PACLIC 2015

29th Pacific Asia Conference on Language, Information and Computation

Proceedings of PACLIC 2015: Oral Papers

Program chair:

Hai Zhao

30 October - 1 November, 2015

Shanghai, China

Sponsors

Department of Computer Science and Engineering, Shanghai Jiao Tong University
Chinese Information Processing Society of Chine (CIPS)
LY Education Technology
Shanghai Computer Federation Artificial Intelligence Committee (SCFAIC)

Preface

Distinguished scholars and colleagues:

The 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 29) is organized by the Department of Computer Science and Engineering, Shanghai Jiao Tong University, October 30 - November 1, 2015. The PACLIC series of conferences emphasize the synergy of theoretical analysis and processing of language, and provide a forum for researchers in different fields of language study in the Pacific-Asia region to share their findings and interests in the formal and empirical study of languages. For the past years since its establishment, the PACLIC conferences have gained more and more interests and participations from linguistic researchers, as evidenced by the increasing number of papers and by the wider range of topics. Organized under the auspices of the PACLIC Steering Committee, it is the latest installment of our long standing collaborative efforts among theoretical and computational linguists in the Pacific-Asia region.

PACLIC conference has received an overwhelming response of 221 papers from 104 countries or regions namely China, Japan, Korea, Hong Kong, Taiwan, France, Israel, New Zealand, Thailand, Tunisia, UK, Vietnam, Algeria, Egypt, Germany, India, Ireland, Singapore (87.50% from 10 regions in Asia, 6.73% from 4 regions in Europe, 3.85% from Africa, 1.92% from New Zealand). To ensure that all accepted papers meet the high quality standard of the PACLIC conference, each submission was reviewed by 2-4 reviewers. As a result, only approximately 63 (28.5%) of top-notch academic papers were accepted for oral presentations and 41 (18.5%) for poster sessions. From these accepted papers, 104 (47.0%) papers were presented and published in this proceedings.

PACLIC-29 thanks for tremendous efforts and contributions from several parties. We congratulate the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Chinese Information Processing Society of Chine (CIPS), LY Education Technology and Shanghai Computer Federation Artificial Intelligence Committee (SCFAIC) for their collaboration towards this significant achievement. We would like to take this opportunity to thank our keynote and invited speakers, namely Dr. Sumita Eiichiro from the National Institute of Information and Communications Technology (NICT, Japan), Professor Guodong Zhou from Soochow University, Dr. Philippe Blache from National Center for Scientific Research (CNRS, France), Professor Renqiang Wang from Sichuan International Studies University and Assistant Professor Yao Yao from Hong Kong Polytechnic University. We are also overwhelmed with a sense of gratitude for the presenters and colleagues for donating your valuable time to attend and enrich this conference.

PC Chair
Hai Zhao

Program Committee:

Wirote Aroonmanakun	Chulalongkorn University
Hailong Cao	HIT
Hee-Rahk Chae	Hankuk University of Foreign Studies
Wanxiang Che	Harbin Institute of Technology
Doris Chen	National Taiwan Normal University
Kuang-Hua Chen	National Taiwan University
Wenliang Chen	Soochow University
Eng-Siong Chng	Nanyang Technological University
Siaw-Fong Chung	National Chengchi University
Beatrice Daille	Laboratoire d'Informatique de Nantes Atlantique
Jing Ding	The Hong Kong Polytechnic University
Amanda Ding	
Chen-Chun E	The Hong Kong Polytechnic University
Guohong Fu	Heilongjiang University
Helena Hong Gao	Nanyang Technological University
Wei Gao	Qatar Computing Research Institute
Yasunari Harada	Waseda University
Choochart Haruechaiyasak	National Electronics and Computer Technology Center (NECTEC)
Munpyo Hong	Sungkyunkwan Univ.
Shu-Kai Hsieh	National Taiwan Normal University
Jong-Bok Kim	Kyung Hee University
Richard Kim	Fudan University
Valia Kordoni	Humboldt University Berlin
Oi-Yee Kwong	The Chinese University of Hong Kong
Huei-Ling Lai	National Chengchi University
Gina-Anne Levow	University of Washington

Shoushan Li	Soochow University
Chao-Lin Liu	National Chengchi University
Jyi-Shane Liu	National Chengchi University
Qing Ma	Ryukoku University
Takafumi Maekawa	Fuculty of Sociology, Ryukoku University
Yuji Matsumoto	Nara Institute of Science and Technology
Mathieu Morey	LPL, Université d'Aix-Marseille & LMS, Nanyang Technological University
Natchanan Natpratan	Department of Linguistics,Faculty of Humanities, Kaset-sart University
Ponrudee Netisopakul	KMAKE LAB
Makoto Okada	Osaka Prefecture University
Chutamanee Onsuwan	Faculty of Liberal Arts, Thammasat University
Ryo Otoguro	Faculty of Law, Waseda University
Jong C. Park	KAIST
Pittayawat Pittayaporn	Chulalongkorn University
Laurent Prévot	Laboratoire Parole et Langage
Haoliang Qi	Heilongjiang Institute of Technology
LongQiu	Institute for Infocomm Research
Bali Ranaivo-Malançon	MALINDO
Samira Shaikh	State University of New York - University at Albany
Melanie Siegel	Hochschule Darmstadt
Pornsiri Singhapreecha	Thammasat University
Simon Smith	Coventry University
Virach Sornlertlamvanich	SIIT, Thammasat University
Keh-Yih Su	Institute of Information Science, Academia Sinica
Weiwei Sun	Peking University
Thepchai Supnithi	NECTEC
Yuen-Hsien Tseng	National Taiwan Normal University
Aline Villavicencio	Universidade Federal do Rio Grande do Sul

Rui Wang	Shanghai Jiao Tong University
Jiajuan Xiong	The University of Hong Kong
Ruifeng Xu	Harbin Institute of Technology
Hongzhi Xu	The Hong Kong Polytechnic University
Cheng-Zen Yang	Dept. of Computer Science and Engineering, Yuan Ze University
Kei Yoshimoto	Tohoku University
Liang-Chih Yu	Yuan Ze University
Jiajun Zhang	Institute of Automation Chinese Academy of Sciences
Qi Zhang	Fudan University
Yu Zhou	CIP, NLPR, CASIA
Ming Zhou	Microsoft
Conghui Zhu	Harbin Institute of Tecnology, China
Michael Zock	CNRS-LIF

Keynote talk

Sumita Eiichiro (NICT, Japan)

Talk Title:Research Activities for Translating Asian Languages

Abstract: This talk will introduce automatic translation projects for Asian languages, wherein we intend to seek greater cooperation.

First, a worldwide speech translation consortium, Universal Speech Translation Advanced Research (U-STAR), is introduced. Speech translation involves the integration of three elements: speech recognition, machine translation, and speech synthesis; therefore, to build a speech translation system that includes many languages including Asian languages, it is a good idea to cooperate with other laboratories that specialize in the languages concerned. The consortium now comprises 32 institutes from 27 different countries/regions. The collaboration has improved the accuracy of the integrated systems and has created new forms of integration. U-STAR is always open and welcomes new participants.

Second, we introduce two projects related to the translation of Asian languages: the Workshop on Asian Translation (WAT) and the Asian Language Treebank (ALT). WAT is an open evaluation campaign focusing on translation among Asian languages. We will outline the workshops conducted in past two years' and touch on our plan for next year. ALT is currently a start-up project that will undertake the task of building a treebank of Asian languages. This will be a valuable language resource, not only as a parser for each language but also as an accurate translation system from one language to another.

Third, we discuss the Global Communication Program (GCP), a Japanese government project announced in April 2014 to develop a multi-lingual speech translation system to bridge the language barrier during the Olympic Games in 2020. It aims to provide real-time machine translation services, by using National Institute of Information and Communications Technology's (NICT) translation technology, in day-to-day situations to help foreigners who may feel hesitant about coming to Japan. It will cover 10 languages, including Asian ones, e.g., Thai, Vietnamese, Indonesian, and Myanmar. At NICT, public and private entities have already begun working together as part of a nationwide collaboration. This talk will explain the current status and future vision.

Finally, we touch on NICT's recent research topics, including an approach to high-quality patent translation and new ideas on neural translation.

Zhou Guodong (Soochow University)

Talk: Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure

Abstract: It is well-known that interpretation of a text requires understanding of its rhetorical relation hierarchy since discourse units rarely exist in isolation. Such discourse structure is fundamental to discourse understanding and many text-based applications. In this talk, we propose a Connective-driven Dependency Tree (CDT) scheme to represent the discourse rhetorical structure in Chinese language, with elementary discourse units as leaf nodes and connectives as non-leaf nodes, largely motivated by the Penn Discourse Treebank and the Rhetorical Structure Theory. In particular, connectives are employed to directly represent the hierarchy of the tree structure and the rhetorical relation of a discourse, while the nuclei of discourse units are globally determined with reference to the dependency theory. Guided by the CDT scheme, we manually annotate a Chinese Discourse Treebank (CDTB) of 500 documents. Preliminary evaluation justifies the appropriateness of the CDT scheme to Chinese discourse analysis and the usefulness of our manually annotated CDTB corpus.

Guodong Zhou is a distinguished professor (Grade II) and a member of the university academic committee in Soochow University, China. He obtained his Ph.D. degree from National University of Singapore in 1999. He joined the Institute of Infocomm Research, Singapore in 1999 and Soochow University in 2006. His research interests include natural language processing and artificial intelligence with more and more focus on fundamental language issues.

Prof Zhou has published over 100 papers in leading NLP and AI conferences and journals such as ACL/EMNLP/COLING/AAAI/IJCAI with over 4000 citations (Google Scholar). He was/is on the editorial board of several international journals, such as Computational Linguistics, ACM TALIP and Chinese Journal of Software, and is a regular PC member of the major conferences in NLP and AI.

Since 2006, Prof Zhou has established the Suda NLP lab with now 16 staff members, including 7 full professors and 7 associate professors.

Philippe Blache (CNRS)

Talk Title:New approaches to sentence processing: a cognitive perspective

Abstract: Sentence processing is usually considered as an incremental mechanism: each new word is integrated into a structure under construction that can be interpreted compositionally. In this architecture, understanding a sentence comes to a step-by-step building of the meaning. I will present in this talk different elements challenging this approach. Starting from works in linguistics, psycholinguistics and natural language processing, we will see that language processing by human can be, depending on the situation, very superficial and incomplete. A more realistic language processing architecture would therefore have to integrate into a unique model different levels of processing: one which is superficial, relying on the recognition of large units with a strong cohesion; and another consisting in a classical incremental word by word integration. This organization corresponds to a double level shallow-and-deep parsing process.

Table of Contents

Two-level Word Class Categorization Model in Analytic Languages and Its Implications for POS Tagging in Modern Chinese Corpora
Renqiang Wang and Changning Huang.....1

A Review of Corpus-based Statistical Models of Language Variation
Yao Yao.....11

Translation of Unseen Bigrams by Analogy Using an SVM Classifier
Hao Wang, Lu Lyu and Yves Lepage.....16

Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus
Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas and Kamel Smail.....26

Surrounding Word Sense Model for Japanese All-words Word Sense Disambiguation
*Kanako Komiya, Yuto Sasaki, Hajime Morita, Minoru Sasaki, Hiroyuki Shinnou and Yoshiyuki Kotan*i.....35

Computing Semantic Text Similarity Using Rich Features
Yang Liu, Chengjie Sun, Lei Lin, Xiaolong Wang and Yuming Zhao.....44

Mechanical Turk-based Experiment vs Laboratory-based Experiment: A Case Study on the Comparison of Semantic Transparency Rating Data
Shichang Wang, Chu-Ren Huang, Yao Yao and Angel Chan.....53

Discourse Relation Recognition by Comparing Various Units of Sentence Expression with Recursive Neural Network
Atsushi Otsuka, Toru Hirano, Chiaki Miyazaki, Ryo Masumura, Ryuichiro Higashinaka, Toshiro Makino and Yoshihiro Matsuo.....63

Bidirectional Long Short-Term Memory Networks for Relation Classification
Shu Zhang, Dequan Zheng, Xinchen Hu and Ming Yang.....73

Distant Supervision for Entity Linking
Miao Fan, Qiang Zhou and Thomas Fang Zheng.....79

Toward Algorithmic Discovery of Biographical Information in Local Gazetteers of Ancient China
Chao-Lin Liu, Chih-Kai Huang, Hongsu Wang and Peter K. Bol.....87

Fast and Large-scale Unsupervised Relation Extraction
Sho Takase, Naoaki Okazaki and Kentaro Inui.....96

Reducing Lexical Features in Parsing by Word Embeddings
Hiroya Komatsu, Ran Tian, Naoaki Okazaki and Kentaro Inui.....106

High-order Graph-based Neural Dependency Parsing
Zhisong Zhang and Hai Zhao.....114

A Dynamic Syntax Modelling of Postposing in Japanese Narratives
Tohru Seraku.....124

Unsupervised and Lightly Supervised Part-of-Speech Tagging Using Recurrent Neural Networks
Othman Zennaki, Nasredine Semmar and Laurent Besacier.....133

Identifying Prepositional Phrases in Chinese Patent Texts with Rule-based and CRF Methods
Hongzheng Li and Yaohong Jin.....143

Japanese Sentiment Classification with Stacked Denoising Auto-Encoder using Distributed Word Representation
Peinan Zhang and Mamoru Komachi.....150

Is Wikipedia Really Neutral? A Sentiment Perspective Study of War-related Wikipedia Articles since 1945
Yiwei Zhou, Alexandra Cristea and Zachary Roberts.....160

A Comprehensive Filter Feature Selection for Improving Document Classification
Nguyen Hoai Nam Le and Bao Quoc Ho.....169

Sentiment Analyzer with Rich Features for Ironic and Sarcastic Tweets
Piyoros Tungthamthiti, Enrico Santus, Hongzhi Xu, Chu-Ren Huang and Shirai Kiyoaki.....178

Thai Stock News Sentiment Classification using Wordpair Features
Ponrudee Netisopakul and Apinan Chattupan.....188

Sentiment Classification of Arabic Documents: Experiments with multi-type features and ensemble algorithms
Amine Bayoudhi, Hatem Ghorbel and Lamia Hadrich Belguith.....196

The Invertible Construction in Chinese
Cong Yan, Lian-Hee Wee and Chu-Ren Huang.....206

Pan's (2001) puzzle revisited
Hyunjun Park.....212

English Right Dislocation
Kohji Kamada.....221

A Comparative Study on Mandarin and Cantonese Resultative Verb Compounds
Helena Yan Ping Lau and Sophia Yat Mei Lee.....231

Complex-NP Islands in Korean: An Experimental Approach
Yong-Hun Lee and Yeonkyung Park.....240

Two Types of Multiple Subject Constructions (MSCs) in Korean
Ji-Hye Kim, Eunah Kim and James Yoon.....250

A Large-scale Study of Statistical Machine Translation Methods for Khmer Language
Ye Kyaw Thu, Vichet Chea, Andrew Finch, Masao Utiyama and Eiichiro Sumita....259

English to Chinese Translation: How Chinese Character Matters
Rui Wang, Hai Zhao and Bao-Liang Lu.....270

Well-Formed Dependency to String translation with BTG Grammar
Xiaoqing Li, Kun Wang, Dakun Zhang and Jie Hao.....281

Large-scale Dictionary Construction via Pivot-based Statistical Machine Translation with Significance Pruning and Neural Network Features
Raj Dabre, Chenhui Chu, Fabien Cromieres, Toshiaki Nakazawa and Sadao Kurohashi.....289

Annotation and Classification of French Feedback Communicative Functions
Laurent Prévot, Jan Gorisch and Sankar Mukherjee.....298

Automatic conversion of sentence-end expressions for utterance characterization of dialogue systems
Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, Toshiro Makino and Yoshihiro Matsuo.....307

Auditory Synaesthesia and Near Synonyms: A Corpus-Based Analysis of sheng1 and yin1 in Mandarin Chinese
Qingqing Zhao, Chu-Ren Huang and Hongzhi Xu.....315

System Utterance Generation by Label Propagation over Association Graph of Words and Utterance Patterns for Open-Domain Dialogue Systems

Hiroshi Tsukahara and Kei Uchiumi.....323

The Cross‐modal Representation of Metaphors
Yutung Chang and Kawai Chui.....332

Writing to Read: the Case of Chinese
Qi Zhang and Ronan Reilly.....341

Design of a Learner Corpus for Listening and Speaking Performance
Katsunori Kotani and Takehiko Yoshimi.....351

Understanding Infants' Language Development in Relation to Levels of Consciousness: An Approach in Building up an Agent-based Model
Helena Hong Gao and Can Guo.....359

Pivot-Based Topic Models for Low-Resource Lexicon Extraction
John Richardson, Toshiaki Nakazawa and Sadao Kurohashi.....369

A Corpus-Based Study of zunshou and Its English Equivalents
Ying Liu.....378

Self Syntactico-Semantic Enrichment of LMF Normalized Dictionaries
Imen Elleuch, Bilel Gargouri and Abdelmajid Ben Hamadou.....387

When Embodiment Meets Generative Lexicon: The Human Body Part Metaphors in Sinica Corpus
Ren-Feng Duann and Chu-Ren Huang.....396

Degree Variables by Choose Degree in Izyooni 'than'-Clauses
Toshiko Oda.....404

Not Voice but Case Identity in VP Ellipsis of English
Myungkwan Park and Sunjoo Choi.....413

A Statistical Modeling of the Correlation between Island Effects and Working-memory Capacity for L2 Learners
Euhee Kim and Myungkwan Park.....422

De-verbalization and Nominal Categories in Mandarin Chinese: A corpus-driven study in both Mainland Mandarin and Taiwan Mandarin
Jiajuan Xiong and Chu-Ren Huang.....431

Zero Object Resolution in Korean
Arum Park, Seunghee Lim and Munpyo Hong.....439

An Improved Hierarchical Word Sequence Language Model Using Directional Information
Xiaoyi Wu and Yuji Matsumoto.....449

Neural Network Language Model for Chinese Pinyin Input Method Engine
Shenyuan Chen, Hai Zhao and Rui Wang.....455

Real-time Detection and Sorting of News on Microblogging Platforms
Wenting Tu, David Cheung, Nikos Mamoulis, Min Yang and Ziyu Lu.....462

Trouble information extraction based on a bootstrap approach from Twitter
Kohei Kurihara and Kazutaka Shimada.....471

Using Twitter Data to Infer Personal Values of Japanese Consumers
Yinjun Hu and Yasuo Tanida.....480

Distant-supervised Language Model for Detecting Emotional Upsurge on Twitter
Yoshinari Fujinuma, Hikaru Yokono, Pascual Martínez-Gómez and Akiko Aizawa..488

Hybrid Method of Semi-supervised Learning and Feature Weighted Learning for Domain Adaptation of Document Classification
Hiroyuki Shinnou, Liying Xiao, Minoru Sasaki and Kanako Komiya.....496

Paraphrase Detection Based on Identical Phrase and Similar Word Matching
Hoang-Quoc Nguyen-Son, Yusuke Miyao and Isao Echizen.....504

Multi-aspects Rating Prediction Using Aspect Words and Sentences
Takuto Nakamuta and Kazutaka Shimada.....513

Understanding Rating Behaviour and Predicting Ratings by Identifying Representative Users
Rahul Kamath, Masanao Ochi and Yutaka Matsuo.....522

Cross-lingual Pseudo Relevance Feedback Based on Weak Relevant Topic Alignment
Xuwen Wang, Qiang Zhang, Xiaojie Wang and Junlian Li 529

Corpus annotation with a linguistic analysis of the associations between event mentions and spatial expressions
Jin-Woo Chung, Jinseon You and Jong C. Park.....535

Recognizing Complex Negation on Twitter
Junta Mizuno, Canasai Kruengkrai, Kiyonori Ohtake, Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer and Kentaro Inui.....544

Topic Model for Identifying Suicidal Ideation in Chinese Microblog
Xiaolei Huang, Xin Li, Tianli Liu, David Chiu, Tingshao Zhu and Lei Zhang.....553

Predicting Sector Index Movement with Microblogging Public Mood Time Series on Social Issues
Yujie Lu, Jinlong Guo, Kotaro Sakamoto, Hideyuki Shibuki and Tatsunori Mori.....563