Lecture Notes in Networks and Systems 990

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Álvaro Rocha · Hojjat Adeli ·
Gintautas Dzemyda · Fernando Moreira ·
Aneta Poniszewska-Marańda
Editors

Good Practices and New Perspectives in Information Systems and Technologies

WorldCIST 2024, Volume 6

Springer

Editors
Álvaro Rocha
ISEG
Universidade de Lisboa
Lisbon, Portugal

Hojjat Adeli
College of Engineering
The Ohio State University
Columbus, OH, USA

Gintautas Dzemyda
Institute of Data Science and Digital Technologies
Vilnius University
Vilnius, Lithuania

Fernando Moreira
DCT
Universidade Portucalense
Porto, Portugal

Aneta Poniszewska-Marańda
Institute of Information Technology
Lodz University of Technology
Łódz, Poland

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-031-60327-3 ISBN 978-3-031-60328-0 (eBook)
https://doi.org/10.1007/978-3-031-60328-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book contains a selection of papers accepted for presentation and discussion at the 2024 World Conference on Information Systems and Technologies (WorldCIST'24). This conference had the scientific support of the Lodz University of Technology, Information and Technology Management Association (ITMA), IEEE Systems, Man, and Cybernetics Society (IEEE SMC), Iberian Association for Information Systems and Technologies (AISTI), and Global Institute for IT Management (GIIM). It took place in Lodz city, Poland, 26–28 March 2024.

The World Conference on Information Systems and Technologies (WorldCIST) is a global forum for researchers and practitioners to present and discuss recent results and innovations, current trends, professional experiences, and challenges of modern Information Systems and Technologies research, technological development, and applications. One of its main aims is to strengthen the drive toward a holistic symbiosis between academy, society, and industry. WorldCIST'23 is built on the successes of: WorldCIST'13 held at Olhão, Algarve, Portugal; WorldCIST'14 held at Funchal, Madeira, Portugal; WorldCIST'15 held at São Miguel, Azores, Portugal; WorldCIST'16 held at Recife, Pernambuco, Brazil; WorldCIST'17 held at Porto Santo, Madeira, Portugal; WorldCIST'18 held at Naples, Italy; WorldCIST'19 held at La Toja, Spain; WorldCIST'20 held at Budva, Montenegro; WorldCIST'21 held at Terceira Island, Portugal; WorldCIST'22 held at Budva, Montenegro; and WorldCIST'23, which took place at Pisa, Italy.

The Program Committee of WorldCIST'24 was composed of a multidisciplinary group of 328 experts and those who are intimately concerned with Information Systems and Technologies. They have had the responsibility for evaluating, in a 'blind review' process, the papers received for each of the main themes proposed for the conference: A) Information and Knowledge Management; B) Organizational Models and Information Systems; C) Software and Systems Modeling; D) Software Systems, Architectures, Applications and Tools; E) Multimedia Systems and Applications; F) Computer Networks, Mobility and Pervasive Systems; G) Intelligent and Decision Support Systems; H) Big Data Analytics and Applications; I) Human-Computer Interaction; J) Ethics, Computers & Security; K) Health Informatics; L) Information Technologies in Education; M) Information Technologies in Radiocommunications; and N) Technologies for Biomedical Applications.

The conference also included workshop sessions taking place in parallel with the conference ones. Workshop sessions covered themes such as: ICT for Auditing & Accounting; Open Learning and Inclusive Education Through Information and Communication Technology; Digital Marketing and Communication, Technologies, and Applications; Advances in Deep Learning Methods and Evolutionary Computing for Health Care; Data Mining and Machine Learning in Smart Cities: The role of the technologies in the research of the migrations; Artificial Intelligence Models and Artifacts for Business Intelligence Applications; AI in Education; Environmental data analytics; Forest-Inspired

Computational Intelligence Methods and Applications; Railway Operations, Modeling and Safety; Technology Management in the Electrical Generation Industry: Capacity Building through Knowledge, Resources and Networks; Data Privacy and Protection in Modern Technologies; Strategies and Challenges in Modern NLP: From Argumentation to Ethical Deployment; and Enabling Software Engineering Practices Via Last Development Trends.

WorldCIST'24 and its workshops received about 400 contributions from 47 countries around the world. The papers accepted for oral presentation and discussion at the conference are published by Springer (this book) in four volumes and will be submitted for indexing by WoS, Scopus, EI-Compendex, DBLP, and/or Google Scholar, among others. Extended versions of selected best papers will be published in special or regular issues of leading and relevant journals, mainly JCR/SCI/SSCI and Scopus/EI-Compendex indexed journals.

We acknowledge all of those that contributed to the staging of WorldCIST'24 (authors, committees, workshop organizers, and sponsors). We deeply appreciate their involvement and support that was crucial for the success of WorldCIST'24.

March 2024

Álvaro Rocha
Hojjat Adeli
Gintautas Dzemyda
Fernando Moreira
Aneta Poniszewska-Marańda

Organization

Conference

Honorary Chair

Hojjat Adeli	The Ohio State University, USA

General Chair

Álvaro Rocha	ISEG, University of Lisbon, Portugal

Co-chairs

Gintautas Dzemyda	Vilnius University, Lithuania
Sandra Costanzo	University of Calabria, Italy

Workshops Chair

Fernando Moreira	Portucalense University, Portugal

Local Organizing Committee

Bożena Borowska	Lodz University of Technology, Poland
Łukasz Chomątek	Lodz University of Technology, Poland
Joanna Ochelska-Mierzejewska	Lodz University of Technology, Poland
Aneta Poniszewska-Marańda	Lodz University of Technology, Poland

Advisory Committee

Ana Maria Correia (Chair)	University of Sheffield, UK
Brandon Randolph-Seng	Texas A&M University, USA

Chris Kimble	KEDGE Business School & MRM, UM2, Montpellier, France
Damian Niwiński	University of Warsaw, Poland
Eugene Spafford	Purdue University, USA
Florin Gheorghe Filip	Romanian Academy, Romania
Janusz Kacprzyk	Polish Academy of Sciences, Poland
João Tavares	University of Porto, Portugal
Jon Hall	The Open University, UK
John MacIntyre	University of Sunderland, UK
Karl Stroetmann	Empirica Communication & Technology Research, Germany
Marjan Mernik	University of Maribor, Slovenia
Miguel-Angel Sicilia	University of Alcalá, Spain
Mirjana Ivanovic	University of Novi Sad, Serbia
Paulo Novais	University of Minho, Portugal
Sami Habib	Kuwait University, Kuwait
Wim Van Grembergen	University of Antwerp, Belgium

Program Committee Co-chairs

Adam Wojciechowski	Lodz University of Technology, Poland
Aneta Poniszewska-Marańda	Lodz University of Technology, Poland

Program Committee

Abderrahmane Ez-zahout	Mohammed V University, Morocco
Adriana Peña Pérez Negrón	Universidad de Guadalajara, Mexico
Adriani Besimi	South East European University, North Macedonia
Agostinho Sousa Pinto	Polytechnic of Porto, Portugal
Ahmed El Oualkadi	Abdelmalek Essaadi University, Morocco
Akex Rabasa	University Miguel Hernandez, Spain
Alanio de Lima	UFC, Brazil
Alba Córdoba-Cabús	University of Malaga, Spain
Alberto Freitas	FMUP, University of Porto, Portugal
Aleksandra Labus	University of Belgrade, Serbia
Alessio De Santo	HE-ARC, Switzerland
Alexandru Vulpe	University Politechnica of Bucharest, Romania
Ali Idri	ENSIAS, University Mohamed V, Morocco
Alicia García-Holgado	University of Salamanca, Spain

Almir Souza Silva Neto	IFMA, Brazil
Álvaro López-Martín	University of Malaga, Spain
Amélia Badica	Universiti of Craiova, Romania
Amélia Cristina Ferreira Silva	Polytechnic of Porto, Portugal
Amit Shelef	Sapir Academic College, Israel
Ana Carla Amaro	Universidade de Aveiro, Portugal
Ana Dinis	Polytechnic of Cávado and Ave, Portugal
Ana Isabel Martins	University of Aveiro, Portugal
Anabela Gomes	University of Coimbra, Portugal
Anacleto Correia	CINAV, Portugal
Andrew Brosnan	University College Cork, Ireland
Andjela Draganic	University of Montenegro, Montenegro
Aneta Polewko-Klim	University of Białystok, Institute of Informatics, Poland
Aneta Poniszewska-Maranda	Lodz University of Technology, Poland
Angeles Quezada	Instituto Tecnologico de Tijuana, Mexico
Anis Tissaoui	University of Jendouba, Tunisia
Ankur Singh Bist	KIET, India
Ann Svensson	University West, Sweden
Anna Gawrońska	Poznański Instytut Technologiczny, Poland
Antoni Oliver	University of the Balearic Islands, Spain
Antonio Jiménez-Martín	Universidad Politécnica de Madrid, Spain
Aroon Abbu	Bell and Howell, USA
Arslan Enikeev	Kazan Federal University, Russia
Beatriz Berrios Aguayo	University of Jaen, Spain
Benedita Malheiro	Polytechnic of Porto, ISEP, Portugal
Bertil Marques	Polytechnic of Porto, ISEP, Portugal
Boris Shishkov	ULSIT/IMI - BAS/IICREST, Bulgaria
Borja Bordel	Universidad Politécnica de Madrid, Spain
Branko Perisic	Faculty of Technical Sciences, Serbia
Bruno F. Gonçalves	Polytechnic of Bragança, Portugal
Carla Pinto	Polytechnic of Porto, ISEP, Portugal
Carlos Balsa	Polytechnic of Bragança, Portugal
Carlos Rompante Cunha	Polytechnic of Bragança, Portugal
Catarina Reis	Polytechnic of Leiria, Portugal
Célio Gonçalo Marques	Polytenic of Tomar, Portugal
Cengiz Acarturk	Middle East Technical University, Turkey
Cesar Collazos	Universidad del Cauca, Colombia
Cristina Gois	Polytechnic University of Coimbra, Portugal
Christophe Guyeux	Universite de Bourgogne Franche Comté, France
Christophe Soares	University Fernando Pessoa, Portugal
Christos Bouras	University of Patras, Greece

Christos Chrysoulas	London South Bank University, UK
Christos Chrysoulas	Edinburgh Napier University, UK
Ciro Martins	University of Aveiro, Portugal
Claudio Sapateiro	Polytechnic of Setúbal, Portugal
Cosmin Striletchi	Technical University of Cluj-Napoca, Romania
Costin Badica	University of Craiova, Romania
Cristian García Bauza	PLADEMA-UNICEN-CONICET, Argentina
Cristina Caridade	Polytechnic of Coimbra, Portugal
Danish Jamil	Malaysia University of Science and Technology, Malaysia
David Cortés-Polo	University of Extremadura, Spain
David Kelly	University College London, UK
Daria Bylieva	Peter the Great St. Petersburg Polytechnic University, Russia
Dayana Spagnuelo	Vrije Universiteit Amsterdam, Netherlands
Dhouha Jaziri	University of Sousse, Tunisia
Dmitry Frolov	HSE University, Russia
Dulce Mourato	ISTEC - Higher Advanced Technologies Institute Lisbon, Portugal
Edita Butrime	Lithuanian University of Health Sciences, Lithuania
Edna Dias Canedo	University of Brasilia, Brazil
Egils Ginters	Riga Technical University, Latvia
Ekaterina Isaeva	Perm State University, Russia
Eliana Leite	University of Minho, Portugal
Enrique Pelaez	ESPOL University, Ecuador
Eriks Sneiders	Stockholm University, Sweden; Esteban Castellanos ESPE, Ecuador
Fatima Azzahra Amazal	Ibn Zohr University, Morocco
Fernando Bobillo	University of Zaragoza, Spain
Fernando Molina-Granja	National University of Chimborazo, Ecuador
Fernando Moreira	Portucalense University, Portugal
Fernando Ribeiro	Polytechnic Castelo Branco, Portugal
Filipe Caldeira	Polytechnic of Viseu, Portugal
Filippo Neri	University of Naples, Italy
Firat Bestepe	Republic of Turkey Ministry of Development, Turkey
Francesco Bianconi	Università degli Studi di Perugia, Italy
Francisco García-Peñalvo	University of Salamanca, Spain
Francisco Valverde	Universidad Central del Ecuador, Ecuador
Frederico Branco	University of Trás-os-Montes e Alto Douro, Portugal
Galim Vakhitov	Kazan Federal University, Russia

Gayo Diallo	University of Bordeaux, France
Gabriel Pestana	Polytechnic Institute of Setubal, Portugal
Gema Bello-Orgaz	Universidad Politecnica de Madrid, Spain
George Suciu	BEIA Consult International, Romania
Ghani Albaali	Princess Sumaya University for Technology, Jordan
Gian Piero Zarri	University Paris-Sorbonne, France
Giovanni Buonanno	University of Calabria, Italy
Gonçalo Paiva Dias	University of Aveiro, Portugal
Goreti Marreiros	ISEP/GECAD, Portugal
Habiba Drias	University of Science and Technology Houari Boumediene, Algeria
Hafed Zarzour	University of Souk Ahras, Algeria
Haji Gul	City University of Science and Information Technology, Pakistan
Hakima Benali Mellah	Cerist, Algeria
Hamid Alasadi	Basra University, Iraq
Hatem Ben Sta	University of Tunis at El Manar, Tunisia
Hector Fernando Gomez Alvarado	Universidad Tecnica de Ambato, Ecuador
Hector Menendez	King's College London, UK
Hélder Gomes	University of Aveiro, Portugal
Helia Guerra	University of the Azores, Portugal
Henrique da Mota Silveira	University of Campinas (UNICAMP), Brazil
Henrique S. Mamede	University Aberta, Portugal
Henrique Vicente	University of Évora, Portugal
Hicham Gueddah	University Mohammed V in Rabat, Morocco
Hing Kai Chan	University of Nottingham Ningbo China, China
Igor Aguilar Alonso	Universidad Nacional Tecnológica de Lima Sur, Peru
Inês Domingues	University of Coimbra, Portugal
Isabel Lopes	Polytechnic of Bragança, Portugal
Isabel Pedrosa	Coimbra Business School - ISCAC, Portugal
Isaías Martins	University of Leon, Spain
Issam Moghrabi	Gulf University for Science and Technology, Kuwait
Ivan Armuelles Voinov	University of Panama, Panama
Ivan Dunđer	University of Zagreb, Croatia
Ivone Amorim	University of Porto, Portugal
Jaime Diaz	University of La Frontera, Chile
Jan Egger	IKIM, Germany
Jan Kubicek	Technical University of Ostrava, Czech Republic
Jeimi Cano	Universidad de los Andes, Colombia

Jesús Gallardo Casero	University of Zaragoza, Spain
Jezreel Mejia	CIMAT, Unidad Zacatecas, Mexico
Jikai Li	The College of New Jersey, USA
Jinzhi Lu	KTH-Royal Institute of Technology, Sweden
Joao Carlos Silva	IPCA, Portugal
João Manuel R. S. Tavares	University of Porto, FEUP, Portugal
João Paulo Pereira	Polytechnic of Bragança, Portugal
João Reis	University of Aveiro, Portugal
João Reis	University of Lisbon, Portugal
João Rodrigues	University of the Algarve, Portugal
João Vidal de Carvalho	Polytechnic of Porto, Portugal
Joaquin Nicolas Ros	University of Murcia, Spain
John W. Castro	University de Atacama, Chile
Jorge Barbosa	Polytechnic of Coimbra, Portugal
Jorge Buele	Technical University of Ambato, Ecuador; Jorge Gomes University of Lisbon, Portugal
Jorge Oliveira e Sá	University of Minho, Portugal
José Braga de Vasconcelos	Universidade Lusófona, Portugal
Jose M. Parente de Oliveira	Aeronautics Institute of Technology, Brazil
José Machado	University of Minho, Portugal
José Paulo Lousado	Polytechnic of Viseu, Portugal
Jose Quiroga	University of Oviedo, Spain
Jose Silvestre Silva	Academia Military, Portugal
Jose Torres	University Fernando Pessoa, Portugal
Juan M. Santos	University of Vigo, Spain
Juan Manuel Carrillo de Gea	University of Murcia, Spain
Juan Pablo Damato	UNCPBA-CONICET, Argentina
Kalinka Kaloyanova	Sofia University, Bulgaria
Kamran Shaukat	The University of Newcastle, Australia
Katerina Zdravkova	University Ss. Cyril and Methodius, North Macedonia
Khawla Tadist	Morocco
Khalid Benali	LORIA - University of Lorraine, France
Khalid Nafil	Mohammed V University in Rabat, Morocco
Korhan Gunel	Adnan Menderes University, Turkey
Krzysztof Wolk	Polish-Japanese Academy of Information Technology, Poland
Kuan Yew Wong	Universiti Teknologi Malaysia (UTM), Malaysia
Kwanghoon Kim	Kyonggi University, South Korea
Laila Cheikhi	Mohammed V University in Rabat, Morocco
Laura Varela-Candamio	Universidade da Coruña, Spain
Laurentiu Boicescu	E.T.T.I. U.P.B., Romania

Lbtissam Abnane	ENSIAS, Morocco
Lia-Anca Hangan	Technical University of Cluj-Napoca, Romania
Ligia Martinez	CECAR, Colombia
Lila Rao-Graham	University of the West Indies, Jamaica
Liliana Ivone Pereira	Polytechnic of Cávado and Ave, Portugal
Łukasz Tomczyk	Pedagogical University of Cracow, Poland
Luis Alvarez Sabucedo	University of Vigo, Spain
Luís Filipe Barbosa	University of Trás-os-Montes e Alto Douro
Luis Mendes Gomes	University of the Azores, Portugal
Luis Pinto Ferreira	Polytechnic of Porto, Portugal
Luis Roseiro	Polytechnic of Coimbra, Portugal
Luis Silva Rodrigues	Polytencic of Porto, Portugal
Mahdieh Zakizadeh	MOP, Iran
Maksim Goman	JKU, Austria
Manal el Bajta	ENSIAS, Morocco
Manuel Antonio Fernández-Villacañas Marín	Technical University of Madrid, Spain
Manuel Ignacio Ayala Chauvin	University Indoamerica, Ecuador
Manuel Silva	Polytechnic of Porto and INESC TEC, Portugal
Manuel Tupia	Pontifical Catholic University of Peru, Peru
Manuel Au-Yong-Oliveira	University of Aveiro, Portugal
Marcelo Mendonça Teixeira	Universidade de Pernambuco, Brazil
Marciele Bernardes	University of Minho, Brazil
Marco Ronchetti	Universita' di Trento, Italy
Mareca María Pilar	Universidad Politécnica de Madrid, Spain
Marek Kvet	Zilinska Univerzita v Ziline, Slovakia
Maria João Ferreira	Universidade Portucalense, Portugal
Maria José Sousa	University of Coimbra, Portugal
María Teresa García-Álvarez	University of A Coruna, Spain
Maria Sokhn	University of Applied Sciences of Western Switzerland, Switzerland
Marijana Despotovic-Zrakic	Faculty Organizational Science, Serbia
Marilio Cardoso	Polytechnic of Porto, Portugal
Mário Antunes	Polytechnic of Leiria & CRACS INESC TEC, Portugal
Marisa Maximiano	Polytechnic Institute of Leiria, Portugal
Marisol Garcia-Valls	Polytechnic University of Valencia, Spain
Maristela Holanda	University of Brasilia, Brazil
Marius Vochin	E.T.T.I. U.P.B., Romania
Martin Henkel	Stockholm University, Sweden
Martín López Nores	University of Vigo, Spain
Martin Zelm	INTEROP-VLab, Belgium

Mazyar Zand	MOP, Iran
Mawloud Mosbah	University 20 Août 1955 of Skikda, Algeria
Michal Adamczak	Poznan School of Logistics, Poland
Michal Kvet	University of Zilina, Slovakia
Miguel Garcia	University of Oviedo, Spain
Mircea Georgescu	Al. I. Cuza University of Iasi, Romania
Mirna Muñoz	Centro de Investigación en Matemáticas A.C., Mexico
Mohamed Hosni	ENSIAS, Morocco
Monica Leba	University of Petrosani, Romania
Nadesda Abbas	UBO, Chile
Narasimha Rao Vajjhala	University of New York Tirana, Tirana
Narjes Benameur	Laboratory of Biophysics and Medical Technologies of Tunis, Tunisia
Natalia Grafeeva	Saint Petersburg University, Russia
Natalia Miloslavskaya	National Research Nuclear University MEPhI, Russia
Naveed Ahmed	University of Sharjah, United Arab Emirates
Neeraj Gupta	KIET group of institutions Ghaziabad, India
Nelson Rocha	University of Aveiro, Portugal
Nikola S. Nikolov	University of Limerick, Ireland
Nicolas de Araujo Moreira	Federal University of Ceara, Brazil
Nikolai Prokopyev	Kazan Federal University, Russia
Niranjan S. K.	JSS Science and Technology University, India
Noemi Emanuela Cazzaniga	Politecnico di Milano, Italy
Noureddine Kerzazi	Polytechnique Montréal, Canada
Nuno Melão	Polytechnic of Viseu, Portugal
Nuno Octávio Fernandes	Polytechnic of Castelo Branco, Portugal
Nuno Pombo	University of Beira Interior, Portugal
Olga Kurasova	Vilnius University, Lithuania
Olimpiu Stoicuta	University of Petrosani, Romania
Patricia Quesado	Polytechnic of Cávado and Ave, Portugal
Patricia Zachman	Universidad Nacional del Chaco Austral, Argentina
Paula Serdeira Azevedo	University of Algarve, Portugal
Paula Dias	Polytechnic of Guarda, Portugal
Paulo Alejandro Quezada Sarmiento	University of the Basque Country, Spain
Paulo Maio	Polytechnic of Porto, ISEP, Portugal
Paulvanna Nayaki Marimuthu	Kuwait University, Kuwait
Paweł Karczmarek	The John Paul II Catholic University of Lublin, Poland

Pedro Rangel Henriques	University of Minho, Portugal
Pedro Sobral	University Fernando Pessoa, Portugal
Pedro Sousa	University of Minho, Portugal
Philipp Jordan	University of Hawaii at Manoa, USA
Piotr Kulczycki	Systems Research Institute, Polish Academy of Sciences, Poland
Prabhat Mahanti	University of New Brunswick, Canada
Rabia Azzi	Bordeaux University, France
Radu-Emil Precup	Politehnica University of Timisoara, Romania
Rafael Caldeirinha	Polytechnic of Leiria, Portugal
Raghuraman Rangarajan	Sequoia AT, Portugal
Radhakrishna Bhat	Manipal Institute of Technology, India
Raiani Ali	Hamad Bin Khalifa University, Qatar
Ramadan Elaiess	University of Benghazi, Libya
Ramayah T.	Universiti Sains Malaysia, Malaysia
Ramazy Mahmoudi	University of Monastir, Tunisia
Ramiro Gonçalves	University of Trás-os-Montes e Alto Douro & INESC TEC, Portugal
Ramon Alcarria	Universidad Politécnica de Madrid, Spain
Ramon Fabregat Gesa	University of Girona, Spain
Ramy Rahimi	Chungnam National University, South Korea
Reiko Hishiyama	Waseda University, Japan
Renata Maria Maracho	Federal University of Minas Gerais, Brazil
Renato Toasa	Israel Technological University, Ecuador
Reyes Juárez Ramírez	Universidad Autonoma de Baja California, Mexico
Rocío González-Sánchez	Rey Juan Carlos University, Spain
Rodrigo Franklin Frogeri	University Center of Minas Gerais South, Brazil
Ruben Pereira	ISCTE, Portugal
Rui Alexandre Castanho	WSB University, Poland
Rui S. Moreira	UFP & INESC TEC & LIACC, Portugal
Rustam Burnashev	Kazan Federal University, Russia
Saeed Salah	Al-Quds University, Palestine
Said Achchab	Mohammed V University in Rabat, Morocco
Sajid Anwar	Institute of Management Sciences Peshawar, Pakistan
Sami Habib	Kuwait University, Kuwait
Samuel Sepulveda	University of La Frontera, Chile
Sara Luis Dias	Polytechnic of Cávado and Ave, Portugal
Sandra Costanzo	University of Calabria, Italy
Sandra Patricia Cano Mazuera	University of San Buenaventura Cali, Colombia
Sassi Sassi	FSJEGJ, Tunisia

Seppo Sirkemaa	University of Turku, Finland
Sergio Correia	Polytechnic of Portalegre, Portugal
Shahnawaz Talpur	Mehran University of Engineering & Technology Jamshoro, Pakistan
Shakti Kundu	Manipal University Jaipur, Rajasthan, India
Shashi Kant Gupta	Eudoxia Research University, USA
Silviu Vert	Politehnica University of Timisoara, Romania
Simona Mirela Riurean	University of Petrosani, Romania
Slawomir Zolkiewski	Silesian University of Technology, Poland
Solange Rito Lima	University of Minho, Portugal
Sonia Morgado	ISCPSI, Portugal
Sonia Sobral	Portucalense University, Portugal
Sorin Zoican	Polytechnic University of Bucharest, Romania
Souraya Hamida	Batna 2 University, Algeria
Stalin Figueroa	University of Alcala, Spain
Sümeyya Ilkin	Kocaeli University, Turkey
Syed Asim Ali	University of Karachi, Pakistan
Syed Nasirin	Universiti Malaysia Sabah, Malaysia
Tatiana Antipova	Institute of Certified Specialists, Russia
TatiannaRosal	University of Trás-os-Montes e Alto Douro, Portugal
Tero Kokkonen	JAMK University of Applied Sciences, Finland
The Thanh Van	HCMC University of Food Industry, Vietnam
Thomas Weber	EPFL, Switzerland
Timothy Asiedu	TIM Technology Services Ltd., Ghana
Tom Sander	New College of Humanities, Germany
Tomasz Kisielewicz	Warsaw University of Technology
Tomaž Klobučar	Jozef Stefan Institute, Slovenia
Toshihiko Kato	University of Electro-communications, Japan
Tuomo Sipola	Jamk University of Applied Sciences, Finland
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Valentim Realinho	Polytechnic of Portalegre, Portugal
Valentina Colla	Scuola Superiore Sant'Anna, Italy
Valerio Stallone	ZHAW, Switzerland
Verónica Vasconcelos	Polytechnic of Coimbra, Portugal
Vicenzo Iannino	Scuola Superiore Sant'Anna, Italy
Vitor Gonçalves	Polytechnic of Bragança, Portugal
Victor Alves	University of Minho, Portugal
Victor Georgiev	Kazan Federal University, Russia
Victor Hugo Medina Garcia	Universidad Distrital Francisco José de Caldas, Colombia
Victor Kaptelinin	Umeå University, Sweden

Viktor Medvedev	Vilnius University, Lithuania
Vincenza Carchiolo	University of Catania, Italy
Waqas Bangyal	University of Gujrat, Pakistan
Wolf Zimmermann	Martin Luther University Halle-Wittenberg, Germany
Yadira Quiñonez	Autonomous University of Sinaloa, Mexico
Yair Wiseman	Bar-Ilan University, Israel
Yassine Drias	University of Algiers, Algeria
Yuhua Li	Cardiff University, UK
Yuwei Lin	University of Roehampton, UK
Zbigniew Suraj	University of Rzeszow, Poland
Zorica Bogdanovic	University of Belgrade, Serbia

Contents

Impact of Preprocessing Using Substitution on the Performance of Selected NER Models - Methodology

Miroslav Potočár(✉) and Michal Kvet

University of Žilina, Žilina, Slovakia
{Miroslav.Potocar,Michal.Kvet}@fri.uniza.sk

Abstract. This paper investigates the effect of preprocessing, specifically word substitution by pseudo words, on the performance of selected named entity recognition (NER) models. The study focuses on explaining the methodology used during the experimental process. The paper comprehensively describes the dataset used, the process of word substitution with pseudo words, the process of model training, the process of executing the test scenario, the performance evaluation criteria and the limitations of the experiment. This paper contributes to the evolving area of Natural Language Processing by providing a comprehensive examination of the impact of preprocessing using substitution strategy on the performance of selected NER models.

Keywords: named entity recognition · preprocessing · substitution · pseudo words

1 Introduction

Named Entity Recognition (NER) plays a key role in solving various Natural Language Processing tasks. It allows the extraction of entities, such as persons, organizations, and places, from unstructured text. New NER models are regularly emerging, which are achieving increasingly better results on specific domains. However, little attention has been paid to text preprocessing, which may be a critical factor in the overall performance of the models. This paper investigates the impact of a particular preprocessing technique - pseudo word substitution - on the performance of selected NER models, namely hidden Markov model (HMM), conditional random fields (CRF), gated recurrent unit (GRU), bidirectional long short-term memory network (BiLSTM) and our Naïve model. Substitution involves replacing a particular word in a sequence with a pseudo word that to some extent reflects one of the features of that word. Such a preprocessing method is intended to improve the model's ability to generalize.

In addition to the approach used and the associated model, the preprocessing of the input data can have a significant impact on the performance of the system [5]. As noted by Hickman et al. [2] certain text preprocessing procedures can

Á. Rocha et al. (Eds.): WorldCIST 2024, LNNS 990, pp. 141–150, 2024.
https://doi.org/10.1007/978-3-031-60328-0_14

help improve the accuracy of subsequent text analysis. Standard preprocessing procedures include stopword removal, lowercase conversion, and stemming. Contractions expansion (converting abbreviations and abbreviated words to their full form) is commonly used in text analysis [3]. Similar preprocessing procedures can be used for the NER task [5]. Despite the potential impact of preprocessing on the resulting performance on various NLP tasks, there has been little attention given to this topic.

This study is designed to explain the methodology used to investigate the impact of substitution on the performance of selected NER models. In the Sect. 2, we will explain the concept of word substitution by pseudo words and also present the main ideas behind the origin of the idea of replacing words by pseudo words representing certain features of the original word. Section 3 focuses on the methodology itself. Here, we describe in detail the selected dataset, the process of replacing words with pseudo words along with the tested scenarios, the process of model training, the process of executing the test scenario, the observed metrics along with the method of performance evaluation, and finally, we conclude with the limitations of the experiment.

By clarifying the methodological background of our experiment, we set the foundation for a deeper understanding of how preprocessing may lead to changes in the performance and robustness of the NER model.

2 Concept of Pseudo Word Substitution

The idea of using word substitution with pseudo words to investigate its impact on the performance of different models arose when reviewing the work of Bikel et al. [1] where pseudo words were used as one of feature that were used to solve a NER task. In our work, we do not use these words as additional features, but use them directly to replace words in sequences.

As stated in the original work [1], the intuition behind the use of these words is clear:

- In Roman languages, a capital letter at the beginning of a word is often good evidence that it is a name. Therefore, it makes sense that if we come across an unknown word that begins with a capital letter, we replace that word with a pseudo word that represents that information.
- If we consider each word consisting of numeric characters as a unique number, we would need an infinitely large vocabulary. Certain forms of numeric characters tend to represent the same information. For example, a four-digit number often represents the year, numbers separated by slashes often represent the date, numbers containing a comma usually represent monetary amounts, and numbers with a period may represent percentages.

We have taken the categories and order of features from the original work. To these features we have assigned a custom pseudo word (tag) to be used in the substitution. We have also defined the rules that must be fulfilled for a word to be replaced by a pseudo word. Almost all of the rules have the form of a regular expression. The only exception is the pseudo word representing the first word in a sentence. Here we needed an index of that word within the sentence when evaluating the condition. The individual pseudo words (tags), the conditions, their order along with an example and intuition can be seen in Table 1 which is a modification of the table from the original work. The meaning of the individual regex symbols can be found at this page. We suggest that pseudo words may help to reduce the vocabulary and increase the model's ability to generalize, making the model better at dealing with unknown words.

Table 1. Word features, pseudo word tag, conditions, examples and intuition behind them

Word feature	Tag	Condition	Example	Intuition
twoDigitNum	[TDN]	^\ d{2}$	90	Two-digit year
fourDigitNum	[FDN]	^\d{4}$	2023	Four-digit year
digitAndAlpha	[CDA]	^(?:.*[a-zA-Z].*\ d.*)| (?:.*\ d.*[a-zA-Z].*)$	A8956-67	Product code
digitAndDash	[CDD]	^(?:[\ d\-]*\ d-[\ d\-]*)$	09-96	Date
digitAndSlash	[CDS]	^(?:[\ d/]*\ d/[\ d\/]*)$	11/9/89	Date
digitAndComma	[CDC]	^(?:-?[\ d]*\ d,[\ d\,]*.?\ d*)$	23,000.00	Monetary amount
digitAndPeriod	[CDP]	^(?:-?\ d+.\ d+)$	1.00	Monetary amount, percentage
otherNum	[ON]	^-?\ d+$	456789	Other number
allCaps	[AC]	^[A-Z]+$	OSN	Organization
capPeriod	[CP]	^([A-Z]([a-z]{0,2}| [a-z][A-Z])\.\ s*)+$	OSN	Organization
firstWord	[FW]	*word index = 0*	*first word of sentence*	No useful capitalization information
initCap	[IC]	^[A-Z][a-zA-Z]*$	Sally	Capitalized word
lowerCase	[LC]	^[a-z]+$	can	Uncapitalized word
other	[OW]	^[\ s\ S]*]+$	,	Punctuation marks, all other words

3 Methodology

3.1 Data

In our research we have used the *CoNLLpp* dataset [6], which is a corrected version of the original *CoNLL2003* dataset [4]. In the *CoNLLpp* version, 5.38% of the sentences in the test set have been manually corrected compared to the original version. *CoNLL2003* is a widely used NER benchmark dataset. The whole dataset is already partitioned into a training set containing 14041 sentences, a validation set consisting of 3250 sentences, and a test set consisting of 3453 sentences. There are 4 types of named entities. The first entity is persons (PER), denoting the names of individuals or groups. The next type of named entities are locations (LOC), where the names of political or geographically defined places such as cities, provinces, states, international regions, bodies of water, mountains, etc. are included. The third group is organizations (ORG), which includes names of companies, agencies, institutions, etc. The last type of named entities is miscellaneous (MISC), which includes names of entities that do not fit into any of the previous three categories. They may include names of events, nationalities, products, artworks, etc.

The version of *CoNLLpp* from the HuggingFace portal that we use contains only data in English. In addition, each word is also given its corresponding part-of-speech tag. However, in our experiment we will not use this knowledge and will only focus on prediction based on word sequences. A single row consists of an array of words and an array of their associated named entity tags. During the experiment we used all available sets. The training set was used to train the models, the validation set was used to tune the hyperparameters, and the test set was used to evaluate the performance of the models on previously unseen data. We kept the individual sets in their original form, i.e., we did not change the order of the sentences during training.

3.2 Replacing Words with Pseudo Words

In our research, we focused on the effect of using pseudo words on the performance of models in a NER task. The sentences and the words occurring in them were sequentially walked through in each dataset. Each word was subject to a series of tests. If a word met any of the conditions, it was replaced in the sentence by the pseudo word corresponding to that condition. A word that already met one of the conditions was excluded from further consideration.

The whole process began with the creation of a dictionary of known words. This dictionary was created based on the training set only. The sentences occurring in the training set are flattened and a single array containing all the words occurring in the corpus is created. In the next stage of dictionary definition, there are two possible scenarios. In the first scenario, words that contain numbers or only consist of punctuations remain in the array. In the second scenario, such words are removed. Independently of the applied scenario, the frequency of occurrences of each word is computed based on the given array. Finally, words with occurrence frequency below a certain threshold are removed. From the remaining unique words, a dictionary of known words is created.

The next stage of the process involved the actual replacement of words in the sets by pseudo words. This phase is applied to all the sets (train, validation, test). The sets are sequentially walked through sentence by sentence and sentence by word. Each word is subjected to a series of conditions. The first condition is the occurrence of the word in a dictionary of known words. If the word is found in this dictionary, its form is kept and it is excluded from further processing. If the word is not found in the dictionary, it is subjected to further testing. In the second step, the conditions are applied to this unknown word in a well-defined order. The individual conditions and their order of application are listed in Table 1. If a word satisfies the corresponding condition, it is replaced by the corresponding pseudo word and is excluded from further processing. If the word does not satisfy the actual condition, the following condition is applied to it in order. If a word does not satisfy any of the conditions, it is placed on the last condition satisfied by each word.

Datasets processed in this way are used to train the model and evaluate its performance. In our experiment, we have tested the following scenarios:

- **No modification** - In this case, we have not applied any changes to the individual datasets and have used them in the format in which we got them from the source.
- **Removal of words containing numbers or consisting only of punctuation marks from the dictionary of known words** - All words from the test set are included in the list of known words, except for words containing digits or consisting only of non-alphanumeric characters. Thus, only these words are replaced by pseudo words in the training set. In the validation and test sets, words that did not appear in the training set are also replaced. However, the model had no opportunity to learn to recognize these pseudo words and thus they will only appear as unknown words for the model.
- **Removal of words from the dictionary of known words where the frequency of occurrences in the training set is less than a threshold** - This scenario contains four sub-scenarios for each frequency of occurrences (1, 2, 3, 4).
- **Remove those words from the dictionary of known words that contain numbers, consist only of punctuation marks, or have a frequency of occurrence less than the threshold** - This scenario is a combination of the two previous scenarios.

3.3 Model Training

The training understandably varied depending on the model. Each model required its specific training data format. Some of the models, namely CRF, GRU and BiLSTM, also required hyperparameter tuning.

The training process of Naïve model looks as follows. The sentences in the training set are flattened into an array of words and an array of associated named entity tags. For each word the most frequently used named tag for that word is defined. Also the most frequent tag in whole dataset is identified. It will be later assigned to unknown words.

As HMM model we have used the *HiddenMarkovModelTagger* implementation, available in the *NLTK* library. This implementation requires a collection of sentences for its training, where each sentence is represented by an array of pairs where the first position contains the word and the second position contains the associated named entity tag. We used the data prepared in this way as an argument to the model's training function.

From the *sklearn-crfsuite* library, we have used the *CRF* implementation. From a training dataframe, individual sentences represented by an array of words and a separate array of corresponding named entity tags are extracted. CRF requires defining a set of features and converting words to these features. We have taken the function that converts a word into a dictionary of features from the documentation page of the *sklearn-crfsuite* library. The original version of the function also produced features based on part-of-speech tags. Such features have been removed from the function to ensure equal conditions and available information across models. The model on its input for training and prediction requires every word in the sentence to be converted into dictionary of features. The CRF contained hyperparameters that needed to be determined. Using restricted grid search, we have tested different combinations of hyperparameters on the validation set. From the measured values, we have found that the best results are given by the combination of hyperparameters shown in Table 2. Remaining hyperparameters were left at their default values.

The training of the GRU and BiLSTM models looks identical in both cases. Since neural networks require only numerical data on their input, it is necessary to convert words and named entity tags to numbers. A special tag [**PAD**] is added to the list of named entity tags, which is used to represent padding. The sentences from the (preprocessed) training dataset are flattened and for each word the frequency of its occurrence within the whole dataset is determined. Since the size of the neural network inputs affects the number of parameters and hence the time required to train them, only a subset of the most frequently occurring words is selected from the list of unique words. The number of these words is determined by the *vocab_size* parameter. From the set of unique words, $vocab\_size - 2$ most frequently occurring words are selected. This set of unique words constitutes our vocabulary. The value 2 is subtracted from the original *vocab_size* parameter, since two values are reserved for special tokens that represent unknown words and padding. Based on the vocabulary, a *Keras StringLookup* layer is created. This layer will provide the conversion of words to numbers. Within all datasets, word arrays representing sentences are converted to number arrays using this layer. Training dataset prepared in this fashion needs to be divided into equally sized mini-batches. The size of a single batch is determined by the *batch_size* parameter. Tensors are created from the training sentences and their associated named entity tags. These tensors are concatenated into equally sized mini batches (the exception is the last batch, which may be smaller). The tensors in each of these mini batches have the same size, which is equal to the number of words in the longest sentence within the mini batch. Shorter sentences are aligned to the required size using a special character, padding. The adjusted data is used as an argument to the fit method, which is used to train the model. Since padding is used, a custom loss function based on the *SparseCategoricalCrossentropy* loss function was created. This loss function only takes into account the error in positions corresponding to the original sentence, so any part with padding is ignored when computing the error. The described loss function as well as the preprocessing of the input data is a modification of an example taken from the official documentation page of the *Keras* library, where the NER task using the transformer model has been solved. During training we used early stopping in order to reduce overtraining. This monitored the loss on the validation set and if the loss increased for two consecutive epochs, training was terminated. The structure of the GRU network can be seen in Table 3 and the structure of the BiLSTM is shown in Table 5. GRU and BiLSTM require hyperparameter tuning for their proper functioning. Using restricted grid search, we tested different combinations of hyperparameters on the validation set. From the measured values, we have found that the best results for GRU model are given by the combination of hyperparameters shown in Table 4 and best hyperparameters combination for BiLSTM is shown in Table 6. The other hyperparameters were left at their default value.

Table 2. CRF hyperparameters

Hyperparameter	Value
algorithm	lbfgs
c1	0.1
c2	0.1
max_iterations	200
all_possible_transitions	True

Table 3. GRU model structure

Layer type	Parameters
Embedding	input_dim=lookup_layer.vocabulary_size()+1, output_dim=100
GRU	units=50, return_sequences=True
Dense	units=10, activation='sigmoid'

3.4 Performing a Test Scenario

The flow of each test scenario consists of several steps. In the first step, a dictionary of known words is generated based on the training data. Based on scenario, words that contain digits or that consist entirely of punctuation marks are retained or removed. Next, the frequency of occurrences for each word is calculated and words that have a frequency lower than a given threshold are removed from the dictionary. In the next step, word replacement with pseudo words is

Table 4. GRU hyperparameters

Hyperparameter	Value	Note
vocab_size	20000	Upper bound for number of words in string lookup layer
output_embedding	100	Each word is converted into 100 dimensional numeric vector
units	50	Number of GRU units
batch_size	32	Each training mini-batch consist of 32 padded sequences (except the last)
epochs	100	Early stopping was used, so this number is the upper limit

Table 5. BiLSTM model structure

Layer type	Parameters
Embedding	input_dim=lookup_layer.vocabulary_size()+1, output_dim=100
Bidirectional(LSTM)	units=100, return_sequences=True
Dense	units=10, activation=‘sigmoid’

Table 6. BiLSTM hyperparameters

Hyperparameter	Value	Note
vocab_size	20000	Upper bound for number of words in string lookup layer
output_embedding	100	Each word is converted into 100 dimensional numeric vector
units	100	Number of LSTM units in one direction
batch_size	32	Each training mini-batch consist of 32 padded sequences (except the last)
epochs	100	Early stopping was used, so this number is the upper limit

handled. Depending on the scenario, this phase can be skipped. If the replacement should be performed, the series of conditions is applied to each dataset (training, validation, test). For each word, it is first checked for its occurrence in the dictionary of known words. If the word occurs in the dictionary, it is left unchanged and excluded from further processing, otherwise, given the condition it satisfied, it is replaced by the corresponding pseudo word. Next comes the initialization of the model. The processed training data is sent to the model initialization method. This initialization step is used by the GRU and BiLSTM models to create the *StringLookup* layer. This is followed by transformation of training data (and, in the case of GRU and BiLSTM, also validation data) into the format needed for model training. Next, all three datasets are transformed into the format required for prediction and performance evaluation. In the last step, training and performance evaluation of the model takes place. This step is performed N times, storing the result in the result list. These N results are then used to calculate the average value of the observed metrics.

3.5 Performance Evaluation

The most commonly used metrics for evaluating NER models are precision, recall, and F1 score. These metrics provide a broad view of model performance. Their use is widespread as they provide a balance between the model’s ability to correctly identify entities (precision) and its ability to not miss any real entities (recall). The F1 score provides a single metric that balances both considerations. Because of this, we provide the F1 score as the main metric.

To evaluate the performance of the models, we used the *seqeval* framework, which is a Python framework available through the *evaluate* library, designed to evaluate labeled sequences. In the context of NER, *seqeval* provides values for metrics such as accuracy, precision, recall, F1 score for the entire dataset and also provides the same metrics for individual named entity categories. The *seqeval* provides two evaluation modes, **default** and **strict**. The **default** mode aims to mimic *conlleval*, while the **strict** mode evaluates inputs based on the specified schema. Since our data uses the IOB2 scheme, we used **strict** mod in our evaluation.

For each scenario we performed 5 runs, i.e. in each run we re-created and retrained the model. Using *seqeval*, we have evaluated the individual metrics and stored them in a list of results. The final result for a given metric is calculated as the average of all runs.

3.6 Experiment Limitations

There are several limitations in our experiment that can be potential sources of error. The first limitation is related to the dataset used. In the experiment, we only used the *CoNLLpp* dataset, containing data from English-language newspaper articles. In order to be able to make general conclusions applicable to different domains and languages, it would be necessary to perform experiments with datasets containing data from different domains and also in different languages.

Another limitation relates to individual sets. For training, validation, and testing, we used prepared sets that were directly available within *CoNLLpp*. For more accurate results, it would be appropriate to combine the individual parts into a whole, which would then be randomly used to create training, validation, and testing sets. We also did not perform random shuffling of the training data as part of the experiment. This is not a problem in case of the Naïve, HMM and CRF models, but in case of neural network based models, training on different mini batches could lead to slightly different results.

Rules that replace words with pseudo words can also be a source of distortions. We create the above regular expressions, and since we are not linguistic experts, we may have created expressions that inadequately capture some of the categories.

Another source of error may be the models themselves. We designed the GRU and BiLSTM models ourselves based on our knowledge and experience. These models are probably not achieving their maximum potential.

Technical limitations were the reason why model tuning was performed only on a subset of all available hyperparameters. Also, due to technical limitations and lack of computational power, we did not conduct model tuning during the experiments themselves, which means that the same hyperparameters are also used for models trained on data in which some words have been replaced by pseudo words.

4 Conclusion

To summarize, the study aimed to systematically investigate the impact of preprocessing based on pseudo word substitution on selected NER models, namely Naïve, HMM, CRF, GRU and BiLSTM. We introduced in detail the concept of pseudo word substitution and provided the Table 1 listing the pseudo words used along with the conditions that must be satisfied for a word to be replaced by a corresponding pseudo word. The remaining part of this work was devoted to a detailed description of the methodology used in the experiment.

The insights gained from this research not only advance our understanding of the interrelationship between preprocessing techniques and NER outcomes, but also have practical relevance for researchers and practitioners who would like to further investigate substitution as a preprocessing technique. We suggest that proper preprocessing techniques may be key to obtaining models capable of better generalization.

As an extension for the future, we propose to extend the set of existing pseudo words with new elements that will allow finer word discrimination and hence more fine-grained feature encoding.

Acknowledgment. It was supported by the Erasmus+ project: Project number: 2022-1-SK01-KA220-HED-000089149, Project title: Including EVERyone in GREEN Data Analysis (EVERGREEN) funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Slovak Academic Association for International Cooperation (SAAIC). Neither the European Union nor SAAIC can be held responsible for them.

Co-funded by the European Union

EverGreen DATA Analytics

References

1. Bikel, D.M., Schwartz, R., Weischedel, R.M.: An algorithm that learns what's in a name. Mach. Learn. **34**, 211–231 (1999)
2. Hickman, L., Thapa, S., Tay, L., Cao, M., Srinivasan, P.: Text preprocessing for text mining in organizational research: review and recommendations. Organ. Res. Methods **25**(1), 114–146 (2022)
3. Naseem, U., Razzak, I., Eklund, P.W.: A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. Multimedia Tools Appl. **80**, 35239–35266 (2021)
4. Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition (2003). arXiv preprint cs/0306050
5. Situmeang, S.: Impact of text preprocessing on named entity recognition based on conditional random field in Indonesian text. Jurnal Mantik **6**(1), 423–430 (2022)
6. Wang, Z., Shang, J., Liu, L., Lu, L., Liu, J., Han, J.: CrossWeigh: training named entity tagger from imperfect annotations (2019). arXiv preprint arXiv:1909.01441