ACL-IJCNLP 2009

ALR-7

7th Workshop on Asian Language Resources

Proceedings of the Workshop

6-7 August 2009
Suntec, Singapore

Production and Manufacturing by
World Scientific Publishing Co Pte Ltd
5 Toh Tuck Link
Singapore 596224

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-932432-56-5 / 1-932432-56-6

Introduction

This volume contains the papers presented at the Seventh Workshop on Asian Language Resources, held on August 6-7, 2009 in conjunction with the joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009).

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 research 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 organized 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 utilizing corpora of various types and languages. In this seventh workshop, we had 37 submissions encompassing the research from NLP community as well as the speech processing community thanks to the contributions from Oriental COCOSDA. The paper selection was highly competitive compared with the last six workshops. The program committee selected 21 papers for regular presentation, 5 papers for short presentation, and one panel discussion to enlightening the information exchange between ALR and FLaReNet initiatives.

This Seventh Workshop on Asian Language Resources would not have been succeeded without the hard work of the program committee. We would like to thank all the authors and the people who attend this workshop to share their research experiences. Last but not least, we would like to express our deep appreciation to the arrangement of the ACL-IJCNLP 2009 organizing committee and secretariat. We deeply hope this workshop further accelerates the already thriving NLP research in Asia.

Hammam Riza
Virach Sornlertlamvanich
Workshop Co-chairs

Organizers

Program Chairs:

Hammam Riza	IPTEKnet-BPPT
Virach Sornlertlamvanich	NECTEC

Program Committee:

Pushpak Bhattacharyya	*IIT-Bombay*
Thatsanee Charoenporn	*NECTEC*
Key-Sun Choi	*KAIST*
Chu-Ren Huang	*Hong Kong Polytechnic University*
Sarmad Hussain	*National University of Computer & Emerging Sciences*
Hitoshi Isahara	*NICT*
Shuichi Itahashi	*NII*
Lin-Shan Lee	*National Taiwan University*
Haizhou Li	*I2R*
Chi Mai Luong	*Institute of Information Technology, Vietnamese Academy of Science and Technology*
Yoshiki Mikami	*Nagaoka University of Technology*
Sakrange Turance Nandasara	*University of Colombo School of Computing*
Thein Oo	*Myanmar Computer Federation*
Phonpasit Phissamay	*NAST*
Oskar Riandi	*ICT Center-BPPT*
Rachel Edita O Roxas	*De La Salle University*
Kiyoaki Shirai	*JAIST*
Myint Myint Than	*Myanmar Computer Federation*
Takenobu Tokunaga	*Tokyo Institute of Technology*
Chiuyu Tseng	*Academia Sinica*
Chai Wutiwiwatchai	*NECTEC*

Book Chair:

Takenobu Tokunaga	*Tokyo Institute of Technology*

Table of Contents

Enhancing the Japanese WordNet
Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki 1

An Empirical Study of Vietnamese Noun Phrase Chunking with Discriminative Sequence Models
Le Minh Nguyen, Huong Thao Nguyen, Phuong Thai Nguyen, Tu Bao Ho and Akira Shimazu . . 9

Corpus-based Sinhala Lexicon
Ruvan Weerasinghe, Dulip Herath and Viraj Welgama 17

Analysis and Development of Urdu POS Tagged Corpus
Ahmed Muaz, Aasim Ali and Sarmad Hussain 24

Annotating Dialogue Acts to Construct Dialogue Systems for Consulting
Kiyonori Ohtake, Teruhisa Misu, Chiori Hori, Hideki Kashioka and Satoshi Nakamura 32

Assas-band, an Affix-Exception-List Based Urdu Stemmer
Qurat-ul-Ain Akram, Asma Naseer and Sarmad Hussain 40

Automated Mining Of Names Using Parallel Hindi-English Corpus
R. Mahesh K. Sinha 48

Basic Language Resources for Diverse Asian Languages: A Streamlined Approach for Resource Creation
Heather Simpson, Kazuaki Maeda and Christopher Cieri 55

Finite-State Description of Vietnamese Reduplication
Le Hong Phuong, Nguyen Thi Minh Huyen and Roussanaly Azim 63

Construction of Chinese Segmented and POS-tagged Conversational Corpora and Their Evaluations on Spontaneous Speech Recognitions
Xinhui Hu, Ryosuke Isotani and Satoshi Nakamura 70

Bengali Verb Subcategorization Frame Acquisition - A Baseline Model
Somnath Banerjee, Dipankar Das and Sivaji Bandyopadhyay 76

Phonological and Logographic Influences on Errors in Written Chinese Words
Chao-Lin Liu, Kan-Wen Tien, Min-Hua Lai, Yi-Hsuan Chuang and Shih-Hung Wu 84

Resource Report: Building Parallel Text Corpora for Multi-Domain Translation System
Budiono, Hammam Riza and Chairil Hakim 92

A Syntactic Resource for Thai: CG Treebank
Taneth Ruangrajitpakorn, Kanokorn Trakultaweekoon and Thepchai Supnithi 96

Part of Speech Tagging for Mongolian Corpus
Purev Jaimai and Odbayar Chimeddorj 103

Interaction Grammar for the Persian Language: Noun and Adjectival Phrases
Masood Ghayoomi and Bruno Guillaume 107

KTimeML: Specification of Temporal and Event Expressions in Korean Text
Seohyun Im, Hyunjo You, Hayun Jang, Seungho Nam and Hyopil Shin 115

CWN-LMF: Chinese WordNet in the Lexical Markup Framework
Lung-Hao Lee, Shu-Kai Hsieh and Chu-Ren Huang 123

Philippine Language Resources: Trends and Directions
Rachel Edita Roxas, Charibeth Cheng and Nathalie Rose Lim 131

Thai WordNet Construction
Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Tan Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich and Hitoshi Isahara 139

Query Expansion using LMF-Compliant Lexical Resources
Takenobu Tokunaga, Dain Kaplan, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Virach Sornlertlamvanich, Thatsanee Charoenporn, Yingju Xia, Chu-Ren Huang, Shu-Kai Hsieh and Kiyoaki Shirai 145

Thai National Corpus: A Progress Report
Wirote Aroonmanakun, Kachen Tansiri and Pairit Nittayanuparp 153

The FLaReNet Thematic Network: A Global Forum for Cooperation
Nicoletta Calzolari and Claudia Soria 161

Towards Building Advanced Natural Language Applications - An Overview of the Existing Primary Resources and Applications in Nepali
Bal Krishna Bal 165

Using Search Engine to Construct a Scalable Corpus for Vietnamese Lexical Development for Word Segmentation
Doan Nguyen 171

Word Segmentation Standard in Chinese, Japanese and Korean
Key-Sun Choi, Hitoshi Isahara, Kyoko Kanzaki, Hansaem Kim, Seok Mun Pak and Maosong Sun 179

Workshop Program

Thursday, August 6, 2009

8:30–9:00	Registration
9:00–9:10	Opening
9:10–9:35	*Enhancing the Japanese WordNet* Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki
9:35–10:00	*An Empirical Study of Vietnamese Noun Phrase Chunking with Discriminative Sequence Models* Le Minh Nguyen, Huong Thao Nguyen, Phuong Thai Nguyen, Tu Bao Ho and Akira Shimazu
10:00–10:30	Break
10:30–10:55	*Corpus-based Sinhala Lexicon* Ruvan Weerasinghe, Dulip Herath and Viraj Welgama
10:55–11:20	*Analysis and Development of Urdu POS Tagged Corpus* Ahmed Muaz, Aasim Ali and Sarmad Hussain
11:20–11:45	*Annotating Dialogue Acts to Construct Dialogue Systems for Consulting* Kiyonori Ohtake, Teruhisa Misu, Chiori Hori, Hideki Kashioka and Satoshi Nakamura
11:45–12:10	*Assas-band, an Affix-Exception-List Based Urdu Stemmer* Qurat-ul-Ain Akram, Asma Naseer and Sarmad Hussain
12:10–13:50	Lunch break
13:50–14:15	*Automated Mining Of Names Using Parallel Hindi-English Corpus* R. Mahesh K. Sinha
14:15–14:40	*Basic Language Resources for Diverse Asian Languages: A Streamlined Approach for Resource Creation* Heather Simpson, Kazuaki Maeda and Christopher Cieri
14:40–15:05	*Finite-State Description of Vietnamese Reduplication* Le Hong Phuong, Nguyen Thi Minh Huyen and Roussanaly Azim
15:05–15:30	*Construction of Chinese Segmented and POS-tagged Conversational Corpora and Their Evaluations on Spontaneous Speech Recognitions* Xinhui Hu, Ryosuke Isotani and Satoshi Nakamura
15:30–16:00	Break
16:00–16:15	*Bengali Verb Subcategorization Frame Acquisition - A Baseline Model* Somnath Banerjee, Dipankar Das and Sivaji Bandyopadhyay
16:15–16:30	*Phonological and Logographic Influences on Errors in Written Chinese Words* Chao-Lin Liu, Kan-Wen Tien, Min-Hua Lai, Yi-Hsuan Chuang and Shih-Hung Wu
16:30–16:45	*Resource Report: Building Parallel Text Corpora for Multi-Domain Translation System* Budiono, Hammam Riza and Chairil Hakim
16:45–17:00	*A Syntactic Resource for Thai: CG Treebank* Taneth Ruangrajitpakorn, Kanokorn Trakultaweekoon and Thepchai Supnithi
17:00–17:15	*Part of Speech Tagging for Mongolian Corpus* Purev Jaimai and Odbayar Chimeddorj

Friday, August 7, 2009

8:45–9:10	*Interaction Grammar for the Persian Language: Noun and Adjectival Phrases* Masood Ghayoomi and Bruno Guillaume
9:10–9:35	*KTimeML: Specification of Temporal and Event Expressions in Korean Text* Seohyun Im, Hyunjo You, Hayun Jang, Seungho Nam and Hyopil Shin
9:35–10:00	*CWN-LMF: Chinese WordNet in the Lexical Markup Framework* Lung-Hao Lee, Shu-Kai Hsieh and Chu-Ren Huang
10:00–10:30	Break
10:30–10:55	*Philippine Language Resources: Trends and Directions* Rachel Edita Roxas, Charibeth Cheng and Nathalie Rose Lim
10:55–11:20	*Thai WordNet Construction* Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Tan Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich and Hitoshi Isahara
11:20–11:45	*Query Expansion using LMF-Compliant Lexical Resources* Takenobu Tokunaga, Dain Kaplan, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Virach Sornlertlamvanich, Thatsanee Charoenporn, Yingju Xia, Chu-Ren Huang, Shu-Kai Hsieh and Kiyoaki Shirai
11:45–12:10	*Thai National Corpus: A Progress Report* Wirote Aroonmanakun, Kachen Tansiri and Pairit Nittayanuparp
12:10–13:50	Lunch break
13:50–14:15	*The FLaReNet Thematic Network: A Global Forum for Cooperation* Nicoletta Calzolari and Claudia Soria
14:15–14:40	*Towards Building Advanced Natural Language Applications - An Overview of the Existing Primary Resources and Applications in Nepali* Bal Krishna Bal
14:40–15:05	*Using Search Engine to Construct a Scalable Corpus for Vietnamese Lexical Development for Word Segmentation* Doan Nguyen
15:05–15:30	*Word Segmentation Standard in Chinese, Japanese and Korean* Key-Sun Choi, Hitoshi Isahara, Kyoko Kanzaki, Hansaem Kim, Seok Mun Pak and Maosong Sun
15:30–16:00	Break
16:00–17:50	Panel discussion “ALR and FLaReNet”
17:50–18:00	Closing