,** Warsaw University, Poland
- **Christozov, Dimitar,** American University in Bulgaria, Bulgaria
- **Jan, Vanthienen,** Katholike Universiteit Leuven, Belgium
- **Mach-Król, Maria,** University of Economics in Katowice, Poland
- **Mercier-Laurent, Eunika,** University Jean Moulin Lyon3, France
- **Sobińska, Małgorzata,** Wroclaw University of Economics, Poland
- **Surma, Jerzy,** Warsaw School of Economics, Poland and University of Massachusetts Lowell, United States
- **Vasiliev, Julian,** University of Economics in Varna, Bulgaria
- **Zhu, Yungang,** College of Computer Science and Technology, Jilin University, China

ORGANIZING COMMITTEE

- **Hołowińska, Katarzyna**
- **Przysucha, Łukasz,** Wroclaw University of Economics

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 63–68
DOI: 10.15439/2018F379
ISSN 2300-5963 ACSIS, Vol. 16

The Potential of the Internet of Things in Knowledge Management System

Artur Rot
Wroclaw University of Economics
ul. Komandorska 118/120
53-345 Wroclaw, Poland
Email: artur.rot@ue.wroc.pl

Malgorzata Sobinska
Wroclaw University of Economics
ul. Komandorska 118/120
53-345 Wroclaw, Poland
Email: malgorzata.sobinska@ue.wroc.pl

***Abstract*—Along with the increasing globalization and development of information and communication technology, business models are changing, and thus the need for innovative knowledge management is growing. Current knowledge management systems very often are not used optimally/effectively for decision-making because of the lack of real-time data. This article draws attention to the current trend in the area of organization management and IT management related to the emergence and growing popularity of the Internet of Things. The authors try to assess the potential of IoT in the context of improving knowledge processes (locating, acquiring, using, sharing and disseminating as well as preserving / coding / archiving / collecting), especially acquiring and sharing data and determining IoT impact on knowledge management and learning of the organization. The possible positive effects of implementing IoT in enterprises of various types, as well as threats and challenges that must be met by organizations that care about increasing their competitive position using IoT will be presented.**

I. Introduction

Success of organisations is, to a large extent, a function of their investment in knowledge and their ability to accumulate and create knowledge. Management of organisations in the 21st century is inextricably linked with creative application of knowledge processes management techniques and the organisational capacity to adapt to changing conditions, based on the rapid development of modern technologies. This approach offers potential of largely improving the decision-making effectiveness, operating capacity, flexibility and employee involvement in organisations.

Organisational processes are the basis for making informed decisions based on knowledge. The development and utilisation of intelligent solutions – such as those involving Big Data processing and the Internet of Things – seems to place strong emphasis on open, systemic approach. For instance, the introduction of fully digitalised services paves the way for the design of new interfaces between service providers and end users.

Proper selection and continuous improvement of ICT-based business models seems particularly important for those organisations which place emphasis on cognisant application of knowledge management techniques and on the practical implementation of the concept of 'the learning organisation" where learning processes are perceived as fundamental for building and maintaining the competitive advantage through continuous advancement and exploration of new paths for organisational development. Organisations of this type are usually more open and receptive to signals from their environment and are more capable of forging those signals into tangible benefits: learning, improvement, and change. As aptly noted by Stańczyk-Hugiet, the most fundamental process in a learning organisation is the one involving transformation of information into knowledge resources which then are used as basis for building a unique selling position on the market [6, p. 136].

Knowledge has become the most important resource of any organisation and a key to its continued growth. Similarly, Kaivo-oja et al. argue that intelligent organisations are those that focus not so much on knowledge production but rather on proper integration of knowledge. In this way, knowledge integration has been elevated to the rank of a key component of a modern management system [11].

The Internet of Things (IoT), defined as the sum of intelligent appliances capable of responding to environmental stimuli as well as storing and processing of digital information and relaying it to other agents (or users) via Internet protocols, seems to offer best potential for supporting the aforementioned task of turning information into knowledge. This context needs careful research, as the subject of IoT and its potential impact on knowledge management systems is, thus far, inadequately represented in professional literature.

This paper places the particular emphasis on capabilities offered by Internet of Things in the context of process optimisation in tasks involving accumulation, transfer, and filtering of data to be used in decision-making scenarios across various operating segments of a contemporary company.

©2018, PTI

II. KNOWLEDGE MANAGEMENT – IT INSTRUMENTS IN SUPPORT OF KM

From a managerial perspective, companies are in urgent need of developing and improving their capacity to operate under the constantly changing, non-linear and largely unforeseeable conditions brought about by the challenges of modern digital evolution. In view of the above, effective knowledge management should focus, most of all, on the following: improving the time of response to environmental stimuli through better access to knowledge and applications; creating and maintaining potential for company growth and development; improving the company value and/or profitability; improving company products and services, stimulating the production of knowledge, reducing the cost of operation, ensuring constant improvement of quality and effectiveness. However, as observed by Uden and He, the traditional knowledge management systems are burdened with certain limitations.

The authors postulate that the next generation of knowledge management systems in the age of IoT should offer the following capabilities [7]:

- Creation of personalised experience and tracking of its various sources;
- Transfer and reorganisation of data from various sources;
- Data optimisation and pre-processing of data for various purposes and applications;
- Provision of on-demand and real-time updates of data used in decision-making contexts;
- Collection of machine data and user activity data for process improvement purposes;
- Ensuring the fluidity and interoperability of the various components of the knowledge management system.

It seems valid to assume that the informative potential of the Internet with respect to the management of large data repositories and knowledge discovery is growing at a rapid pace, with the advance of such technologies as the IoT, intelligent computing, machine learning, big data, cloud computing and sensor technologies. Those and other modern IT solutions may largely improve the effectiveness of data management and knowledge discovery processes in a modern organisation. One of the prime objectives of knowledge management is the effective filtering of data originated from unreliable sources.

At present, the use of IoT in modern knowledge management systems offers potential for not only real-time updates and synthetic evaluation of data from various sources, but also the capacity for knowledge discovery. With the growing popularity of the Internet as a medium of communication, modern companies are more inclined to redefine their existing business models, i.e. the company prime objectives and approaches to value creation. By basing their business models on the IoT, organisations may benefit from the wealth of new channels and means for value exchange. At the same time, the IoT may be perceived as an innovation, since it involves a creative use of the existing services and protocols for the development of new, more responsive and dynamic products and services. When viewed from the context of an innovation, the IoT represents a broad spectrum of intelligent technologies capable of communicating directly with the user or with other machines and agents to offer a brand new level of interaction or information retrieval in any environment populated by such objects [12].

III. NEW CAPABILITIES FOR THE RETRIEVAL OF INFORMATION AND KNOWLEDGE FOR DECISION-MAKING PURPOSES

Organisations employ information and knowledge both for the improvement of their decision-making processes and for the legitimisation of decisions (including those issued on the basis of unreliable information). In addition, it seems valid to assume that modern organisations retrieve and store much more information than needed for their decision-making purposes, and still call for more. The overabundant supply of information may come as a result of various forces and trends at play. For instance, organisations tend to underestimate the cost of information retrieval in proportion to the benefits offered. Usually, decisions made with respect to information are delegated to the authority of units other than those directly involved in its retrieval. By maintaining such clear separation between information use and information retrieval/storage, decision makers are able to initiate information retrieval procedures they deem of potential value for them, but without taking heed of the real cost-benefit ratio as perceived from the viewpoint of the organisation as a whole. The rationality of information oversupply can also be viewed in the context of strategic value of information, particularly when such information is employed to enforce certain response rather than serve as basis for rational decisions. In organisational settings, information is rarely of neutral character. Most of the information feed is prone to misinterpretation. From the viewpoint of decision makers, it is better to have access to information deemed unproductive (for decision-making purposes) than to bear the risk of missing information that may ultimately prove valuable [11].

Both the Internet of Things and cloud computing offer potential to simplify the identification of requirements to be addressed by knowledge management systems and to improve their effectiveness by providing real-time access to data/information ("just in time" and "right on point").

The Internet of Things is also perceived as another important factor to affect the transaction processes, most specifically: the methods for identifying and satisfying customer needs and requirements. Based on different recent reports and surveys, the number of objects connected to the Internet, including smartphones, is already in excess of the total Earth population and, according to the Gartner report, will reach 26 billion wireless devices by the end of 2020 [14].

IV. The Internet of Things – selected areas of application

The Internet of Things can be defined as the sum of intelligent appliances capable of responding to environmental stimuli as well as storing and processing of digital information and relaying it to other agents (or users) via Internet protocols [15]. It represents a networking structure of independent devices designed to enable and facilitate decentralised communication between each and every node. The concept is based on continued technological advancement and utilisation of the existing structures of communication to serve as a medium for unrestrained exchange and sharing of information between numerous devices, appliances and sensors.

It must be noted at this point that the Internet of Things can also be perceived as the latest fashion and may be prone to various interpretations. The term itself may relate to any clearly identifiable object capable of – directly or indirectly – storing and processing data. The number of such objects grows exponentially, along with the spectrum of their potential applications. The IoT is closely associated with the Big Data concept, since the effective provision and processing of large amounts of data is deemed to be the core benefit offered by IoT solutions [13].

There are many potential areas of IoT application and the solution may be adopted with success in many aspects of modern life. According to the Gartner report, the IoT will soon generate revenue in excess of USD 300bn, mostly in services [14]. Similarly, the McKinsey&Company [16] report suggests that the IoT may soon produce global economic profit between USD 2.7 and 6.2 trillion by the year 2025. IoT will find its use in many areas of economic activity and in many sectors, including energy, production, logistics, health care, IT. With further rapid expansion (which seems fairly inevitable), the IoT will manifest its potential in other areas, such as intelligent construction, intelligent vehicles and cities, and in industrial automation (often referred to as Industry 4.0). Table I presents one of the most popular categorisations of potential IoT applications, as postulated by Beecham Research [17].

There are certain areas of the socio-economic life which offer particular benefits from the use of IoT solutions, in terms of value creation. According to the report by McKinsey&Company consulting agency, nine such areas can be distinguished [18]: human, home, retail environments, office, factories, worksites (e.g. oil rigs), vehicles, cities, and outside areas (i.e. spaces between urbanised areas). The Internet of Things will find many uses in various segments of business and services sector.

At the same time, the potential economic impact of the IoT will vary from segment to segment. The McKinsey forecast suggests that, by the year 2025, the highest revenues will be observed in production (1.2–3.7 trillion USD), and the lowest – in office environments (70–150 billion USD) [18].

Table I.
The Internet of Things – areas of application

No.	Sector	Examples of applications
1	Construction	Automated heating, ventilation, air-conditioning, access control, lighting, security systems, etc.
2	Energy	Resource extraction and mining, exploration of alternative (renewable) energy sources, electric generators.
3	Household and consumer goods	Household security (alarms, monitoring of seniors and children), control and regulation of devices (including energy supply and lighting), home entertainment systems.
4	Health care and natural sciences	Telemedicine, remote monitoring of patients and seniors (e.g. patients with cardiac pacemakers), research and development of drugs and medical equipment,
5	Industry/production	Monitoring and tracking of assets, machinery and industrial products, location analyses for a wide range of industrial and factory processes.
6	Transport	Fleet management (navigation systems, distribution management), passenger information systems, automated toll/fee collection systems (motorways, parking).
7	Retail	Supply chain management, product/client information management, stock management, vending machines (foods, beverages), parking meters, information display systems (billboards, digital displays).
8	Public safety	Monitoring of environmental hazards, meteorological and climate information, tracking of persons/animals/packages, military security.
9	IT sector	Office appliances, mobile communication infrastructure, data relay centres (automated energy and air-conditioning relay systems), e-commerce, etc.

To sum up, it may be stated that the Internet of Things offers potential support in nearly all areas of human life and has already become an indispensable element of modern existence in an information society.

Lastly, by connecting many of their home or office devices to the Internet, individual consumers and organisations may receive additional benefits, such as energy savings or new business opportunities.

V. The organisational potential of the IoT

With the rapid changes observed in modern economic environment, companies face more and more fierce competition and are expected to meet the mounting expectations of their customers. To oppose this trend, modern companies place great emphasis on improving the effectiveness of their day-to-day operations. For this purpose, company management needs constant and real-time

access to reliable and verifiable information on the present condition of their enterprise or unit, as a prerequisite for making accurate and informed business decisions. Lambert believes that the world at present is witness to a global technological revolution, and that the Internet of Things is only a part of a large-scale phenomenon.

It is true that the IoT has become an indispensable element of most any modern technological solution employed in business. By making good use of constant connectivity between integrated devices, human resources and data, companies will be better equipped to analyse and respond to the changing needs of their customers and effect prompt changes in their supply chains or innovative implementations. The IoT may also improve the quality of life (for customers, consumers) by offering remote and automated fee collection, monitoring of health, etc. In addition, the IoT may soon dominate such areas of everyday life as waste management, energy management or traffic control [20].

Data presented by T-Mobile suggests that investment in the IoT brings tangible returns in terms of employee effectiveness, utilisation of material resources, cost of operation, monitoring effectiveness, employee safety, integrated supply chain management, and creation of value from new products and services [21].

As suggested by the authors of the IDC report, companies and consumers as well as the whole ecosystem of IoT solutions suppliers may soon see many benefits from the continued advancement of the associated technologies. These may include the following [22]:

– New business models. Faced with a highly competitive markets, both sellers and enterprises constantly seek new ways of conducting their business and new methods for effective creation of customer value. Companies operating within the B2B model may utilise IoT solutions to automate some of their business processes to gain fast access to the market and to better respond to changes in customer needs.
– Mission-critical processes and products. IoT solutions may help companies gain more data on own processes and products. In some cases, such supplementary data may provide crucial and valuable information of corporate character, to allow for operational changes and real-time decision-making. In other scenarios, such supplementary data may be used to support curative and diagnostic decisions in patient care or improve the accuracy of shopping suggestions and offers designed for specific customer segments.
– Diversification of revenues. The Internet of Things may offer potential for innovations, particularly to sellers and service providers, which may result in new products and services.
– For some types of enterprises, IoT may prove a valuable source of supplementary income, attached to the servicing of products in circulation. For instance, providers of vending machines may offer a range of associated services, such as the automated monitoring and resupply of stock.
– Global players. By making good use of the IoT functionality, organisations may gain improved access to their business structures and units (including remote locations), with potential to monitor, manage, and track the operation of each independent node. For some enterprises, such a transition may prove of particular value, by offering remote resource and product management capabilities.
– Efficient intelligent operations. The Internet of Things will greatly affect the way companies make their decisions, as optimisation and productivity improvements may now be managed in real time. With instantaneous access to relevant information on independent nodes – based on the use of own or customer resources – and with potential to promptly address any problem revealed by the data feed, organisations may find reasons to redesign their business processes to improve their long-term effectiveness and productivity.
– Increased expectations. With more consumers, organisations and institutions being aware of the life quality improvements and savings offered by products and services on the market, their expectations towards new products and services will naturally rise. This will stimulate further increase of market competition between various actors and suppliers involved in the IoT segment (ecosystem).

It goes without saying that the utilisation of the full potential offered by the IoT solutions will not be easy, since organisations need new sets of skills and technological tools to ensure proper implementation and protection of large systems responsible for the servicing of hundreds of thousands, if not millions, of individual, spatially dispersed and independent nodes [23].

VI. THE BENEFITS OF INTERNET OF THINGS IN SELECTED AREAS OF KNOWLEDGE MANAGEMENT

The Internet of Things also brings a number of benefits in the area of the knowledge management system. The most important of them, combined with potential risk areas, are presented in the Table II.

Knowledge management today has also the opportunity and the capability to synthesise data from diverse sources and to create new knowledge. Because of the IoT, businesses today are forced to rethink their goals. With the IoT, products can be tracked anytime, making it possible to respond to customer behaviour. It is also now possible for products to connect with other products, leading to new analytics and new services for more effective forecasting, process optimisation and customer service experiences [24] [28].

TABLE II .
THE MAIN BENEFITS AND RISKS OF IMPLEMENTATION OF INTERNET OF THINGS IN SELECTED AREAS OF KNOWLEDGE MANAGEMENT

Areas/objectives of knowledge management	The potential of Internet of Things	The risk associated with the IoT implementation
Strategy/development of knowledge and competence resources	Access to external expertise, the latest technology; providing tools to enable access to organizational and technological knowledge; providing tools for support of the transfer and sharing of knowledge.	Need of better control over the IoT environment; Dependence on the suppliers of IoT.
Human resources management/ development of intellectual capital	Developing new skills and competences in the field of IT; Improved cooperation and transfer of knowledge between organizations.	Possible loss of intellectual capital in case of the dismissal of IT staff as a result of the adaptation of IoT solutions; The need to improve knowledge and to develop new skills to use the new technology (IoT)
Process management	Supporting innovation; Reducing the time of process implementation.	
Marketing/business intelligence/innovation	Possibly unlimited access to internal and external data. Ensuring the possibility of direct integration of external entities with the company; Supporting market-driven innovation (for products/services).	The possibility of losing competitive advantage by acquiring strategic information or sensitive data by the competitors.
Information technology/Information security management	Improved use of IT resources; Possibility of collecting data from diverse products, company assets or the operating environment; Possibility of the transmission of real-time data from wireless network	Loss of control over the IT environment; System failures; Services unfit for the actual needs of the organization; The possible decrease in safety; The possibility of loss and/or unauthorized use of sensitive data.

Sensors and embedded technology now enable the transmission of real-time data from wireless networks which will lead to the co-creation of new real-time knowledge among customers and vendors. Companies which use the IoT can gather data about how their products behave and interact, and can then use it to understand and predict future behaviours [7].

Companies can also use the data collected from sensors to change the way that they design, upgrade and maintain devices in the field. The result is not merely greater efficiency, but entirely new functionality and levels of service. Real-time monitoring and analysis of physical assets allow companies to understand and act on a variety of real-time insights. The real-time data can be accessible by all the stakeholders, thereby facilitating knowledge sharing. Organisations using KM have the ability to interpret the real-time data and convert them into real-time knowledge for competitive advantage [7].

IoT can also help KM to capture data to be used in organisations. The world of connected devices today presents the opportunity for organisations to create new product experiences.

The increasing use of internet-enabled devices with sensors provides more opportunities both to improve the way services are delivered and to harness those data to gain faster insights into whether interventions are working. Manufacturers are using data obtained from sensors embedded in products to create innovative after-sales service offerings such as proactive maintenance to avoid failures in new products [25] [26].

The data also enable companies to design better products in the future. KM today has the opportunity and the capability to synthesise data from diverse sources and arrive at new knowledge. The benefits of IoT for KM can be summarised as follows [7] [27]:

- IoT enables a user to collect data from diverse products, company assets or the operating environment. It allows for the generation of better information and analysis, which can significantly enhance decision-making.
- When products are embedded with sensors, companies can track the movements of these products and can monitor interactions with them. Businesses can take advantage of customers' behavioural data to make appropriate decisions.
- Data from large number of sensors, deployed in infrastructure (such as roads and buildings), can give decision makers a heightened awareness of real-time events, particularly when the sensors are used with advanced display or visualization technologies [27].
- The IoT can support longer-range and more complex human planning and decision-making.
- The IoT can raise productivity, because the IoT can help systems adjust automatically to complex situations, which can make a number of human interventions unnecessary.
- The IoT enables rapid, real-time sensing of unpredictable conditions and instantaneous responses guided by automated systems.

VII. CONCLUSION

New, breakthrough technologies, in particular the Internet of Things, change the way knowledge is managed in organizations, forcing the implementation of innovative knowledge management systems and an open approach to supporting knowledge flows. Bearing in mind their importance already today, IoT and Big Data most definitely are key factors affecting societal development in the future. Private and public organizations have begun to gain critical insights from the Big Data and ubiquitous technology through various management systems. Basically, the issue at stake here is the fact that it is not just the question how to manage and control the technological possibilities. The development also concern leadership functions. Namely, taking seriously Internet of Things and ubiquitous technology may lead towards the revolution of digitalization which effects on knowledge management processes in organizations.

The authors intend to conduct further, more detailed research on the impact of IoT on knowledge management processes in organizations, such as: locating knowledge, acquiring knowledge, developing knowledge, sharing, using and protecting knowledge.

REFERENCES

[1] A. Rot, M. Sobińska M., "Knowledge management in the cloud computing model - challenges, opportunities and risks", *in Position Papers of the 2017 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, M. Paprzycki (eds.), *Annals of Computer Science and Information Systems*, vol. 12, 2017, Polskie Wydawnictwo Informatyczne, Warsaw, 2017, pp. 177-181.

[2] M. Dolińska, *Innovations in knowledge based economy*, (In Polish) Warsaw, Polskie Wydawnictwo Ekonomiczne, 2010.

[3] E. Stańczyk-Hugiet, *Strategic context of knowledge management*, (In Polish), Wroclaw University of Economics Publishing House, Wroclaw, 2007.

[4] L. Uden, W. He, "How the Internet of Things can help knowledge management: a case study from the automotive domain", *Journal of Knowledge Management*, Vol. 21 Issue: 1, 2017, https://www.emeraldinsight.com/doi/pdfplus/10.1108/JKM-07-2015-0291, pp.57-70.

[5] U. Ganesh, *Internet of things: Impact on learning and knowledge management*, Financial Express 22-nd of December 2014, https://www.financialexpress.com/industry/technology/internet-of-things-impact-on-learning-and-knowledge-management/21816/

[6] G. Santoro, D. Vrontis, A. Thrassou, L. Dezi , "The Internet of Things: Building a knowledge management system for open innovation and knowledge management capacity", *Technological Forecasting and Social Change*, Available online 16 March 2017, Elsevier, 2017, https://www.sciencedirect.com/science/article/pii/S0040162517302846?via%3Dihub

[7] M. Krysiński, "The Internet of Things - innovative tool for companies", (In Polish) *Ekonomiczne Problemy Usług* no. 122, University of Szczecin, Szczecin 2016, http://www.wzieu.pl/zn/epu/122/epu_122.pdf, pp. 281-290.

[8] J. Kaivo-oja, P. Virtanen, H. Jalonen, J. Stenvall, "The Effects of the Internet of Things and Big Data to Organizations and Their Knowledge Management Practices", *Conference Paper in Lecture Notes in Business Information Processing*, August 2015.

[9] A. Murray, A. Papa, B. Cuozzo, G. Russo, "Evaluating the innovation of the Internet of Things: Empirical evidence from the intellectual capital assessment", *Business Process Management Journal*, Vol. 22 Issue: 2/2016, Emerald Insight, https://www.emeraldinsight.com/doi/abs/10.1108/BPMJ-05-2015-0077, pp. 341-356.

[10] J. Holler, V. Tsiatsis, *From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence*, Elsevier, 2014.

[11] P. Middleton, P. Kjeldsen, J. Tully, *Forecast: The Internet of Things, Worldwide 2013*, Gartner, November 2013, www.gartner.com/doc/2625419/forecast-internet-things-worldwide-.

[12] W. Nowakowski, "Closer to the cloud, computational services in the fog" (In Polish), *Elektronika – konstrukcje, technologie, zastosowania* nr 5/2015, Instytut Maszyn Matematycznych, Warsaw 2015, http://www.imm.org.pl/imm/plik/pliki-do-pobrania-elektronika52015_nn358.pdf

[13] McKinsey&Company, *The Internet Of Things: Mapping The Value Beyond The Hype*, McKinsey Global Institute, 2015, http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/the-internet-of-things-the-value-of-digitizing-the-physical-world.

[14] Beecham Research, *IoT Sector Map*, 2016, http://www.beechamresearch.com/article.aspx?id=4.

[15] H. Bauer, M. Patel, J. Veira, *The Internet of Things: Sizing up the opportunity*, http://www.mckinsey.com/industries/high-tech/our-insights/the-internet-of-things-sizing-up-the-opportunity, McKinsey&Company, 2014.

[16] M Van Rijmenam M., *Three Major Challenges for the Internet of Things*, http://www.datafloq.com/read/three-major-challenges-internet-of-things/83

[17] EY, *Insights on governance, risk and compliance: Cybersecurity and the Internet of Things*, Ernst & Young 2015, www.ey.com/Publication/vwLUAssets/EY-cybersecurity-and-the-internet-of-things/$FILE/EY-cybersecurity-and-the-internet-of-things.pdf

[18] D. Niyato, *Economics of Internet of Things (IoT): An Information Market Approach*, Computer Science, 2012.

[19] D. Lund, C. MacGillivray, V. Turner, M. Morales, *Worldwide and Regional Internet of Things (IoT) 2014–2020 Forecast: A Virtuous Circle of Proven Value and Demand*, Worldwide Semiannual Internet of Things Spending Guide IDC, 2014, https://www.business.att.com/content/article/IoT-worldwide_regional_2014-2020 -forecast.pdf.

[20] A. Rot, B. Blaicke B., "The Internet of Things Security. Selected threats and countermeasures. Manufacturing systems case study" [In Polish], *Management*, Czestochowa Technical University, no. 26, 2017, pp. 188-198.

[21] G. Hui, *How the Internet of Things Changes Business Models*, Harvard Business Review, 29 July 2014, https://hbr.org/2014/07/how-the-internet-of-things-changes-business-models

[22] L. Xu, W. He, S. Li, *Internet of Things in industries: a survey*, IEEE Transactions on Industrial Informatics, Vol. 10 No. 4, 2014, pp. 2233-2243.

[23] W. He, L. Xu, "A state-of-the-art survey of cloud manufacturing", *International Journal of Computer Integrated Manufacturing*, Vol. 28 No. 3, 2015, pp. 239-250.

[24] Z. Bi, G. Wang, L.D. Xu, "A visualization platform for Internet of Things in manufacturing applications", *Internet Research*, Vol. 26 No. 2, 2016, pp. 377-401.

[25] K. H. Badimo, *The Increasing Importance of Knowledge Management in the Digital World through the Internet of Things*, https://www.linkedin.com/pulse/increasing-importance-knowledge-management-digital-world-kgabo-badimo/

[26] A. Rot, M. Sobińska, "IT security threats in cloud computing sourcing model", *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, M. Paprzycki (eds.), 2013, IEEE, pp. 1141-1144.

Software Systems Development & Applications

SSD&A is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the discipline of software engineering. The SSD&A area emphasizes the issues relevant to developing and maintaining software systems that behave reliably, efficiently and effectively. This area investigates both established traditional approaches and modern emerging approaches to large software production and evolution. Events that constitute SSD&A are:

- MDASD'18—5th Workshop on Model Driven Approaches in System Development
- MIDI'18- 6th Conference on Multimedia, Interaction, Design and Innovation
- LASD'18—2nd International Conference on Lean and Agile Software Development
- SEW-38 & IWCPS-5—Joint 38th IEEE Software Engineering Workshop (SEW-38) and 5th International Workshop on Cyber-Physical Systems (IWCPS-5)

AREA SUPERVISORY COMMITTEE

- Hinchey, Mike, SEW-38-IWCPS-5
- Kornecki, Andrew J., MMAP'18
- Luković, Ivan, MDASD'18
- Marasek, Krzysztof, MIDI'18
- Przybyłek, Adam, LASD'18

5th Workshop on Model Driven Approaches in System Development

FOR many years, various approaches in system design and implementation differentiate between the specification of the system and its implementation on a particular platform. People in software industry have been using models for a precise description of systems at the appropriate abstraction level without unnecessary details. Model-Driven (MD) approaches to the system development increase the importance and power of models by shifting the focus from programming to modeling activities. Models may be used as primary artifacts in constructing software, which means that software components are generated from models. Software development tools need to automate as many as possible tasks of model construction and transformation requiring the smallest amount of human interaction.

A goal of the proposed workshop is to bring together people working on MD languages, techniques and tools, as well as Domain Specific Languages (DSL) and applying them in the requirements engineering, information system and application development, databases, and related areas, so that they can exchange their experience, create new ideas, evaluate and improve MD approaches and spread its use. The intention is to target an interdisciplinary nature of MD approaches in software engineering, as well as research topics expressed by but not limited to acronyms such as Model Driven Software Engineering (MDSE), Model Driven Software Development (MDSD), Domain Specific Modeling (DSM), and OMG's Model Driven Architecture (MDA).

1st Workshop on MDASD was organized in the scope of ADBIS 2010 Conference, held in Novi Sad, Serbia. From 2012, MDASD becomes a regular bi-annual FedCSIS event.

Topics

- MD Approaches in System Design and Implementation – Problems and Issues
- MD Approaches in Software Process Models
- MD Approaches in Databases and Information Systems
- MD Approaches in Software Quality and Standards
- Metamodeling, Modeling and Specification Languages
- Model Transformation Languages
- Model-to-Model, Model-to-Text, and Model-to-Code Transformations in Software Process
- Transformation Techniques and Tools
- Domain Specific Languages (DSL) and Domain Specific Modeling (DSM) in System Specification and Development
- Design of Metamodeling and Modeling Languages and Tools
- MD Approaches in Requirements Engineering and Business Process Modeling
- MD Approaches in System Reengineering and Reverse Engineering
- MD Approaches in HCI development
- MD Approaches in GIS development
- MD Approaches in Document Engineering
- Model Based Software Verification
- Theoretical and Mathematical Foundations of MD Approaches
- Organizational and Human Factors, Skills, and Qualifications for MD Approaches
- Teaching MD Approaches in Academic and Industrial Environments
- MD Applications and Industry Experience

Event Chairs

- **Luković, Ivan,** University of Novi Sad, Serbia

Steering Committee

- **Gray, Jeff,** University of Alabama, United States
- **Mernik, Marjan,** University of Maribor, Slovenia
- **Ristić, Sonja,** University of Novi Sad, Faculty of Technical Sciences, Serbia
- **Tolvanen, Juha-Pekka,** MetaCase, Finland

Program Committee

- **Amaral, Vasco,** The New University of Lisbon, Portugal
- **Bryant, Barrett,** University of North Texas, United States
- **Budimac, Zoran,** Faculty of Sciences, Univ. of Novi Sad, Serbia
- **Chen, Haiming,** Chinese Academy of Sciences, China
- **Erradi, Mohammed,** ENSIAS, Mohammed-V University, Morocco
- **Fertalj, Krešimir,** University of Zagreb, Croatia
- **Gray, Jeff,** University of Alabama, United States
- **Härting, Ralf-Christian,** Hochschule Aalen, Germany
- **Ivanović, Mirjana,** University of Novi Sad, Serbia
- **Janousek, Jan,** Czech Technical University, Czech Republic
- **Karagiannis, Dimitris,** University of Vienna, Austria
- **Kardaş, Geylani,** Ege University International Computer Institute, Turkey
- **Kern, Heiko,** University of Leipzig, Germany
- **Kollár, Ján,** Technical University of Kosice, Slovakia
- **Kosar, Tomaž,** University of Maribor, Slovenia

- **Krdzavac, Nenad,** Michigan State University, United States
- **Liu, Shih-Hsi Alex,** California State University, United States
- **Maćoš, Dragan,** Beuth University of Applied Sciences, Germany
- **Melo de Sousa, Simão,** University of Beira Interior, Portugal
- **Mernik, Marjan,** University of Maribor, Slovenia
- **Milosavljević, Gordana,** University of Novi Sad, Serbia
- **Porubän, Jaroslav,** Technical University of Kosice, Slovakia
- **Rangel Henriques, Pedro,** Universidade do Minho, Portugal
- **Ristić, Sonja,** University of Novi Sad, Faculty of Technical Sciences, Serbia
- **Seidl, Martina,** Johannes Kepler University, Austria
- **Selic, Bran,** Malina Software Co., Canada
- **Sierra Rodríguez, José Luis,** Universidad Complutense de Madrid, Spain
- **Slivnik, Boštjan,** University of Ljubljana, Slovenia
- **Suvajdžin-Rakić, Zorica,** University of Novi Sad, Serbia
- **Tolvanen, Juha-Pekka,** MetaCase, Finland
- **Vangheluwe, Hans,** University of Antwerp, Belgium
- **Varanda Pereira, Maria João,** Instituto Politecnico de Braganca, Portugal
- **Wimmer, Manuel,** Vienna University of Technology, Austria

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 73–80
DOI: 10.15439/2018F370
ISSN 2300-5963 ACSIS, Vol. 16

A Model-Driven Approach to Microservice Software Architecture Establishment

Branko Terzić, Vladimir Dimitrieski, Slavica Kordić, Ivan Luković
University of Novi Sad, Faculty of Technical Sciences,
Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
Email: {branko.terzic, dimitrieski, slavica, ivan}@uns.ac.rs

***Abstract*—In this positional paper we propose a model-driven approach which addresses challenges related to modeling, development and deployment of software applications that follow the microservice architecture (MSA) design principles. We argue in favor of a model-driven tool which can be used to resolve challenges from the MSA establishment domain by providing a domain-specific language for MSA modeling and code generators for producing: (i) program and configuration code for MSA implementation; and (ii) program procedures for MSA building, packaging and installation. We give a brief description of two approaches to software application development which emerged in the last decade: the monolithic architecture approach and the MSA approach. We focus on challenges related to MSA establishment and argue that our model-driven approach could be suitable for their resolution. We also propose a plan of research activities aimed at improving our approach and which will lead to the final implementation of a model-driven tool to support such an approach**

I. Introduction

IN THE past decade two approaches to software application development became dominant among the majority of engineers: (i) software application that follows the Monolithic Software Architecture (MTA) design principles; and (ii) software application that follows the Microservice Software Architecture (MSA) design principles [1].

MTA is composed of software modules (SM) that mainly cannot exist and run independently from the core application they belong to [2]. Therefore, the whole business logic layer of the application typically runs within a single operating system process and all SMs execute within that process. Since all SMs are tightly coupled, development of individual SMs is hard to strictly divide between engineering teams. Accordingly, MTA software solutions are harder to develop, test and maintain [3]. Also, there is no possibility to choose different software technologies for individual SM development, so engineers are forced to make a final selection of technologies at the beginning of the development process. Sometimes, such decisions, which were made in the past, may turn out wrong after the years of development, leading to a great waste of time and even to the project failure. The configuration of MTA must be done at the level of the whole application rather than at the level of an individual SM. Therefore, there is a great possibility that an individual SM requires a usage of certain software libraries which are incompatible with the libraries in other SMs. Nevertheless, there are specialized modularization techniques and frameworks, for some programming languages, that can be used to overcome MTA configuration challenges. For example, Open Service Gateway Initiative (OSGi) [4] is the Java programming language framework which can be used for developing modular SMs within MTA. Horizontal scaling of an MTA must be done at the application level also, without the opportunity to scale individual SMs. Nevertheless, these types of applications are usually scaled vertically by increasing the infrastructure resources such as processing power and memory [5]. Thus, resources of an execution platform infrastructure cannot be adjusted in accordance with the requirements of individual SMs. Accordingly, engineers are forced to build "one size fits all" execution platforms, which result in irrational resource consumption and maintenance cost increase [6]. Deployment of MTA implies procedures for building, packaging and installation of the complete MTA, without a possibility to deploy individual SMs [7].

On the other hand, MSA was introduced as a suite of loosely coupled SMs, called microservices [8]. Each microservice exists and runs within a separate operating system process, independently from the other microservices. A microservice has well defined set of responsibilities and functionalities exposed through its application programming interface (API) [9]. Accordingly, engineers are able to group up into development teams in charge of developing different microservices, choosing technologies vendors and technical

The research presented in this paper was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, Grant III-44010

©2018, PTI

characteristics which are the most suitable for their needs. MSA testing comes down to testing of individual microservices, which eases locating bugs and bug fixing. The configuration of an MSA can be done separately for each microservice, eliminating possibilities for software library incompatibility. Horizontal scaling becomes a natural procedure to increase the MSA availability and it is done by running additional instances of a required microservices [10]. Since microservices are independent SMs, the adaptation of infrastructure resources to the needs of the individual microservices becomes easier and resource consumption becomes more rational [11]. MSA deployment procedure implies building, packaging and installation of the individual microservices, rather than the whole ecosystem. Thus, multiple versions of the same microservice can be run in order to compare them in production.

The MSA-specific infrastructure and the large number of microservices introduce several challenges to:

- MSA modeling, as there is a need for the MSA modeling framework which should provide a formal modeling technique and modeling tool which will ensure a higher abstraction viewpoint to engineers while specifying microservice business logic, microservice API, microservice configuration and inter-microservice communication patterns. The usage of such a modeling framework should decrease the ecosystem complexity in early phases of MSA specification, while enabling the usage of the MSA model specifications in the later development and deployment stages;
- MSA development, as there is a need for the implementation of mechanisms and infrastructure for: user request acceptance and routing, microservice auto-discovery and registry, microservice frontend and backend load-balancing, microservice fault-tolerance and health check; and
- MSA deployment, in regard to provisioning automated procedures for the MSA ecosystem building, packaging, monitoring, horizontal scaling and installation to the dedicated or cloud execution platforms.

Since the MSA approach has become dominant in the past several years [12], [13], there are plenty of development frameworks introduced by the large software companies and the open-source software community [14]. These frameworks address the majority of the aforementioned development and deployment challenges by introducing well-defined software libraries which wrap-up the core functionality of a framework. On the other hand, the usage of the aforementioned frameworks requires redundant program and configuration code to be written for different layers of the MSA ecosystem. Since the MSA ecosystem usually comprises a lot of microservices, this can lead to mistakes as engineers unintentionally introduce errors to repetitive code constructs. Also, workflow procedures used for ecosystem deployment need to be written repeatedly for individual microservices within the MSA ecosystem. Such a procedure development is often harder for average engineers and needs to be done by engineers which are specialized for the MSA deployment tasks.

Therefore, the first goal of this research is to provide a formal procedure for MSA specification in order to address the MSA modeling challenges. The second goal is to address the development and deployment challenges by using the MSA model specification for generation of program, configuration and infrastructure program code constructs. In this way, first, we want to ease the usage of MSA development frameworks and to generate all the repetitive code constructs in order to eliminate potential errors. Second, we want to generate all required procedures for the MSA ecosystem deployment in order to ease this process, make it less dependent from the specialized engineering teams and therefore less time consuming.

In order to achieve the aforementioned goals, it could be beneficial for engineers to have a domain-specific language (DSL) which will provide a formal technique for MSA modeling, as well as a set of code generators which will generate all the aforementioned artifacts based on a MSA specification written in the DSL. In order to enable usage of a DSL and code generators, we plan to develop a model-driven software tool which will support this approach.

Apart from introduction and conclusion, this paper is divided into 3 sections. In Section 2, we discuss in detail all of the challenges caused by the large number of microservices within MSA and MSA-specific infrastructure. We also propose a model-driven approach as a possible solution to these challenges. In Section 3, we present our previous research efforts to the MSA establishment, alongside the plan of research activities that should lead to its improvement and implementation of a model-driven tool which will support the realization of such an approach. In Section 4, we give an overview of related works.

II. MSA Ecosystem Establishment Challenges

In this section, we present challenges that engineers typically face during modeling, development and deployment of the MSA ecosystem. We also propose a model-driven approach as a possible solution to these challenges.

Since MSA was introduced in 2011 [15], this software development style has become the fundamental for many engineers [16] as it overcomes the most of the challenges encountered in the monolithic software application development. Large companies have adopted MSA and developed many open-source frameworks which are constantly maintained and improved by the large MSA developer community. For example, An American media streaming service Netflix has developed a set of open-source frameworks called Netflix OSS [17]. Netflix OSS was used by Netflix to divide their monolithic software system into a MSA. Today, Netflix software system consists of over 900 microservices [18]. Currently, the MSA development frameworks are introduced in almost all mainstream

programming languages [14], with a strong development and maintenance support by the community.

On the other hand, MSA have introduced the new challenges to software development process, particularly caused by the large number of microservices within the MSA ecosystem and MSA-specific infrastructure features which are required for the ecosystem establishment.

A. MSA Modeling Challenges

The first challenge is related to the MSA modeling process. Usually, at the beginning of the MSA development, it is hard for engineers to have the complete overview over the individual layers of the MSA ecosystem. The situation is even more complex if existing monolithic software is required to be migrated to MSA. Therefore, engineers are trying to decrease the MSA ecosystem complexity by specifying different types of MSA models. These models usually comprise microservice business entity models, microservice API models, inter-microservice communication pattern models and deployment strategy specification. Thus, MSA models are often specified by using an informal modelling techniques and, at the end, used just for documentation purposes. On the other hand, Model-Driven Software Engineering (MDSE) practitioners argue in favor of models as a formal way to describe the entities from the specific domain and use of such specifications as primary artifacts in the development process [19]. Domain entities, their attributes and relationships are described in a form of a meta-model which represents the abstract syntax of a DSL [20]. In order to use such a DSL for the specification of meta-model concept instances, called models, a concrete DSL syntax must be developed [20]. Therefore, in order to address MSA modeling challenges, the application of MDSE should introduce a DSL as a formal way for the MSA ecosystems modeling. The MSA DSL should provide an abstraction level which is high enough to decrease MSA modeling complexity, but which provides enough information that can be used for automation of the MSA development and deployment process. In order to use the MSA DSL in practice, a model-driven tool should be developed. Such a tool should provide a SM which will support the usage of the MSA DSL concrete syntax, used for the MSA model specification (MDM), and an appropriate file format for the MDM storage and representation.

B. MSA Development Challenges

The second challenge is related to the MSA development process. After the end of the MSA modelling process, usually begins the MSA development process which consists of: (i) development of the user-defined microservice (UMS) layer, i.e. microservices which implement the MSA ecosystem business logic; (ii) development of the infrastructure microservice (IMS) layer, i.e. microservices which ensure accessibility, availability, durability and monitoring of individual microservices within the UMS layer; and (iii) development of the inter-microservice communication patterns (MSC), i.e. selection and implementation of microservice communication patterns and the message distribution infrastructure.

The first step in the UMS layer development comprises configuration of individual microservices, including: (i) specification of the UMS API settings, such as microservice name, host name, port number and data persistence layer; and (ii) specification of a software library list, necessary for using the chosen software framework. Therefore, UMS which use the same technology stack, have common configuration properties with specific values for each UMS. As engineers try to reduce the development time, by copying repeatable configuration code to the different UMS specifications, they are unintentionally introducing errors by skipping values for common configuration parameters. For example, UMS name misconfiguration can cause microservice auto-discovery and registry inconsistent behavior within the IMS layer which is usually hard to understand and debug. Also, data persistence layer misconfiguration can cause the inconsistency and data collisions for UMS using the common database management systems. In this case, engineers usually forget to change database connection profile settings for database-specific object names. In such a situation, different UMS can try to use the same database objects, such as database tables, for storing different business model objects, or different UMS try to create their own database objects with the same name. Further, since there is a certain set of programming libraries which are required for the usage of a chosen MSA development framework, engineers easy forget some of them or misconfigure their versions. This type of the UMS misconfiguration leads to unintuitive error messages in runtime and results in a great waste of time. Accordingly, it can be beneficial for engineers if configuration and technology stack settings can be specified during the MSA modeling process, within a single MDM specification. In this way, first, engineers are able to write an in-place UMS configuration specifications without writing any boilerplate or redundant code. Second, engineers do not need to specify individual software libraries within the MDM. It is enough to specify which development framework they want to use and that is enough information which software libraries need to be included in the UMS configuration. Thus, such a MDM specification further can be used as an entry artifact for the generation of configuration code required for the individual UMS.

The second step in the UMS development is the specification of the UMS business layer (BL) which comprises the UMS business entity models and implementation of the UMS business logic. The UMS business logic functionalities are exposed to the end user or the other UMS in a form of the UMS API. The UMS API is typically developed applying the REST API design principles [21] and using the HTTP application protocol [22]. The UMS REST API method specifications have a

similar structure, depending on the REST method type [21], but with parameters specific to the individual UMS. The "copy-paste" problem is even more conspicuous in this case, because developers usually forget to change microservice-specific API settings, such as the REST method name, type or HTTP request content type for example. Therefore, it can be beneficial if engineers could use a DSL in order to specify the BL API within the same MDM, avoiding the need for the repetitive code constructs and potential mistakes. Such a specification can be then used in order to generate the BL API program code templates for the chosen technology stack. The generated code templates then can be manually filled out with program code which implements the concrete business logic for the specific UMS API.

In order to resolve the aforementioned UMS development challenges in practice, a model-driven tool should provide a separate SMs, which can be used for the UMS configuration and business logic code generation, using a MDM as its input.

The IMS layer is the heart of the MSA ecosystem as it provides the following infrastructure features:

1. user request acceptance and routing, i.e. exposing a unified access interface and a single entry point to the whole MSA ecosystem,
2. microservice auto-discovery and registering, i.e. providing a single point for microservice instances monitoring and microservice name, host and port registry,
3. frontend load-balancing, i.e. providing an improvement of workload distribution during the inter-microservice communication,
4. backend load-balancing; i.e. providing an improvement of workload distribution for incoming user requests across the MSA ecosystem,
5. microservice fault tolerance and circuit-breaking, i.e. providing a mechanism for microservice failure resistance,
6. the MSA ecosystem monitoring, i.e. providing procedures for acquisition and presentation of the MSA ecosystem metrics of interest, and
7. the MSA ecosystem scaling, i.e. increasing the availability of the ecosystem by provisioning the additional microservice instances in the UMS layer.

According to the aforementioned features, we can argue that implementation of the IMS layer is crucial to the MSA ecosystem establishment in practice. Depending on a chosen technology stack, there are different requirements which are not so trivial to fulfill and require repetitive procedures to be performed for each of the microservices from the UMS layer. Thus, configuration of the IMS layer depends on the configuration parameters of the individual microservices from the UMS layer, such as microservice names, host addresses and port numbers. Since the record about the aforementioned setting can be obtained from the MDM specification, automation of the IMS configuration and development can be achieved. Such and automation can reduce engineering efforts and radically decrease development time since the infrastructure microservice development requires knowledge of framework specifics, which depends on a chosen technology stack. For example, in order to enable microservice auto-discovery for the newly specified UMS, engineer can set one additional parameter within the existing MDM specification. This parameter can be a Boolean flag which determines if certain UMS should be added to the IMS auto-discovery settings. On the other hand, in order to achieve the same goal using the Netflix OSS framework, for example, engineer needs to write program and configuration code separately for all, the newly created UMS and the auto-discovery microservice from the IMS layer. Therefore, it could be beneficial if another code generator module could be built within a model-driven tool. This module should be dedicated to generation of configuration and program code for the IMS layer, so no significant manual and repeatable configuration or development is needed.

The development of the MSC layer implies specification and development of the communication patterns which enable inter-microservice communication and message exchange. Synchronous inter-microservice communication happens when microservice which initiates communication (client) consumes the functionality of the other microservice (server) using its API [23]. The client microservice API is blocked while it waits for the server microservice API to answer. On the other hand, asynchronous communication is done through messages sent to mediator (message provider) rather than directly to the server microservice [23]. The Client microservice API is not blocked while waiting for the server microservice API answer. In situations when a large number of microservices exist within the ecosystem, it is hard for engineers to have a clear overview on the individual microservice communication links, their type, message format and message content. This is especially pronounced in early phases of the MSA ecosystem establishment, when engineers have not developed individual microservice APIs yet, but they have to specify how microservices will communicate and which type of communication technique they will use in order to determine microservice roles and responsibilities. Therefore, the usage of a DSL can be beneficial in this situation since it can provide higher abstraction level for the specification of the inter-mircoservice communication templates, avoiding the need for the complete MSA API existence. Thus, such a specification can provide enough information that can be used to generate program code templates which implement the basic nutshell for communication infrastructure, business rules, and message format. The separate code generator module should be developed within a model-driven tool in order to generate required code templates using the MDM as its entry artifact. Later, as a MSA development moves on, these code templates should be filled-out with program code which implements required communication business roles and the message content.

C. MSA Deployment Challenges

The third challenge to MSA establishment is related to the MSA deployment process which comprise: (i) the MSA ecosystem building; (ii) the MSA ecosystem packaging; and (iii) the MSA ecosystem installation to the target execution platform.

The MSA building procedure differs depending on the chosen technology stack. Building procedure utilize a set of commands which need to be executed over the microservice program code, using program code build engine. For example, if Java and Netflix OSS are chosen, then Maven [24] or Gradle [25] build engines could be used for ecosystem building. Anyhow, the procedure is the same and repeatable for each microservice utilizing the same technology stack, no matter if it belongs to the UMS or the IMS layer.

The MSA packaging procedure depends on the chosen technology stack, as well as on the target execution platform type and configuration. For example, if Java and Netflix OSS are chosen, microservices could be packaged to JAR (Java Archive) [26] files, or could be packaged in a form of the Docker image in order to be run as isolated Docker containers [27]. However, the packaging procedure is also repeatable for microservices which share common packaging settings.

The MSA ecosystem installation to target execution platform comprise the specification of a blueprint which describes the structure of the MSA ecosystem and its desired state. For example, if an Amazon Web Service (AWS) and Docker packaging are chosen, the "Dockerrun.aws.json" file needs to be specified [28]. This file typically comprises specification of an individual microservice names, hosts, ports, storage volume settings, allocation of infrastructure resources and path to repository which keeps microservice Docker images.

Based on what was previously stated, it is obvious that an engineer needs to be familiar with many different fields of software engineering in order to complete the MSA deployment tasks. In practice, separate teams of engineers are dedicated to these tasks. However, it could be beneficial if the MSA deployment could be automated to certain extent. This can reduce the time needed for such a procedure development, and enable engineers from other teams to be less dependent on the deployment team. This is particularly important in situations when the MSA ecosystem, or some parts of it, needs to be deployed on different type of execution platforms [29]. In this case, it is crucial for the deployment procedure to be flexible and adaptive in order to provide rapid MSA migration and reduce the time needed for its customization.

In practice, MSA building, packaging and deployment procedures usually comprise well defined set of steps which mutually stem from one another. For example, if the MSA ecosystem is developed using the Java programming language and the Netflix OSS framework, then Maven can be used for a MSA building, Docker containers can be used for a MSA packaging and AWS can be used as target execution platform. In order to develop deployment procedure which supports the aforementioned technology stack and target execution platform, engineers need general microservice settings such as microservice name, host name, port number, desired number of microservice replicas, amount of memory that needs to be reserved for the microservice and so on. All these settings then need to be packed within the Dockerrun.aws.json file, so MSA is able to be installed to the AWS instance and to work correctly. Thus, engineers need to be familiar with specific format and individual settings of target execution platform blueprint. Therefore, there is an opportunity to build deployment templates, for different execution platforms, which consists of common configuration parameters with specific values for individual microservices. Further, using a DSL engineers do not need to be familiar with all the configuration parameters from specific deployment templates. Engineers just need to specify build engine, packaging strategy and target execution platform names as individual parameter within the same MDM specification. Accordingly, code generators can use the aforementioned general microservice settings from the MDM in order to fill-out the appropriate deployment script and blueprint setting parameters. Thus, changes in technology stack or target execution platform type require minor interventions in the MDM specification, re-generation and re-execution of deployment procedures in order to apply these changes in production.

In order to support the aforementioned MSA deployment requirements in practice, a separate code generator module within a model-driven tool can be developed. This module should use the MDM specification as its input and generate all required deployment procedures on the output.

III. Model-Driven Tool Prototype and Research Activity Plan

During our previous research [30] we have developed MicroBuilder, a model-driven tool for the specification of software applications that follow Representational State Transfer (REST) microservice software architecture design principles. MicroBuilder comprises two modules: (i) MicroDSL, a module that provides a DSL used for the specification of the REST microservice software architecture, and (ii) MicroGenerator, a module which comprises a set of code generators which implement series of model-to-text transformations (M2T). The M2T transformations are used to generate executable program and configuration code based on the model specification made using MicroDSL. We have supported generation of the Java program code for the implementation of the UMS layer and REST-based synchronous inter-microservice communication. We also generate the UMS configuration code with no manual configuration needed. Talking about UMS business logic code generation, we generate the Java

program code for: (i) implementation of the UMS business models; and (ii) create, update and delete (CRUD) operations for data manipulation over a business models. For generation of custom business logic, we generate API templates which should be manually filled with program code by engineers. We have also supported the generation of the Java program code which applies the Netflix OSS framework in implementation of the IMS layer. For the MSA ecosystem monitoring, we have used Netflix Turbine [31], for acquisition of the MSA ecosystem metrics, and Spring Cloud Dashboard [32] for metric visualization.

We have also presented a detailed case study where we have used the MicroBuilder tool in order to establish the web shop MSA. The structure of generated Java code is also discussed in order to explain all benefits of the MicroDSL language usage. We have compared the number of lines of code needed to specify the web shop MSA using MicroDSL to the number of manually written lines of code needed to specify the same MSA.

In order to understand MicroBuilder strengths and shortcomings, we have performed the evaluation of the MicroBuilder tool. We have applied two types of evaluation approach: (i) evaluation by example in which have used the MicroBuilder tool in order to specify various real-world examples of the microservice software architectures in order to iteratively improve the MicroDSL language and code generators; and (ii) evaluation by questionnaire in which we were using a series of questions in order to perform an objective assessment of the MicroBuilder tool. Based on the obtained results we have concluded that MicroDSL satisfies the following DSL quality characteristics: functional stability, usability, reliability, expressiveness, and productivity.

To develop the MicroBuilder tool, we have used Eclipse Modelling Framework (EMF) [33]. The MicroDSL abstract syntax concepts conform to Ecore meta-meta-model [34]. The MicroDSL textual concrete syntax was developed using the Xtext framework [35], while graphical concrete syntax was developed using the Sirius framework [36]. Individual code generators within the MicroGenerator module were developed using the Xtend framework [37].

Since challenges related to the MSA deployment and asynchronous inter-microservice communication were not considered during the aforementioned research, in the research proposed in this paper we plan to: (i) extend the MicroDSL meta-model in order to enable specification of missing MSA concepts and settings; (ii) implement the new code generators which will generate required programcode; and (iii) improve the existing code generators in order to support the additional MSA development languages and frameworkds

In order to support the asynchronous inter-microservice communication, we plan to extend the MicroDSL meta-model by adding concepts and attributes which will be used for the specification of MSA events and event messages. Events will comprise a list of event messages used for relevant data exchange between the microservices.

In order to support automated deployment procedures, we plan to add concepts describing basic building, packaging and installation strategies within the MSA core concept. We also plan to add the container concept, as additional microservice resource type, in order to support the MSA container configuration and packaging.

In Figure 1 we present the architecture of the model-driven tool prototype which comprises two main modules: the MsaDSL module and the MsaCodeGen module. The MsaDSL module will provide a DSL which will be the improved version of the MicroDSL language. The new version of a DSL should provide additional concepts for specification of: (i) asynchronous inter-microservice communication patterns; and (ii) building, packaging and installation settings.

The MicroGenerator module [30] will be transformed to the MsaCodeGen module and divided into to several submodules:

1. the MsaUMS submodule, used for the generation of program and configuration code which implements the UMS layer. MsaUMS supports the generation of the Java executable program code,
2. the MsaIMS submodule, used for the generation of program and configuration code which implements the IMS layer. MsaIMS supports the generation of the Java executable code which utilizes the Netflix OSS framework,
3. the MsaIMC submodule, used for the generation of program code which implements the synchronous and asynchronous inter-microservice communication patterns. For synchronous communication patterns, the Java program code which uses the Hypertext Transfer Protocol (HTTP) communication protocol is generated. For asynchronous communication MsaIMC will support generation of the Java program code which implements the Apache Kafka message provider [38], and
4. the MsaDPY submodule which will provide a set of code generators for generation of program code which implements the MSA deployment procedures. We plan to support generation of provisioning scripts for the IBM Cloud Container services [39] and the AWS EC2 Multicontainer Docker Environments [40]. We also plan to use the Netflix Spinnaker [41] platform to support the MSA ecosystem continuous integration and continuous

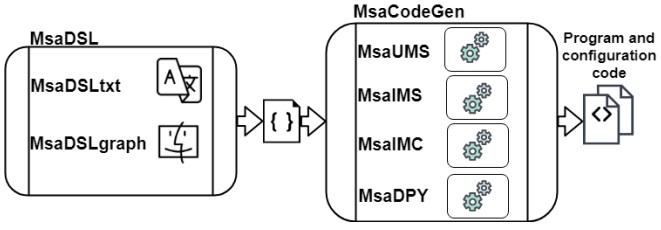

Fig. 1 A Model-Driven Tool Prototype Architecture

delivery.

IV. Related Work

While surveying the state-of-the-art literature in this area, we have found several research papers that deal with the specification of different MSA layers, using the MDSE approach. In the rest of the section we discuss the individual approaches and compare them with our approach.

In [42], the authors present an automated approach for the selection and configuration of cloud providers for multi-cloud microservices-based applications. They have developed a DSL which can be used for the specification of the application's multi-cloud requirements. Authors also provide a systematic method for obtaining proper configurations that comply with the application's requirements and the cloud providers' constraints. Comparing to our approach, authors were focused just on one aspect of the MSA ecosystem deployment, which refers to specification of installation settings for different cloud providers. On the other hand, their approach provides an opportunity for more fine-grained specifications, since they have developed a DSL that is used just for this particular use-case.

In [43], the authors try to answer the question if and to what extent MSA might build upon existing findings of Service-Oriented Architecture (SOA) research. They try to find the answer to the aforementioned question in the area of Model-driven Development (MDD), whose application to SOA has been intensively studied. The presented meta-model is divided into the three viewpoints Data, Service and Operation, each of which encapsulates concepts related to a certain aspect of MSA. The meta-model aims to support DevOps-based MSA development and automatic transformation of meta-model instances into MSA implementations. Comparing to our approach, the authors were focused on the MSA meta-model development by utilizing the deduction procedure based on the several SOA modeling approaches with the goal to identify the modeling concepts which can be used for MSA specification. Therefore, the main goal of the aforementioned research is more related on the MSA meta-model specification procedure, rather than to the MSA ecosystem establishment in practice. Nevertheless, the presented meta-model and approach seems to be still a work-in-progress towards a tool which can be used in practice.

In [44], the authors present the Aji Modeling Language (AjiL) which can be used for the MSA ecosystem specification. The AjiL abstract syntax was derived from several public MSA examples and is depicted as a Unified Modeling Language (UML) class diagram. The AjiL graphical concrete syntax was developed using the Sirius framework. Comparing to our approach the aforementioned authors have developed a DSL which can be used for the basic specification of MSA. There is still no support for specification of inter-microservice communication patterns and deployment settings. Nevertheless, we have utilized the similar set of techniques and technologies for the specification and development of a DSL concrete syntax.

V. Conclusion

In this paper we argue in favor of MDSE utilization in resolution of challenges related to the MSA ecosystem establishment in practice. We propose a DSL as a formal technique for the MSA ecosystem modeling in order to: (i) decrease the system complexity in early phases of the MSA ecosystem development; and (ii) use such a formal specification as entry artifact to process of the MSA program code generation.

Our goal is to improve our model-driven approach established during previous research efforts [30] in order to address all remaining challenges related to MSA modeling, development and deployment.

In order to achieve this goal, we plan to improve MicroBuilder, a model-driven tool which we have developed during our previous research [30]. MicroBuilder has addressed the majority of challenges related to MSA modeling and development, including the automation of the UMS and the IMS layer development and the REST-based synchronous inter-microservice communication specification. In order to address the challenges related to the MSA deployment and asynchronous inter-microservice communication, first, we plan to extend the MicroDSL meta-model with additional concepts, attributes and constraints. We also need to update textual and graphical concrete syntax specifications in order to support the new concepts. Second, we plan to improve existing code generators and build new ones in order to support the generation of missing program code constructs. In this way, we want to fulfill all the prerequisites, so the new version of the MicroBuilder tool can be used for MSA establishment in practice.

Since we have supported the generation of the Java program code which uses the Netflix OSS framework, in our future research we plan to extend technology stack by implementing addition code generators for other programming languages and frameworks. Since there is an effort [45] in development and improvement of the Netflix OSS framework for the Node.js language [46], we plan to develop code generators which will support the Node.js code generation. Also, we plan to support the usage of Zookeeper [47] as an alternative for the Netflix Eureka [48].

After the completion of the model-driven tool, which we propose in this research, we expect it to be used by software engineers in real-world projects. The tool can be used in order to develop the MSA ecosystem from scratch and deploy it to different production environments. On the other hand, the tool can be also used in situations when existing MTA should be migrated to MSA. Anyhow, the usage of the proposed model-driven tool should ease the process of the MSA establishment in production and significantly reduce development time and engineering effort.

REFERENCES

[1] A. Balalaie, H. Abbas, and J. Pooyan. "Migrating to cloud-native architectures using microservices: an experience report," In European Conference on Service-Oriented and Cloud Computing, pp. 201-215. Springer, Cham, 2015.

[2] A. Levcovitz, R. Terra, and M. Tulio Valente. "Towards a technique for extracting microservices from monolithic enterprise systems," arXiv preprint arXiv:1605.03175 (2016).

[3] J.P. Gouigoux, and D. Tamzalit. "From Monolith to Microservices: Lessons Learned on an Industrial Migration to a Web Oriented Architecture," In Software Architecture Workshops (ICSAW), 2017 IEEE International Conference on, pp. 62-65. IEEE, 2017.

[4] "Open Services Gateway initiative" [Online], Available: https://en.wikipedia.org/wiki/OSGi [Accessed: 27-Jun-2018].

[5] M. Villamizar, O. Garcés, H. Castro, M. Verano, L. Salamanca, R. Casallas, and S. Gil. "Evaluating the monolithic and the microservice architecture pattern to deploy web applications in the cloud," In Computing Colombian Conference (10CCC), 2015 10th, pp. 583-590. IEEE, 2015.

[6] M. Villamizar, O. Garces, L. Ochoa, H. Castro, L. Salamanca, M. Verano, R. Casallas. "Infrastructure cost comparison of running web applications in the cloud using AWS lambda and monolithic and microservice architectures," In Cluster, Cloud and Grid Computing (CCGrid), 2016 16th IEEE/ACM International Symposium on, pp. 179-182. IEEE, 2016.

[7] E. Daniel, D. Cárdenas, R. Amarillo, E. Castro, K. Garcés, C. Parra, and R. Casallas. "Towards the understanding and evolution of monolithic applications as microservices," In Computing Conference (CLEI), 2016 XLII Latin American, pp. 1-11. IEEE, 2016.

[8] N. Dragoni, G. Saverio, A. L. Lafuente, M. Mazzara, F. Montesi, R. Mustafin, and L. Safina. "Microservices: yesterday, today, and tomorrow," In Present and Ulterior Software Engineering, pp. 195-216. Springer, Cham, 2017.

[9] T. Johannes. "Microservices." IEEE Software 32, no. 1 (2015): 116-116.

[10] N. Dragoni, I. Lanese, S. T. Larsen, M. Mazzara, R. Mustafin, and L. Safina. "Microservices: How to make your application scale," In International Andrei Ershov Memorial Conference on Perspectives of System Informatics, pp. 95-104. Springer, Cham, 2017.

[11] D. S. Linthicum. "Practical use of microservices in moving workloads to the cloud." IEEE Cloud Computing 3, no. 5 (2016): 6-9.

[12] K. Bakshi. "Microservices-based software architecture and approaches," In Aerospace Conference, 2017 IEEE, pp. 1-8. IEEE, 2017.

[13] J. Bogner, and A. Zimmermann. "Towards integrating microservices with adaptable enterprise architecture," In Enterprise Distributed Object Computing Workshop (EDOCW), 2016 IEEE 20th International, pp. 1-6. IEEE, 2016.

[14] "Awesome Microservices" [Online], Available: https://github.com/mfornos/awesome-microservices. [Accessed: 03-Jun-2018].

[15] "Microservices Martin Fowler and James Levis" [Online], Available: https://www.martinfowler.com/articles/microservices.html [Accessed: 03-Jun-2018].

[16] G. Kecskemeti, A. C. Marosi, and A. Kertesz. "The ENTICE approach to decompose monolithic services into microservices." In High Performance Computing & Simulation (HPCS), 2016 International Conference on, pp. 591-596. IEEE, 2016.

[17] "Netflix OSS" [Online], Available: https://netflix.github.io/ [Accessed: 03-Jun-2018].

[18] "Adopting Microservices at Nerflix" [Online], Available: https://netflix.github.io/ [Accessed: 03-Jun-2018].

[19] I. Lukovic, S. Ristic, S. Aleksic, A. Popovic. "An application of the MDSE principles in IIS* Case," Model Driven Software Engineering-Transformations and Tools (2008): 85.AS

[20] J. Porubän, M. Sabo, J. Kollár, and M. Mernik. "Abstract syntax driven language development: Defining language semantics through aspects." In Proceedings of the International Workshop on Formalization of Modeling Languages, p. 2. ACM, 2010.

[21] A. Rodriguez. "Restful web services: The basics. IBM developerWorks". 2008 Nov 6:33.

[22] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. "Hypertext transfer protocol--HTTP/1.1". No. RFC 2616. 1999.

[23] "Communication between the microservices" [Online], Available: https://dzone.com/articles/communicating-between-microservices [Accessed: 03-Jun-2018].

[24] "Maven" [Online], Available: https://maven.apache.org/ [Accessed: 03-Jun-2018].

[25] "Gradle" [Online], Available: https://gradle.org/ [Accessed: 03-Jun-2018].

[26] "JAR file" [Online], Available: https://docs.oracle.com/javase/tutorial/deployment/jar/basicsindex.html [Accessed: 03-Jun-2018].

[27] "Docker" [Online], Available: https://www.docker.com/what-docker [Accessed: 03-Jun-2018].

[28] "Aws Multicontainer Docker Configuration" [Online], Available: https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/create_deploy_docker_v2config.html [Accessed: 03-Jun-2018].

[29] C. Esposito, A. Castiglione, and K.-K Raymond Choo. "Challenges in delivering software in the cloud as microservices." IEEE Cloud Computing 3, no. 5 (2016): 10-14.

[30] B. Terzić, V. Dimitrieski, S. Kordić, G. Milosavljević, and I. Luković. "Development and evaluation of MicroBuilder: A Model-Driven tool for the specification of REST Microservice Software Architectures," Enterprise Information Systems (2018): 1-24.

[31] "Netflix Turbine" [Online], Available: https://github.com/Netflix/Turbine [Accessed: 03-Jun-2018].

[32] "Spring Cloud Dashobard" [Online], Available: https://github.com/VanRoy/spring-cloud-dashboard [Accessed: 03-Jun-2018].

[33] "EMF" [Online], Available: http://www.eclipse.org/modeling/emf/ [Accessed: 03-Jun-2018].

[34] "Ecore" [Online], Available: http://download.eclipse.org/modeling/emf/emf/javadoc/2.7.0/org/eclipse/emf/ecore/package-summary.html [Accessed: 03-Jun-2018].

[35] Eysholdt, Moritz, and Heiko Behrens. "Xtext: implement your language faster than the quick and dirty way." In Proceedings of the ACM international conference companion on Object oriented programming systems languages and applications companion, pp. 307-309. ACM, 2010.

[36] "Sirius" [Online], Available: https://www.eclipse.org/sirius/ [Accessed: 03-Jun-2018].

[37] "Xtend" [Online], Available: http://www.eclipse.org/xtend/ [Accessed: 03-Jun-2018].

[38] "Apache Kafka" [Online], Available: https://kafka.apache.org/ [Accessed: 03-Jun-2018].

[39] "IBM Container Service" [Online], Available : https://www.ibm.com/cloud/container-service [Accessed: 03-Jun-2018].

[40] "Aws Multicontainer Docker Configuration" [Online], Available: https://www.ibm.com/cloud/ [Accessed: 03-Jun-2018].

[41] "Netflix Spinnaker" [Online], Available: https://www.spinnaker.io/ [Accessed: 03-Jun-2018].

[42] G. Sousa, W. Rudametkin, and L. Duchien. "Automated setup of multi-cloud environments for microservices applications," In Cloud Computing (CLOUD), 2016 IEEE 9th International Conference on, pp. 327-334. IEEE, 2016.

[43] F. Rademacher, J. Sorgalla, S. Sachweh, and A. Zündorf. "Towards a Viewpoint-specific Metamodel for Model-driven Development of Microservice Architecture," arXiv preprint arXiv:1804.09948 (2018).

[44] J. Sorgalla. "Ajil: A graphical modeling language for the development of microservice architectures," In Extended Abstracts of the Microservices 2017 Conference. 2017.

[45] "Slaying Monoliths at Netflix with Node.js" [Online], Available: https://www.linux.com/news/event/nodejs/2017/3/slaying-monoliths-netflix-nodejs/ [Accessed: 03-Jun-2018].

[46] "Node.js" [Online], Available: https://nodejs.org/en/ [Accessed: 03-Jun-2018].

[47] "Apache Zookeeper" [Online], Available: https://nodejs.org/en/ [Accessed: 03-Jun-2018].

[48] "Nerflix Eureka" [Online], Available: https://github.com/Netflix/eureka [Accessed: 03-Jun-2018].

6th Conference on Multimedia, Interaction, Design and Innovation

MIDI Conference provides an interdisciplinary forum for academics, designers and practitioners to discuss the challenges and opportunities for enriching human interaction with digital products and services.

The main focus of MIDI Conference is exploring design methods for creating novel human-system interaction, developing user interfaces and implementing innovations in user-centred development of advanced IT systems and on-line services.

Topics

Topics of interest include (but are not limited to) the following areas:

- interactive multimedia and multimodal interaction design
- novel interaction techniques, voice interfaces, interactive multimedia
- ubiquitous, multimodal, pervasive and mobile interaction, wearable computing
- novel information visualization and presentation techniques, Augmented/Virtual Reality
- design methods for usability, accessibility and outstanding user experience
- prototyping of user interfaces and interactive services
- human-centred design practices, methods and tools, user interface design
- unfolding trends in HCI research and practice, customer experience, Service Design
- advances in user-centred interaction design
- understanding people and interactions: theory, concepts, models and methods
- understanding people and interactions: contextual, ethnographical and field studies
- critique and evolution of methods, processes, theories and tools for human-computer interaction
- novel methodologies for conceptualization, design and evaluation of interactive products and services

Event Chairs

- **Marasek, Krzysztof,** Polish-Japanese Academy of Information Technology, Poland
- **Romanowski, Andrzej,** Lodz University of Technology, Poland
- **Sikorski, Marcin,** Polish-Japanese Academy of Information Technology, and Gdansk University of Technology, Poland

Program Committee

- **Biele, Cezary,** Information Processing Institute, Poland
- **Brocki, Łukasz,** Polish-Japanese Academy of Information Technology
- **Forbrig, Peter,** University of Rostock
- **Grudzień, Krzysztof**
- **Guttormsen, Sissel,** University of Bern, Institute of Medical Education, Switzerland
- **Kaptelinin, Victor,** Umea University
- **Koržinek, Danijel,** Polish-Japanese Academy of Information Technology, Poland
- **Kołakowska, Agata,** Faculty of Electronics, Telecommunications And Informatics, Gdansk University of Technology, Poland
- **Landowska, Agnieszka,** Gdansk University of Technology, Poland
- **Marti, Patrizia,** University of Siena, Italy
- **Masoodian, Masood,** Aalto University
- **Miler, Jakub,** Faculty of Electronics, Telecommunications And Informatics, Gdansk University of Technology, Poland
- **Pribeanu, Costin,** National Institute for Research and Development in Informatics - ICI Bucuresti
- **Satalecka, Ewa,** Polish-Japanese Academy of Information Technology
- **Slavik, Pavel,** Czech Technical University
- **Wichrowski, Marcin,** Polish-Japanese Academy of Information Technology, Poland
- **Wieczorkowska, Alicja,** Polish-Japanese Academy of Information Technology, Poland
- **Winkler, Marco,** University Paul Sabatier
- **Wojciechowski, Adam,** Institute of Inf. Techn., Lodz Univ. of Techn.
- **Woźniak, Paweł W.,** University of Stuttgart, Germany
- **Wołk, Krzysztof,** Polish-Japanese Academy of Information Technology, Poland
- **Ziegler, Juergen,** University of Duisburg-Essen

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 83–89
DOI: 10.15439/2018F372
ISSN 2300-5963 ACSIS, Vol. 16

Reading is Vital, but will it be Invisible? Screens vs. Paper on Our Way to the Naturalized Technology of Reading

Jarosław Kowalski
Laboratory of Interactive Technologies, National Information Processing Institute, al. Niepodległości 188b, 00-608, Warsaw, Poland
Email: jkowalski@opi.org.pl

Piotr Toczyski
Faculty of Applied Social Sciences, Maria Grzegorzewska University
Szczęśliwicka 40, 02-353 Warsaw, Poland
Email: ptoczyski@aps.edu.pl

Cezary Biele, Aldona Zdrodowska
Laboratory of Interactive Technologies, National Information Processing Institute, al. Niepodległości 188b, 00-608, Warsaw, Poland
Email: {cbiele, azdrodowska}@opi.org.pl}

***Abstract*— This analytical position paper aims to open a discussion on the future of the naturalized technology of reading. Our analysis contributes to the discussion that we think scholars in human-computer interaction should adopt from other disciplines. We begin with a seven-level technology pyramid which ends with naturalized technology. We look for the place of paper and digital books in this pyramid, currently somewhere between invisible and vital technologies. We discuss scenarios for the book, using current theories of reading from both philosophical and neuropsychological viewpoints. Finally, we show inspiring quantitative and qualitative data gathered during "total research" into the literary culture. They illustrate the ongoing change in the reading ecosystem.**

I. Introduction

THE main question we posit in the paper is whether the digital transformation can upgrade the status of books from vital to invisible in the technology pyramid. Secondly, we ask which features of the current paper and digital book technology we would want to bring into the future naturalized technology of reading. Asking these questions is more than justified in the context of the current media ecosystem change. In 1992, the New York Times Book Review published Robert Coover's article titled "The End of Books" [1]. In the early 1990s, it was difficult to assess the direction of changes that the Internet would evoke. The digitization process was just beginning. Coover's essay undoubtedly contributed to the mainstream discourse by offering both an observation and a prophecy that literature can be and will be read on electronic screens. The technology of reading became part of the interdisciplinary research agenda. The pace of change, which results from rapid technological progress, causes various reactions among the research community, but voices of those concerned about the fate of the printed word are heard particularly loudly.

Press titles that have been in existence for decades, as well as new entrants to the market, are desperately struggling for survival. As sales of newspapers and magazines plunge, publishers are forced to look for new business models. Despite these efforts, some media scholars expect a sudden death of the press. Not only textual media undergo a transformation: audiovisual ones are affected as well. Television channels are losing their audience while niche thematic channels multiply to meet the needs of fragmented audiences. TV channels become increasingly specialized, often focusing on one topic (i.e., cooking, interior design, healthcare).

A parallel trend accompanies this fragmentation: Internet increasingly incorporates other media. One can watch television, listen to the radio, and make video calls via a web connection. The Internet has become an inseparable part of everyday life and, according to Lee Rainie and Barry Wellmann, social media have become an operating system of the contemporary society [2]. The popularization of smartphones and tablets makes this system constantly present and constantly indispensable, and a large number of young adults can no longer imagine their lives without the Web and phones. Studies show that college students generally perceive their cell phones and digital resources as an integral part of who they are, a significant "extension" of themselves [3].

On the other hand, some researchers say that nothing has contributed to the massification of reading activities like the Internet. While this medium is increasingly audiovisual, for the past three decades it has been largely read (since Web 1.0) and even written (since Web 2.0). In as early as 2008, researchers at the University of San Diego, California, estimated that the average American "consumed" over one hundred thousand words per day [4]. They also estimated that compared to 1960, the percentage of words consumed from paper fell from 26% to 9%, but it was outweighed by reading from the computer screen: at that time, 27% of the daily balance of words came from a computer, according to that research. Therefore, the Web and its networks supported reading. Never in history had people read so much. Thus reading is undoubtedly vital, although the digital revolution we are witnessing is changing the media landscape. The

This work was not supported by any organization

©2018, PTI

digital media ecosystem is much more demanding for publishers. It raises the following question: can digital transformation upgrade the status of books from vital to invisible in the technology pyramid?

II. OUR QUESTIONS: ARE BOOKS ONLY A VITAL OR AN ALMOST INVISIBLE TECHNOLOGY? HOW CAN BOOKS CONTRIBUTE TO THE EMERGING NATURALIZED TECHNOLOGY OF READING?

Before we focus on the voices and scenarios for the future of reading, let us ask a question about technology levels. Where exactly do books lie in the technology pyramid?

According to Koert van Mensvoort, the author of the pyramid of technology concept [5], every form of technology falls into one of seven hierarchically arranged categories. The higher the level, the more integrated with the human species the technology is. Some technologies evolve, "climbing" to the next levels. Others stop evolving at some level or even slide down to previous phases in their evolution. First, let us look at these seven levels.

Level 1: Envisioned. The first level is about imagination and fantasy. Technology is not here; it functions rather as an emanation of human needs and imagined ideas. Before becoming materialized, inventions, discoveries, and innovations appeared in stories and dreams (e.g., many elements of the twentieth-century technique were described in his novels by Jules Verne; other examples include Leonardo da Vinci and many others).

Level 2: Operational. The next stage is operationalization, i.e. developing prototypes in laboratory conditions, testing, improving and discussing their form.

Level 3: Applied. The third level is the implementation stage: when innovation functions in the natural environment, it is used, initially experimentally, among the innovators.

Level 4: Accepted. This is followed by social acceptance; the invention spreads. To achieve this, time and infrastructure are needed. Technology must adapt to the needs of users and be "tame".

Level 5: Vital. The further life of technology is a gradual transition to, and integration with, social life. At this (fifth) stage, the elimination of the technology would have a dramatic and negative impact on the lifestyle of people and societies because it has become an indispensable element of their existence. One example is the city: it is difficult to imagine modern civilization or banking without cities. It seems that innovations such as telephones, the Internet or smartphones are currently at this level.

Level 6: Invisible. The next, sixth level is the level of "invisibility", where technology becomes indistinguishable from nature, as, for example, in the case of clothing, agriculture or writing.

Level 7: Naturalized. Naturalization is the peak of development, with technologies blending with the nature of *homo sapiens*, and becoming part of it. Some examples here include thermal processing of food or spoken language.

Now, where are printed books and the press in this hierarchy at present? It seems that printed books, newspapers and magazines are at the fifth stage, i.e. vital. The question is: what next? Will they become invisible? As a consequence of digital transformation and digital access to books, we are witnessing a process of ongoing rejection of paper, but not a rejection of reading as such. What does this mean from the technology pyramid perspective? From this viewpoint, paper is only a specific "larval" form of textual distribution for ideas, concepts, data, information or even wisdom. With the rejection of paper, our text processing will be able to evolve and move to the sixth level, which could take such forms as content displayed directly on the retina.

III. THEORIES AND IDEAS: CURRENT READING PHENOMENA IN THREE OVERLAPPING SCENARIOS

Although our search for an answer begins with literature review, we can begin to answer our central question by asking what happened to horses when cars were introduced. They certainly were vital, but—disrupted by automobiles and cars—horses had never been an invisible technology in transport. It seems that computers, smartphones, tablets and all kinds of other digital screens have captured the time spent reading books, we could assume that books are descending in van Mensvoort's pyramidal hierarchy. Does the digital transformation push books to "descend" from the level of "vital" to the level of "accepted" technology in van Mansvoort's technology pyramid? Is it at all possible to descend from level 5 to level 4? We can imagine this process in many scenarios.

There are at least three scenarios for paper books (and, to a degree, also for electronic books). They can be summarized as overlapping scenarios for the future of book reading.

Scenario 1: No paper books, no human need. Paper books are sentenced to gradual extinction in the digital West because Westerners no longer feel that they need to read them.

Scenario 2: Even an invisible technology does not fulfill some human needs which were easily fulfilled in the age of paper. Maybe we will see the developments described by Umberto Eco: "When reading a book, we need to remember what was said on page 20 when we reach page 200, we need to activate our memory and use our ability to navigate through the space that we have shaped in our imagination while reading. (...) New technologies inspire new forms of creativity (...). And yet, these new forms do not replace the old ones, which we still need dearly." [6].

Scenario 3. The new digital genres and communication forms influence traditional genres and make them superficial. It seems that this third way is also possible. Many scholars point to the renaissance of writing, citing as evidence the gigantic distribution and social circulation of such novel

series as "Harry Potter", "Game of Thrones", the "Twilight" saga or "The Hunger Games." Is their form the same as that of a twentieth century novel? As Jacek Dukaj observes [7], a new form of storytelling is being born before our eyes, drawing on the digital model. He calls this form a "google novel." As Dukaj writes: "In this way, the novel (and a TV series) artificially satisfies the need, imprinted in the "googled" brain, of continuously hopping to new streams of data: "I can not open another browser window on the pages of the book or enter another chat channel, but before this unconditional reflex tears me away from reading, I'm already in a different story, in a different set design, with other heroes. And again. And again. And again…" [7].

The thick volumes that are making it to bestseller lists are written according to new conventions. Their fictional structure and storytelling manner are very different from the way in which Tolstoy's "War and Peace" or Conrad's "Nostromo" were narrated. They are characterized by three phenomena. The first one are fragmented narratives. The Google-type literature—as Dukaj observes—uses small portions of narratives. The second characteristic is a multiplicity of threads, making it difficult to identify the main character of the story, as in the "Game of Thrones". The third phenomenon within Google-like literature is that talking replaces describing. The dominance of dialogues over descriptions is effort-saving since everyone in the pop-culture reality knows what an archetypical jungle, a big city or a police station look like, and there is no need to describe them.

Three abovementioned scenarios and some related phenomena can be identified in the interdisciplinary literature review we conducted. According to technological determinism, technology has a formatting effect on humans. New information processing technologies have such a multi-dimensional and profound impact on users that researchers view this process as progressive formatting of humans through technology. Neil Postman was a prophet of such technological formatting [8]. Technology becomes the operating system of everyday life but it can also cause dysfunctions. Trend researchers and sociologists, among them John and Nana Naisbitt along with Douglas Phillips [9], formulate a diagnosis that people in America are still in a state of technological intoxication. They describe six signs of such a state of mind: 1) Favoring ad hoc solutions, from religion to food; 2) Worship of technology and fear of it; 3) Bridging the differences between what is real and what is pretended; 4) Taking violence as a normal thing; 5) Loving technology, as one loves a toy; 6) Experiencing one's life, being distanced and distracted.

The "man of numbers" replaces the "typographic man". After 25 years, the pessimistic voices expecting an apocalypse of books entailed by the popularization of electronic screens, sound ever louder. Gołębiewski [10] proposes an alarming vision, in which the impact of digital screens will cause the traditional reading of literature (i.e. from paper) to degenerate. Reading from old-fashioned paper "interfaces" [11], as envisioned by Gołębiewski, will be pushed into the ghetto of anachronistic connoisseurs, which will exist alongside the mainstream culture in the same way as the current ghettos of cassette or vinyl record lovers. He points out, after Marshall McLuhan [12], that Gutenberg's invention of print created a "typographic man". The world of the "typographic man" facilitated the dissemination of education, invented and disseminated the press and gave broad access to literature through libraries. It also created the novel which throughout the nineteenth century and until the invention of the motion pictures was the default form of storytelling. According to Gołębiewski, the world of the "typographic man" is currently undergoing a process of degeneration.

Researchers worry about the future of in-depth reading. There is a growing concern that immersion in digitally mediated modes of knowledge acquisition might hinder or even eliminate mental activities that emerged from interactions with traditional media, and which are still tremendously valuable. The leading example here is the concern about the demise of "deep reading", a practice traditionally associated with interacting with long, linear texts. It is understood as the practice of, and ability to concentrate on, reading long texts for extended periods of time, with the aim of enhanced comprehension, construction of meanings, and facilitation of "deep learning" [13]. Research results suggest that the cognitive performance of even skilled comprehenders might be hindered if critical sections of digital media learning material are only scanned [14]. The skill of deep reading, trained in interaction with traditional books and longer printed texts, is argued to be essential for developing the faculties of critical thinking, reasoning, and insight. The logic of hypertext and multitasking, vital for proficient navigation in the digital world, is supposed to shape brains and habits in a way that makes deep reading practices less viable [15, 16]. Maryanne Wolf [17] notes that reading is something that the human brain has not evolved into. Among various forms of communing with culture, reading is the one to which nature has not prepared humans through the process of adaptation.

From the brain neuroplasticity perspective, reading is changeable, not given. Since we do not have a genetic instruction on how to read (while we have instructions on how to look, listen or smell), reading skills require training that is not easy and takes time. It is worth considering how digital developments affect activities such as reading. The Internet, a new medium originally conceived as a text medium, is becoming increasingly pictorial and visual. Also, the popularity of new smartphone applications suggests that young people increasingly need to express themselves through images (using Snapchat, Instagram, or Pinterest). In other words, extensive training in reading books and longer texts (which children are subjected to at school) is supplemented (perhaps even gradually dominated) by

training in the use of random, short texts, accompanied by graphics.

When humans are learning to read, their brain must adapt to the process physically and internal resources must be managed. This "development" has the most physical character: new connections are being created in the brain. The existing ones (i.e. shape recognition, sound recognition) are harnessed to new functions. Resources and created connection systems must then be fixed and trained, which requires long hours of interaction with texts, and arduous learning. Finally, the brain becomes adapted to reading and becomes compatible with the type of information that it will process. Consequently, the brains of people using different types of writing are physically different from one another. When reading the Latin alphabet, Westerners activate the left hemisphere while reading. In contrast, Chinese (logographic) script requires the use of both hemispheres. Interesting in this light is that Japanese readers simultaneously use their own form of logographic writing and a type of syllabic writing, consisting of 46 characters (the so-called kana script), and used to write their own names.

Neuroimaging studies have demonstrated [17] how the brain operation changes depending on whether it decodes logographic or syllabic writing. Maryanne Wolff's analysis shows that the brains of people using different types of writing function differently and they also process information differently.

According to Nicholas Carr [18], about four hours of Surfing the internet is enough for the brain to switch to a slightly different mode of operation. Reading a text such as a book is (and must be) analog in nature – this is a continuous process where we accompany the author throughout the whole argument: from the beginning, through subsequent stages up to the final. This is how an essay, a novel or a scientific lecture is structured. Using an Internet browser is closer to scanning, where the gaze scans the screen in search of interesting content, and anchors on those points that are most likely to solve the problem or offer something interesting. A new website is scanned, an item is found and clicked on. Then a new page is opened and the process begins again.

There is also a new phenomenon at the doorstep: we are moving from codices to scrolls. Dissemination of smartphones and tablets has slightly redefined the use of the Web, introducing the use of applications (known as apps). Technological limitations (small screens and the inability to use the cursor, inherently targeting small objects on the screen more accurately than a finger) have put virtual content navigation into the hands of virtual image scrolling. The so-called "scrolling" is reminiscent of scrolling a large (sometimes almost endless) physical scroll. This is an interesting sign of our times where, after centuries of dominance of books in the form of "codices" (with sheets stapled or glued on one side), we are witnessing a gradual return to the scroll format (this time a virtual one). In antiquity, the codex format had an undoubted advantage, giving users the possibility of free and quick access to any place of the text, whereas a scroll required arduous physical scrolling (also performed through technically when using a music cassette or a VHS tape). Today, Facebook or Twitter users do not seem to need such a feature: a continuous stream of notifications from friends (and strangers who are tagged as "friends") is just a stream, a flow, something that is happening here and now.

In his works, Łukasz Gołębiewski claims that the digital revolution that changes our habits of media consumption will

TABLE I.
TEXTUAL "DIET" OF POLES [21]

	yesterday	**a year ago**
Texts on paper		
Articles in printed daily newspapers or general-interest magazines	37%	35%
Books – novels, stories, biographies, non-fiction	18%	26%
Short texts such as advertisements, informational messages, inscriptions on walls, leaflets, etc.	18%	16%
Manuals, training materials	9%	11%
Articles in industry and specialist magazines	7%	9%
Tutorials and guidebooks	6%	8%
Poetry	5%	6%
Comic books, albums	1%	2%
Text on screen		
Short text messages: texts, tweets, single-sentence posts	38%	32%
Short entries and messages on the Internet: on blogs, forums, social media, e-mails	31%	26%
Longer texts on the Internet such as articles on blogs, websites, etc.	18%	17%

change our civilization. "There has been a terrible thing that happened to readership. It has been reduced to mindless activities. We live in a time of breakthrough, the generation of those raised on books is still statistically large enough to give the illusion that the world is still normal. But this is an illusion. A new generation, called Digital Natives, has grown up in front of a computer monitor. They have other needs and are effectively pushing them." [19]

So can reading disappear at all? Can we imagine a world without books? No libraries or press? At this point, it is worth considering how much these information processing and gathering technologies are fused with our civilization. Our civilization does not want us to withdraw from linear reading. One of the eminent fruits of digital transformation is Amazon Kindle.

As we mentioned in this literature review, according to some voices the digital revolution is expected to eliminate the "typographic men" and replace them with the "men of numbers" because exposure to information is now mediated through electronic screens and digital media. The future world is painted by extrapolation as a place with falling circulation of magazines and newspapers, and declining book spending. For the "man of numbers", cursory cognition is the default mode of operation. These humans do not go deeply into the content, and live in constant inattention in multiple channels, among applications, screens, programs and chats. The "men of numbers" also consume culture differently: they listen to mp3s rather than CDs or vinyl records, Spotify playlists rather than music albums, they search for information in Google or Wikipedia rather than in an encyclopedia, they draw on a tablet app rather than using a pencil. Such descriptions evoke the image of humans affected by all five symptoms listed by John and Nana Naisbit.

Taking into account the context of readership, it is worth considering whether the metaphor of the "man of numbers" against MacLuhan's "typographic man" is correct. When reading a book purchased from Amazon on an electronic reader, the reader does not consume digits but letters. It is the same when surfing the Internet and reading characters displayed on the LCD screen. The fact that these characters are digitally coded is irrelevant to human perceptual apparatus. When they read, the hypothetical "men of numbers" read letters. Doubts also arise when we try to equalize all types of electronic screens, as Gołębiewski would like.

Meanwhile, reading from an electronic Kindle reader (or many of its varieties produced and sold under other brands) has completely different qualities than reading from a screen of a smartphone or a tablet. The functionalities of an ebook reader are undoubtedly narrower than those of a smartphone. The latter has many functions that can disrupt concentration. For example, they may generate the need to check e-mails or respond to Facebook conversation threads. The paper versus electronic screen dichotomy is probably too simple and, in effect, it obscures the real image of the phenomenon of digital reading. From this perspective, one can try to categorize "reading interfaces" regarding their potential ability to narrow attention. The screen of an electronic reader (e-reader) would be much closer to paper than to a computer screen. We liken it with cuneiform script rather than a scroll or a codex.On the other hand, it is beyond doubt that people consume ever more information from computer screens (which includes learning from digital media), so it is reasonable to consider whether a change in the medium causes triggers a change in reading habits, i.e. in what, how and why we read and how we obtain information. In this context, the ability to use screens for reading is a prerequisite for proficient usage of digital resources and for taking part in electronically mediated activities, with learning being the most important one. Moreover, Kindle is not the same as Google. It is the post-digital hybrid of cuneiform script, a scroll and a codex. Kindle is technology that enables deep reading. Symbolically speaking, Google and Kindle are two

TABLE II.
EVALUATION OF BOOK AND NON-BOOK TEXTS ON FIVE DIMENSIONS OF THE RESPONDENTS' FEELINGS. DATA PRESENTATION BY MICHAŁ FELIKSIAK AND BARBARA BADORA, PUBLIC OPINION RESEARCH CENTER [21]

Type of text read yesterday	**For me, this text was**				
Average results on a scale from 1 to 7, where 1 means "does not apply at all" and 7 means "applies completely"	Significant, important	Touching, moving	Thought provoking	Initiating a conversation	Useful, helpful
Books	4.90	3.75	4.98	4.90	5.12
Articles in trade press and specialist magazines	5.14	2.30	4.43	4.55	5.43
Short text messages: texts, tweets, single-sentence posts	3.88	2.14	2.71	2.94	4.06
Short entries and messages on the Internet: on blogs, forums, social media, e-mails	3.82	2.21	3.39	3.48	4.21
Longer texts on the Internet such as articles on blogs, websites, etc.	4.33	2.59	4.21	4.23	4.58
Short texts such as advertisements, informational messages, inscriptions on walls, leaflets, promotional leaflets, etc.	2.43	1.58	2.30	2.48	3.15
Articles in printed daily newspapers or general-interest magazines	4.28	3.38	4.27	4.37	4.56
Comic books, albums	3.06	3.13	3.19	3.50	3.90

contradictory faces of the same process.

IV. EMPIRICAL OBSERVATIONS: WHAT DOES AN ATTEMPT OF TOTAL RESEARCH INTO LITERARY CULTURE TELL US ABOUT PAPER AND DIGITAL READING?

What is similar and what is different in digital reading and paper reading? Is digital reading almost the same as analog reading on paper? Do we think of the same kind of reading? We draw our answers from a research project where a national representative random sample of Poles was asked about their reading habits in the digital era a few years ago.

The project entitled "Directions and forms of reading transformation in Poland" consisted of data from personal interviews (CAPI and CAWI), short interviews with visitors to a book fair, qualitative data from household ethnography and expert discussion, which were then all merged with readership research and book market statistics. Some of us co-authored the research led by Warsaw-based state-funded National Information Processing Institute, with Krzysztof Krejtz as its scientific coordinator. We refer to this research not only because we co-authored it and can be proud of its "enormous cognitive potential" [20], noted as a step towards "total research into the literary culture" [20]. We believe that its results illustrate and deepen our insights into reading technology [21, 22].

Results 1: Retrospective panel data on how respondents read yesterday and a year ago. Part of the survey was conducted using the so-called retrospective panel method. It is not a comparison of two measurements taken in separate moments in time but, rather, a kind of synthetic self-reflection obtained from the respondents.

The respondents were asked what and where they had read on the day before the survey and one year before the survey. Among answers, the respondents could select not only books but also different types of text. This is visible in the answers shown in Table 1. Respondents admit that during the preceding year they read fewer books than shorter messages and entries on the Internet and mobile devices such as smartphones or tablets.

Results 2: Declarative data on how respondents felt about different kinds of texts. While looking for ways of getting an insight into almost intimate issues of readers' relations to a book (as well as other textual) "interface", in one set of survey questions we asked how the respondents felt after reading books and non-book texts on the previous day. The respondents were asked to evaluate previous-day texts using five seven-point scales, where 1 meant that the statement did not apply at all to the text they read and 7 meant that the statement applied completely to the readers' experience.

We checked the five dimensions of the assessment:

Dimension 1: Importance. First of all, we asked if the text was "significant, important" for the respondents.

Dimension 2: Emotional impact. Secondly, we were interested in how much the text was "moving, touching" for them.

Dimension 3: Thoughtfulness. The third dimension was the extent to which the text prompted the respondents to think.

Dimension 4: Sociability. We were also interested in the social dimension, i.e. to what extent the text encouraged the respondents to talk about it.

Dimension 5: Usefulness. The fifth dimension was the extent to which the text turned out to be "useful" for the respondents.

After conducting statistical tests, we learnt that books are rated significantly higher than all other texts read on the day preceding the survey, on each dimension tested. Articles in trade press and specialist magazines performed similarly: in terms of four dimensions, they were ranked close to books, but they failed to move readers emotionally. Detailed results are presented in Table 2.

While juxtaposing these results with his idea of literacy culture total research, Maciej Maryl [20] noted: "The concept of "textual diet", i.e. the diversity of the types of texts that the recipients are acquainted with, proposed by Toczyski and Krejtz, may prove useful. However, it would be necessary to refine this catalog in such a way as to avoid mixing up formal determinants (e.g. length of the text, paper vs. electronic format) with genres (e.g. novels, poetry, guides).

We find this criticism justified. However, we still recommend separating digital and analog forms, so that we have a better idea about the place of Web versions, and the place of print versions in readers' textual diet. Fortunately for the present paper, even despite the mix of genres, lengths, and formats, we can see the dominance of books on almost every dimension, which proves their position.

Results 3: Ethnographic interviewing and observing readers as users. The analysis of ethnographic interviews conducted under the same project shows that reading from an LCD screen and paper are two different worlds. Reading from a screen is primarily done for work (e-mails, electronic documents, work with software) and entertainment (short "fillers of time" when there is a break at work). These fillers can be social media posts, funny drawings (memes), information on portals and other things [22]. The reader does not delve into the topic, which is a result of performing many activities at once. Everything works in parallel mode on the Internet. Running a conversations in a chat box does not prevent users from holding a second conversation in parallel, making online purchases, listening to music or playing a game.

Reading from paper, be it books or newspapers, belongs to the realm of luxury and pleasure. So, what are the barriers that prevent paper-based reading from becoming part of everyday life? We found three types of barriers.

Barrier 1: Time. The main limitation is connected with time. In order to delve into a narrative, we need to find more free time. Paper texts have ceased to function as "gap fillers".

Barrier 2: Distractors. Other barriers mentioned in the study included the barrier of concentration: even if you do have some free time, if you want to immerse in reading, you should devote yourself to it and exclude yourself from other activities. This is difficult due to the excessive burden of everyday duties. Also, many media and leisure activities, often more attractive (TV, computer, smartphone), are fighting for people's attention. Worth noting is that a smartphone with apps generating a continuous stream of interrupts ("you have a message", "you have an email", "your post has been commented on") becomes a distracting trap and a generator of constant temptations to check what is going on without our participation.

Barrier 3: Mode-switching. The third type of barrier that emerged in ethnographic interviews was the "transition barrier". By this we mean the process of switching attention to a focused mode. For many people, this is so unobvious and difficult that they need various kinds of rituals, such as a favorite chair, tea, absolute silence (hence, the natural time for book reading comes in the evening, at bedtime).

What emerges from this research is an image where an easy, pleasant and entertaining text (usually read from the screen) displaces more ambitious content read on paper. The victims of this process are not only the shrinking book expenditures but also printed newspapers and magazines (however, they seem more accessible due to their volume: in a busy schedule, it is easier to accommodate a single article than a section of a book).

V. OPENING REMARKS: INSTEAD OF A CONCLUSION

The whole technology-related and work-related social system makes people read, but in a distracted way, probably not fulfilling many of their needs which could be met by the written word. Within the scope of the three elements discussed here, i.e. technology, biology and culture, the former one seems to dominate at present, suppressing the other two. All three create a triad where each element dynamically shapes the other two. The example of reading is illustrative in showing that currently, i.e. since the 1990s, peaking in the 2010s, technology is in the lead while the other two elements are only forced to adapt.

We began this analysis with technology levels shown as a seven-level pyramid. We asked about the position of books in this pyramid. We discussed some theories to find out whether books could become an invisible technology. We showed some encouraging data illustrating our approach (both quantitative and qualitative). Finally, with the above work, we intend to open a discussion, not even trying to reach final conclusions on any part. The ongoing digital transformation is leading us towards many fundamental questions about civilization founded on letters, which require further interdisciplinary studies involving the high tech industry, human-computer interaction experts, and social theory readers. One such question is what the invisible technology of reading will be like. Which qualities of traditional books will this new technology use? Which aspects of books do we want to transfer to the invisible (and ultimately the naturalized) technology of reading?

REFERENCES

[1] R. Coover, *The End of Books*, New York Times Book Review, 1992.

[2] H. Rainie and B. Wellman, *Networked: The new social operating system*. Cambridge, Mass: MIT Press, 2012.

[3] R. Belk. "Extended Self in a Digital World" *Journal of Consumer Research*, 2013, vol. 40(3), pp. 477–500.

[4] How Much Information? Report on American Consumers, San Diego, 2009.

[5] K. van Mensvoort, *Pyramid of Technology: How Technology Becomes Nature in Seven Steps*. Eindhoven: Technische Universiteit Eindhoven, 2013.

[6] U. Eco. Nowe środki przekazu a przyszłość książki. Warsaw: Państwowy Instytut Wydawniczy, 1996 (translation of the quote by the authors)

[7] J. Dukaj, "Za długie, nie przeczytam…" *Tygodnik Powszechny* (Oct. 17, 2010).

[8] N. Postman, *Technopoly: Surrender of Culture to Technology*, New York: Vintage Books, 1993.

[9] J. Naisbitt, N. Naisbitt, D. Philips. *High tech, high touch: Technology and our search of Meaning,* New York: Broadway, 1999.

[10] L. Gołębiewski. *Śmierć książki. No Future Book*, Warsaw: Biblioteka analiz, 2008

[11] L. Manovich. *The language of new media*. Cambridge, Mass: MIT Press, 2001.

[12] M McLuhan, *The Gutenberg Galaxy: The making of typographic man, Toronto*, University of Toronto Press, 1962.

[13] J. C. Roberts and K. A. Roberts. "Deep Reading, Cost/Benefit, and the Construction of Meaning: Enhancing Reading Comprehension and Deep Learning in Sociology Courses" *Teaching Sociology,* vol. 36(2), 2008, pp. 125–140.

[14] L. Salmeron, J. Naumann, V. Garcia and I. Fajardo. "Scanning and deep processing of information in hypertext: a cued retrospective think aloud study" *Journal of Computer Assisted Learning*, vol. 33, 2017, pp. 222–233.

[15] M. Wolf, Our 'Deep Reading' Brain: Its Digital Evolution Poses Questions. *Nieman Reports,* vol. 64(2), 2010, pp. 7–8.

[16] M. Wolf and M. Barzillai. "The Importance of Deep Reading" *Educational Leadership,* vol. 66(6), 2009, pp. 3–37.

[17] M. Wolf, *Proust and the Squid: The Story and Science of the Reading Brain,* New York: Harper Perrenial, 2007.

[18] N. Carr, *The Shallows: What the Internet Is Doing to Our Brains*, New York: W. W. Norton & Company, 2011.

[19] L. Gołębiewski. *Gdzie jest czytelnik?* Warsaw: Biblioteka Analiz, 2012.

[20] M. Maryl. "Kulturowa piśmienność: projekt badań totalnych kultury literackiej" *Teksty Drugie,* 2015, vol. 3, pp. 9–29.

[21] P. Toczyski and K. Krejtz. "Dieta tekstowa Polaków. Rola i znaczenie książki na tle innych tekstów" in *Raport Polskiej Izby Książki z projektu badawczego Kierunki i formy transformacji czytelnictwa w Polsce 2013–2014*, P. Toczyski, J. Kowalski, A. Niedzielska, K. Krejtz, P. Kobyliński. Ed. Warsaw: Polska Izba Książki, 2014. pp. 30–41.

[22] J. Kowalski. "Od papieru do LCD. Wnioski z wywiadów etnograficznych" in *Raport Polskiej Izby Książki z projektu badawczego Kierunki i formy transformacji czytelnictwa w Polsce 2013-2014*, P. Toczyski, J. Kowalski, A. Niedzielska, K. Krejtz, P. Kobyliński. Ed. Warsaw: Polska Izba Książki, 2014. pp. 42–49.

Joint 38th IEEE Software Engineering Workshop (SEW-38) and 5th International Workshop on Cyber-Physical Systems (IWCPS-5)

THE IEEE Software Engineering Workshop (SEW) is the oldest Software Engineering event in the world, dating back to 1969. The workshop was originally run as the NASA Software Engineering Workshop and focused on software engineering issues relevant to NASA and the space industry. After the 25th edition, it became the NASA/IEEE Software Engineering Workshop and expanded its remit to address many more areas of software engineering with emphasis on practical issues, industrial experience and case studies in addition to traditional technical papers. Since its 31st edition, it has been sponsored by IEEE and has continued to broaden its areas of interest.

One such extremely hot new area are Cyber-physical Systems (CPS), which encompass the investigation of approaches related to the development and use of modern software systems interfacing with real world and controlling their surroundings. CPS are physical and engineering systems closely integrated with their typically networked environment. Modern airplanes, automobiles, or medical devices are practically networks of computers. Sensors, robots, and intelligent devices are abundant. Human life depends on them. CPS systems transform how people interact with the physical world just like the Internet transformed how people interact with one another.

The joint workshop aims to bring together all those researchers with an interest in software engineering, both with CPS and broader focus. Traditionally, these workshops attract industrial and government practitioners and academics pursuing the advancement of software engineering principles, techniques and practices. This joint edition will also provide a forum for reporting on past experiences, for describing new and emerging results and approaches, and for exchanging ideas on best practice and future directions.

TOPICS

The workshop aims to bring together all those with an interest in software engineering. Traditionally, the workshop attracts industrial and government practitioners and academics pursuing the advancement of software engineering principles, techniques and practice. The workshop provides a forum for reporting on past experiences, for describing new and emerging results and approaches, and for exchanging ideas on best practice and future directions.

Topics of interest include, but are not limited to:

- Experiments and experience reports
- Software quality assurance and metrics
- Formal methods and formal approaches to software development
- Software engineering processes and process improvement
- Agile and lean methods
- Requirements engineering
- Software architectures
- Design methodologies
- Validation and verification
- Software maintenance, reuse, and legacy systems
- Agent-based software systems
- Self-managing systems
- New approaches to software engineering (e.g., search based software engineering)
- Software engineering issues in cyber-physical systems
- Real-time software engineering
- Safety assurance & certification
- Software security
- Embedded control systems and networks
- Software aspects of the Internet of Things
- Software engineering education, laboratories and pedagogy
- Software engineering for social media

EVENT CHAIRS

- **Bowen, Jonathan,** Museophile Ltd., United Kingdom
- **Hinchey, Mike**(Lead Chair), Lero-the Irish Software Engineering Research Centre, Ireland
- **Szmuc, Tomasz,** AGH University of Science and Technology, Poland
- **Zalewski, Janusz,** Florida Gulf Coast University, United States

PROGRAM COMMITTEE

- **Ait Ameur, Yamine,** IRIT/INPT-ENSEEIHT, France
- **Banach, Richard,** University of Manchester, United Kingdom
- **Bensalem, Saddek,** VERIMAG, France
- **Broy, Manfred,** Technische Universitaet Muenchen, Germany
- **Čaplinskas, Albertas,** Vilnius University, Lithuania
- **Carter, John,** University of Guelph, Canada
- **Cicirelli, Franco,** Universita della Calabria, Italy
- **Denney, Ewen,** SGT/NASA Ames, United States

- **Derrick, John,** Unversity of Sheffield
- **Ehrenberger, Wolfgang,** Hochschule Fulda, Germany
- **Eleftherakis, George,** The University of Sheffield International Faculty, CITY College, Greece
- **Fantechi, Alessandro,** DSI - Universita' di Firenze, Italy
- **Fidge, Colin,** Queensland University of Technology, Australia
- **Forbrig, Peter,** University of Rostock
- **Fortiers, Stephen,** George Washington University
- **Friesel, Anna,** Technical University of Dennmark, Denmark
- **Fujita, Masahiro,** University of Tokyo, Japan
- **Golatowski, Frank,** University of Rostock, Germany
- **Gomes, Luis,** Universidade Nova de Lisboa, Portugal
- **Gracanin, Denis,** Virginia Tech, United States
- **Grega, Wojciech,** AGH University of Science and Technology, Poland
- **Gumzej, Roman,** Faculty of Logistics, University of Maribor, Slovenia
- **Havelund, Klaus,** Jet Propulsion Laboratory, California Institute of Technology, United States
- **Hsiao, Michael,** Virginia Tech, United States
- **Kornecki, Andrew J.,** Embry Riddle Aeronautical University, United States
- **Laplante, Phillip A.,** PennState University, United States
- **Letia, Tiberiu,** Technical University of Cluj-Napoca, Romania
- **Li, Jianwen,** Iowa State University, United States
- **Liu, Zhiming,** Southwest University, China
- **Lopezo, Oscar Pastor,** Valencia
- **Malloy, Brian,** Clemson University, United States
- **Marwedel, Peter,** Technische Universität Dortmund, Germany
- **Minchev, Zlatogor,** Bulgarian Academy of Sciences, Bulgaria
- **Monostori, László,** Hungarian Academy of Sciences, Hungary
- **Nesi, Paolo,** DSI-DISIT, University of Florence, Italy
- **Obermaisser, Roman,** Universität Siegen, Germany
- **Palanque, Philippe,** ICS-IRIT, University Toulouse 3, France
- **Pu, Geguang,** East China Normal University
- **Pullum, Laura,** Oak Ridge National Laboratory, United States
- **Qin, Shengchao,** Teesside University, United Kingdom
- **Reeves, Steve,** University of Waikato, New Zealand
- **Roman, Dumitru,** SINTEF / University of Oslo, Norway
- **Rouff, Christopher,** Lockheed Martin, United States
- **Rozier, Kristin Yvonne,** NASA Ames Research Center
- **Ryan, Kevin,** Lero-the Irish Software Research Centre, Ireland
- **Rysavy, Ondrej,** Brno University of Technology, Czech Republic
- **Sachenko, Anatoly,** Ternopil National Economic University, Ukraine
- **Sanden, Bo,** Colorado Technical University, United States
- **Seceleanu, Cristina,** Mälardalen University, Västerås, Sweden
- **Sekerinski, Emil,** McMaster University, Canada
- **Selic, Bran,** Simula Research Lab, Norway
- **Sojka, Michal,** Czech Technical University, Czech Republic
- **Sun, Jing,** The University of Auckland, New Zealand
- **Taguchi, Kenji,** AIST, Japan
- **Trybus, Leszek,** Rzeszow University of Technology, Poland
- **van Katwijk, Jan,** Delft University of Technology, The Netherlands
- **Vardanega, Tullio,** University of Padova, Italy
- **Velev, Miroslav,** Aries Design Automation, United States
- **Vilkomir, Sergiy,** East Carolina University, United States
- **Waeselynck, Hélène,** LAAS-CNRS Toulouse, France
- **Zhu, Huibiao,** Software Engineering Institute - East China Normal University
- **Zoebel, Dieter,** University Koblenz-Landau, Germany

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 93–97
DOI: 10.15439/2018F375
ISSN 2300-5963 ACSIS, Vol. 16

Simulation Driven Development of Distributed Systems – Coupling of virtual and real system components

Tommy Baumann
Andato GmbH & Co. KG
Ehrenbergstraße 11
98693 Ilmenau, Germany
tommy.baumann@andato.com

Bernd Pfitzinger
Toll Collect GmbH
Linkstraße 4
10785 Berlin, Germany
bernd.pfitzinger@toll-collect.de

***Abstract*—Looking at the end-to-end processing, typical software-intensive systems are built as a system-of-systems where each sub-system specializes according to both the business and technology perspective. One challenge is the integration of all systems into a single system – crossing technological and organizational boundaries as well as functional domains. To facilitate the successful integration, we apply the system design process Simulation Driven Development (SDD). The basic idea is the application of realistic simulation models in parallel to existing software engineering procedures to enable testing and validation from day one. In this article we discuss the coupling of sub-systems implemented as virtual simulation models with sub-systems implemented as real sub-systems to enable unit tests against system level. Two approaches are presented: loose coupling and tight coupling.**

I. Introduction

THE ever-increasing complexity of software-intensive systems is an ongoing challenge to the tools and methods used in system engineering [1]. The building blocks of software-intensive systems or even systems-of-systems are often already systems themselves – either as commercial-off-the-shelf software used with 'minor' adaptations or custom-made software. This reuse of existing and proven solutions allows for the rapid construction of large systems. In the system development process, the engineering practices typically address – successfully – the scope of a single, independent system: The requirements are (and can be) defined, the system is modeled, implemented, tested and put into operations.

What started as the 'waterfall model' [2] has been adapted to mitigate the risks of failing to deliver a system in time, in budget and in quality. The software engineering models address the risk of failure in several ways: Using formal rather than natural language reduces misunderstandings; a small project scope reduces the 'value at risk' and testing assures that the intended level of quality is reached. These different approaches can be seen in a software engineering meta-model, like the V-Model XT [3] [4], a model in widespread use in Germany. The formal, document-oriented waterfall model is represented by the left side of the V, proceeding through the typical phases towards the implementation. In particular the model requires for each phase to define a dedicated test phase at a later time (the right-hand side of the 'V'). Depending on the focus of a particular project the meta-model allows for modern approaches e.g. test-driven development [5] or agile methods [6].

Yet methods proven to deliver successful implementations at the level of a single system seem to be inadequate for large-scale systems. [7] argues that implicit assumptions – the development process can be controlled, decisions are rational, the problem can be defined and delimited – break down when independent systems interact. Often the behavior of a coalition of systems is emergent and it is difficult or impossible to predict. An unintended consequence could be the complete failure of the overall system without a clear software bug within any contributing system ([8] gives the 2010 stock market flash crash as a prominent example).

We propose extending common software engineering practices through the use of simulation models. At any stage during the system development process the current level of detail needs to be captured in form of an executable specification. The system design process SDD follows exactly this axiom [9]. A major challenge of SDD is the coupling of different simulation models, simulators and already implemented system components to enable unit tests within the expected operational context. Each requirement and their test cases depend on the context, i.e. reusing a component or integrating a system into another system will most probably invalidate many assumptions and test cases [10]. Features, performance and operational aspects need to be proven before a sub-system is ready for integration. The development of such simulation-based test environments is a challenging task: sub-systems can be interfaced with simulation models where the remaining sub-systems operate at a higher degree of abstraction and the technical performance of the simulators can be insufficient.

In this article we discuss the possibilities and technical challenges regarding the coupling of simulation models with real world implementations. In section two we first introduce the SDD approach and show how modeling and simulation technologies can be applied to leverage specification quality and speed in the development of complex distributed systems. Section three explains the idea of loose coupling

©2018, PTI

and applies it in the example of the German toll system. In section four we explain the idea of tight coupling and apply it on the example of testing an electronic control unit (ECU) of a car. In the last section, we summarize and provide an outlook about future work.

II. Simulation Driven Development

Simulation Driven Development (SDD) is an efficient methodology for developing complex distributed systems. In contrast to conventional methodologies simulation models of the system to be developed are used during the entire development process [11]. These simulation models allow description of static and dynamic system properties which enormously improves specification quality and speed. Furthermore, they are used for analysis, evaluation, validation, verification and optimization during the course of the developing process. In addition to the simulation model of the system to be developed, SDD uses a simulation model of the development process itself. This simulation model allows automation of development steps in the form of workflows like optimization cycles, test cycles, revision controls and document sharing with component manufacturers. Fig. 1 shows SDD in the form of an extended V-Model.

User requirements of the system to be developed are determined and transferred into an initial simulation model during the requirement analysis. It represents the dynamic user behavior as well as the user interaction with the system in form of application scenarios / use cases. The simulation model is the basis for future acceptance tests and allows the first analyses in developing the design of the system.

In the subsequent phase of system specification, developer requirements (functional and not-functional) are derived and transferred into an executable specification. This is a simulation model of the whole system encompassing technical functions, architecture and user behavior. Each element of the simulation model is related to the user's or developer's requirements to ensure their compliance. The simulation model allows an early validation of the whole system against the application scenarios and allows locating integration problems (dynamic coupling effects between system components). Furthermore, it permits system-level optimizations and is the basis for later system integration tests.

Within the component specification each system component is further refined in form of simulation models. Starting point is the present executable system specification with the component related requirements and parameters. Simultaneously the executable system specification serves as development environment for component validation and optimization against the holistic system. As a consequence of the refinement, the components and thus also the system parameters become more specific. The component specification forms the basis for the later unit tests and component integration tests.

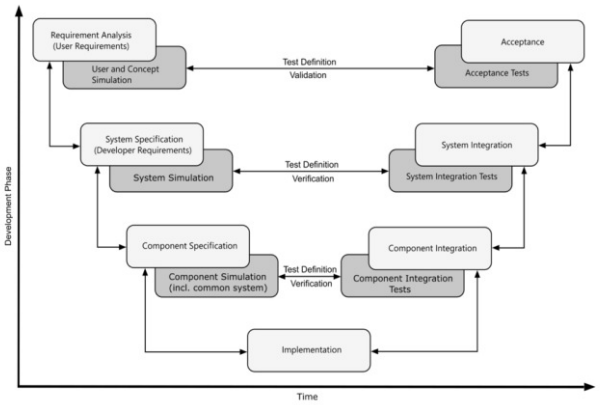

Fig. 1: Simulation Driven Development as a V-Model

After the specification of the system and its components the development process proceeds with the implementation. Based on the previously specified test cases the implementation will be verified and validated (performance tests and endurance tests). First, component integration tests are performed. In this step the functionality of each component is viewed in isolation. Afterwards components are integrated to a system followed by system integration tests. Finally, the system is tested against the initially defined user requirements by an acceptance test. SDD allows the reuse of the previously defined test cases which exist in form of simulation models. To that extent, simulation models are attached to the system implementation by Hardware/Software in the Loop (SiL/HiL).

To implement SDD a formal definition of the overall process is required. Figure 1 depicts the SDD approach as a V process model enriched by simulation-based test activities throughout the whole process. As in the classical V process model the level of detail increases when going downwards in the process and it is possible to go back to upstream process steps if it emerges that changes are required. The result of each process step is captured by one or more simulation models which provide test drivers for integration steps like "in-the-loop", integration or even acceptance tests. It is therefore similar to the W model in software development [12]. Both models share the concept of "testing starts at day one", i.e. that tests performed during the design stages prepare the integration tests and that test teams and system designers work together from the very beginning. A key improvement is that SDD features dynamic testing by the use of simulation models from the very beginning which directly supports dynamic testing in the integration phase in contrast to the classical W model where only static analysis is performed at early design phases.

Test failures or emerging new requirements can lead to changes in upstream (more general) designs which in turn might invalidate downstream (more detailed) specifications and simulation models. It is therefore critical to maximize the overall degree of formalism, automation and requirements traceability within and between each of the

steps in the SDD process model. This in turn requires detailed workflows for which individual working steps can then eventually be executed automatically by a workflow engine.

As mentioned above, SDD follows the idea of "test starts at day one". A prerequisite is the presence of an executable system model throughout all design stages. To enable unit tests, the system models need to be coupled to existing hardware or software components (HiL and SiL) which can be implemented in two different ways: loose coupling and tight coupling [13] [14]. While loose coupling doesn't consider the dynamic interaction between the sub-systems, tight coupling does.

III. Loose Coupling

Loose coupling simplifies the implementation of unit tests of relatively independent sub-systems. The prerequisite is a mostly unidirectional exchange of data. e.g. lacking feedback or disregarding feedback that deviates from the simulation. This enables technically simpler test benches, since a sub-system can be tested without real-time connectivity to the rest of the system. One possible implementation is to simply take the events passing through the simulation model and export those events that are relevant to a particular domain-specific interface. The event list is a flat file with a chronological list of points in time and additional domain specific information.

In the example of the German automatic toll system we use this approach to provide the temporal behavior of 1.1 million on-board units communicating with the central system [15]. The event list contains all TCP/IP connections and their domain specific payload, e.g. the number of journey data files transmitted. Operational testing takes this temporal data, adds the domain specific content and uses another more technical simulator to drive the real-world data center applications. The example illustrates two challenges encountered in this approach:

- The abstraction gap – when a high-level simulation model encounters a real-world system – is overcome by additional simulation models. Loose coupling of the simulation models immediately translates into loose coupling of the technical teams involved.
- The simulation performance – where a real-world system must not be impacted by the simulation performance. Loose coupling allows to precompute at least parts of an operational test.

IV. Tight Coupling

Tight coupling is used to develop and test sub-systems that work closely together, e.g. as parts of an embedded system. The entire logic of such systems is typically distributed across several sub-systems. These sub-systems need to be connected dynamically to deliver the desired logic/process. Each sub-system is characterized by interconnectivity and inter-processing to the other sub-systems and depends strongly on them.

Two possibilities of technical implementation can be distinguished: direct tight coupling using data files, shared memories or network protocols and indirect tight coupling using a central data distribution service (DDS). Fig. 2 shows direct coupling using TCP/IP sockets. Each participating simulator needs to implement an adapter to each connected simulator – a time-consuming and expensive task. All requirements regarding quality of service (real-time capabilities, distribution of simulators across multiple nodes), standardization of data types (different data types of simulators) and time synchronization (global clock as simulation time) need to be considered. These issues are solved using a central data management entity like DDS.

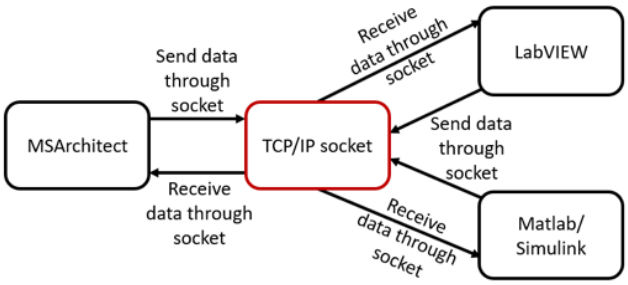

Fig. 2: Direct tight coupling using TCP/IP

DDS is a networking middleware that simplifies complex network programming and distribution of data. It implements a publish–subscribe pattern for sending and receiving data, events, and commands among the nodes. Nodes that produce information (publishers) create "topics" (e.g., temperature or pressure) and publish "samples". DDS delivers the samples to subscribers that declare an interest in that topic. DDS handles transfer chores: message addressing, data marshalling and de-marshalling (so subscribers can be on different platforms from the publisher), delivery, flow control, retries, etc. Any node can be a publisher, subscriber, or both simultaneously.

The DDS publish-subscribe model virtually eliminates complex network programming for distributed applications. The key benefit is that applications that use DDS for their communications are decoupled. Little design time needs be spent on handling their mutual interactions. In particular, the applications never need information about the other participating applications, including their existence or locations [16] [17]. Fig. 3 shows the basic architecture of a DDS solution.

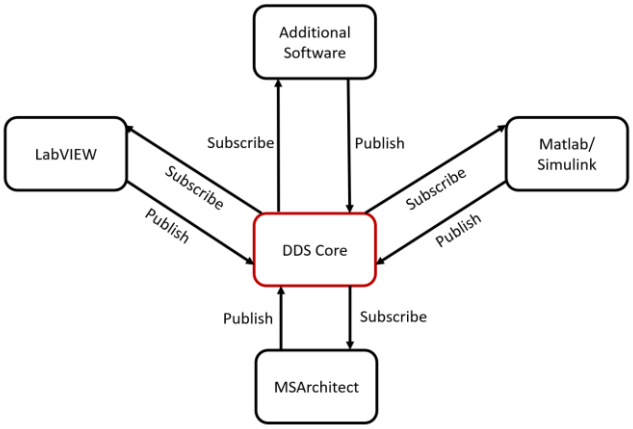

Fig. 3: Indirect tight coupling using DDS

We have developed an ECU test environment which couples three different simulation tools (MSArchitect, Matlab/Simulink and LabVIEW). It is used to validate and test the function and performance of the interface between the three simulation tools. The coupling between those systems is accomplished by using TCP/IP sockets achieving a high data transfer rate.

In this example MSArchitect takes the server role while Matlab/Simulink and LabVIEW are the clients. During the initialization phase of the simulation the server is waiting for incoming connections. Once all clients have been connected to the TCP/IP sockets the simulation starts.

The tight coupling example caused several challenges which have to be solved. The data transfer through TCP/IP sockets is limited to character arrays. Therefore, the different data types of the several simulation tools could cause issues because of the conversion in character arrays and the conversion back to the data types. For this reason, a definition of supported data types and a standardization of those data types need to be developed. Furthermore, the time of the connected simulators need to be synchronized, for example by implementing a global clock, controlled and distributed by one simulator.

Fig. 4 shows the user interfaces of the test environment. The whole car is modelled in MSArchitect and parts of the model are implemented in LabVIEW or Matlab/Simulink. The LabVIEW model contains the entire logic of the ECU which is built in the car model. Matlab/Simulink provides the logic and behavior of the combustion engine. The performance of the ECU test environment is limited by the hardware resources of the test computer.

In our case the desired maximum performance of the implemented interface could not be accomplished since all three models were running on the same system. In future work, we want to split the simulation models over three processing nodes to reach a significant higher performance of the entire simulation. At this point, the ECU test environment is running without any data loss or synchronization issues, which in fact is a great starting point for further development.

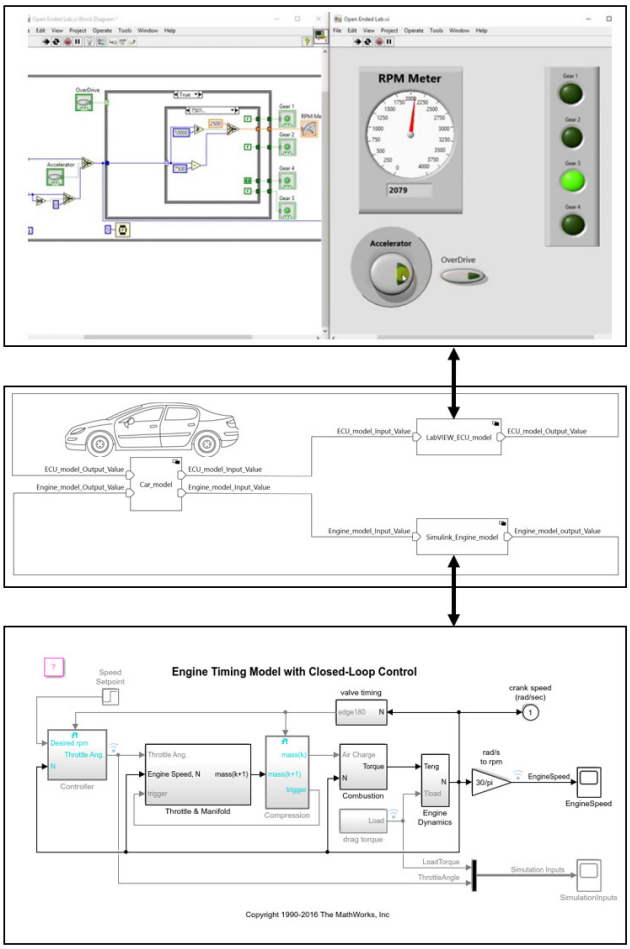

Fig. 4: ECU test environment

V. Conclusion

We have shown the applicability of simulation models along the development process using a simulation driven design process. From the onset the simulation model provides the operational context and over time the level-of-detail can be increased to the point where real-world sub-systems interface with the simulation model. In the article we discussed how simulation models can interface with real-world hardware or software to enable realistic unit tests. Two possibilities have been introduced: loose coupling and tight coupling.

Next, we're planning to extend our DDS solution and connect the examples mentioned above to this solution. Above all, securing a defined quality of service within DDS will be a challenging task. In addition, we're preparing a guideline to select a proper coupling mechanism.

References

[1] B. Boehm: A view of 20th and 21st century software engineering. Proceedings of the 28th international conference on Software engineering, ACM, 2006, pp. 12-29. doi: 10.1145/1134285.1134288.

[2] W. W. Royce: Managing the development of large software systems. Proceedings of IEEE WESCON, Los Angeles, Aug. 1970, pp. 1-9.

[3] Verein zur Weiterentwicklung des V-Modell XT e.V. (Weit e.V.): V-Modell XT version 2.0. [accessed 21-Jan-16]. Available: http://www.v-modell-xt.de/

[4] S. Biffl, D. Winkler, R. Höhn, and H. Wetzel: Software process improvement in Europe: Potential of the new V- Modell XT and research issues. Software Process: Improvement and Practice, vol. 11, no. 3, pp. 229-238, Jun. 2006. doi: 10.1002/spip.266

[5] D. Janzen and H. Saiedian: Test-Driven Development: Concepts, taxonomy, and future direction. Computer, vol. 38, no. 9, pp. 43-50, Sep. 2005. doi: 10.1109/MC.2005.314

[6] Manifesto for agile software development, [accessed 29- Jul-2015]. Available: http://www.agilemanifesto.org/.

[7] I. Sommerville, D. Cliff, R. Calinescu, J. Keen, T. Kelly, M. Kwiatkowska, J. McDermid, and R. Paige: Large-scale complex IT systems. Communications of the ACM, vol. 55, no. 7, pp. 71-77, Jul. 2012, doi: 10.1145/2209249.2209268

[8] D. Cliff and L. Northrop: The global financial markets: An ultra-large-scale systems perspective. In R. Calinescu and D. Garlan (Eds.): Large-Scale Complex IT Systems. Development, Operation and Management. Ser. Lecture Notes in Computer Science, vol. 7539, pp. 29-70, Berlin: Springer, 2012. doi: 10.1007/978-3-642-34059-8_2.

[9] Baumann, T, "Simulation-driven design of distributed systems". In SAE International, SAE Technical Paper, pp. 1-7, 2011. doi:10.4271/2011-01-0458.

[10] E. J. Weyuker: Testing component-based software: A cautionary tale. IEEE Software, vol. 15, no. 5, pp. 54-59, Sep. 1998. doi:10.1109/52.714817

[11] B. Pfitzinger, T. Baumann, and T. Jestädt, "Simulation driven development - validation of requirements in the early design stages of complex systems - the example of the German toll system," in Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., ser. Annals of Computer Science and Information Systems, vol. 11, IEEE, Sep. 2017, pp. 1127-1134. doi: 10.15439/2017F133.

[12] Gerrard, P. and N. Thompson: „Risk-Based E-Business Testing", Artech House INC 2002, ISBN-13: 9781580533140, ISBN-10: 1580533140

[13] A. Salkintzis, C. Fors, and R. Pazhyannur, "WLAN-GPRS Integration for Next-generation Mobile Data Networks," IEEE Wireless Communications, vol. 9, no. 5, pp. 112-124, October 2002. doi: 10.1109/MWC.2002.1043861

[14] Beck, F. and Diehl, S., "On the Congruence of Modularity and Code Coupling", in 19th ACM SIGSOFT Symposium on the Foundations of Software Engineering and 13rd European Software Engineering Conference (ESEC/FSE '11), Szeged, Hungary, September 2011. doi:10.1145/2025113.2025162

[15] Pfitzinger, B.; Baumann, T.; Jestädt, T.: Network Resource Usage of the German Toll System: Lessons from a Realistic Simulation Model. In: 46th Hawaii International Conference on System Sciences (HICSS) (2013), pp. 5115–5122. doi: 10.1109/HICSS.2013.41

[16] Rekik, R. and Hasnaoui, S., "Application of a CAN BUS Transport for DDS Middleware"; 2009 Second International Conference on the Applications of Digital Information and Web Technologies, London, 2009, pp. 766-771.doi: 10.1109/ICADIWT.2009.5273919

[17] Deniz, E. et al., "DDS Based MIL-STD-1553B Data Bus Interface Simulation"; The Journal of Defense Modeling and Simulation, vol 12, issue 2, pp. 179 – 188. doi: 10.1177/1548512914530534

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 99–102
DOI: 10.15439/2018F378
ISSN 2300-5963 ACSIS, Vol. 16

Challenges in Causal Inference from Personal Monitoring Devices

Tomasz Wiktorski
Department of Electrical and Computer Science
University of Stavanger
Stavanger, Norway
Email: tomasz.wiktorski@uis.no

***Abstract*—Personal Monitoring Devices (PMDs) collect immense amount of data about health and wellness of hundreds of millions of people. One of the obstacles of the prevailing data analytics approaches to PMDs' data is limited value of correlation-based conclusions in a health context. Causal inference seems a natural solution, but general causal inference methodologies are difficult to apply to PMDs data due to size and complexity of observational data. Some methods, such as randomized trials, are largely infeasible in PMDs' context due to lack of control over the investigated population. In this paper, we overview existing approaches to causal inference including recent works that attempt to take advantage of time series data to automatically derive causality using extended difference-in-deference or Granger methods. We then outline challenges and opportunities for causal inference in the health context. Finally, we propose a following challenge: can we establish a new standard of evidence and a study design process that: (1) allows for drawing causal conclusions from large observational datasets and (2) can suggest interventions to enforce causal links discovered in the data.**

I. Introduction

OFTEN repeated phrase "correlation is not causation" became with time an offhand apology for providing only correlation-based conclusions for research questions, rather than motivating efforts to develop improved methods for determining causation. Even though, it might require a considerable extra effort to arrive at causal relations, the utility of causal results is far greater than correlation results, even if causal link would be of limited precision.

Many researchers and practitioners settle on correlation only, but this trend appears to change recently. Highly publicized cases, such as flops of Google Flu Trends[1], show dangers of ignoring proper causal analysis. If we want to plan and evaluate interventions or make safe long-term predictions, a tool stronger than correlation is necessary.

Pearl, originally known for development of Bayesian Networks and probability-centered approach to AI, suggests in his new book [2] that current developments in machine learning and artificial intelligence focus too much on improved curve fitting, that is correlation and probability. To lead to the next breakthrough machines have to reason beyond probability. He brings an example of malaria and fever. It is important to understand that malaria causes fever and not only that they are correlated. Introducing causality language into existing data analyses could help us reason about possible results of interventions using observational data.

Correlation prevails as for now in big data publications, as documented by Ekibia et al. review [3]. Other authors, e.g., George et al. [4] also notice that causality does not inform the design of big data as a domain. Grimmer [5] points out that large amounts of data will not solve the problem of variable selection in causal inference. Scientists require more training to better understand causal inference. With time, problems, such as those with Google Flu Trends, uncovered issues with such approach. Lazer et al. in their Science paper [6] suggest changes to approaches to data analysis based on the problems with Google Flu prediction. They show that analysis based on like Google Correlate is difficult to reproduce. Even Google Flu cannot be reproduced using Correlate. Analyses that base on sources produced by users are susceptible to blue team and red team problem. Blue team problem is the unintended influence on algorithm results due to changes to data generation as a result of changing company policies and goals. Red team problem is the unintended influence on algorithm results due to attack on data generation mechanism (generating fake data based on knowledge of algorithm functioning).

More and more authors stress importance of causal analysis for big data and data science. A good example is an article by Provost and Fawcett [7]. They stress, in particular, the importance of careful analysis of assumptions on confounding factors, causal links are directly dependent on these assumptions. Confounding factor is characterized by associations both with outcome and evaluated cause variable, but does not lie in the cause path between them. Kitchin [8] describes his reservations about the shift from knowledge-driven science to data-driven science. At the same time he notices benefits of big data and suggests that they could lead to creation of more sophisticated models.

After Introduction, in Section II we explain basics of causal inference including most common models and methods, such as Neyman-Rubin Causal Model, Structural Equation Modeling, Structural Causal Model, and Difference in Differences. In Section III we present current trends in causal inference in big data, healthcare, and time series leading to automation. We conclude in Section IV where we propose a challenges.

II. Basics of Causal Inference

Gelman [9] organizes causal reasoning in two general types of questions one can ask: forward and reverse. These two

©2018, PTI

questions are sometimes formulated in a shorthand form as "effects of causes vs. causes of effect", which traces back to Mill [10]. Forward question asks about results of a certain intervention one might perform. Reverse question asks about causes of a certain observed effect. The consensus based on Gelman's overview is that effects of causes can be usually traced, so forward question can be answered. Typical methods used to answer forward questions include: Neyman-Rubin causal model (RCM), Structural Equation Modeling (SEM), and Structural Causal Model (SCM).

Causes of effects are more difficult to deduce, especially if we are facing complex processes involving, e.g., politics, such as causes of war. Nevertheless, some reverse questions, even involving social matters, might still be answerable. It typically requires expert reasoning to pre-identify meaningful factors for actual analysis. The analysis would then account for these factors, which should lead to disappearance of differences in outcomes. This confirms causes of effects under assumption that factors are meaningful. This assumption naturally weakens such causal argument.

A. Model vs. Data

There are several general viewpoints on what is necessary to derive causal relation. These views are summarized and organized by Gelman. The main question that Gelman asks is how permissive a particular view is in allowing correlation and observational data as a basis for causal relations. Most strict approaches are based exclusively on strict models and randomized studies. Some scientists, such as Pearl, hold a view that observational data under strong models may lead to causal conclusions. In some domains it is also common to derive causality from covariance matrices, but it is unclear how widely it can be applied. The most permissive view covers automatic derivation of causality by computer from observational data. Gelman's organization can be seen as a scale that balances between model and data. All methods require data for confirmation of the actual relation, but few see data as a sole basis for such relation.

RCM is the method that often unconsciously guides people's thinking about causal inference. It introduced concept that are now taken for granted, such as: counterfactual, treatment, and control. Main methods that base on structure are SEM and SCM. They use both graphical and purely computational techniques to a different extent, particularly there are many different methods under SEM. Main method which uses minimal structure is DD (Difference in Differences), it also heavily focuses on use of data. It can be used as a basic calculation technique for some other methods that employ structural approach. An interesting example of a method with minimal expert structure is Google's CasualImpacts [11], which we mention in greater detail later in this paper. We will now shortly describe the four main methods: Neyman-Rubin causal model (RCM), structural equation modeling (SEM), structural causal model (SCM), and Difference in Differences (DD).

B. Neyman-Rubin Causal Model

The basis for Neyman-Rubin causal model was first formulated by Neyman [12] for randomized (alternatively called controlled) experiments and extended by Rubin [13] and others for both observational and experimental studies. This model introduced basic tools considered now a standard in any causal analysis. The model defines counterfactuals used to specify potential outcomes and then comparing outcomes of alternative exposures. In practice, it is usually impossible to check more than one intervention on the same subject. This problem is often named the fundamental problem of causal inference. The goal of randomized experiments is to solve this problem by creating two comparable groups that collectively can form a basis for causal reasoning.

C. Structural Equation Modeling

SEM is, as Kline [14] describes, a grouping of several methods used to verify a proposed causal model. Methods include path analysis, confirmatory factor analysis, structural regression models, latent growth models, covariance and correlation structure models. Sometimes path analysis is called causal modeling, which is considered an anachronism. Results of SEM cannot be assumed to be causal per se. Causality is dependent on the design of experiment which SEM methods then verify.

Confirmatory Factor Analysis is a form of Factor Analysis. It is used to compare researchers understanding of factors and their relation against the measurement. CFA is sometimes called measurement model in SEM and it does not specify structure. Factor analysis (Exploratory Factor Analysis, EFA) uses unobserved variables, called factors, to reduce amount of variables necessary to described variability in observed data. Principal Component Analysis is a simple version of EFA

Path analysis, is a method to specify directed dependencies between variables in the model. It only provides structure, no measurement. Latent growth modeling adds a time component to explanation of dependent variables, this way it can describe longitudinal change (change over time). Most applications limit themselves to slope and starting point only, but higher order methods are also used.

D. Structural Causal Model

SCM proposed by Pearl [15] attempts to integrate other existing approaches to causal inference into a unified model, in particular it subsumes SEM and various graphical models. It also claims to subsume RCM, but Aliprantis [16] indicates that an important difference in modeling approaches exist between these two methods. Even though, it is possible to find an SCM generating any RCM contingency tables, there is no unique SCM that generates a particular solution. It may result in critical discrepancies between each approach. However, this limitation does not necessarily take away all usability of SCM. It is still important because the integrative approach is helpful to clarify formal meaning of important concepts, such as: d-separation, interpretation of counterfactuals, and confounding.

E. Difference in Differences

Difference in differences method uses observational data to imitate experimental design. It calculates normal difference in

the outcome, using control group for normalization. Measurements need to be done at least once before and once after the intervention. The control group by assumption should: (1) show similar trends over time, (2) not be exposed to intervention enacted on or experienced by treatment group, (3) nothing other than treatment changes in only one group. These assumption can be collectively called a parallel trend assumption.

In the context of DD one might talk about natural experiment, which is non-randomized experiment. DD allows to estimate counterfactual from observational data from natural experiments. There preferably should be two groups with similar experience, but differently affected by an experiment or intervention. Intervention should be well planned, it should be exogenous that is not a reaction to a specific behavior, because it could cause people to change their behavior unpredictably to *game the system*. Resulting in inferior or even useless results.

III. Current Trends

A. Big Data

Emerging sources of data, often coming from sensors, are characterized by large volume (especially accumulative), sometimes high velocity, and wide variety of actual data sources and related data formats. These properties are commonly refereed as 3 Vs of big data. Moreover, such data that reflect various natural phenomena are often observational. By natural phenomena in this context we mean ones that actually occur due to normal human or other activity, in contrast to, e.g., simulations. While it is in theory possible to attempt randomized experimentation in such setting, it is unfortunately usually infeasible. Reasons are multiple, but major ones include: cost, ethical considerations, and general feasibility of recruitment to control groups.

Despite these obvious problems, new data sources provide important opportunities due to their granular and longitudinal character. Progress in sensor and storage technologies provides the possibility to capture data over long periods of time, large areas, and with ample frequency. These properties are not yet well explored in the existing research, with several exceptions which we mention here.

The most notable example might be work of Brodersen et al. [11] on CausalImpacts. The motivation of this approach was to infer causal impact for marketing campaigns, but it is not necessarily limited just to this one domain. In principle, it bases on difference in differences (DD) approach, but uses a state-space model of time series to predict the counterfactual. This is to compensate for many of the limitations of basic DD. It allows considering synthetic control, inclusion of Bayesian priors for parameters, and measuring evolution of impact of an intervention over time.

Three types of sources are used to construct the synthetic control. The first are properties of outcome time series before intervention. The second are properties of time series that could be used to predict outcome time series before intervention. The third source, if available, are parameter values from older related studies used as Bayesian priors.

Zigler and Dominici [17] propose a new Bayesian method to select Propensity Score variable, which is supposed to help causal inference in i.a. big data scenario. Beyond these examples, we have not found other explicit attempts to address data scale in causal analysis.

B. Time Series and Automation

Gelman in his aforementioned review suggests that time can be considered from two different non-exclusive perspectives in causal inference. First perspective considers contamination of results due to exposure to more than one treatment over time. Such contamination requires special attention in observational data, where exposure to more than one treatment is difficult to avoid and not always direct. For instance mere knowledge about alternative treatment might impact the results. Placebo method, which corrects for this issue, is difficult to deploy outside controlled randomized experiment setting.

Second perspective includes time as system variable to express changing effect of one treatment over time. Gelman only considers it for limited amount of future time points. Such limited treatment presumably results from combination of data available in traditional causal research and related methods. Some most resent developments address this limitation.

Granger causality test [18] has been more and more commonly used in the recent years. It provides well defined tool to determine if one time series is a predictor of another. In many cases this might be considered close to causality, but in principle it is a correlation with time precedence. The method relies on time series and allows for automation of the causality inference process. It defines determinate formulas eliminating (to some extent) necessity for human involvement, what makes it easier to scale with growing data, as long as we consider some form of windowing or sampling.

C. Health and Welfare Monitoring

Krumholz in his widely cited paper on "Big Data and New Knowledge in Medicine" [19] observes that improvements in any service (be it either health or movie rental) are possible despite lack of conceptual models, hypotheses testing, or randomized trials, but can come directly from observational data. Historically empirical insights derived from existing data were considered inferior to insights based on theory and experiments. However, many research questions can be answered based on observational data without understanding the underlining mechanism. In some case, these data might contribute then to understanding of the mechanism. Author provides an example of aspirin that was successfully used without physicians at the time knowing why it produced the results.

The paper opens for development of causal inference from non experimental studies. It is particularly important when experimentation might be unethical. Influence of smoking on development of cancer is widely accepted despite relying largely on observational data. Author identifies development of widely accepted criteria for evaluating causal conclusions from large observational datasets as crucial for further work, but does not attempt that in the paper.

Visvanathan et al. in the Research Statement of American Society of Clinical Oncology [20] notice that observational

studies can be complimentary to Randomized Controlled Trials (RCTs) by generating new hypotheses, revealing patterns and answering questions that cannot be answered by RCTs. To achieve this it is necessary to use a rigorous methodology and transparent reporting, in addition to ensuring data quality, interoperability and privacy.

What emerges through this Research Statement is a need for a new standard of evidence to take advantage of observational data and compensate for the lack of randomness and controlled groups. It remains an open question whether elements of quasi-randomness could be effectively introduced to observational studies on large populations.

Janke et al. [21] discuss the potential of primitive analytics and big data in emergency car. The potential that they consider unexplored. They notice that observational data could improve patient care but weak causal inference is a limiting factor at the moment. One of the approaches they suggest is that systems could prompt for additional information after discovering patterns of interest. It could be done both on population and individual level.

They believe that observational data might be used to derive and validate new models, it cannot be used to evaluate the effects of implementing these models. Example of influence of smoking on cancer development shows that the mentioned limitation should not preclude wider use of observational approaches. Moreover, possible advancements in causal reasoning could at least partially reduce that limitation.

IV. Conclusions - The Challenge

In order to take advantage of large observational dataset coming, in the context of underdevelopment methodologies of causal inference from such datasets, propose the following challenge: can we establish a new standard of evidence and a study design process that: (1) allows for drawing causal conclusions from large observational datasets and (2) can suggest interventions to enforce causal links in these data.

We propose to focus on Personal Monitoring Devices, at least in the beginning. Large portion of population is already in possession of such device, data are already collected (though in proprietary systems) and causal analysis has potential to lead to meaningful health and well-being recommendations both on individual and societal level.

Based on the existing literature we suggest that the starting point should be causal inference based on time series properties as preliminary explored in, e.g., aforementioned Google's library CausalImpacts, possibly in combination with some formal causality formulation method, such as Pearl's Do-Calculus [22].

These developments could later also be applied wider to smart technologies and observational data, but we think that proposed initial limitation to would in fact stimulate faster development of an accepted methodology.

References

[1] "Google's flu project shows the failings of big data | time," http://time.com/23782/google-flu-trends-big-data-problems/, (Visited on 12/29/2015).

[2] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. Penguin UK, 2018.

[3] H. Ekbia, M. Mattioli, I. Kouper, G. Arave, A. Ghazinejad, T. Bowman, V. R. Suri, A. Tsou, S. Weingart, and C. R. Sugimoto, "Big data, bigger dilemmas: A critical review," *Journal of the Association for Information Science and Technology*, 2015.

[4] G. George, M. R. Haas, and A. Pentland, "Big data and management," *Academy of Management Journal*, vol. 57, no. 2, pp. 321–326, 2014.

[5] J. Grimmer, "We are all social scientists now: How big data, machine learning, and causal inference work together," *PS: Political Science & Politics*, vol. 48, no. 01, pp. 80–83, 2015.

[6] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of google flu: traps in big data analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[7] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big Data*, vol. 1, no. 1, pp. 51–59, 2013.

[8] R. Kitchin, "Big data, new epistemologies and paradigm shifts," *Big Data & Society*, vol. 1, no. 1, p. 2053951714528481, 2014.

[9] A. Gelman, "Causality and statistical learning," *American Journal of Sociology*, vol. 117, no. 3, pp. 955–966, 2011.

[10] J. S. Mill, *A System of Logic Ratiocinative and Inductive: Boeing a Connected View of the Principales of Evidence and the Methods of Scientific Investigation*. Bombay, 1906.

[11] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, S. L. Scott *et al.*, "Inferring causal impact using bayesian structural time-series models," *The Annals of Applied Statistics*, vol. 9, no. 1, pp. 247–274, 2015.

[12] J. Neyman, "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes," *Roczniki Nauk Rolniczych*, vol. 10, pp. 1–51, 1923.

[13] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[14] K. A. Markus, "Principles and practice of structural equation modeling by rex b. kline," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 19, no. 3, pp. 509–512, 2012.

[15] J. Pearl *et al.*, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, pp. 96–146, 2009.

[16] D. Aliprantis, "A distinction between causal effects in structural and rubin causal models," 2015.

[17] C. M. Zigler and F. Dominici, "Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 95–107, 2014.

[18] C. W. Granger, "Causality, cointegration, and control," *Journal of Economic Dynamics and Control*, vol. 12, no. 2-3, pp. 551–559, 1988.

[19] H. M. Krumholz, "Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system," *Health Affairs*, vol. 33, no. 7, pp. 1163–1170, 2014.

[20] K. Visvanathan, L. A. Levit, D. Raghavan, C. A. Hudis, S. Wong, A. Dueck, and G. H. Lyman, "Untapped potential of observational research to inform clinical decision making: American society of clinical oncology research statement," *Journal of Clinical Oncology*, vol. 35, no. 16, pp. 1845–1854, 2017.

[21] A. T. Janke, D. L. Overbeek, K. E. Kocher, and P. D. Levy, "Exploring the potential of predictive analytics and big data in emergency care," *Annals of emergency medicine*, vol. 67, no. 2, pp. 227–236, 2016.

[22] J. Pearl and E. Bareinboim, "External validity: From do-calculus to transportability across populations," *Statistical Science*, pp. 579–595, 2014.

5th Doctoral Symposium on Recent Advances in Information Technology

THE aim of this meeting is to provide a platform for exchange of ideas between early-stage researchers, in Computer Science and Information Systems, PhD students in particular. Furthermore, the symposium will provide all participants an opportunity to get feedback on their studies from experienced members of the IT research community invited to chair all DS-RAIT thematic sessions. Therefore, submission of research proposals with limited preliminary results is strongly encouraged.

Besides receiving specific advice for their contributions all participants will be invited to attend plenary lectures on conducting high-quality research studies, excellence in scientific writing and issues related to intellectual property in IT research. Authors of the two most outstanding submissions will have a possibility to present their papers in a form of short plenary lecture.

TOPICS

- Automatic Control and Robotics
- Bioinformatics
- Cloud, GPU and Parallel Computing
- Cognitive Science
- Computer Networks
- Computational Intelligence
- Cryptography
- Data Mining and Data Visualization
- Database Management Systems
- Expert Systems
- Image Processing and Computer Animation
- Information Theory
- Machine Learning
- Natural Language Processing
- Numerical Analysis
- Operating Systems
- Pattern Recognition
- Scientific Computing
- Software Engineering

EVENT CHAIRS

- **Kowalski, Piotr Andrzej,** Systems Research Institute, Polish Academy of Sciences; AGH University of Science and Technology, Poland
- **Lukasik, Szymon,** Systems Research Institute, Polish Academy of Sciences, AGH University of Science and Technology, Poland

PROGRAM COMMITTEE

- **Arabas, Jaroslaw,** Warsaw University of Technology, Poland
- **Atanassov, Krassimir T.,** Bulgarian Academy of Sciences, Bulgaria
- **Balazs, Krisztian,** Budapest University of Technology and Economics, Hungary
- **Bronselaer, Antoon,** Department of Telecommunications and Information at Ghent University, Belgium
- **Castrillon-Santana, Modesto,** University of Las Palmas de Gran Canaria, Spain
- **Charytanowicz, Malgorzata,** Catholic University of Lublin, Poland
- **Corpetti, Thomas,** University of Rennes, France
- **Courty, Nicolas,** University of Bretagne Sud, France
- **De Tré, Guy,** Faculty of Engineering and Architecture at Ghent University, Belgium
- **Fonseca, José Manuel,** UNINOVA, Portugal
- **Fournier-Viger, Philippe,** University of Moncton, Canada
- **Gil, David,** University of Alicante, Spain
- **Herrera Viedma, Enrique,** University of Granada, Spain
- **Hu, Bao-Gang,** Institute of Automation, Chinese Academy of Sciences, China
- **Koczy, Laszlo,** Szechenyi Istvan University, Hungary
- **Kokosinski, Zbigniew,** Cracow University of Technology, Poland
- **Krawiec, Krzysztof,** Poznan University of Technology, Poland
- **Kulczycki, Piotr,** Systems Research Institute, Polish Academy of Sciences, Poland
- **Kusy, Maciej,** Rzeszow University of Technology, Poland
- **Lilik, Ferenc,** Szechenyi Istvan University, Hungary
- **Lovassy, Rita,** Obuda University, Hungary
- **Malecki, Piotr,** Institute of Nuclear Physics PAN, Poland
- **Mesiar, Radko,** Slovak University of Technology, Slovakia
- **Mora, André Damas,** UNINOVA, Portugal
- **Noguera i Clofent, Carles,** Institute of Information Theory and Automation (UTIA), Academy of Sciences of the Czech Republic, Czech Republic
- **Pamin, Jerzy,** Institute for Computational Civil Engineering, Cracow University of Technology, Poland
- **Petrik, Milan,** Czech University of Life Sciences Prague, Faculty of Engineering, Department of Mathematics, Czech Republic
- **Ribeiro, Rita A.,** UNINOVA, Portugal

- **Sachenko, Anatoly,** Ternopil State Economic University, Ukraine
- **Samotyy, Volodymyr,** Lviv State University of Life Safety, Ukraine
- **Szafran, Bartlomiej,** Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Poland
- **Tormasi, Alex,** Szechenyi Istvan University, Hungary
- **Wei, Wei,** School of Computer science and engineering, Xi'an University of Technology, China
- **Wysocki, Marian,** Rzeszow University of Technology, Poland
- **Yang, Yujiu,** Tsinghua University, China
- **Zadrozny, Slawomir,** Systems Research Institute, Poland
- **Zajac, Mieczyslaw,** Cracow University of Technology, Poland

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 105–107
DOI: 10.15439/2018F240
ISSN 2300-5963 ACSIS, Vol. 16

Static typing and dependency management for SOA

Nikita Gerasimov
Saint Petersburg University
Mathematics and Mechanics Faculty
Universitetsky prospekt, 28, Peterhof, St. Petersburg, Russia
Email: n.gerasimov@2015.spbu.ru

***Abstract*—Several problems related to work reliability appear while building service-oriented systems. The first problem consists in lack of static typing and lack of inter-service data type checking. The second one consists in high services connectivity. The article shows an example of strong and static polymorphic type system and a type check algorithm. Type system syntax and service-contract concept are described. Theoretic results were realized in a service form and were applied in practice in a real system, which improved its reliability. Also, technical realization decreased services connectivity which promoted system quality increase.**

I. Introduction

A DEVELOPMENT of convenient multi-logic systems bases often on service or microservice oriented architectures (SOA). SOA means that application logic is divided into several self-sufficient components, providing separate tasks realization [8]. Every component has the single responsibility. The advantages of SOA are simplified maintenance, independence of single technology or programming language. Every logic change requires modifying only in-component realization and implementation of the current external interface.

However, SOA has also disadvantages, an inequality of providing and using interfaces and type checking in the whole system [5]. Various frameworks and approaches suggest the ways of system decomposition but don't suggest any ways of types consistency statically checking. Building analogy from dynamic-typing language this fact leads to in-production system instability increase.

RPC-frameworks like Google Protobuf or Apache Thrift partially solve static type checking by providing client and server code generation based on API definition. Mentioned solutions allow ensuring at development stage that client and server would use the identical protocol. However code generation becomes less trivial while using JSON/XML-PRC, REST or using event-driven architectures.

Next problem is less critical. Detection of outdated API usage can be nontrivial in complex systems with various components. Lack of automatized control over API usage leads to possibility of important component disabling. The real case is that an outdated service A provides statistics collecting once per month by some mailing service. Logs analysis proves that there were no API calls during last 3 weeks that's why the service can be disabled.

Finally, we have 2 main problems:

- the absence of strong type system with static checking for SOA that leads to potential stability decrease
- the absence of dependency control for SOA leading to possible breaking system in runtime after disabled outdated APIs

Therefore the main goal of this research is to increase stability of SOA-based systems and decrease runtime errors.

There are two stages to reach the goal:

- improve the existing approach to service API typification to statically check types
- develop service that should control API dependencies in the SOA system

Much of the work presented here is connected with the description of a new tool providing the achievement of formulated goals. New service purpose is close to service-discovery systems purpose: to detect suitable components over the network automatically [4]. The main objective of the new service is checking of type consistency for providing and using API definitions. We call it "contract discovery".

The next section defines the proposed type system and type checking algorithm to be realized in contract discovery service. Section 3 surveys the concept of contract and how contract-discovery service provides client to service linking. Section 4 describes our contract discovery service realization details. Section 5 illustrates our experience of application such service to the real microservice-oriented event-driven system.

II. Type system

Data can be encoded with custom binary or text format: with XML or JSON while interoperation. Encoded data satisfies restrictions of communication protocol: SOAP, XML-RPC (XML); REST, JSON-RPC (JSON); Protobuf, Thrift (binary) and so on. APIs based on the communication protocols can be described with formal specifications: WSDL for SOAP, OpenAPI for REST, etc. Event-driven SOA usually uses JSON as a data format and JSON Schema standard for validating and describing data structures.

All mentioned protocols are limited by using simple (integer, boolean, etc.) or complex (arrays and records) types [7][9]. For example, simplified JSON Schema type system [1] can be expressed as presented at the figure 1.

Described grammar is simplified because it does not cover complex predicates containing boolean logic. Also, the grammar does not cover specific type formats.

According to standardization and popularity we took JSON Schema as a base for our type system. To improve compatibility of services we suppose the described type system to be

©2018, PTI

⟨*t*⟩ ::= ⟨*arr*⟩ | ⟨*obj*⟩ | ⟨*num*⟩ | ⟨*str*⟩ | **boolean** | ⟨*p*⟩ ⟨*t*⟩

⟨*arr*⟩ ::= {⟨*t*⟩} | ⟨*ap*⟩ ⟨*arr*⟩

⟨*ap*⟩ ::= **additionalItems** | **maxItems** | **minItems**
| **uniqueItems** | **contains**

⟨*obj*⟩ ::= ⟨*t*⟩ ⟨*t*⟩ | ⟨*op*⟩ ⟨*obj*⟩

⟨*op*⟩ ::= **maxProperties** | **minProperties** | **required**
| **properties** | **patternProperties**
| **additionalProperties** | **dependencies**
| **propertyNames**

⟨*num*⟩ ::= **integer** | **real** | ⟨*np*⟩ ⟨*int*⟩

⟨*np*⟩ ::= **multipleOf** | **maximum** | **minimum**

⟨*str*⟩ ::= **string** | ⟨*sp*⟩ ⟨*str*⟩

⟨*sp*⟩ ::= **maxLength** | **minLength** | **pattern**

⟨*p*⟩ ::= **const** | **enum**

Fig. 1. Simplified grammar of JSON Schema

structural one [2]. This statement allows us to assert that B is a subtype of A in $A <: B$ if for every future from A can be found equal one from B (1). We assert that types predicates are equal if their names and parameters conform. An induction rule is used to specify the subtype with predicates relation (2).

$$\begin{aligned} &\Gamma \vdash A \\ &\Gamma \vdash B \\ &\Gamma \vdash A <: B \end{aligned} \tag{1}$$

$$\frac{\Gamma \vdash AP_1 \quad \Gamma \vdash BP_2 \quad \Gamma \vdash P_1 = P_2}{\Gamma \vdash AP_1 <: BP_2} \tag{2}$$

Finally, we did not change JSON Schema syntax for compatibility with existing software purposes.

A. Algorithm of subtype checking

Our type checking algorithm 1 verifies that every field from the type A is equal to the same one from the type B. Record $type.p$ returns all predicates from the type $type$. Code $type2[field]$ takes from $type2$ subfield with the name $field$ and code $type2.f$ takes all fields from $type2$. The algorithm does not try to analyse predicates, it just checks identity of the name and the parameter. Types of the JSON Schema object are checking recursively. List of subtype required fields must be equal to the parent type one.

III. Description of contract concept

We introduce the concept of a contract to describe communication between services. Service contract is an analogue of communication specification which describes one remote call or one session of information transfer. List of contracts forms

Algorithm 1 Type checking

Require: $type1$, $type2$
  $subtype \leftarrow true$;
  if $type1$ is scalar **then**
    $subtype \leftarrow type1! = type2 || type1.p! = type2.p$;
  else {$type1$ is object}
    for all $type1.f$ **do**
      $subtype \leftarrow subtype \&\& self(type1.f, type2[type1.f])$;
    end for
  end if

regular communication protocol (like OpenAPI or WSDL) if every item of the list is provided by the same service or the same endpoint.

Interoperation of services divides into two categories: a synchronous and nonsynchronous one. The synchronous communication (RPC, REST) requires a protocol to define the way of call, the way of response and optionally an error definition. Custom protocols can specify complex sequences of data units passing to inter-service channel. The nonsynchronous one (event-driven design) requires a protocol to define only type of transmitting data.

In order to level differences between the methods we define contract as a sequence of message types. Thus HTTP call would be a chain of two messages while an event would be the chain consisting of the one element. A contract also contains:

- an endpoint of service which provides contract realization (provider)
- an address to check the contract provider urgency
- a direction of every chain unit - is the message incoming or outgoing for provider

In opposite to provider of contract, the user one claims that a service needs any provider to work correctly. User contract has the same format as the provider one but does not specify endpoint. User contracts ensure that the system's services have all dependencies and work correctly.

User contract is compatible to provider contract if the chains of the first one occur to be subtypes of the second one. This means that if $A <: B$ is valid judgement than provider takes type A at input when client can call it with type B. The provider service must be ready for input data with type B and must process it like data with type A.

A. Description of conceptual contract discovery service

Services must declare their requirements themselves because they contain all related API information. There are two targets for pushing declarations:

- all other services (e.g. broadcast notification)
- central service delegated to manage contracts

Notification of all other services requires broadcast messaging and storing information about the whole system in each one. Moreover, broadcast notification would require implementation of type and contract checking in every service. Therefore central control is preferable.

Services which collect information about system components, provide their addresses and watch for their state are called “service discovery”. Since our tool manages contract providers we call it “contract discovery” service. Prospective realization must have following features:

1) register contract provider
2) register contract user
3) watch for providers and users to be alive
4) deliver on demand information about contract providers for contract users
5) verify that all dependencies are resolved and show dependency problems
6) warn after disabling all providers of the contract that is still used

Providers send information to the service at their startup moment or at their deploy moment. Users get their dependencies also at the start by registering their dependencies or by separate call.

IV. Realization and testing

We implemented the first version of the contract discovery service as a proof-of-concept PHP daemon built on top of ReactPHP [6]. The daemon was used within a test suite containing stub services. After proving the idea we made the second realization with Golang. Service implements all requirements and all described functions. Daemon registers contract providers and users, performs regular alive checks and type checking.

We used described service for managing dependencies and for type checking in existing event-driven system. Services in this system register their contracts at their start. They also gain their own requirements via contract discovery. While services use message broker and do not expect any result of the call all registered contracts consist of no more than one schema. Users obtain routing keys for dispatching messages from matched provider contracts.

Though the proposed approach does not suppose improvement of some specific algorithm or data passing technique we have no ability to present any numeric metrics. However, after registering automatization had been made we noticed that the process of adding new services to the system became easier. Advances that we found are:

- inter-service integration became easier as the result of inter-service strong typing - service would not start while dependencies are not resolved
- contract-first development makes positive influence on service building speed
- developers do not need to keep track of the service dependencies in configuration

We also noticed several complicities:

- maintenance of all types consistency is complicated - there is no one place to store all actual contracts. Contract discovery stores only registered at present time items.
- lack of information about actual data routes
- all system depends on the central component
- since contract discovery checks only online services it does not provide real static type system

V. Conclusion and the future work

As the result we have replaced direct static services linking with detection of the most suitable contract provider. This kind of interaction allows us to ensure that enabled service would work correctly and have all required dependencies. Also usage of strong polymorphic typing allows us to ensure that APIs of interacting services are compatible. Contract discovery service ensures that a system does not have any dependency problems at the moment.

From the other side presented approach sophisticates control over the current interaction of the system components. It also does not provide real static type checking for the communication of the elements.

Finally we did not gain the main goal: we did not strongly increase stability of the system.

Therefore we have new ideas on how to provide strong control over the services interaction. We suggest to specify all data types in a single file or project with a description of whole services communication design. Moreover such definition is expected to resemble a source code on any functional programming language and can also introduce instructions for deploying services. We expect that such source code would be assembled into container configuration files and the translator would perform static type checking. The concept that we are developing now recalls behavioural and session types [3]. Replacing dynamic contract discovery with service definition compiler would save listed advantages and decrease described disadvantages.

References

[1] A. Wright, H. Andrews, G. Luff “JSON Schema Validation: A Vocabulary for Structural Validation of JSON”, http://json-schema.org/latest/json-schema-validation.html
[2] B. Pierce “Types and Programming Languages”, London: MIT Press, 2002, pp. 251-254.
[3] H. Kohei, V. T. Vasconcelos, M. Kubo “Language primitives and type discipline for structured communication-based programming” Programming Languages and Systems, Lecture Notes in Computer Science, vol 1381, Springer, Berlin, 1998
[4] L. Sun, H. Dong, F. Hussain, O. Hussain, E. Chang “Cloud service selection: State-of-the-art and future research directions” Journal of Network and Computer Applications, Austria, October 2014.
[5] N. Dragoni et al “Microservices: Yesterday, Today, and Tomorrow” Present and Ulterior Software Engineering, pp. 195-216, Springer, Cham, 2017
[6] Contract checker, http://github.com/tariel-x/cc
[7] OpenAPI Specification, https://github.com/OAI/OpenAPI-Specification/blob/master/versions/3.0.1.md
[8] R. Rodger “The tao of microservices”, Manning publications, 2017, unpublished, pp. 17-19.
[9] Web Services Description Language (WSDL), https://www.w3.org/TR/wsdl

Position Papers of the Federated Conference on Computer Science and Information Systems pp. 109–114
DOI: 10.15439/2018F255
ISSN 2300-5963 ACSIS, Vol. 16

Acoustic Model Training, using Kaldi, for Automatic Whispery Speech Recognition

Piotr Kozierski[1,2)], Talar Sadalla[2)], Szymon Drgas[1)], Adam Dąbrowski[1)], Joanna Ziętkiewicz[2)], Wojciech Giernacki[2)]
Poznan University of Technology, Piotrowo street 3a, 60-965 Poznan, Poland[1,2)]
Faculty of Computing, Institute of Automation and Robotics, Division of Signal Processing and Electronic Systems[1)]
Faculty of Electrical Engineering, Institute of Control, Robotics and Information Engineering, Division of Control and Robotics[2)]
Email: piotr.kozierski@gmail.com, szymon.drgas@put.poznan.pl

***Abstract*—The article presents research on the automatic whispery speech recognition. The main task was to find dependences between a number of triphone classes (number of leaves in decision tree) and the total number of Gaussian distributions and therefore, to determine optimal values, for which the quality of speech recognition is best. Moreover, it was found, how these dependences differ between normal and whispery speech, what was not done earlier, and this is the innovative part of this work. Based on the performed experiments and obtained results one can say that the number of triphone classes (number of leaves) for whispered speech should be significantly lower than for normal speech.**

I. Introduction

WHISPERS are relatively rarely used in comparison to normal speech. Usually people whisper in specific environment or during private communication [1]. However, for persons after laryngectomy operation, the whispered speech is the only way to communicate with others without special prosthesis [2].

The largest companies (such as Microsoft or Apple) are interested in whispered speech recognition [3]-[4], and also in military domain one of research directions is focused on Automatic Speech Recognition (ASR) systems [5]-[6].

One can find studies on ASR systems for whispered speech [7], even a whispering speaker identification [8]; however, there is still very little research in this area. And even if some studies are provided about whispered speech, very small corpora are used (in latter two references corpora contain less than 500 sentences in sum).

In this paper the authors were focused on the acoustic model training. The most common approach, in which Gaussian Mixture Models (GMMs) are used for Probability Density Functions (PDFs) of features vector values modeling, was taken into account; however, one can find also other approaches, such as Decision Tree-based Acoustic Models (DTAM), in which decision trees are used instead of of GMMs [9].

This work was not supported by any organization

The research task of this paper was to find the optimal value of Gaussian distributions for the given number of leaves, and the optimal number of triphone classes for given number of Gaussian distributions simultaneously. Moreover, the differences between normal and whispery speech were investigated, because such research has not been performed before.

In the second section, one can find a description of software which was used during studies. Information about operation principle of automatic speech recognition are given in Section III. Section IV contains details about speech corpus, which was used in research. In fifth section, one can find description of quality index and obtained results. In the last section, drawn conclusions are presented.

II. Used Software – Kaldi

During studies the Kaldi toolkit [10] was used, which contains scripts and programs for speech recognition task. This software is available under Apache v2.0 license, is easy to change and is still being developed. In Kaldi, two external libraries are used, i.e. BLAS/LAPACK for linear algebra calculations (library is available on the website www.netlib.org) and OpenFST [11]. The latter is used due to the fact that in Kaldi Finite State Transducers (FSTs) are used as representation of most of data [12].

Moreover, SRILM package [13] and Sequitur [14] programs were used. The first one was used for Language Model (LM) preparation (including Witten-Bell smoothing [15]) – it contains information, how .

The second one was used for graphemes-to-phonemes (G2P) conversion – the tool is language independent. The studies were performed for Polish, hence the G2P model was trained (based on the International Phonetic Alphabet – IPA – pronunciation available in Wiktionary). Extended SAMPA notation [16]-[17] with 39 phonemes was used instead of SAMPA notation with 37 phonemes [18]. For comparison, there are 55 phonemes in Russian and 49 phonemes in American English [19].

SAMPA (Speech Assessment Methods Phonetic Alphabet) is a standard for describing specific phonemes

©2018, PTI

(and allophones) by signs from 7-bit ASCII. In this approach it is assumed that there are 37 phonemes (sounds) in Polish.

Extended SAMPA notation was proposed by authors to improve speech recognition of Polish, hence that standard was used in performed research. Differences between SAMPA and Extended SAMPA notations are presented in Table I (for whole list of Polish phonemes see [16] or [20]).

III. Automatic Speech Recognition

The main task of ASR systems is to decode the most probable word sequence (or only word, if isolated words are recognized instead of continuous speech), based on the audio signal with speech. The audio signal is divided into very short (usually 16-25 ms [21], but 25 ms default value was used) overlapping parts (the frame shift default value was equal to 10 ms), so called frames. From each frame, a feature vector is obtained – for Mel Frequency Cepstral Coefficients (MFCC) it is usually 13 values (twelve based on windowing, FFT, Mel scaling and DCT transformations and 13-th – signal energy). Next, first and second derivatives are calculated (Δ+ΔΔ) obtaining 39 MFCC features.

Each feature vector is associated with *i*-th signal frame and is treated as observation o_i. The whole recording is a sequence of such observations $O=\{o_1, o_2, ..., o_M\}$. Among all possible utterances (word sequences) w one must find the most likely sentence, and it can be written as

$$\hat{w} = \arg\max_w \left(p(w \mid O)\right) = \arg\max_w \left(\frac{p(w)p(O \mid w)}{p(O)}\right) = \\ = \arg\max_w \left(p(w)p(O \mid w)\right), \quad (1)$$

where $p(w|O)$ is a posterior PDF, which must be maximized, $p(w)$ is the prior PDF, which informs, what is the probability that sentence w occurs, $p(O|w)$ is the conditional probability that word sequence w occurs for observations O, and $\hat{w}$ is the most probably words sequence.

The whole ASR model is constructed as the Hidden Markov Model (HMM), in which each state is associated with some phoneme (or triphone). Based on the subsequent observations, ASR system should estimate sequence of these states, and this can be solved using the Viterbi algorithm [22].

The prior probability $p(w)$ is represented by the LM in ASR model. Similarly, Acoustic Model (AM) represents probability $p(O|w)$. Therefore, one can see that appropriate LM and AM are fundamental for good ASR working. Language model contains information about "word connections", one can say "grammar" of specific language. Very helpful may be corpus which contains a huge number of works (articles, novels, poems, blog entries, etc.), e.g. [23]. However, in practice, only utterances available in speech corpus are taken into account.

Preparation of LM is relatively fast. Much more difficult is AM training. Acoustic model is also constructed in HMM form; hence, transition probabilities are calculated during training, but also probabilistic distributions associated with specific states (triphones) must be estimated. These distributions are usually represented by the Gaussian mixture models; however, one can find also other approaches [9]. GMM is a distribution which is created from the combination (addition) of two or more Gaussian distributions.

The all states (associated with triphones – three consecutive phonemes) are clustered and not all possible triphones are modeled. It does not mean that different states are treated as the same, but at this processing stage few (or dozen) triphones from the same class are unrecognizable. Based on the lexicon or LM a specific triphone will be recognized later. In Kaldi toolkit this classification is done by the Decision Tree (DT), i.e., one specific class is chosen based on the series of comparisons (which parameters are compared and boundary values are chosen during AM training).

The size of DT (number of classes) and the whole number of Gaussian distributions (each GMM, which described one class, is composed of few or dozen ones) are the main two parameters in AM training in Kaldi.

TABLE I.
DIFFERENCES BETWEEN SAMPA AND EXTENDED SAMPA NOTATIONS [16]

Word in Polish	SAMPA		Ext. SAMPA	
	notation	transcription	notation	transcription
typ	I	t **I** p	y	t **y** p
gęś	e~	g **e~** s'	–	–
wąs	o~	v **o~** s	–	–
kat	k	**k** a t	k	**k** a t
gen	g	**g** e n	g	**g** e n
kiedy	–	–	c	**c** j e d y
giełda	–	–	J	**J** j e w d a
cyk	ts	**ts** I k	t^s	**t^s** y k
dzwon	dz	**dz** v o n	d^z	**d^z** v o n
czyn	tS	**tS** I n	t^S	**t^S** y n
dżem	dZ	**dZ** e m	d^Z	**d^Z** e m
ćma	ts'	**ts'** m a	t^s'	**t^s'** m a
dźwig	dz'	**dz'** v i k	d^z'	**d^z'** v i k
ciąża	–	–	w~	t^s' o **w~** Z a
więź	–	–	j~	v j e **j~** s'

IV. Speech Corpus

The authors were focused on the Automatic Whispery Speech Recognition (AWSR), and hence a specific speech corpus was needed. However there are very few of databases with whispery speech. One of such is CHAINS corpus [24] which contains about 1200 sentences in whisper. The second available corpus – Audiovisual (AVW) [25] contains over 1300 whispered sentences. Unfortunately, both are too small for AWSR research.

The largest corpus with whispery speech, which was used in [26], contains about 14,000 whispered sentences and theoretically could be used in AWSR task; however, this is Japanese corpus, and in such languages like Mandarin or Japanese it is important to model the accent, which can change the word meaning [27].

Due to the lack of required database, the authors decided to prepare corpus with Polish (normal and whispered) speech (see Table II, and for comparison also older version of corpus is described in Table III). The sentences come mainly from Andersen's fairy tales (like "The Toad", "The Nightingale", "The Ugly Duckling"); however, one can find also fragments from Grimm brothers' fairy tales.

All utterances were recorded in 48 kSps sampling rate and 16-bit quantization depth. Every speaker (there are over 50 different speakers) recorded sentences on his/her own device, so the recordings quality widely vary between speakers.

V. Experiments and Results

All experiments were repeated two times for different test speakers group (the choice of testing speakers was widely described in [28]) and presented results are mean values. Each time ASR system was trained on recordings from all speakers (except the testing ones).

The quality of speech recognition is described by the Word Error Rate (WER) index

$$\mathrm{WER} = \frac{\mathrm{Del} + \mathrm{Ins} + \mathrm{Subs}}{N_{utt}} \cdot 100\%, \qquad (2)$$

TABLE III.
PROPERTIES OF THE USED SPEECH CORPUS (2ND VERSION)

Property	Normal speech	Whispered speech
Number of sentences	9,522	8,753
Number of words	108,038	95,305
Number of different words	5,094	4,763
Total recordings length	988.5 min (16.5 h)	942.9 min (15.7 h)
Number of speakers	56	

TABLE III.
PROPERTIES OF THE OLD SPEECH CORPUS (1ST VERSION) – ALL THESE RECORDING ARE ALSO CONTAINED IN 2ND VERSION

Property	Normal speech	Whispered speech
Number of sentences	5,935	5,411
Number of words	61,964	53,335
Number of different words	3,556	3,427
Total recordings length	547.5 min (9.1 h)	548.5 min (9.1 h)
Number of speakers	33	

where Del is the number of deletions (cases where word from reference is not present in output sentence), Ins is the number of insertions (cases where word in recognized sentence does not occur in reference), Subs is the number of substitutions (cases where one word from reference is confused with another one from output sentence), and N_{utt} is the number of words in a reference sentence. The sum in numerator is a minimum edit distance on words between obtained output from ASR system and the reference utterance [12].

The lexicon was extended from 5,000 words, which occur in speech corpus, to 50,000 words to obtain large vocabulary ASR system (based on the classification in [29]). It was done by adding new sentences during LM creation. Moreover, the differences in results and directions of changes are better visible for higher WER level.

During research the AM training path mono → tri1 → tri2a (designations from Kaldi) was used. The choice was dictated by the previous studies [30], where it was concluded that this training path provides satisfying speech recognition quality and quite short training computation time simultaneously.

The experiments were performed for different number of leaves and number of Gaussian distributions in AM training. In each case values were the same for tri1 and tri2a steps – it was caused by the preliminary research. In both tri1 and tri2a steps the same scripts are used, and the only difference between tri1 and tri2a is their order. The authors tried to add third step tri3a (again – the same script run third time) and the impact of leaves number and Gauss number in first two runs on speech recognition quality was studied (in all cases numbers of leaves and Gauss were the same for third – tri3a – step).

The results of preliminary research showed that only parameters in the last step have significant influence on the speech recognition quality. Based on this, it was decided to set the same parameter values for both tri1 and tri2a steps. Obtained results are presented in Fig. 1-4.

VI. Conclusions

Based on the obtained results presented in Fig. 1-2 one can see that with increasing the total number of Gaussian distributions in AM model the speech recognition quality improves. This is caused by better modeling of the probability distribution of phones' features vectors (more sum components in GMMs).

However, for the number of leaves (number of different triphones classes – see Fig. 3-4) one can see that there is an optimal value, for which WER index is the lowest. For normal speech, number of Gaussian distributions should be increased together with the number of classes. And the default values (2,000 leaves and 11,000 Gaussian distributions) are quite good for normal speech (if one wants to improve acoustic model, and quality of ASR system at the same time, both values should be increased).

On the contrary, in whispery speech one can see the difference – the optimal value of decision tree leaves is lower in all cases. The default value (2000) should be decreased, or the number of Gaussian distributions should be greatly increased; however, in all cases, further increase of class number has a negative impact on speech recognition quality. This is probably caused by the fact that in whispered speech there is no more than 2,000 different triphones (theoretically it could be $39^3 \approx 60{,}000$; however, in whispers many phonemes sounds the same).

In the future research the authors plan to use neural networks for whispery speech recognition, and also sharing of prepared speech corpus is planned.

REFERENCES

[1] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 10, pp. 2448–2458, 2010.

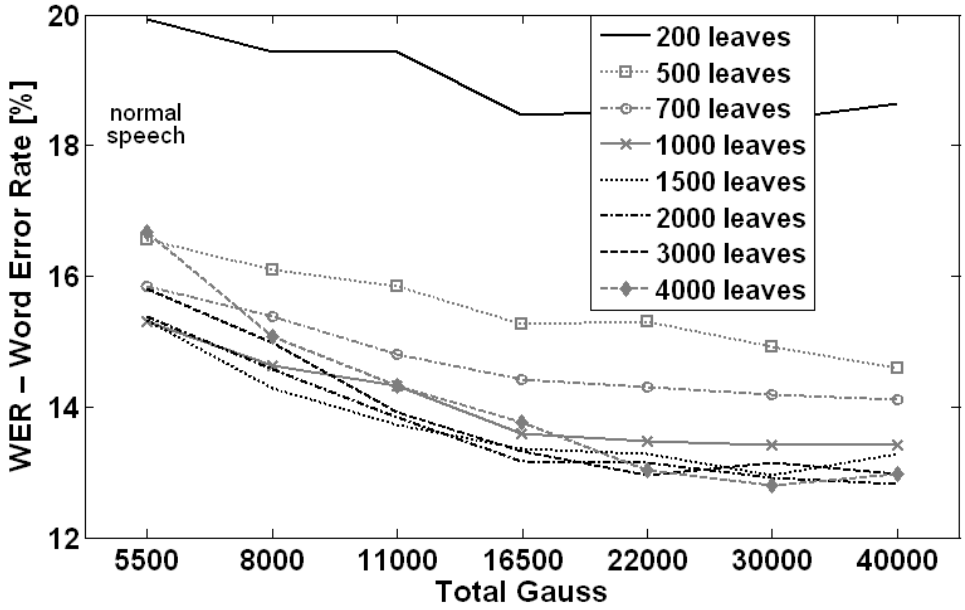

Fig. 1 Quality of speech recognition in function of total Gaussian distributions number for normal speech

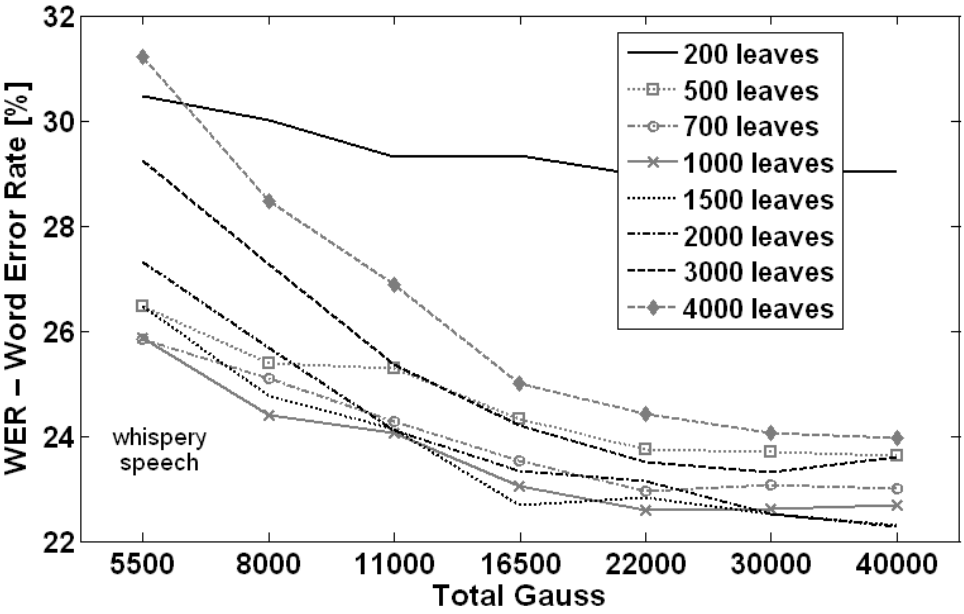

Fig. 2 Quality of speech recognition in function of total Gaussian distributions number for whispery speech

[2] H. F. Nijdam, A. A. Annyas, H. K. Schutte, and H. Leever, "A new prosthesis for voice rehabilitation after laryngectomy," *Archives of Otorhinolaryngology*, vol. 237, no. 1, pp. 27–33, 1982.

[3] X. Huang, A. Acero, F. Alleva, M. Y. Hwang, L. Jiang, and M. Mahajan, "Microsoft Windows highly intelligent speech recognizer: Whisper," in *Acoustics, Speech, and Signal Processing, 1995 International Conference on (ICASSP-95)*, vol. 1, pp. 93–96.

[4] T. J. Raitio, M. J. Hunt, H. B. Richards, and M. Chinthakunta, "Digital assistant providing whispered speech," U.S. Patent 15/266,932, December 14, 2017.

[5] D. T. Williamson, M. H. Draper, G. L. Calhoun, and T. P. Barry, "Commercial speech recognition technology in the military domain: Results of two recent research efforts," *International Journal of Speech Technology*, vol. 8, no. 1, pp. 9–16, 2005.

[6] S. Pigeon, C. Swail, E. Geoffrois, G. Bruckner, D. Van Leeuwen, C. Teixeira, et al., *Use of speech and language technology in military environments*, Montreal, Canada, North Atlantic Treaty Organization, 2005.

[7] S. C. S. Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *ICASSP*, March 2005, pp. 1009–1012.

[8] Q. Jin, S. C. S. Jou, and T. Schultz, "Whispering speaker identification," in *Multimedia and Expo, 2007 IEEE International Conference on*, pp. 1027–1030.

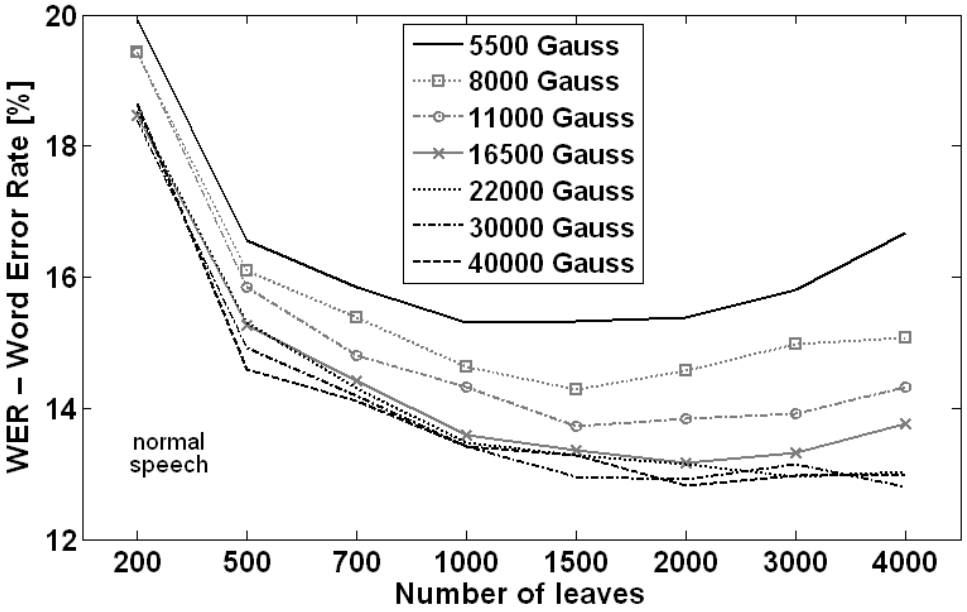

Fig. 3 Quality of speech recognition in function of leaves number for normal speech

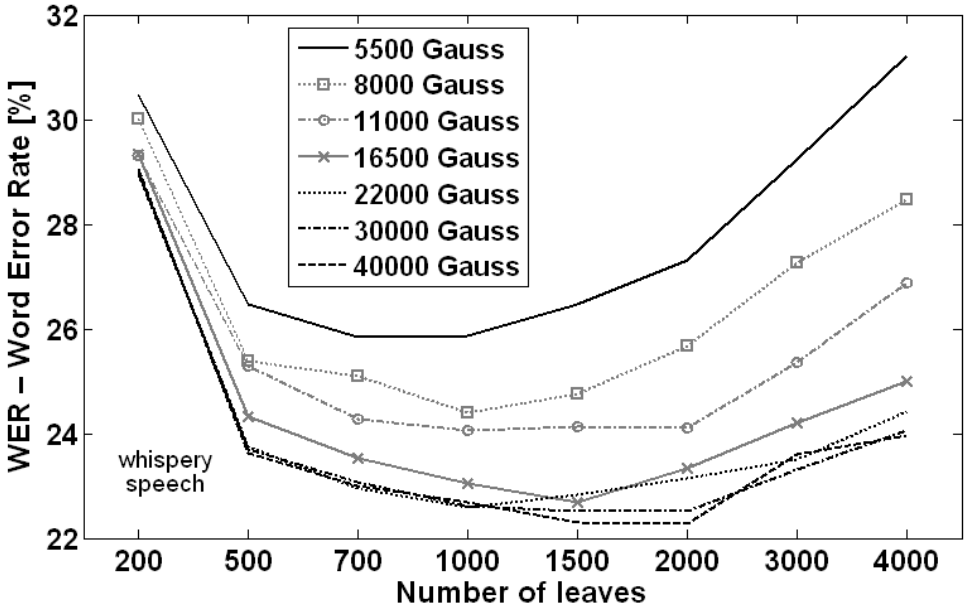

Fig. 4 Quality of speech recognition in function of leaves number for whispery speech

[9] M. Akamine, and J. Ajmera, "Decision tree-based acoustic models for speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2012, art. no. 10, p. 8, 2012.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, No. EPFL-CONF-192584, 2011.

[11] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *Implementation and Application of Automata*, J. Holub and J. Žďárek, Ed. Berlin: Springer Heidelberg, 2007, pp. 11–23.

[12] O. Platek, "Speech recognition using KALDI," M.S. thesis, Inst. Form. Appl. Ling., Charles Univ., Prague, Czech Republic, 2014.

[13] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Proc. Intl. Conf. Spoken Language Processing (INTERSPEECH)*, Denver, Colorado, September 2002, pp. 901–904.

[14] M. Bisani, and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[15] I. H. Witten, and T. C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1085–1094, 1991.

[16] G. Demenko, M. Wypych, and E. Baranowska, "Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis," *Speech and Language Technology*, vol. 7, pp. 79–97, 2003.

[17] M. Wypych, E. Baranowska, and G. Demenko, "A grapheme-to-phoneme transcription algorithm based on the SAMPA alphabet extension for the Polish language," in *Phonetic Sciences, 15th International Congress of (ICPhS)*, Barcelona, August 2003, pp. 2601–2604.

[18] P. Kłosowski, "Improving speech processing based on phonetics and phonology of Polish language," *Przeglad Elektrotechniczny*, vol. 89, no. 8, pp. 303–307, 2013.

[19] A. Karpov, K. Markov, I. Kipyatkova, D. Vazhenina, and A. Ronzhin, "Large vocabulary Russian speech recognition using syntactico-statistical language modeling," *Speech Communication*, vol. 56, pp. 213–228, 2014.

[20] P. Kozierski, T. Sadalla, S. Drgas, A. Dąbrowski, "Allophones in automatic whispery speech recognition," *in Methods and Models in Automation and Robotics (MMAR), 21st International Conference on*, 2016, pp. 811-815. DOI: 10.1109/MMAR.2016.7575241

[21] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, "Design and evaluation of a smart home voice interface for the elderly: Acceptability and objection aspects," *Personal and Ubiquitous Computing*, vol. 17, no. 1, pp. 127–144, 2013.

[22] K. Szostek, "Optimization of HMM models and their usage in speech recognition (in Polish)," *Elektrotechnika i Elektronika*, vol. 24, no. 2, pp. 172–182, 2005.

[23] B. Lewandowska-Tomaszczyk, M. Bańko, R. L. Górski, P. Pęzik, and A. Przepiórkowski, *National corpus of Polish language (in Polish)*, Warszawa: Wydawnictwo Naukowe PWN, 2012.

[24] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The Chains corpus: Characterizing individual speakers," *in Proc. of SPECOM*, vol. 6, 2006, pp. 431–435.

[25] T. Tran, S. Mariooryad, and C. Busso, "Audiovisual corpus to analyze whisper speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 8101–8105.

[26] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.

[27] C. Huang, E. Chang, J. Zhou, K. and F. Lee, "Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition," in *INTERSPEECH*, October 2000, pp. 818–821.

[28] P. Kozierski, T. Sadalla, S. Drgas, A. Dąbrowski, and J. Zietkiewicz, "The impact of vocabulary size and language model order on the Polish whispery speech recognition," in *Methods and Models in Automation and Robotics (MMAR), 22nd International Conference on*, 2017, pp. 616–621. DOI: 10.1109/MMAR.2017.8046899

[29] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol 56, pp. 85–100, 2014.

[30] P. Kozierski, T. Sadalla, S. Drgas, A. Dąbrowski, and D. Horla, "Kaldi toolkit in Polish whispery speech recognition," *Przeglad Elektrotechniczny*, vol. 92, no. 11, pp. 301–304, 2016. DOI: 10.15199/48.2016.11.70

Author Index

Aleithe, Michael 51

Baumann, Tommy 93
Biele, Cezary 83

Dąbrowski, Adam 109
Dimitrieski, Vladimir 73
Drgas, Szymon 109

Esche, Marko 43

Garncarek, Łukasz 9
Gerasimov, Nikita 105
Giernacki, Wojciech 109

Hagihara, Atsushi 3
Hubl, Marvin 51

Jakubik, Jan 19

Kajiwara, Yusuke 3
Kordić, Slavica 73
Koval, Martin 43
Kowalski, Jarosław 83
Kozierski, Piotr 109

Luković, Ivan 73

Nurzyński, Marek 23

Pfitzinger, Bernd 93
Podlodowski, Łukasz 23
Prus-Zajączkowski, Bartłomiej 9

Rot, Artur 63
Roziewski, Szymon 23

Sadalla, Talar 109
Shimakawa, Hiromitsu 3
Skowron, Philipp 51
Sobińska, Małgorzata 63
Stanisławek, Tomasz 9

Terzic, Branko 73
Toczyski, Piotr 83
Toro, Federico Grasso 43

Víta, Martin 33

Wiktorski, Tomasz 99
Wróblewska, Anna 9

Zdrodowska, Aldona 83
Ziętkiewicz, Joanna 109