©2008 Asian Federation of Natural Language Processing

**Sponsor**

Special Coordination Funds for Promoting Science and Technology, Ministry of Education, Culture, Sport, Science and Technology, MEXT Japan.
Preface

This volume contains the papers presented at the sixth workshop on Asian Language Resources, held on 11–12 January 2008 in conjunction with the third International Joint Conference on Natural Language Processing (IJCNLP 2008).

Language resources have played an essential role in empirical approaches to natural language processing (NLP) for the last two decades. Previous concerted efforts on construction of language resources, particularly in the US and European countries, have laid a solid foundation for the pioneering NLP researches in these two communities. In comparison, the availability and accessibility of many Asian language resources are still very limited except for a few languages. Moreover, there is a greater diversity in Asian languages with respect to character sets, grammatical properties and the cultural background.

Motivated by such a context, we have organised a series of workshops on Asian language resources since 2001. This workshop series has contributed to the activation of the NLP research in Asia particularly of building and utilising corpora of various types and languages. In this sixth workshop, we had 31 submissions encompassing 13 languages. The paper selection was highly competitive compared with the last five workshops. The program committee selected 10 regular papers, 3 short papers and 8 resource reports for presentation at the workshop.

The workshop is comprised of two parts, technical sessions and a session devoted to reporting activities related to language resources in several languages. Following the resource report session, we have an open discussion on the collaboration in building, standardising and exchanging language resources in Asia. We hope this workshop further accelerates the already thriving NLP research in Asia.

Chu-Ren Huang
Mikami Yoshiki
Workshop Co-chairs

Hasida Kōiti
Tokunaga Takenobu
Program Co-chairs
Organiser

Workshop chairs
Huang, Chu-Ren  
Mikami, Yoshiki

Program chairs
Hasida, Kōiti  
Tokunaga, Takenobu

Program Committee

<table>
<thead>
<tr>
<th>Name</th>
<th>Institutional Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bhattacharyya, Pushpak</td>
<td>IIT, Bombay</td>
</tr>
<tr>
<td>Fang, Alex Chengyu</td>
<td>City University of Hong Kong</td>
</tr>
<tr>
<td>Riza, Hamman</td>
<td>IPTEKnet–BPPT</td>
</tr>
<tr>
<td>Hasida, Kōiti</td>
<td>National Institute of Advanced Industrial Science and Technology</td>
</tr>
<tr>
<td>He, Tingting</td>
<td>Huazhong Normal University</td>
</tr>
<tr>
<td>Huang, Chu-Ren</td>
<td>Academia Sinica</td>
</tr>
<tr>
<td>Hussain, Sarmad</td>
<td>National University of Computer &amp; Emerging Sciences</td>
</tr>
<tr>
<td>Itahashi, Shuichi</td>
<td>National Institute of Informatics</td>
</tr>
<tr>
<td>Lu, Qin</td>
<td>Hong Kong Polytechnic University</td>
</tr>
<tr>
<td>Luong, Chi Mai</td>
<td>National Center for Sciences and Technologies of Vietnam</td>
</tr>
<tr>
<td>Mikami, Yoshiki</td>
<td>Nagaoka University of Technology</td>
</tr>
<tr>
<td>Nandasara, Shakrange Turrance</td>
<td>University of Colombo, School of Computing</td>
</tr>
<tr>
<td>Nguyen, Thi Minh Huyen</td>
<td>Hanoi University of Sciences</td>
</tr>
<tr>
<td>Oo, Thein</td>
<td>Myanmar Computer Federation</td>
</tr>
<tr>
<td>Rau, Victoria</td>
<td>Providence University</td>
</tr>
<tr>
<td>Rim, Hae-Chang</td>
<td>Korea University</td>
</tr>
<tr>
<td>Roxas, Rachel Edita O</td>
<td>De La Salle University, Manila</td>
</tr>
<tr>
<td>Shirai, Kiyouki</td>
<td>Japan Advanced Institute of Science and Technology</td>
</tr>
<tr>
<td>Sornlertlamvanich, Virach</td>
<td>Thai Computational Linguistics Laboratory, NICT</td>
</tr>
<tr>
<td>Sui, Zhifang</td>
<td>Peking University</td>
</tr>
<tr>
<td>Tokunaga, Takenobu</td>
<td>Tokyo Institute of Technology</td>
</tr>
<tr>
<td>Vikas, Om</td>
<td>Indian Institute of Information Technology and Management</td>
</tr>
<tr>
<td>Zhao, Jun</td>
<td>Chinese Academy of Sciences</td>
</tr>
</tbody>
</table>

This workshop is supported by Special Coordination Funds for Promoting Science and Technology, Ministry of Education, Culture, Sport, Science and Technology, MEXT Japan.
Workshop Program
11-12 January 2008
Indian School of Business, Hyderabad, India

Day 1 (11 January)

9:00 Registration
9:20 Opening
9:30 Development of Bengali Named Entity Tagged Corpus and its Use in NER Systems
Asif Ekbal and Sivaji Bandyopadhyay
9:55 Gazetteer Preparation for Named Entity Recognition in Indian Languages
Sujan Kumar Saha, Sudeshna Sarkar and Pabitra Mitra
10:20 Preliminary Chinese Term Classification for Ontology Construction
Gaoying Cui, Qin Lu and Wenjie Li
10:45 Break
11:05 Technical Terminology in Asian Languages: Different Approaches to Adopting Engineering Terms
Makiko Matsuda, Tomoe Takahashi, Hiroki Goto, Yoshikazu Hayase, Robin Lee Nagano and Yoshiki Mikami
11:30 Selection of XML tag set for Myanmar National Corpus
Wunna Ko Ko and Thin Zar Phyo
11:55 Myanmar Word Segmentation using Syllable level Longest Matching
Hla Hla Htay and Kavi Narayana Murthy
12:20 Lunch
13:50 The Link Structure of Language Communities and its Implication for Language-specific Crawling
Rizza Caminero and Yoshiki Mikami
14:15 A Multilingual Multimedia Indian Sign Language Dictionary Tool
Tirthankar Dasgupta, Sambit Shukla, Sandeep Kumar, Synny Diwakar and Anupam Basu
14:40 A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus
Deniz Zeyrek and Bonnie Webber
15:05 Towards an Annotated Corpus of Discourse Relations in Hindi
Rashmi Prasad, Samar Husain, Dipti Sharma and Aravind Joshi
15:30 Break
15:50 A Semantic Study on Yami Ontology in Traditional Songs
Yin-Sheng Tai, D. Victoria Rau and Meng-Chien Yang
16:05 Assessment and Development of POS Tag Set for Telugu
Rama Sree R.J., Uma Maheswara Rao G and Madhu Murthy K.V.
16:20 Designing a Common POS-Tagset Framework for Indian Languages
Sankaran Baskaran, Kalika Bali, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, Rajendran S, Saravanan K, Sobha L and Subbarao K V
Day 2 (12 January)

9:00 Resources Report on Languages of Indonesia
Hammam Riza
9:15 Confirmed Language Resource for Answering How Type Questions Developed by Using Mails Posted to a Mailing List
Ryo Nishimura, Yasuhiro Watanabe and Yoshihiro Okada
9:30 Corpus building for Mongolian language
Purev Jaimai and Odbayar Chimeddorj
9:45 Resources for Urdu Language Processing
Sarmad Hussain
10:00 Balanced Corpus of Contemporary Written Japanese
Kikuo Maekawa
10:15 Break
10:35 A Basic Framework to Build a Test Collection for the Vietnamese Text Categorization
Viet Hoang-Anh, Thu Dinh-Thi-Phuong and Thang Huynh-Quyet
10:50 Enhanced Tools for Online Collaborative Language Resource Development
Virach Sornlertlamvanich, Thatstanee Charoenporn, Suphanut Thayaboon, Chumpol Mokarat and Hitoshi Isahara
11:05 Japanese Effort Toward Sharing Text and Speech Corpora
Shuichi Itahashi and Kōiti Hasida
11:20 Open Discussion
12:20 Closing
# Table of Contents

(Regular papers)

**Development of Bengali Named Entity Tagged Corpus and its Use in NER Systems**
Asif Ekbal and Sivaji Bandyopadhyay ................................................................. 1

**Gazetteer Preparation for Named Entity Recognition in Indian Languages**
Sujan Kumar Saha, Sudeshna Sarkar and Pabitra Mitra ......................................... 9

**Preliminary Chinese Term Classification for Ontology Construction**
Gaoying Cui, Qin Lu and Wenjie Li ................................................................. 17

**Technical Terminology in Asian Languages: Different Approaches to Adopting Engineering Terms**
Makiko Matsuda, Tomoe Takahashi, Hiroki Goto, Yoshikazu Hayase, Robin Lee Nagano and Yoshiki Mikami ................................................................. 25

**Selection of XML tag set for Myanmar National Corpus**
Wunna Ko Ko and Thin Zar Phyo ................................................................. 33

**Myanmar Word Segmentation using Syllable level Longest Matching**
Hla Hla Htay and Kavi Narayana Murthy ......................................................... 41

**The Link Structure of Language Communities and its Implication for Language-specific Crawling**
Rizza Caminero and Yoshiki Mikami ................................................................. 49

**A Multilingual Multimedia Indian Sign Language Dictionary Tool**
Tirthankar Dasgupta, Sambit Shukla, Sandeep Kumar, Synny Diwakar and Anupam Basu .... 57

**A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus**
Deniz Zeyrek and Bonnie Webber ................................................................. 65

**Towards an Annotated Corpus of Discourse Relations in Hindi**
Rashmi Prasad, Samar Husain, Dipti Sharma and Aravind Joshi ................................ 73

(Short papers)

**A Semantic Study on Yami Ontology in Traditional Songs**
Yin-Sheng Tai, D. Victoria Rau and Meng-Chien Yang ........................................ 81

**Assessment and Development of POS Tag Set for Telugu**
Rama Sree R.J., Uma Maheswara Rao G and Madhu Murthy K.V. ............................. 85

**Designing a Common POS-Tagset Framework for Indian Languages**
Sankaran Baskaran, Kalika Bali, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, Rajendran S, Saravanan K, Sobha L and Subbarao K V ................................................ 89

(Resource reports)

**Resources Report on Languages of Indonesia**
Hammam Riza ................................................................. 93

**Confirmed Language Resource for Answering How Type Questions Developed by Using Mails Posted to a Mailing List**
Ryo Nishimura, Yasuhiro Watanabe and Yoshihiro Okada ...................................... 95
Corpus building for Mongolian language
Purev Jaimai and Odbayar Chimeddorj ................................................................. 97

Resources for Urdu Language Processing
Sarmad Hussain ................................................................. 99

Balanced Corpus of Contemporary Written Japanese
Kikuo Maekawa ................................................................. 101

A Basic Framework to Build a Test Collection for the Vietnamese Text Categorization
Viet Hoang-Anh, Thu Dinh-Thi-Phuong and Thang Huynh-Quyet ......................... 103

Enhanced Tools for Online Collaborative Language Resource Development
Virach Sornlertlamvanich, Thatsanee Charoenporn, Suphanut Thayaboon, Chumpol Mokarat and Hitoshi Isahara ................................................................. 105

Japanese Effort Toward Sharing Text and Speech Corpora
Shuichi Itahashi and Kōiti Hasida ................................................................. 107