

Lecture Notes in Artificial Intelligence 4285

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Yuji Matsumoto Richard Sproat
Kam-Fai Wong Min Zhang (Eds.)

Computer Processing of Oriental Languages

Beyond the Orient:
The Research Challenges Ahead

21st International Conference, ICCPOL 2006
Singapore, December 17-19, 2006
Proceedings

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Yuji Matsumoto
Nara Institute of Science and Technology, Japan
E-mail: matsu@is.naist.jp

Richard Sproat
University of Illinois at Urbana-Champaign
Dept. of Linguistics, Dept. of Electrical Engineering, USA
E-mail: rws@xoba.com

Kam-Fai Wong
The Chinese University of Hong Kong
Department of Systems Engineering and Engineering Management
Shatin, N.T., Hong Kong
E-mail: kfwong@se.cuhk.edu.hk

Min Zhang
Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
E-mail: mzhang@i2r.a-star.edu.sg

Library of Congress Control Number: 2006937162

CR Subject Classification (1998): I.2.6-7, F.4.2-3, I.2, H.3, I.7, I.5

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN	0302-9743
ISBN-10	3-540-49667-X Springer Berlin Heidelberg New York
ISBN-13	978-3-540-49667-0 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springer.com

© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11940098 06/3142 5 4 3 2 1 0

Message from the President

“Beyond the Orient: The Research Challenges Ahead”

The International Conference on Computer Processing of Oriental Languages (ICCPOL) is a regular conference series of the Chinese and Oriental Languages Computer Society, COCLS (formerly known as the Chinese Language Computer Society, CLCS), which was established 30 years ago, to be exact on June 9, 1976. The society's name change was made in the winter of 2005 in response to the growing international demand in Chinese and oriental languages research and applications. The new vision of the society was also launched at the same time. COLCS was set "to be the international computer society driving the advancement and globalization of the science and technology in Chinese and oriental languages processing." On this front, the society's conference, ICCPOL, and journal, namely, the *International Journal on Computer Processing of Oriental Languages* (IJCPOL) provide two effective platforms.

This year marked the 21st meeting of the ICCPOL conference. I was delighted that despite his heavy workload, Kim-Teng Lua kindly accepted my invitation to host ICCPOL 2006 in Singapore. He put together one of the most energetic organizing committees: Minghui Dong, who looked after the local organization including the conference Website, Hui Wang the registration and Min Zhang the publication. Without their dedication and professionalism, ICCPOL 2006 would not have been so successful.

I am grateful to the Department of Systems Engineering and Engineering Management at The Chinese University of Hong Kong not only for allowing me to take up the conference chairmanship but, even more importantly, for providing financial aid for students in need. I am thankful to my colleague, Chris Yang, for working closely with me to assess every application in detail.

I would also like to thank the program Co-chairs Yuji Matsumoto and Richard Sproats, who jointly worked out an inspiring program. The combination of Asian and American scientists supported our theme of "Beyond the Orient." Many high-quality papers from all around the world were received and unfortunately due to limited space, only a few were accepted for publication in this year's proceedings. The accepted papers truly highlighted "The Research Challenges Ahead" in Chinese and Oriental language processing.

Kam-Fai Wong
Conference Chair ICCPOL 2006
and
President Colcs

Message from the Program Co-chairs

As the Co-chairs of the technical program of the 21st International Conference on Computer Processing of Oriental Languages (December 17-19, Singapore) we are delighted and honored to write the introduction to these proceedings.

The subtitle of this year's convocation was "Beyond the Orient: The Research Challenges Ahead," the goal being to broaden the scope of ICCPOL beyond its origins in East Asia, to become a truly international event. We believe that we made a very successful first step in this direction both in the composition of the Program Committee, which is made up of members from countries around the world and the accepted papers, which include a large number of contributions from outside ICCPOL's traditional home base.

We received 169 submissions from a variety of countries. We initially accepted only 38 full papers (30 oral presentations and 8 poster presentations determined by the authors' preference), but since this was a fairly small set, we decided to accept another 20 papers as short papers, all of which were presented as posters. Thus, the acceptance rate of full papers is 23% (38/169), and that of all accepted papers is 34% (58/169). Since two papers were withdrawn after the notification, this volume includes 56 papers (36 full papers and 20 short papers) presented at the conference.

As Technical Co-chairs we can claim but a small amount of credit for the success of the technical program. Our main thanks go to the Program Committee members who worked diligently to give fair reviews of the submitted papers, and most of whom spent additional time coming to a consensus on papers where there was a wide amount of disagreement.

We also thank the many authors who submitted their work for inclusion in the conference. Needless to say, the conference would not exist were it not for the technical presentations. We are mindful of the fact that there are many computational linguistics conferences and workshops available, and we are therefore happy that so many papers were submitted to ICCPOL 2006.

We would also like to thank our invited keynote speakers Gerald Penn, Claire Cardie and Hwee-Tou Ng for agreeing to present their work at ICCPOL.

In this year's conference, we used the START Conference Manager System for most of the paper handling process, that is, paper submission, paper reviews and discussion, notification of acceptance/rejection of papers, and uploading of final manuscripts, all of which went very smoothly. Thanks are especially due to Rich Gerber, the maintainer of the system, who was always quick to answer our queries, and even modified the system to handle the specific needs of our conference. We would also like to thank the committee members of ICCPOL, especially Kam-Fai Wong for his continuous support and timely advice, Minghui Dong for preparing very beautiful Web pages, and Min Zhang and Wai Lam for handling all the final manuscripts that are included in this volume.

December 2006

Yuji Matsumoto and Richard Sproat
ICCPOL 2006 Program Co-chairs

Organization

Conference Committee

Honorary Conference Co-chairs

Shi-Kuo Chang, University of Pittsburgh, USA (Co-founder, COLCS)
Benjamin Tsou, City University of Hong Kong, China (President, AFNLP)
Jun'ichi Tsujii, University of Tokyo, Japan (President, ACL)

Conference Chair

Kam-Fai Wong, The Chinese University of Hong Kong, China

Conference Co-chair

Jong-Hyeok Lee, POSTECH, Korea

Organization Chair

Kim-Teng Lua, COLIPS, Singapore

Program Co-chairs

Yuji Matsumoto, Nara Institute of Science and Technology, Japan
Richard Sproat, University of Illinois at Urbana-Champaign, USA

General Secretary

Minghui Dong, Institute for Infocomm Research, Singapore

Publication Co-chairs

Min Zhang, Institute for Infocomm Research, Singapore
Wai Lam, Chinese University of Hong Kong, China

Finance Co-chairs

Chris Yang, Chinese University of Hong Kong, China
Hui Wang, National University of Singapore, Singapore

Program Committee

Galen Andrew, Microsoft Research, USA
Masayuki Asahara, Nara Institute of Science and Technology, Japan

Hsin-Hsi Chen, National Taiwan University, Taiwan
Keh-Jiann Chen, Academia Sinica, Taiwan
David Chiang, ISI, USA
Lee-Feng Chien, Academia Sinica, Taiwan
Key-Sun Choi, KAIST, Korea
Susan Converse, University of Pennsylvania, USA
Robert Dale, Macquarie University, Australia
Guohong Fu, Hong Kong University, China
Pascale Fung, Hong Kong University of Science and Technology, China
Niyu Ge, IBM T. J. Watson Research Center, USA
Julia Hockenmaier, University of Pennsylvania, USA
Liang Huang, University of Pennsylvania, USA
Kenji Imamura, NTT, Japan
Kentarō Inui, Nara Institute of Science and Technology, Japan
Martin Jansche, Columbia University, USA
Donghong Ji, Institute for Infocomm Research, Singapore
Gareth Jones, Dublin City University, Ireland
Genichiro Kikui, NTT, Japan
Sadao Kurohashi, University of Tokyo, Japan
Kui-Lam Kwok, City University of New York, USA
Olivia Oi Yee Kwong, City University of Hong Kong, China
Gary Geunbae Lee, POSTECH, Korea
Gina-Anne Levow, University of Chicago, USA
Roger Levy, University of Edinburgh, UK
Haizhou Li, Institute for Infocomm Research, Singapore
Hang Li, Microsoft Research Asia, China
Mu Li, Microsoft Research Asia, China
Wenjie Li, Polytechnic University of Hong Kong, China
Chin-Yew Lin, ISI, USA
Qin Lu, Polytechnic University of Hong Kong, China
Bin Ma, Institute for Infocomm Research, Singapore
Qing Ma, Ryukoku University, Japan
Helen Meng, Chinese University of Hong Kong, China
Tatsunori Mori, Yokohama National University, Japan
Hwee Tou Ng, National University of Singapore, Singapore
Cheng Niu, Microsoft
Douglas Oard, University of Maryland, USA
Kemal Oflazer, Sabanci University, Turkey
Manabu Okumura, Tokyo Institute of Technology, Japan
Martha Palmer, University of Colorado, USA
Hae-Chang Rim, Korea University, Korea
Laurent Romary, LORIA, France
Tetsuya Sakai, Toshiba, Japan
Rajeev Sangal, International Institute of Information Technology, India
Jungyun Seo, Sogang University, Korea

Kiyoaki Shirai, Japan Advanced Institute of Science and Technology, Japan
Dawei Song, Open University, UK
Virach Sornlertlamvanich, Thai Computational Linguistics Lab., Thailand
Keh-Yih Su, Behavior Design Corporation, Taiwan
Jian Su, Institute for Infocomm Research, Singapore
Maosong Sun, Tsinghua University, China
Kumiko Tanaka-Ishii, University of Tokyo, Japan
Takenobu Tokunaga, Tokyo Institute of Technology, Japan
Kiyotaka Uchimoto, NICT, Japan
Takehito Utsuro, Tsukuba University, Japan
Hui Wang, National University of Singapore, Singapore
Patrick Wang, Northeastern University, USA
Andi Wu, Microsoft, GrapeCity Inc., USA
Fei Xia, University of Washington, USA
Yunqing Xia, The Chinese University of Hong Kong, China
Bo Xu, Chinese Academy of Sciences, China
Jie Xu, National University of Singapore, Singapore and Henan University, China
Nianwen (Bert) Xue, University of Colorado, USA
Tianshun Yao, Northeastern University, China
Zaharin Yusoff, Malaysian Institute of Micro-Electronics, Malaysia
Min Zhang, Institute for Infocomm Research, Singapore
Guodong Zhou, Institute for Infocomm Research, Singapore
Ming Zhou, Microsoft Research Asia, China

Hosted by the Chinese and Oriental Languages Computer Society (COLCS)

Organized by the Chinese and Oriental Languages Information Processing Society (COLIPS)

Supported by

Asian Federation of Natural Language Processing (AFNLP)

Department of Systems Engineering and Engineering Management (SEEM), The Chinese University of Hong Kong, China

Publisher Springer

Table of Contents

Information Retrieval/Document Classification/QA/ Summarization I

Answering Contextual Questions Based on the Cohesion with Knowledge	1
<i>Tatsunori Mori, Shinpei Kawaguchi, Madoka Ishioroshi</i>	
Segmentation of Mixed Chinese/English Document Including Scattered Italic Characters	13
<i>Yong Xia, Chun-Heng Wang, Ru-Wei Dai</i>	
Using Pointwise Mutual Information to Identify Implicit Features in Customer Reviews	22
<i>Qi Su, Kun Xiang, Houfeng Wang, Bin Sun, Shiwen Yu</i>	
Using Semi-supervised Learning for Question Classification	31
<i>Nguyen Thanh Tri, Nguyen Minh Le, Akira Shimazu</i>	
Query Similarity Computing Based on System Similarity Measurement	42
<i>Chengzhi Zhang, Xiaoqin Xu, Xinning Su</i>	

Machine Translation I

An Improved Method for Finding Bilingual Collocation Correspondences from Monolingual Corpora	51
<i>Ruifeng Xu, Kam-Fai Wong, Qin Lu, Wenjie Li</i>	
A Syntactic Transformation Model for Statistical Machine Translation	63
<i>Thai Phuong Nguyen, Akira Shimazu</i>	
Word Alignment Between Chinese and Japanese Using Maximum Weight Matching on Bipartite Graph	75
<i>Honglin Wu, Shaoming Liu</i>	
Improving Machine Transliteration Performance by Using Multiple Transliteration Models	85
<i>Jong-Hoon Oh, Key-Sun Choi, Hitoshi Isahara</i>	

Information Retrieval/Document Classification/ QA/Summarization II

Clique Percolation Method for Finding Naturally Cohesive and Overlapping Document Clusters	97
<i>Wei Gao, Kam-Fai Wong, Yunqing Xia, Ruifeng Xu</i>	
Hybrid Approach to Extracting Information from Web-Tables	109
<i>Sung-won Jung, Mi-young Kang, Hyuk-chul Kwon</i>	
A Novel Hierarchical Document Clustering Algorithm Based on a kNN Connection Graph	120
<i>Qiaoming Zhu, Junhui Li, Guodong Zhou, Peifeng Li, Peide Qian</i>	

Poster Session 1

The Great Importance of Cross-Document Relationships for Multi-document Summarization	131
<i>Xiaojun Wan, Jianwu Yang, Jianguo Xiao</i>	
The Effects of Computer Assisted Instruction to Train People with Reading Disabilities Recognizing Chinese Characters	139
<i>Wan-Chih Sun, Tsung-Ren Yang, Chih-Chin Liang, Ping-Yu Hsu, Yuh-Wei Kung</i>	
Discrimination-Based Feature Selection for Multinomial Naïve Bayes Text Classification	149
<i>Jingbo Zhu, Huizhen Wang, Xijuan Zhang</i>	
A Comparative Study on Chinese Word Clustering	157
<i>Bo Wang, Houfeng Wang</i>	
Populating FrameNet with Chinese Verbs Mapping Bilingual Ontological WordNet with FrameNet	165
<i>Ian C. Chow, Jonathan J. Webster</i>	
Collecting Novel Technical Terms from the Web by Estimating Domain Specificity of a Term	173
<i>Takehito Utsuro, Mitsuhiro Kida, Masatsugu Tonoike, Satoshi Sato</i>	
Building Document Graphs for Multiple News Articles Summarization: An Event-Based Approach	181
<i>Wei Xu, Chunfa Yuan, Wenjie Li, Mingli Wu, Kam-Fai Wong</i>	
A Probabilistic Feature Based Maximum Entropy Model for Chinese Named Entity Recognition	189
<i>Suxiang Zhang, Xiaojie Wang, Juan Wen, Ying Qin, Yixin Zhong</i>	

Correcting Bound Document Images Based on Automatic and Robust Curved Text Lines Estimation	197
<i>Yichao Ma, Chunheng Wang, Ruwei Dai</i>	
Cluster-Based Patent Retrieval Using International Patent Classification System	205
<i>Jungi Kim, In-Su Kang, Jong-Hyeok Lee</i>	
Word Error Correction of Continuous Speech Recognition Using WEB Documents for Spoken Document Indexing	213
<i>Hiromitsu Nishizaki, Yoshihiro Sekiguchi</i>	
Extracting English-Korean Transliteration Pairs from Web Corpora	222
<i>Jong-Hoon Oh, Hitoshi Isahara</i>	
Word Segmentation/Chunking/Abbreviation Expansion/Writing-System Issues	
From Phoneme to Morpheme: Another Verification Using a Corpus	234
<i>Kumiko Tanaka-Ishii, Zhihui Jin</i>	
Chinese Abbreviation Identification Using Abbreviation-Template Features and Context Information	245
<i>Xu Sun, Houfeng Wang</i>	
Word Frequency Approximation for Chinese Using Raw, MM-Segmented and Manually Segmented Corpora	256
<i>Wei Qiao, Maosong Sun</i>	
Identification of Maximal-Length Noun Phrases Based on Expanded Chunks and Classified Punctuations in Chinese	268
<i>Xue-Mei Bai, Jin-Ji Li, Dong-Il Kim, Jong-Hyeok Lee</i>	
A Hybrid Approach to Chinese Abbreviation Expansion	277
<i>Guohong Fu, Kang-Kuong Luke, Min Zhang, GuoDong Zhou</i>	
Category-Pattern-Based Korean Word-Spacing	288
<i>Mi-young Kang, Sung-won Jung, Hyuk-chul Kwon</i>	
An Integrated Approach to Chinese Word Segmentation and Part-of-Speech Tagging	299
<i>Maosong Sun, Dongliang Xu, Benjamin K. Tsou, Huaming Lu</i>	
Kansuke: A Kanji Look-Up System Based on a Few Stroke Prototypes	310
<i>Kumiko Tanaka-Ishii, Julian Godon</i>	

Modelling the Orthographic Neighbourhood for Japanese Kanji	321
<i>Lars Yencken, Timothy Baldwin</i>	

Reconstructing the Correct Writing Sequence from a Set of Chinese Character Strokes	333
<i>Kai-Tai Tang, Howard Leung</i>	

Machine Translation II

Expansion of Machine Translation Bilingual Dictionaries by Using Existing Dictionaries and Thesauruses	345
<i>Takeshi Kutsumi, Takehiko Yoshimi, Katsunori Kotani, Ichiko Sata, Hitoshi Isahara</i>	

Feature Rich Translation Model for Example-Based Machine Translation	355
<i>Yin Chen, Muyun Yang, Sheng Li, Hongfei Jiang</i>	

Dictionaries for English-Vietnamese Machine Translation	363
<i>Le Manh Hai, Nguyen Chanh Thanh, Nguyen Chi Hieu, Phan Thi Tuoi</i>	

Poster Session 2

Translation Selection Through Machine Learning with Language Resources	370
<i>Hyun Ah Lee</i>	

Acquiring Translational Equivalence from a Japanese-Chinese Parallel Corpus	378
<i>Yujie Zhang, Qing Ma, Qun Liu, Wenliang Chen, Hitoshi Isahara</i>	

Deep Processing of Korean Floating Quantifier Constructions	387
<i>Jong-Bok Kim, Jaehyung Yang</i>	

Compilation of a Dictionary of Japanese Functional Expressions with Hierarchical Organization	395
<i>Suguru Matsuyoshi, Satoshi Sato, Takehito Utsuro</i>	

A System to Indicate Honorific Misuse in Spoken Japanese	403
<i>Tamotsu Shirado, Satoko Marumoto, Masaki Murata, Kiyotaka Uchimoto, Hitoshi Isahara</i>	

A Chinese Corpus with Word Sense Annotation	414
<i>Yunfang Wu, Peng Jin, Yangsen Zhang, Shiwen Yu</i>	

Multilingual Machine Translation of Closed Captions for Digital Television with Dynamic Dictionary Adaptation	422
<i>Sanghwa Yuh, Jungyun Seo</i>	

Acquiring Concept Hierarchies of Adjectives from Corpora	430
<i>Kyoko Kanzaki, Qing Ma, Eiko Yamamoto, Tamotsu Shirado, Hitoshi Isahara</i>	
Pronunciation Similarity Estimation for Spoken Language Learning	442
<i>Donghyun Kim, Dongsuk Yook</i>	
A Novel Method for Rapid Speaker Adaptation Using Reference Support Speaker Selection	450
<i>Jian Wang, Zhen Yang, Jianjun Lei, Jun Guo</i>	
Using Latent Semantics for NE Translation	457
<i>Boon Pang Lim, Richard W. Sproat</i>	
Chinese Chunking with Tri-training Learning	466
<i>Wenliang Chen, Yujie Zhang, Hitoshi Isahara</i>	
Binarization Approaches to Email Categorization	474
<i>Yunqing Xia, Kam-Fai Wong</i>	
Investigating Problems of Semi-supervised Learning for Word Sense Disambiguation	482
<i>Anh-Cuong Le, Akira Shimazu, Le-Minh Nguyen</i>	
Developing a Dialog System for New Idea Generation Support	490
<i>Masahiro Shibata, Yoichi Tomiura, Hideki Matsumoto, Tomomi Nishiguchi, Kensei Yukino, Akihiro Hino</i>	
Parsing/Semantics/Lexical Resources	
The Incremental Use of Morphological Information and Lexicalization in Data-Driven Dependency Parsing	498
<i>Gülşen Eryiğit, Joakim Nivre, Kemal Oflazer</i>	
Pattern Dictionary Development Based on Non-compositional Language Model for Japanese Compound and Complex Sentences	509
<i>Satoru Ikehara, Masato Tokuhisa, Jin'ichi Murakami, Masashi Saraki, Masahiro Miyazaki, Naoshi Ikeda</i>	
A Case-Based Reasoning Approach to Zero Anaphora Resolution in Chinese Texts	520
<i>Dian-Song Wu, Tyne Liang</i>	
Building a Collocation Net	532
<i>GuoDong Zhou, Min Zhang, GuoHong Fu</i>	
Author Index	543