COLING 2010

23rd International Conference on Computational Linguistics

Proceedings of the

8th Workshop on Asian Language Resources

21-22 August 2010
Beijing, China

Produced by
Chinese Information Processing Society of China

To order the CD of Coling 2010 and its Workshop Proceedings, please contact:

Chinese Information Processing Society of China
No.4, Southern Fourth Street
Haidian District, Beijing, 100190
China
Tel: +86-010-62562916
Fax: +86-010-62562916
cips@iscas.ac.cn

Preface

Language resources play a central role in statistical and learning-based approaches to natural language processing. Thus, recent research has put great emphasis in building these resources for target languages. Parallel resources across various languages are also being developed for multilingual processing. These include lexica and corpora with multiple levels of annotations. Though significant progress has been achieved in modeling few of the Asian languages, with the wider spread of ICT use across the region, there is a growing interest in this field from other linguistic communities. As research in the field matures across Asia, there is a growing need for developing language resources. However the region is not only short in the linguistic resources for more than 2200 language spoken in the region, there is also lack of experience in the researchers to develop these resources. As the efforts to develop the linguistic resources increases, there is also need to coordinate the efforts to develop common frameworks and processes so that these resources can be used by various groups of researchers equally effectively.

The workshop is organised by the Asian Language Resources Committee (ALRC) of Asian Federation for Natural Language Processing (AFNLP). The aim are to chart and catalogue the status of Asian Language Resources, to investigate and discuss the problems related to the standards and specification on creating and sharing various levels of language resources, to promote a dialogue between developers and users of various language resources in order to address any gaps in language resources and practical applications, and to nurture collaboration in their development and use, to provide opportunity for researchers from Asia to collaborate with researchers in other regions.

This is the eighth workshop in the series, and has been representative, with 35 submissions for Asian languages, including Bahasa Indonesia, Chinese, Dzongkha, Hindi, Japanese, Khmer, Sindhi, Sinhala, Thai, Turkish and Urdu, of which 22 have been finalized for presentation. We would like to thank the authors for their submissions and the Program Committee for their timely reviews. We hope that ALR workshops will continue to encourage researchers to focus on developing and sharing resources for Asian languages, an essential requirement for research in NLP.

ALR8 Workshop Organizers

Organizers:

Sarmad Hussain, CLE-KICS, UET, Pakistan
Virach Sornlertlamvanich, NECTEC, Thailand
Hammam Riza, BPPT, Indonesia
(on behalf of ALRC, AFNLP)

Program Committee:

Mirna Adriani
Pushpak Bhatacharyya
Francis Bond
Miriam Butt
Thatsanee Charoenporn
Key-Sun Choi
Ananlada Chotimongkol
Jennifer Cole
Li Haizhou
Choochart Haruechaiyasak
Hitoshi Isahara
Alisa Kongthon
Krit Kosawat
Yoshiki Mikami
Cholwich Nattee
Rachel Roxas
Dipti Sharma
Kiyoaki Shirai
Thepchai Supnithi
Thanaruk Theeramunkong
Takenobu Tokunaga
Ruvan Weerasinghe
Chai Wutiwiwatchai
Yogendra Yadava

Table of Contents

Conference Program

Saturday August 21, 2010

Semantics

9:00–9:25 *A Thesaurus of Predicate-Argument Structure for Japanese Verbs to Deal with Granularity of Verb Meanings*
Koichi Takeuchi, Kentaro Inui, Nao Takeuchi and Atsushi Fujita

9:25–9:50 *Collaborative Work on Indonesian WordNet through Asian WordNet (AWN)*
Chairil Hakim, Budiono Budiono and Hammam Riza

9:50–10:15 *Considerations on Automatic Mapping Large-Scale Heterogeneous Language Resources: Sejong Semantic Classes and KorLex*
Heum Park, Ae sun Yoon, Woo Chul Park and Hyuk-Chul Kwon

10:15–10:40 *Sequential Tagging of Semantic Roles on Chinese FrameNet*
Jihong LI, Ruibo WANG and Yahui GAO

10:40–11:00 Coffee Break

Semantics, Sentiment and Opinion

11:00–11:25 *Augmenting a Bilingual Lexicon with Information for Word Translation Disambiguation*
Takashi Tsunakawa and Hiroyuki Kaji

11:25–11:50 *Construction of bilingual multimodal corpora of referring expressions in collaborative problem solving*
Takenobu Tokunaga, Ryu Iida, Masaaki Yasuhara, Asuka Terai, David Morris and Anja Belz

11:50–12:15 *Labeling Emotion in Bengali Blog Corpus A Fine Grained Tagging at Sentence Level*
Dipankar Das and Sivaji Bandyopadhyay

12:15–12:40 *SentiWordNet for Indian Languages*
Amitava Das and Sivaji Bandyopadhyay

12:40–14:10 Lunch Break

Saturday August 21, 2010 (continued)

Opinion and Information Retrival

14:10–14:35 *Constructing Thai Opinion Mining Resource: A Case Study on Hotel Reviews*
Choochart Haruechaiyasak, Alisa Kongthon, Pornpimon Palingoon and Chatchawal Sangkeettrakarn

14:35–15:00 *The Annotation of Event Schema in Chinese*
Hongjian Zou, Erhong Yang, Yan Gao and Qingqing Zeng

15:00–15:25 *Query Expansion for Khmer Information Retrieval*
Channa Van and Wataru Kameyama

15:25–16:00 Coffee Break

Text Corpus

16:00–16:25 *Word Segmentation for Urdu OCR System*
Misbah Akram and Sarmad Hussain

16:25–16:50 *Dzongkha Word Segmentation*
Sithar Norbu, Pema Choejey, Tenzin Dendup, Sarmad Hussain and Ahmed Muaz

16:50–17:15 *Building NLP resources for Dzongkha: A Tagset and A Tagged Corpus*
Chungku Chungku, Jurmey Rabgay and Gertrud Faaß

17:15–17:40 *Unaccusative/Unergative Distinction in Turkish: A Connectionist Approach*
Cengiz Acarturk and Deniz Zeyrek

Sunday August 22, 2010

	Grammars and Parsing
9:00–9:25	*A Preliminary Work on Hindi Causatives* Rafiya Begum and Dipti Misra Sharma
9:25–9:50	*A Supervised Learning based Chunking in Thai using Categorial Grammar* Thepchai Supnithi, Chanon Onman, Peerachet Porkaew, Taneth Ruangrajitpakorn, Kanokorn Trakultaweekoon and Asanee Kawtrakul
9:50–10:15	*A hybrid approach to Urdu verb phrase chunking* Wajid Ali and Sarmad Hussain
10:15–10:40	*Development of the Korean Resource Grammar: Towards Grammar Customization* Sanghoun Song, Jong-Bok Kim, Francis Bond and Jaehyung Yang
10:40–11:00	Coffee Break
	Grammars and Applications
11:00–11:25	*An Open Source Urdu Resource Grammar* Shafqat Mumtaz Virk, Muhammad Humayoun and Aarne Ranta
11:25–11:50	*A Current Status of Thai Categorial Grammars and Their Applications* Taneth Ruangrajitpakorn and Thepchai Supnithi
11:50–12:15	*Chained Machine Translation Using Morphemes as Pivot Language* Li Wen, Chen Lei, Han Wudabala and Li Miao
12:15	Concluding Remarks and Discussion