Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)

Edited by Helena Hong Gao and Minghui Dong

16–18 December
Nanyang Technological University, Singapore

ISBN 978-4-905166-02-3

Published by the Institute for Digital Enhancement of Cognitive Development, Waseda University.

Acknowledgments

PACLIC 25, *the 25th Pacific Asia Conference on Language, Information and Computation*, is organized by the Centre for Liberal Arts and Social Sciences (CLASS), Nanyang Technological University (NTU) and Chinese and Oriental Languages Information Processing Society (COLIPS), under the auspices of PACLIC Steering Committee. We gratefully acknowledge the support from the Lee foundation for students from developing countries.

The workshop *The Fourth International Symposium on Languages in Biology and Medicine (LBM 2011)* is organized by the LBM committee and sponsored by the Systems Biomedical Informatics National Core Research Centre (Korea).

Preface

Welcome to Singapore and welcome to Nanyang Technological University (NTU)! This is the third time that the PACLIC conference comes to Singapore after its two previous visits in 1998 and 2003. This year, the PACLIC 25 is organized by Nanyang Technological University (NTU) and the Chinese and Oriental Languages Information Processing Society (COLIPS).

The PACLIC conference has a long history, dating back to 1982. Over the years, the conference has developed into one of the leading conferences in the research community. Like previous PACLIC conferences, PACLIC 25 has received a wide range of interesting research papers in the fields of theoretical and computational linguistics. The specific research topics that the papers focus on can be classified into the following: cognitive linguistics, corpus linguistics, discourse analysis, formal grammar theory, grammar and parsing, human and machine language processing, information extraction, information retrieval, language acquisition, language resources, language technology and its application, machine translation, morphology, natural language processing, phonology, pragmatics, semantics, speech processing, syntax, and typology.

The paper submissions we received are from 21 countries or regions, including Australia, Canada, China, Czech Republic, Denmark, France, Germany, Hong Kong, India, Indonesia, Italy, Japan, Kingdom of Saudi Arabia, Korea, Malaysia, Myanmar, Philippines, Singapore, Taiwan, Thailand, and the United States of America. To ensure that all papers accepted meet the high quality standard of the PACLIC conference, we had each submission reviewed by three to four experts in the field. As a result, of the 125 valid submitted papers, we accepted 50 (40%) for oral presentations and 25 (20%) for poster sessions.

With such a wide variety of research topics to be presented and participants from all over the world, we believe that this conference will be a truly stimulating scholarly forum. New research findings, new approaches, new ideas for tackling technical challenges and further explorations of a particular research topic, and new information on current research trends will be discussed and shared both within and across disciplines. This scenario is the sign of a most successful conference, which will benefit all the participants as well as the development of the research fields.

A successful conference is the result of many people's efforts and contributions. The quality papers included in the program are the results of our participants' on-going contributions as well as the tremendous efforts made by the program committee members in their paper reviews. Besides the oral and poster paper presentations, the program is enriched by the keynote and invited speakers, Professor Paul Kay from the University of California at Berkeley, Professor Key-Sun Choi from KAIST, and Dr Jian Su, Senior Scientist from the Institute for Infocomm Research. We have also scheduled a panel discussion on multilingualism by several distinguished experts. The panel will be chaired by Prof James German from Nanyang Technology University and the discussion will be led by presentations from Prof Chu-Ren Huang from Hong Kong Polytechnic University, Prof Laurent Prevot from Aix en Provence, and Prof Francis Bond from Nanyang Technology University. Their expertise in their respective fields will provide us with new insights for the current and future research. On behalf of the program committee, we express our heartfelt thanks to them all.

We would like to thank the steering committee for their guidance, and the local organizing committee chaired by Professor Francis Bond at Nanyang Technological University and Dr. Min Zhang at the Institute for Infocomm Research, Singapore, for their dedicated efforts and their excellent coordination with all parties, which has ensured that this conference will be a successful event.

Finally, we wish that you will all enjoy the conference presentations, discussions, and exchanges between old and new friends on the beautiful campus of NTU.

Helena Hong Gao (Nanyang Technological University, Singapore)
Minghui Dong (Institute for Infocomm Research, Singapore)

PACLIC 25 Program Committee Chairs

Organizers

Steering Committee

Jae-Woong Choe, Korea University
Yasunari Harada, Waseda University
Chu-Ren Huang, Hong Kong Polytechnic University
Kim Teng Lua, Chinese and Oriental Languages Information Processing Society
Rachel Edita Roxas, De La Salle University-Manila
Maosong Sun, Tsinghua University
Benjamin T'sou, City University of Hong Kong

Program Committee Chairs

Helena Hong Gao, Nanyang Technological University
Minghui Dong, Institute for Infocomm Research (I^2R)

Co-Chairs

Olivia Kwong, City University of Hong Kong
Shu-Kai Hsieh, National Taiwan Normal University
Donghong Ji, Wuhan University
Seungho Nam, Seoul National University
Rachel Edita Roxas, De La Salle University-Manila

Members

Doug Arnold, University of Essex
Wirote Aroonmanakun, Chulalongkorn University
Timothy Baldwin, University of Melbourne
Francis Bond, Nanyang Technological University
Hee-Rahk Chae, Hankuk University of Foreign Studies
Jason Chang, National Tsing Hua University
Jing-Shin Chang, National Chi-Nan University
Hsin-Hsi Chen, National Taiwan University
Eng Siong Chng, Nanyang Technological University
Siaw-Fong Chung, National Chengchi University
Beatrice Daille, University of Nantes
Danilo Dayag, De La Salle University
Shirley Dita, De La Salle University
Minghui Dong, Institute for Infocomm Research
Alex Fang, City University of Hong Kong
Maria Flouraki, SOAS, University London
Guohong Fu, Heilongjiang University
Helena Hong Gao, Nanyang Technological University
Wei Gao, Chinese University of Hong Kong
Yi Guan, Harbin Institute of Technology
Yasunari Harada, Waseda University
Lin He, Wuhan University
Munpyo Hong, Sungkyunkwan University
Shu-Kai Hsieh, National Taiwan Normal University
Xuanjing Huang, Fudan University
Kentoro Inui, Nara Institute of Science and Technology
Donghong Ji, Wuhan University
Mun-Yen Kan, National University of Singapore
Jong-Bok Kim, Kyung Hee University
Chunyu Kit, City University of Hong Kong
Valia Kordoni, Saarland University
Sadao Kurohashi, Kyoto University
Olivia Kwong, City University of Hong Kong
Tom Lai, City University of Hong Kong

Paul Law, City University of Hong Kong
Hae-yun Lee, Hankuk University of Foreign Studies
Yae-Sheik Lee, Kyungpook National University
Alessandro Lenci, University of Pisa
Gina-Anne Levow, University of Manchester
Haizhou Li, Institute for Infocomm Research
Mu Li, Microsoft
Wenjie Li, The Hong Kong Polytechnic University
Qun Liu, Chinese Academy of Sciences
Qing Ma, Ryukoku University
Yanjun Ma, Baidu
Takafumi Maekawa, Hokusei Gakuen University Junior College
Ruli Manurung, University of Indonesia
Yuji Matsumoto, Nara Institute of Science and Technology
Yoshiki Mori, University of Tokyo
Vincent Ng, University of Texas at Dallas
Jian-Yun Nie, Universit de Montral
Toshiyuki Ogihara, University of Washington
David Y. Oshima, Nagoya University
Ryo Otoguro, Waseda University
Haihua Pan, City University of Hong Kong
Prashant Pardeshi, National Institute for Japanese Language and Linguistics
Ceile Paris, Commonwealth Scientific and Industrial Research Organisation
Jong C. Park, Korea Advanced Institute of Science and Technology
Laurent Prevot, Universite de Provence
Long Qiu, Institute for Infocomm Research
Rachel Edita Roxas, De La Salle University-Manila
Rajeev Sangal, International Institute of Information Technology
Samira Shaikh, State University of New York - University at Albany
Sachiko Shudo, Waseda University
Pornsiri Singhapreecha, Thammasat University
Virach Sornlertlamvanich, Thai Computational Linguistics Laboratory, NICT
Andrew Spencer, University of Essex
I-Wen Su, National Taiwan University
Jian Su, Institute for Infocomm Research
Keh-Yih Su, Behavior Design Corporation
Zhifang Sui, Peking University
Le Sun, Chinese Academy of Sciences
Thanaruk Theeramunkong, Sirindhorn International Institute of Technology
Takenobu Tokunaga, Tokyo Institute of Technology
Shu-Chuan Tseng, Academia Sinica
Josef van Genabith, Dublin City University
Aline Villavicencio, Universidade Federal do Rio Grande do Sul
Haifeng Wang, Baidu
Houfeng Wang, Peking University
Hui Wang, National University of Singapore
Chung-Hsien Wu, National Cheng Kung Univeristy
Dekai Wu, The Hong Kong University of Science and Technology
Jiun-Shiung Wu, National Chiayi University
Deyi Xiong, Institute for Infocomm Research
Jae II Yeom, Hongik University
Satoru Yokoyama, Tohoku University
Aesun Yoon, Pusan National University
Hao Yu, Fujitsu China Research Center
Min Zhang, Institute for Infocomm Research
Yujie Zhang, Beijing Jiaotong University
Jun Zhao, Chinese Academy of Sciences
Tiejun Zhao, Harbin Institute of Technology
Guodong Zhou, Soochow University
Qiang Zhou, Tsinghua University
Jingbo Zhu, Northeast University, China

Michael Zock, Laboratoire d'Informatique Fondamentale de Marseille, C.N.R.S.
Chengqing Zong, Chinese Academy of Sciences

Local Organizing Committee

Chairs

Francis Bond, Nanyang Technological University
Min Zhang, Institute for Infocomm Research (I^2R)

Members

Ma Bin, COLIPS, I^2R (Sponsorship)
Gao Cong, Nanyang Technological University (Publications)
Chuancong Gao, Nanyang Technological University (Publications)
Petter Haugereid, Nanyang Technological University (Finance)
Jung-Jae Kim, Nanyang Technological University (Workshops)
Tse Min Lua, COLIPS (Committee Member)
Nicole Wong, COLIPS (Committee Member)
Ryo Otoguro, Waseda University (PACLIC 24 Liaison)
Ruli Manurung, University of Indonesia (PACLIC 26 Liaison)

Table of Contents

Invited Talk 1

A Theory of Idioms

Paul Kay, ICSI, UCLA, USA

Abstract

The beginnings of a theory of multi-word expressions (MWEs, AKA idioms) will be sketched. The basic insight on which this approach to MWEs is based holds that many or all MWEs can be analyzed as composed exclusively of special idiom lexemes, frequently homophonous with canonical (non-idiom) lexemes; these interact with the familiar phrasal constructions of the grammar to license idiomatic expressions. The original insight is due to Nunberg et al. (1994). The framework in which the theory is expressed is Sign-Based Construction Grammar (SBCG), which combines some of the insights of Berkeley Construction Grammar (Fillmore et al. 1988, Kay and Fillmore 1999) with recent developments in HPSG (Sag 2010, to appear). Part of the purpose of the talk will be to introduce some of the workings of SBCG via informal exemplification.

Biodata

Paul Kay is Emeritus Professor of Linguistics at the University of California, Berkeley. His research interests, within the broad area of cognitive science, center around language, its structure, and its relation to thought and perception. Much of his research concerns either color naming or grammatical structure.

Invited Talk 2

Linked open data: for NLP or by NLP?

Key-Sun Choi, KAIST, Korea

Abstract

If we call Wikipedia or Wiktionary as "web knowledge resource", the question is about whether they can contribute to NLP itself and furthermore to the knowledge resource for knowledge-leveraged computational thinking. Comparing with the structure inside WordNet from the view of its human-encoded precise classification scheme, such web knowledge resource has category structure based on collectively generated tags and structures like infobox. They are called also as "Collectively Generated Content" and its structuralized content based on collective intelligence. It is heavily based on linking among terms and we also say that it is one member of linked data. The problem is in whether such collectively generated knowledge resource can contribute to NLP and how much it can be effective.

The more clean primitives of linked terms in web knowledge resources will be assumed, based on the essential property of Guarino (2000) or intrinsic property of Mizoguchi (2004). The number of entries in web knowledge resources increases very fast but their inter-relationships are indirectly calculated by their link structure. We can imagine that their entries could be mapped to one of instances under some structure of primitive concepts, like synsets of WordNet. Let's name such primitives to be "intrinsic tokens" that are derived from collectively generated knowledge resource under the principles of intrinsic properties. The procedure could be approximately proven and it will be a kind of statistical logic. We then go to the issues about what area of NLP can be solved by the so-called intrinsic tokens and their relations, a resultant approximately generated primitives.

Can NLP contribute to the user generation process of content? Consider the structure of infobox in Wikipedia more closely. It will be discussed about how NLP can help the population of relevant entries, like the social network mechanism for multi-lingual environment and information extraction purpose.

The traditional NLP starts from words in text but now also works have been undergoing on the web corpus with hyperlinks and html markups. In web knowledge resources, the words and chunks have underlying URIs, a kind of annotation. It signals a new paradigm of NLP.

Biodata

Dr. Key-Sun Choi is a tenured full professor and the head of Computer Science Department, KAIST, Korea. He founded and directed National Language Resource Bank and Korea Terminology Research Center for Language and Knowledge Engineering. He had been a researcher in NEC C&C Lab of Japan, CSLI of Stanford University, and NHK Broadcasting Lab. His areas of expertise are natural language processing, ontology and knowledge engineering, semantic web and linked data, and their infrastructure. He has served as the President (2009-2010) of AFNLP (Asia Federation of Natural Language Processing) and the secretary of ISO/TC37/SC4 for language resource management standards since 2002. His recent works include the creative web such as semantic hierarchy mining from data webs, temporal and spatial entity mining, multilingual information synchronization in Wikipedia context, Korean language processing and Semantic Word Net for CJK called CoreNet. The main issue is about how to gather the human knowledge encoded in the documents, analyze its knowhow, and provide the customized answers depending on the users and their cultural and regional weights.

Invited Talk 3

A Unified Event Coreference Resolution Using Multiple Resolvers

Jian Su, Institute for Infocomm Research, Singapore

Abstract

Event coreference is an important and complicated task in event template extraction and other natural language processing applications. Despite its importance, it was merely discussed in previous studies. In this talk, I present our recent work on a globally optimized coreference resolution system dedicated to various sophisticated event coreference phenomena. Seven resolvers for both event and object coreference cases are utilized, which include 5 event coreference resolvers for event NP-NP, Verb-NP, Verb-Pron, NP-Pron as well as Verb Verb cases with quite some linguistic features different from the object counterparts. Three enhancements are further proposed at both mention pair detection and chain formation levels. First, the object coreference resolvers are used to effectively reduce the false positive cases for event coreference. Second, a revised instance selection scheme is proposed to improve link level mention-pair model performances. Last but not least, an efficient and globally optimized graph partitioning model is employed for coreference chain formation using spectral partitioning which allows the incorporation of pronoun coreference information. The three techniques contribute to a significant improvement of 8.54% in B3 F-score for event coreference resolution on OntoNotes 2.0 corpus.

Biodata

Jian Su is a senior scientist and group leader at the Institute for Infocomm Research, Singapore. Her research interests include information extraction, discourse analysis, text mining, language resources and evaluation, machine translation, segmentation, tagging and chunking, machine learning for natural language.