

**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**

© 2011 The PACLIC 25 Organizing Committee
and PACLIC Steering Committee

All rights reserved. Except as otherwise expressly permitted under copyright law, no part of this publication may be reproduced, digitized, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, Internet or otherwise, without the prior permission of the publisher.

Copyright of contributed papers reserved by respective authors

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 (I²R)

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 (I²R)

Members

Ma Bin, COLIPS, I²R (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 Otaguro, Waseda University (PACLIC 24 Liaison)
Ruli Manurung, University of Indonesia (PACLIC 26 Liaison)

Table of Contents

<i>Web based English-Chinese OOV term translation using Adaptive rules and Recursive feature selection</i>	1
Jian Qu	1
<i>English-Chinese Name Transliteration with Bi-Directional Syllable-Based Maximum Matching</i>	11
Oi Yee Kwong	11
<i>Language Model Weight Adaptation Based on Cross-entropy for Statistical Machine Translation</i>	20
Yinggong Zhao, Yangsheng Ji, Ning Xi, Shujian Huang and Jiajun Chen	20
<i>A grammar design accommodating packed argument frame information on verbs</i>	31
Petter Haugereid	31
<i>Quantification and the Garden Path Effect Reduction: The Case of Universally Quantified Subjects</i>	41
Akira Ohtani and Takeo Kurafuji	41
<i>A Simple Surface Realizer for Filipino</i>	51
Ethel Ong, Stephanie Abella, Lawrence Santos and Dennis Tiu	51
<i>Measuring Concept Concreteness from the Lexicographic Perspective</i>	60
Oi Yee Kwong	60
<i>Logical Information Processing of Possibility and Negation: Cases from Taiwanese Hakka</i>	70
Chiou-shing Yeh and Huei-ling Lai	70
<i>The Syntax-Semantics Interface of Resultative Constructions in Mandarin Chinese and Cantonese</i>	80
Pui Lun Chow	80
<i>Automatic Wrapper Generation and Maintenance</i>	90
Yingju Xia, Yuhang Yang, Shu Zhang and Hao Yu	90
<i>Evaluation via Negativa of Chinese Word Segmentation for Information Retrieval</i>	100
Mike Tian-Jian Jiang, Cheng-Wei Shih, Richard Tzong-Han Tsai and Wen-Lian Hsu	100
<i>Factual or Satisfactory: What Search Results Are Better?</i>	110
Yu Hong, Jun Lu and Shiqi Zhao	110
<i>A Graph-based Bilingual Corpus Selection Approach for SMT</i>	120
Wenhan Chao and Zhoujun Li	120
<i>Myanmar Phrases Translation Model with Morphological Analysis for Statistical Myanmar to English Translation System</i>	130
Thet Thet Zin, Khin Mar Soe and Ni Lar Thein	130
<i>Context Resolution of Verb Particle Constructions for English to Hindi Translation</i>	140
Niladri Chatterjee and Renu Balyan	140
<i>Improving Sampling-based Alignment by Investigating the Distribution of N-grams in Phrase Translation Tables</i>	150
Juan Luo, Adrien Lardilleux and Yves Lepage	150
<i>Predicting Linguistic Difficulty by Means of a Morpho-Syntactic Probabilistic Model</i>	160
Philippe Blache and Stéphane Rauzy	160
<i>Tibetan Word Segmentation as Syllable Tagging Using Conditional Random Field</i>	168
Huidan Liu, Minghua Nuo, Longlong Ma, Jian Wu and Yeping He	168
<i>Plural Problems in the Nominal Morphology of Marathi</i>	178
Shalmalee Pitale	178
<i>The L1 Acquisition of the Imperfective Aspect markers in Korean: a Comparison with Japanese</i>	186
Ju-Yeon Ryu	186
<i>Semi-Automatic Identification of Bilingual Synonymous Technical Terms from Phrase Tables and Parallel Patent Sentences</i>	196
Bing Liang, Takehito Utsuro and Mikio Yamamoto	196
<i>Automatic Error Analysis Based on Grammatical Questions</i>	206
Tomoki Nagase, Hajime Tsukada, Katsunori Kotani, Nobutoshi Hatanaka and Yoshiyuki Sakamoto	206

<i>Maximum Entropy Based Lexical Reordering Model for Hierarchical Phrase-based Machine Translation</i>	
Zhongguang Zheng, Yao Meng and Hao Yu	216
<i>A Bare-bones Constraint Grammar</i>	
Eckhard Bick	226
<i>Spring Cleaning and Grammar Compression: Two Techniques for Detection of Redundancy in HPSG Grammars</i>	
Antske Fokkens, Yi Zhang and Emily M. Bender	236
<i>Developing a Chunk-based Grammar Checker for Translated English Sentences</i>	
Nay Yee Lin, Khin Mar Soe and Ni Lar Thein	245
<i>Creating the Open Wordnet Bahasa</i>	
Nurril Hidfana Bte Mohamed Noor, Suerya Sapuan and Francis Bond	255
<i>Automatic identification of words with novel but infrequent senses</i>	
Paul Cook and Graeme Hirst	265
<i>Annotating the Structure and Semantics of Fables</i>	
Oi Yee Kwong	275
<i>Verbal Inflection in Hindi: A Distributed Morphology Approach</i>	
Smriti Singh and Vaijayanthi M Sarma	283
<i>Word classes in Indonesian: A linguistic reality or a convenient fallacy in natural language processing?</i>	
Meladel Mistica, Timothy Baldwin and I Wayan Arka	293
<i>Automated Proof Reading of Clinical Notes</i>	
Jon Patrick and Dung Nguyen	303
<i>Modelling Word Meaning using Efficient Tensor Representations</i>	
Mike Symonds, Peter Bruza, Laurianne Sitbon and Ian Turner	313
<i>A Study of Sense-Disambiguated Networks Induced from Folksonomies</i>	
Hans-Peter Zorn and Iryna Gurevych	323
<i>Unsupervised Word Sense Disambiguation Using Neighborhood Knowledge</i>	
Heyan Huang, Zhizhuo Yang and Ping Jian	333
<i>Dependency-based Analysis for Tagalog Sentences</i>	
Erlyn Manguilimotan and Yuji Matsumoto	343
<i>Case study of BushBank concept</i>	
Marek Grac	353
<i>Building and Annotating the Linguistically Diverse NTU-MC (NTU-Multilingual Corpus)</i>	
Liling Tan and Francis Bond	362
<i>In Situ Text Summarisation for Museum Visitors</i>	
Timothy Baldwin, Patrick Ye, Fabian Bohnert and Ingrid Zukerman	372
<i>Iteratively Estimating Pattern Reliability and Seed Quality With Extraction Consistency</i>	
Yi-Hsun Lee, Chung-Yao Chuang and Wen-Lian Hsu	382
<i>The Effect of Answer Patterns for Supervised Named Entity Recognition in Thai</i>	
Nutchra Tirasaroj and Wirete Aroonmanakun	392
<i>A Listwise Approach to Coreference Resolution in Multiple Languages</i>	
Oanh Thi Tran, Bach Xuan Ngo, Minh Le Nguyen and Akira Shimazu	400
<i>Combining Dependency and Constituent-based Syntactic Information for Anaphoricity Determination in Coreference Resolution</i>	
Fang Kong and Guodong Zhou	410
<i>Sentiment Classification in Resource-Scarce Languages by using Label Propagation</i>	
Yong Ren, Nobuhiro Kaji, Naoki Yoshinaga, Masashi Toyoda and Masaru Kitsuregawa	420
<i>A Hybrid Extraction Model for Chinese Noun/Verb Synonymous bi-gram Collocations</i>	
Wanyin Li and Qin Lu	430
<i>Word-order and argument-marking: Japanese vs Chinese vs Nazi</i>	
Paul Law	440
<i>Classification of Filipino Speech Rhythm Using Computational and Perceptual Approach</i>	
Timothy Israel Santos and Rowena Cristina Guevara	450

<i>Disfluencies in Consecutive Interpreting among Undergraduates in the Language Lab Environment in the final version</i>	
Kexiu Yin	459
<i>An English-Chinese Cross-lingual Word Semantic Similarity Measure Exploring Attributes and Relations</i>	
Lin Dai and Heyan Huang	467
<i>Learning-to-Translate Based on the S-SSTC Annotation Schema</i>	
Enya Kong Tang, Zaharin Yusoff and Christian Boitet	477
<i>Translating English Names to Arabic Using Phonotactic Rules</i>	
Faisal Alshuwaier and Ali Areshey	485
<i>Translating Common English and Chinese Verb-Noun Pairs in Technical Documents with Collocational and Bilingual Information</i>	
Yi-Hsuan Chuang, Chao-Lin Liu and Jing-Shin Chang	493
<i>System for Flexibly Judging the Misuse of Honorifics in Japanese</i>	
Tamotsu Shirado, Satoko Marumoto, Masaki Murata and Hitoshi Isahara	503
<i>Compound Event Nouns of the Modifier-head Type in Mandarin Chinese</i>	
Shan Wang and Chu-Ren Huang	511
<i>The Co-occurrence of Two Delimiters: An Investigation of Mandarin Chinese Resultatives</i>	
Jingxia Lin and Chu-Ren Huang	519
<i>The Order of Mandarin Chinese Motion Morphemes and the Scalar Specificity Constraint</i>	
Jingxia Lin	525
<i>Verbs and (sub)Event Structure: A Case Study from Italian</i>	
Francesca Strik Lievers	533
<i>The Effects of EFL Learners Awareness and Retention in Learning Metaphoric and Metonymic Expressions</i>	
Yi-chen Chen and Huei-ling Lai	541
<i>NERSIL - the Named-Entity Recognition System for Iban Language</i>	
Yong Soo Fong, Bali Ranaivo Malanon and Alvin Yeo Wee	549
<i>Improving PP Attachment Disambiguation in a Rule-based Parser</i>	
Yoon-Hyung Roh, Ki-Young Lee and Young-Gil Kim	559
<i>Fully-Automatic Marker-based Chunking in 11 European Languages and Counts of the Number of Analogies between Chunks</i>	
Kota Takeya and Yves Lepage	567
<i>Extraction of Broad-Scale, High-Precision Japanese-English Parallel Translation Expressions Using Lexical Information and Rules</i>	
Qing Ma, Shinya Sakagami and Masaki Murata	577
<i>Analyzing the characteristics of academic paper categories by using an index of representativeness</i>	
Takafumi Suzuki, Kiyoko Uchiyama, Ryota Tomisaka and Akiko Aizawa	587
<i>Exploring Emotional Words for Chinese Document Chief Emotion Analysis</i>	
Yunong Wu, Kenji Kita, Fuji Ren, Kazuyuki Matsumoto and Xin Kang	597
<i>A Construction Grammar Approach to Prepositional Phrase Attachment: Semantic Feature Analysis of V NP1 into NP2 Construction</i>	
Liyin Chen, Siaw-Fong Chung and Chao-Lin Liu	607
<i>Supervised and Semi-supervised Methods based Organization Name Disambiguity</i>	
Shu Zhang and Hao Yu	615
<i>Study and Implementation of Monolingual Approach on Indonesian Question Answering for Factoid and Non-Factoid Question</i>	
Alvin Andhika Zulen and Ayu Purwarianti	622

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.