Lecture Notes in Artificial Intelligence 12598

Subseries of Lecture Notes in Computer Science

More information about this subseries at http://www.springer.com/series/1244

Zygmunt Vetulani · Patrick Paroubek ·
Marek Kubis (Eds.)

Human Language Technology

Challenges for Computer Science and Linguistics

8th Language and Technology Conference, LTC 2017
Poznań, Poland, November 17–19, 2017
Revised Selected Papers

Springer

Editors
Zygmunt Vetulani
Adam Mickiewicz University
Poznań, Poland

Marek Kubis
Adam Mickiewicz University
Poznań, Poland

Patrick Paroubek
Laboratoire d'Informatique pour la Méca
Orsay, France

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-030-66526-5 ISBN 978-3-030-66527-2 (eBook)
https://doi.org/10.1007/978-3-030-66527-2

LNCS Sublibrary: SL7 – Artificial Intelligence

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 present a selection of the refereed papers of the 8th Language and Technology Conference: Challenges for Computer Science and Linguistics, LTC 2017, held in Poznań, Poland, in November 2017, in memoriam Alain Colmerauer (1941–2017), pioneer of Logic Programming in natural language processing.

People started to use rules to describe language several thousand years ago, going back to Plato in the western tradition, and it was only much later that computer scientists joined them. Among the numerous contributions at the crossing of the two communities, Prolog is one of the most noticeable. For Prolog people 2017 is a special year since it is the year of the demise of Alain Colmerauer, one of the two inventors of Prolog and a member of the LTC Program Committee in 2005. Naturally, the 2017 LTC conference was dedicated to him. It is remarkable that this edition of LTC was also the first one to have the word "deep" in some of the paper titles, not associated with "parsing" as it is usually but with "neural network" or used in reference to deep learning, acknowledging the fact that computational linguistics is entering a new era.

The selection of updated papers from the LTC 2017 proceedings that we present in this book therefore offers a view of our domain where the classical approaches lie next to the most recent developments in computational linguistics.

In this book the reader will find a selection of 25 revised and in most cases substantially extended and updated versions of papers presented at the 8th Language and Technology Conference in 2017. The reviewing process was done by the international jury, composed of the program committee members, or experts nominated by them. The selection was made among 97 contributions presented at the conference and is indicative of the preferences of the reviewers. Totaling 72, the authors of the selected contributions represent research institutions from 14 countries: Canada, Czech Republic, France, India, Ireland, Japan, Nigeria, Norway, Poland, Spain, Switzerland, UK, Ukraine.

What are the presented papers about?

To try to make the presentation of the papers transparent we have organized them into seven parts. These are:

1. Language Resources, Tools, and Evaluation (8)
2. Less-Resourced-Languages (LRL) (2)
3. Speech Processing (4)
4. Morphology (2)
5. Computational Semantics (3)
6. Machine Translation (1)
7. Information Retrieval and Information Extraction (5)

The clustering of the articles is approximate as many papers address more than one thematic area. The ordering of the chapters has no "deep" meaning: it roughly

approximates the order in which humans proceed in natural language production and processing: starting with language resources, speech and text processing, to LT applications. Within these parts we ordered contributions in alphabetical order with respect to the family name of the first author.

We start this volume with the **Language Resources, Tools, and Evaluation** part, containing eight contributions. In the paper "Creating Norwegian valence resources from a deep grammar" the authors (Lars Hellan, Dorothee Beermann, Tore Bruland, Tormod Haugland, and Elias Aamot) propose a procedure for generating valence resources for the Norwegian language from a deep grammar, which was intended to make grammatical information encoded in a deep parser more accessible for humans and for further processing. The aim of the paper "How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine – Final Notes on Development and Evaluation" (Mika Koistinen, Kimmo Kettunen, and Jukka Kervinen) is to present experiments aiming at improvement of optical character recognition (OCR) technology applied to a 500,000 word sample from the historical Finnish newspaper collection for the period from 1771 to 1910. In the next paper "Fine-tuning Tree-LSTM for phrase-level sentiment classification on a Polish dependency treebank" the authors (Tomasz Korbak and Paulina Żak) describe a variant of Child-Sum Tree-LSTM deep neural networks fine-tuned for working with dependency trees and morphologically rich languages using the example of Polish (presented at the LTC evaluation challenge PolEval). The fourth contribution of this part is "Supervised Transfer Learning for Sequence Tagging of User-Generated-Content in Social Media" (Sara Meftah, Nasredine Semmar, Othmane Zennaki, and Fatiha Sadat). In this work, the authors analyse the impact of supervised sequential transfer learning to overcome the sparse data problem in the *Tweets-domain* by leveraging the huge annotated data available for the *Newswire-domain*. What follows is the paper "Investigating the Lack of Consensus among Sentiment Analysis Tools" (Marco A. Palomino, Aditya Padmanabhan Varma, Gowriprasad Kuruba Bedala, and Aidan Connelly), which contains a survey on the current state of the art in the field of sentiment analysis in which the authors compare the performance of several Sentiment Analysis systems on a Twitter-based corpus. In the sixth contribution "Automated normalization and analysis of historical texts" (Paweł Skórzewski, Krzysztof Jassem, and Filip Graliński) the authors introduce a method for processing historical texts that applies a list of diachronic pairs found with the aid of word distribution vectors in historical corpora. The seventh article is "PADI-web: an event-based surveillance system for detecting, classifying and processing online news" (Sarah Valentin, Elena Arsevska, Alize Mercier, Sylvain Falala, Julien Rabatel, Renaud Lancelot, and Mathieu Roche). Here the authors present a platform for automated extraction of animal disease information from the Web (PADI-web) which is a multilingual text mining tool for automatic detection, classification, and extraction of disease outbreak information from online news articles. In the last paper of the first part "KRNNT: Polish Recurrent Neural Network Tagger Extended", another contribution to the PolEval evaluation challenge, its author (Krzysztof Wróbel) presents the morphosyntactic tagger KRNNT for Polish based on recurrent neural networks, and argues for the superiority of neural bidirectional approaches over other existing tagging methods.

Less-Resourced Languages are considered of special interest for the LTC community and since 2009 the LRL workshop has constituted an integral part of LTC meetings. The Less-Resourced-Languages (LRL) part contains two papers. The first one "Experiments with automatic and semi-automatic detection of sparse word forms in Old Braj" (Rafał Jaworski and Krzysztof Stroński) presents the authors' work on automatic converb detection in Old Braj poetry from the 15th-17th centuries and is a part of research on non-finite verbal forms in early New Indo-Aryan (NIA) language corpora comprising data from Old Rajasthani, Awadhi, Braj, Dakkhini, and Pahari. In the second one, titled "Towards Better Text Processing Tools for the Ainu Language" (Karol Nowakowski, Michał Ptaszyński, and Fumito Masui), the authors present their research devoted to the development of Natural Language Processing technologies for the Ainu language, a critically endangered language isolate spoken by the Ainu people, the native inhabitants of northern parts of the Japanese archipelago.

The **Speech Processing** part contains four papers. The contribution "The Harmonia Corpus – a Dialogue Corpus for Automatic Analysis of Phonetic Convergence" (Jolanta Bachan, Mariusz Owsianny, and Grażyna Demenko) describes the Harmonia spoken dialogue corpus created for analysis and objective evaluation of phonetic convergence in human-human communication with the goal to build convergence models which could be implemented in spoken dialogue systems. The authors of the second article in this part, "Resources and tools for Automated Speech Segmentation of the African Language Naija (Nigerian Pidgin)" (Brigitte Bigi, Oyelere S. Abiola, and Bernard Caron), present the development of HLT resources and tools for the African language Naija (Nigerian Pidgin), spoken in Nigeria, focusing on language resources for a tokenizer, an automatic speech system for predicting the pronunciation of words and their segmentation. In the next paper, "Speaker Variability for Emotions Classification in African Tone Languages" (Moses Ekpenyong, Udoinyang Inyang, Nnamso Umoh, Temitope Fakiyesi, Okokon Akpan, and Nseobong Uto), the authors examine the effect of speaker variability on emotions and languages, and propose a classification system based on the study of speech characteristics such as fundamental frequency and intensity for two languages, Ibibio (New Benue Congo, Nigeria) and Yoruba (Niger Congo, Nigeria), from voice recordings of native speakers of these languages. The last paper of this part "Analysis of Polish nasalized vowels based on spatial energy distribution and formant frequency measurement" (Anita Lorenc, Katarzyna Klessa, Daniel Król, and Łukasz Mik) offers the results of the analysis of F1 and F2 frequency measurements in Polish nasalized vowels represented in writing by the graphemes *e* and *a* (realized before voiceless fricatives).

The **Morphology** part is composed of two papers. The authors of the first one, "RNN Language Model Estimation for Out-of-Vocabulary Words" (Irina Illina and Dominique Fohr), propose new approaches to out-of vocabulary proper noun probability estimation using a Recurrent Neural Network Language Model. The second paper, "Automatic Pairing of Perfective and Imperfective Verbs in Polish" (Zbigniew Kaleta) presents an algorithm that automatically detects morphological dependencies between verbs in Polish and uses them to match corresponding perfective and imperfective verbs.

The **Computational Semantics** part is composed of three papers. This part opens with the text "Transforming Syntactic Relations in Attributive Groups" (Iuliia

Romaniuk, Nina Suszczańska, and Przemysław Szmal). In their paper on the Thetos translation system from Polish into sign language, the authors present recent translation quality improvements resulting from deepened syntactic and semantic analysis of the source text. Then the paper "Syntactic-Semantic Classes of Context-Sensitive Synonyms Based on a Bilingual Corpus" (Zdenka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič) summarizes findings of a three-year study on verb synonymy in Czech-English translation based on both syntactic and semantic criteria on the basis of existing CL resources, including the Prague Dependency Treebank-style valency lexicons, FrameNet, VerbNet, PropBank, WordNet, and the parallel Prague Czech-English Dependency Treebank. In the third contribution, "Towards the evaluation of feature embedding models of the fusional languages" (Alina Wróblewska, Katarzyna Krasnowska-Kieraś, and Piotr Rybak), the authors investigate features to be used for estimating Neural-Networks-based NLP models of the fusional languages.

The **Machine Translation** section contains one contribution: "Syntactic and Semantic Impact of Prepositions in Machine Translation: An Empirical Study of French-English Translation of Prepositions 'à', 'de' and 'en'" (Violaine Prince), where the author presents a study about ambiguous French prepositions, stressing their role as dependency-introducers in order to derive some French-English MT translation heuristics, based on a French-English set of parallel texts.

The **Information Retrieval and Information Extraction** part contains five contributions. The first one, "Sentence Answer Selection for Open Domain Question Answering via Deep Word Matching" (Fabrizio Ghigi, Diana Turcsany, Thomas Kaltenbrunner, and Maurizio Cibelli), proposes an unsupervised approach for sentence answer selection (called Deep Word Matching) that uses both the string form and distributed representations of words, thereby capturing their hidden semantic relatedness. In the second paper "On the contribution of specific entity detection in comparative constructions to automatic spin detection in biomedical scientific publications" (Anna Koroleva and Patrick Paroubek), the authors address the problem of providing automated aid for the detection of misrepresentation ("spin") of research results in scientific publications from the biomedical domain. In the next paper "Automatic Taxonomy Generation: A Use-Case in the Legal Domain" (Cécile Robin, James O'Neill, and Paul Buitelaar), the authors describe a methodology for generating a taxonomy of legal concepts based on the analysis of a collection of official legal texts from Great Britain and Northern Ireland. The next paper, "Title Categorization based on Category Granularity" (Kazuya Shimura and Fumiyo Fukumoto), focuses on a problem of short-text categorization (newspaper titles), and presents a method that maximizes the impact of informative words due to the titles' sparseness. This part ends with the article "Identification of Domain-Specific Senses based on Word Embedding Learning" (Attaporn Wangpoonsarp and Fumiyo Fukumoto). This paper is about the domain-specific meaning of a word and proposes a machine-learning approach for detecting the main meaning of a word given the domain.

We wish you all interesting reading.

Zygmunt Vetulani
Patrick Paroubek

Organization

Organizing Committee Chair

Zygmunt Vetulani	Adam Mickiewicz University, Poland

Organizing Committee

Jolanta Bachan	Adam Mickiewicz University, Poland
Marek Kubis	Adam Mickiewicz University, Poland
Jacek Marciniak	Adam Mickiewicz University, Poland
Tomasz Obrębski	Adam Mickiewicz University, Poland
Hanna Szafrańska	Adam Mickiewicz University Foundation, Poland
Marta Witkowska	Adam Mickiewicz University, Poland
Mateusz Witkowski	Adam Mickiewicz University, Poland

Program Committee Chairs

Zygmunt Vetulani	Adam Mickiewicz University, Poland
Patrick Paroubek	LIMSI-CNRS, France

Program Committee

Victoria Arranz	ELRA, France
Jolanta Bachan	Adam Mickiewicz University, Poland
Núria Bel	Universitat Pompeu Fabra, Spain
Krzysztof Bogacki	Warsaw University, Poland
Christian Boitet	IMAG, France
Gerhard Budin	University of Vienna, Austria
Nicoletta Calzolari	ILC/CNR, Italy
Nick Campbell	Trinity College Dublin, Ireland
Christopher Cieri	LDC, USA
Khalid Choukri	ELRA, France
Adam Dąbrowski	Poznań University of Technology, Poland
Elżbieta Dura	University of Skövde, Sweden
Katarzyna Dziubalska-Kołaczyk	Adam Mickiewicz University, Poland
Moses Ekpenyong	University of Uyo, Nigeria
Cedrick Fairon	UCLouvain, Belgium
Christiane Fellbaum	Princeton University, USA
Piotr Fuglewicz	TiP Sp. z o.o., Poland
Maria Gavrilidou	ILSP, Greece
Dafydd Gibbon	Bielefeld University, Germany

Marko Grobelnik	Jožef Stefan Institute, Slovenia
Eva Hajičová	Charles University, Czech Republic
Krzysztof Jassem	Adam Mickiewicz University, Poland
Girish Nath Jha	Jawaharlal Nehru University, India
Katarzyna Klessa	Adam Mickiewicz University, Poland
Cvetana Krstev	University of Belgrade, Serbia
Eric Laporte	Université Paris-Est Marne-la-Valée, France
Yves Lepage	Waseda University, Japan
Gerard Ligozat	LIMSI/CNRS, France
Natalia Loukachevitch	Research Computing Center of Moscow State University, Russia
Wiesław Lubaszewski	AGH, Poland
Bente Maegaard	Centre for Language Technology, Denmark
Bernardo Magnini	ITC IRST, Italy
Jacek Marciniak	Adam Mickiewicz University, Poland
Joseph Mariani	LIMSI-CNRS, France
Jacek Martinek	Poznań University of Technology, Poland
Gayrat Matlatipov	Urgench State University, Uzbekistan
Keith J. Miller	MITRE, USA
Asunción Moreno	UPC, Spain
Agnieszka Mykowiecka	IPI PAN, Poland
Jan Odijk	Utrecht University, The Netherlands
Maciej Ogrodniczuk	IPI PAN, Poland
Karel Pala	Masaryk University, Czech Republic
Pavel S. Pankov	National Academy of Sciences, Kyrgyzstan
Adam Pease	IPsoft, USA
Maciej Piasecki	Wrocław University of Science and Technology, Poland
Stelios Piperidis	ILSP, Greece
Gabor Proszeky	MorphoLogic, Hungary
Georg Rehm	DFKI, Germany
Michał Ptaszyński	Hokkaido University, Japan
Rafał Rzepka	Hokkaido University, Japan
Kepa Sarasola Gabiola	Universidad del País Vasco , Spain
Frèdèrique Segond	VISEO Group, France
Sanja Seljan	University of Zagreb, Croatia
Zhongzhi Shi	Institute of Computing Technology, Chinese Academy of Sciences, China
Janusz Taborek	Adam Mickiewicz University, Poland
Ryszard Tadeusiewicz	AGH, Poland
Marko Tadić	University of Zagreb, Croatia
Dan Tufiş	RCAI, Romania
Hans Uszkoreit	DFKI, Germany
Tamás Váradi	RIL, Hungary
Andrejs Vasiljevs	Tilde, Latvia
Cristina Vertan	University of Hamburg, Germany

Dusko Vitas	University of Belgrade, Serbia
Piek Vossen	Vrije Universiteit Amsterdam, The Netherlands
Jan Węglarz	Poznań University of Technology, Poland
Bartosz Ziółko	AGH, Poland
Mariusz Ziółko	AGH, Poland
Richard Zuber	CNRS, France
Andrzej Zydroń	XTM-INTL, UK

Reviewers

Alladin Ayesh	De Montfort University, UK
Bogdan Babych	University of Leeds, UK
Jolanta Bachan	Adam Mickiewicz University, Poland
Esha Banerjee	Google, Tokyo
Dorothee Beermann	Norwegian University of Science and Technology, Norway
Delphine Bernhard	LiLPa, University of Strasbourg, France
Laurent Besacier	Laboratoire d'Informatique de Grenoble, équipe GETALP, France
Krzysztof Bogacki	University of Warsaw, Poland
Christian Boitet	IMAG, France
Tiberiu Boros	RACAI, Romania
Nicoletta Calzolari	ILC/CNR, Italy
Subhash Chandra	Delhi University, India
Narayan Choudhary	CIIL, India
Monojit Choudhary	MSRI, India
Khalid Choukri	ELRA, France
Adam Dąbrowski	Poznań University of Technology, Poland
Damien De Meyere	UCLouvain, Belgium
Tomasz Dwojak	Adam Mickiewicz University, Poland
Katarzyna Dziubalska-Kolaczyk	Adam Mickiewicz University, Poland
Moses Ekpenyong	University of Uyo, Nigeria
Cedrick Fairon	UCLouvain, Belgium
Karen Fort	Sorbonne University, France
Piotr Fuglewicz	TiP Sp. z o.o., Poland
Maria Gavrilidou	ILSP, Greece
Filip Graliński	Adam Mickiewicz University, Poland
Eva Hajičová	Charles University, Czech Republic
Dai Hasegawa	Aoyama Gakuin University, Japan
Lars Hellan	Norwegian University of Science and Technology, Norway
Magdalena Igras-Cybulska	AGH, Poland
Krzysztof Jassem	Adam Mickiewicz University, Poland
Girish Nath Jha	Jawaharlal Nehru University, India
Yasutomo Kimura	Otaru University of Commerce, Japan

Katarzyna Klessa	Adam Mickiewicz University, Poland
Łukasz Kobyliński	IPI PAN, Poland
Cvetana Krstev	University of Belgrade, Serbia
Marek Kubis	Adam Mickiewicz University, Poland
Ritesh Kumar	Dr. Bhimrao Ambedkar University, India
Sachin Kumar	C-DAC, India
Eric Laporte	Université Paris-Est Marne-la-Valée, France
Yves Lepage	Waseda University, Japan
Gérard Ligozat	LIMSI/CNRS, France
Natalia Loukachevitch	Research Computing Center of Moscow State University, Russia
Paweł Lubarski	Poznań University of Technology, Poland
Wiesław Lubaszewski	AGH, Poland
Bente Maegaard	Centre for Language Technology, Denmark
Bernardo Magnini	ITC IRST, Italy
Jacek Marciniak	Adam Mickiewicz University, Poland
Małgorzata Marciniak	IPI PAN, Poland
Joseph Mariani	LIMSI-CNRS, Orsay, France
Jacek Martinek	Poznań University of Technology, Poland
Fumito Masui	Kitami Institute of Technology, Japan
Gayrat Matlatipov	Urgench State University, Uzbekistan
Diwakar Mishra	EZDI, India
Massimo Moneglia	University of Florence, Italy
Asunción Moreno	UPC, Spain
Mikołaj Morzy	Poznań University of Technology, Poland
Adeline Muller	UCLouvain, Belgium
Koji Murakami	Rakuten, USA
Agnieszka Mykowiecka	IPI PAN, Poland
Hubert Naets	UCLouvain, Belgium
Pinky Nainwani	Optimum InfoSystem Pvt Ltd, India
Tomasz Obrebski	Adam Mickiewicz University, Poland
Jan Odijk	Utrecht University, The Netherlands
Maciej Ogrodniczuk	IPI PAN, Poland
Noriyuki Okumura	National Institute of Technology, Akashi College, Japan
Karel Pala	Masaryk University, Czech Republic
Pavel S. Pankov	National Academy of Sciences, Kyrgyzstan
Michał B. Paradowski	University of Warsaw, Poland
Patrick Paroubek	LIMSI-CNRS, France
Maciej Piasecki	Wrocław University of Technology, Poland
Stelios Piperidis	ILSP, Greece
Prokopis Prokopidis	ILSP, Greece
Gábor Prószéky	MorphoLogic, Hungary
Delyth Prys	Bangor University, UK
Michał Ptaszyński	Hokkaido University, Japan
Georg Rehm	DFKI, Germany

Tyson Roberts	Google, Japan
Piotr Rychlik	IPI PAN, Poland
Rafał Rzepka	Hokkaido University, Japan
Kevin Scannell	Saint Louis University, USA
Sanja Seljan	University of Zagreb, Croatia
Elizabeth Sherley	IIITM-Kerala, India
Zhongzhi Shi	Institute of Computing Technology, Chinese Academy of Sciences, China
Marcin Skowron	Johannes Kepler University Linz, Austria
Claudia Soria	CNR-ILC, Italy
Virach Sornlertlamvanich	NECTEC, Thailand
Janusz Taborek	Adam Mickiewicz University, Poland
Ryszard Tadeusiewicz	AGH, Poland
Dan Tufiş	RACAI, Romania
Yuzu Uchida	Hokkai-Gakuen University, Japan
Tamás Váradi	RIL, Hungary
Andrejs Vasiljevs	Tilde, Latvia
Zygmunt Vetulani	Adam Mickiewicz University, Poland
Dusko Vitas	University of Belgrade, Serbia
Aleksander Wawer	IPI PAN, Poland
Katarzyna Węgrzyn-Wolska	Efrei/Esigetel, France
Adam Wierzbicki	Polish-Japanese Academy of Information Technology, Poland
Rodrigo Wilkens	Universidade Federal do Rio Grande do Sul, Brasil
Marcin Woliński	IPI PAN, Poland
Bartosz Ziółko	AGH, Poland
Mariusz Ziółko	AGH, Poland
Andrzej Zydroń	XTM-INTL, UK

EDO Workshop Organizers

Michał Ptaszyński	Kitami Institute of Technology, Japan
Rafał Rzepka	Hokkaido University, Japan
Paweł Dybała	Jagiellonian University, Poland

EDO Workshop Program Committee

Alladin Ayesh	De Montfort University, UK
Karen Fort	Sorbonne University, France
Dai Hasegawa	Aoyama Gakuin University, Japan
Magdalena Igras-Cybulska	AGH, Poland
Yasutomo Kimura	Otaru University of Commerce, Japan
Paweł Lubarski	Poznań University of Technology, Poland
Fumito Masui	Kitami Institute of Technology, Japan
Mikołaj Morzy	Poznań University of Technology, Poland
Koji Murakami	Rakuten, USA

Noriyuki Okumura	National Institute of Technology, Akashi College, Japan
Michał B. Paradowski	University of Warsaw, Poland
Tyson Roberts	Google, Japan
Marcin Skowron	Johannes Kepler University Linz, Austria
Yuzu Uchida	Hokkai-Gakuen University, Japan
Zygmunt Vetulani	Adam Mickiewicz University, Poland
Katarzyna Węgrzyn-Wolska	Efrei/Esigetel, France
Adam Wierzbicki	Polish-Japanese Academy of Information Technology, Poland
Bartosz Ziółko	AGH, Poland

LRL Workshop Organizers

Girish Nath Jha	JNU, India
Claudia Soria	CNR-ILC, Italy

PolEval Workshop Organizers

Maciej Ogrodniczuk	Polish Academy of Sciences, Poland
Łukasz Kobyliński	Polish Academy of Sciences, Poland
Aleksander Wawer	Polish Academy of Sciences, Poland

Contents

Speech Processing

Morphology

Computational Semantics

Machine Translation

Information Retrieval and Information Extraction