

Preface

As predicted, the demand for language technology applications has kept growing. The explosion of valuable information and knowledge on the Web is accompanied by the evolution of hardware and software powerful enough to manage this flood of unstructured data. The spread of smart phones and tablets is accompanied by higher bandwidth and broader coverage of wireless Internet connectivity. We find language technology in software for search, user interaction, content production, data analytics, learning, and human communication.

Our world has changed and so have our needs and expectations. Whatever we call the new form of technology-supported life and work – information society, digital society, or knowledge society – it is not going to stay the same since it is just the transitional phase on the way to a reality in which all these contemporary mega-trends – ubiquitous computing, big data, Internet of Things, industry 4.0, artificial intelligence – have organically merged. There is only one vision in which this breathtaking universal transformation of our world will not eventually overwhelm the mental capacity and nature of the human individual and not crush the volatile cultural fabric of our civilization, a vision in which the machinery will neither dwarf nor replace their masters.

In this vision, the powerful technology will be a much appreciated extension of our limited capacities, augmenting our cognition and serving those parts of our nature that are not possessed by machines such as desires, creativity, curiosity, and passion. In such a set-up, every human individual will feel central – and actually be central. There is no way to realize this vision without human language technology. If the technology does not master the human medium for communication and thinking, the human masters will feel like aliens in their own universe.

Technology that can understand and produce human language cannot only improve our daily life and work, it can also help us to solve life-threatening problems, for example, through applications in medical research and practice that exploit research texts and patient records. Of similar importance are software systems for safety and security that help recognize and manage natural and manmade disasters and that guard technology against abuse. The instability of the political situation at the global level is evidence of the dangers and challenges connected with the new information technologies that may easily degenerate into redoubtable arms in the hands of international terrorists or totalitarian or fanatical administrations.

The challenges that lie between us and the benevolent vision of human-centered IT are the complexity and versatility of human language and thought, the range of languages, dialects, and jargons, and the different modes of using language such as speaking, writing, signing, listening, reading, and translating. But we do not only face problems. In the last few years, powerful new generic methods of machine learning have been developed that combine well with corpus work and dedicated techniques from computational linguistics. Together with the increased computing power and means for handling big data, we now have much better tools for tackling the

complexity of language. Finding appropriate combination of methods, data, and tools for each task and language creates an additional layer of challenges.

The research reported in this volume cannot cover all these challenges but each of the selected papers addresses one or several major problems that need to be solved before the vision can be turned into reality.

In the volume the reader will find the revised and in many cases substantially extended versions of 31 selected papers presented at the 6th Language and Technology Conference. The selection was made among 103 conference contributions and basically represents the preferences of the reviewers. The reviewing process was made by the international jury composed of the Program Committee members or experts nominated by them. Finally, the 90 authors of selected contributions represent research institutions from the following countries: Austria, Croatia, Ethiopia France, Germany, Hungary, India, Italy, Japan, Nigeria, Poland, Portugal, Russia, Serbia, Slovakia, Tunisia, UK, USA.¹

What the papers are about?

The papers selected for this volume belong to various fields of human language technologies and illustrate the large thematic coverage of the LTC conferences. The papers are “structured” into nine chapters. These are:

1. Speech Processing (6)
2. Morphology (2)
3. Parsing-Related Issues (4)
4. Computational Semantics (1)
5. Digital Language Resources (4)
6. Ontologies and Wordnets (3)
7. Written Text and Document Processing (7)
8. Information and Data Extraction (2)
9. Less-Resourced Languages (2)

Clustering the articles is approximate, as many addressed more than one thematic area. The ordering of the chapters does not have any “deep” significance, it approximates the order in which humans proceed in natural language production and processing: starting with (spoken) speech analysis, through morphology, (syntactic) parsing, etc. To follow this order, we start this volume with the Speech Processing chapter containing six contributions. In the paper “Boundary Markers in Spontaneous Hungarian Speech” (András Beke, Mária Gósy, and Viktória Horváth) an attempt is made at capturing objective temporal properties of boundary marking in spontaneous Hungarian, as well as at characterizing separable portions of spontaneous speech (thematic units and phrases). The second contribution concerning speech, “Adaptive Prosody Modelling for Improved Synthetic Speech Quality” (Moses E. Ekpenyong, Udoinyang G. Inyang, and EmemObong O. Udoh), is on an intelligent framework for modelling prosody in tone languages. The proposed framework is fuzzy logic based (FL-B) and is adopted to offer a flexible, human reasoning approach to the imprecise

¹ This list differs from the list of countries represented at the conference, as we identified a number of PhD students (e.g., from Iran and Mali) affiliated temporarily at foreign institutes.

and complex nature of prosody prediction. The authors of “Diacritics Restoration in the Slovak Texts Using Hidden Markov Model” (Daniel Hládek, Ján Staš, and Jozef Juhár) present a fast method for correcting diacritical markings and guessing original meaning of words from the context, based on a hidden Markov model and the Viterbi algorithm. The paper “Temporal and Lexical Context of Diachronic Text Documents for Automatic Out-Of-Vocabulary Proper Name Retrieval” (Irina Illina, Dominique Fohr, Georges Linarès, and Imane Nkairi) focuses on increasing the vocabulary coverage of a speech transcription system by automatically retrieving proper names from diachronic contemporary text documents.

In the paper “Advances in the Slovak Judicial Domain Dictation System” (Milan Rusko, Jozef Juhár, Marian Trnka, Ján Staš, Sakhia Darjaa, Daniel Hládek, Róbert Sabo, Matúš Pleva, Marian Ritomský, and Stanislav Ondáš), the authors discuss recent advances in the application of speech recognition technology in the judicial domain. The investigations on performance of Polish taggers in the context of automatic speech recognition (ASR) is the main issue of the last paper of the Speech section, “A Revised Comparison of Polish Taggers in the Application for Automatic Speech Recognition” (Aleksander Smywiński-Pohl and Bartosz Ziółko).

The Morphology section contains two papers. The first one, “Automatic Morpheme Slot Identification Using Genetic Algorithm” (Wondwossen Mulugeta, Michael Gasser, and Baye Yimam), introduces an approach to the grouping of morphemes into suffix slots in morphologically complex languages, such as Amharic, using a genetic algorithm. The second paper, “From Morphology to Lexical Hierarchies and Back” (Krešimir Šojat and Matea Srebačić), deals with language resources for Croatian – a Croatian WordNet and a large database of verbs with morphological and derivational data – and discusses the possibilities of their combination in order to improve their coverage and density of structure.

Parsing-Related Issues are presented in four papers. The chapter opens with the text “System for Generating Questions Automatically from Given Punjabi Text” (Vishal Goyal, Shikha Garg, and Umrinderpal Singh) that introduces a system for generating questions automatically for Punjabi and transforming declarative sentences into their interrogative counterparts. The next article, “Hierarchical Amharic Base Phrase Chunking Using HMM with Error Pruning” (Abeba Ibrahim and Yaregal Assabie), presents an Amharic base phrase chunker that groups syntactically correlated words at different levels (using HMM). The main goal of the authors of the paper “A Hybrid Approach to Parsing Natural Languages” (Sardar Jaf and Allan Ramsay) is to combine different parsing approaches and produce a more accurate, hybrid, grammatical rules guided parser. The last paper in the chapter is an attempt at creating a probabilistic constituency parser for Polish: “Experiments in PCFG-like Disambiguation of Constituency Parse Forests for Polish” (Marcin Woliński and Dominika Rogozińska).

The Computational Semantics chapter contains one paper, “A Method for Measuring Similarity of Books: A Step Towards an Objective Recommender System for Readers” (Adam Wojciechowski and Krzysztof Gorzynski), in which the authors propose a book comparison method based on descriptors and measures for particular properties of analyzed text.

The first of the four papers of the Digital Language Resources chapter, “MCBF: Multimodal Corpora Building Framework” (Maria Chiara Caschera, Arianna D’Ulizia,

Fernando Ferri, and Patrizia Grifoni), presents a method of dynamic generation of a multimodal corpora model as a support for human–computer dialogue. The paper “Syntactic Enrichment of LMF Normalized Dictionaries Based on the Context-Field Corpus” (Imen Elleuch, Bilel Gargouri, and Abdelmajid Ben Hamadou) describes Arabic corpora processing and proposes to the reader an approach for identifying the syntactic behavior of verbs in order to enrich the syntactic extension of the LMF-normalized Arabic dictionaries. A multilingual annotation toolkit is presented in the paper “An Example of a Compatible NLP Toolkit” (Krzysztof Jassem and Roman Grundkiewicz). The article “Polish Coreference Corpus” (Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawislawska) describes a composition, annotation process and availability of the Polish Coreference Corpus.

The Ontologies and Wordnets part comprises three papers. The contribution “GeoDomainWordNet: Linking the Geonames Ontology to WordNet” (Francesca Frontini, Riccardo Del Gratta, and Monica Monachini) demonstrates a wordnet generation procedure consisting in transformation of an ontology of geographical terms into a WordNet-like resource in English and its linking to the existing generic wordnets of English and Italian. The second article, “Building Wordnet Based Ontologies with Expert Knowledge” (Jacek Marciniak) presents the principles of creating wordnet-based ontologies that contain general knowledge about the world as well as specialist expert knowledge. In “Diagnostic Tools in plWordNet Development Process” (Maciej Piasecki, Łukasz Burdka, Marek Maziarsz, and Michał Kaliński), the third of the contributions in this chapter, the authors describe formal, structural, and semantic rules for seeking errors within plWordNet, as well as a method of automated induction of the diagnostic rules.

The largest chapter, Written Text and Document Processing, presents seven contributions of which the first is “Simile or Not Simile?: Automatic Detection of Metonymic Relations in Japanese Literal Comparisons” (Paweł Dybala, Rafał Rzepka, Kenji Araki, and Kohichi Sayama). Its authors propose how to automatically distinguish between two types of formally identical expressions in Japanese: metaphorical similes and metonymical comparisons. The issues of diacritic error detection and restoration – tasks of identifying and correcting missing accents in text – are addressed in “Spanish Diacritic Error Detection and Restoration—A Survey” (Mans Hulden and Jerid Franc com). The article “Identification of Event and Topic for Multi-document Summarization” (Fumiyo Fukumoto, Yoshimi Suzuki, Atsuhiko Takasu, and Suguru Matsuyoshi) is a contribution in which the authors investigate continuous news documents and conclude with a method for extractive multi-document summarization. The next paper, “Itemsets-Based Amharic Document Categorization Using an Extended *A Priori* Algorithm” (Abraham Hailu and Yaregal Assabie), presents a system that categorizes Amharic documents based on the frequency of itemsets obtained from analyzing the morphology of the language. In the paper “NERosetta for the Named Entity Multi-lingual Space” (Cvetana Krstev, Anđelka Zečević, Duško Vitas, and Tita Kyr-iacopoulou) the authors present a Web application, NERosetta, that can be used to compare various approaches to develop named entity recognition systems. In the study “A Hybrid Approach to Statistical Machine Translation Between Standard and Dialectal Varieties” (Friedrich Neubarth, Barry Haddow, Adolfo Hernández Huerta,

and Harald Trost), the authors describe the problem of translation between the standard Austrian German and the Viennese dialect. From the last paper of the Text Processing chapter, “Evaluation of Uryupina’s Coreference Resolution Features for Polish” (Bartłomiej Nitoń), the reader will get familiar with an evaluation of a set of surface, syntactic, and anaphoric features proposed for coreference resolution in Polish texts.

The Information and Data Extraction chapter contains two studies. In the first one, “Aspect-Based Restaurant Information Extraction for the Recommendation System” (Ekaterina Pronoza, Elena Yagunova, and Svetlana Volskaya), a method for Russian reviews corpus analysis aimed at future information extraction system development is proposed. In the second article, “A Study on Turkish Meronym Extraction Using a Variety of Lexico-Syntactic Patterns” (Tuğba Yıldız, Savaş Yıldırım, and Banu Diri), lexico-syntactic patterns to extract meronymy relation from a huge corpus of Turkish are presented.

The Less-Resourced Languages are considered of special interest for the LTC community and were presented at the LRL conference workshop. We decided to place the two selected LRL papers in a separate chapter, the last in this volume. The first paper, “A Phonetization Approach for the Forced-Alignment Task in SPPAS” (Brigitte Bigi), presents a generic approach for text phonetization, concentrates on the aspects of phonetizing unknown words, and is tested for less resourced languages, for example, Vietnamese, Khmer, and Pinyin for Taiwanese. The final paper in the volume, “POS Tagging and Less Resources Languages Individuated Features in CorpusWiki” (Maarten Janssen), explores the hot topic of the lack of corpora for LRL languages and proposes a Wikipedia-based solutions with particular attention paid to the POS annotation.

We wish you all interesting reading.

March 2016

Zygmunt Vetulani
Hans Uszkoreit

Medical Computer Vision: Algorithms for Big Data
International Workshop, MCV 2015, Held in Conjunction
with MICCAI 2015, Munich, Germany, October 9, 2015,
Revised Selected Papers
Menze, B.; Langs, G.; Montillo, A.; Kelm, B.M.; Müller, H.;
Zhang, S.; Cai, W.; Metaxas, D. (Eds.)
2016, XV, 182 p. 70 illus., Softcover
ISBN: 978-3-319-42015-8