IJCNLP-05

Fourth SIGHAN Workshop on Chinese Language Processing

Proceedings of the Workshop

14-15 October 2005
Jeju Island, Korea

Preface

We would like welcome you to the Fourth SIGHAN Workshop on Chinese Language Processing held on Jeju Island, Republic of Korea, October 14-15, 2005. Building on previous workshops in Sapporo and Barcelona, SIGHAN4 brings together researchers from Asia and around the world to report recent developments in many aspects of Chinese language processing. This year for the first time the workshop is being held under the auspices of the Asia Federation of Natural Language Processing, in conjunction with the Second International Joint Conference on Natural Language Processing. We hope that this collaboration will encourage the participation of a broad group of researchers.

This year's workshop has assembled a diverse general technical program. With ten regular papers and six poster presentations, the general session spans topics from segmentation and part-of-speech tagging to semantics, discourse, and dialogue. Presenters represent institutions in Asia, North America, and Europe. We hope that these papers will engage the participants and promote wide-ranging discussion.

This year's workshop also presents the results of the Second International Chinese Word Segmentation Bakeoff sponsored by SIGHAN. The First Bakeoff, held in conjunction with SIGHAN2 at ACL 2003 in Sapporo, quickly became the standard evaluation for segmentation systems, and we hope this new evaluation will prove to be a worthy successor. Twenty-three groups from eight countries participated in the evaluation, and their results as well as an overview comprise the Bakeoff section of the workshop.

We thank all who submitted papers. We greatly appreciate the time and insightful reviews provided by the members of the Program Committee. We also thank Tom Emerson for his great efforts in bringing together the Segmentation Bakeoff and all the bakeoff participants for their contribution to the success of the event. Finally, we would also like to thank the SIGHAN Board, the IJCNLP-05 workshop chairs, Yuji Matsumoto and Laurent Romary, and the publications chairs, Olivia Kwong and Jian Su, for their guidance and support.

Chu-Ren Huang, Gina-Anne Levow , SIGHAN4 Co-chairs

Organizers

Co-Chairs:

Chu-Ren Huang - Academia Sinica, Taiwan
Gina-Anne Levow - University of Chicago, USA

Segmentation Bakeoff Organizer:

Tom Emerson - Basis Technology Corp, USA

Program Committee

Joyce Chai - Michigan State Univ, USA
Keh-Jiann Chen - Academia Sinica, Taiwan
Zheng Chen - Microsoft Research Asia, China
Tom Emerson - Basis Technology Corp, USA
Pascale Fung - Hong Kong University of Science and Technology, Hong Kong
Tingting He - Huazhong Normal University, China
Chu-Ren Huang - Academia Sinica, Taiwan
K.L.Kwok - Queens College, USA
Tom Lai - City Univ. of Hong Kong, Hong Kong
Gina-Anne Levow - University of Chicago, USA Mingjing Li - Microsoft Research Asia, China
Mu Li - Microsoft Research Asia, China
Qin Lu - The Hong Kong Polytechnic University, Hong Kong
Qing Ma - Ryukoku University, Japan
Masaki Murata - Communications Research Laboratory, Japan
Shimei Pan - IBM, USA
Richard Sproat - University of Illinois, Urbana-Champaign, USA
Keh-Yih Su - Behavior Design Corporation, Taiwan
Maosong Sun - Tsinghua University, China
Shu-chuan Tseng - Academia Sinica, Taiwan
Benjamin Tsou - City University of Hong Kong, Hong Kong
Dekai Wu - Hong Kong University of Science and Technology, Hong Kong
Yunfang Wu - Peking University, China
Nianwen Xue - University of Pennsylvania, USA
Qiang Zhou - Tsinghua University, China

Further Information

Gina-Anne Levow
Department of Computer Science
University of Chicago
1100 E 58th St.
Chicago, IL 60637 USA

Fourth SIGHAN Workshop on Chinese Language Processing
Workshop Program

October 14

08:30 Opening Remarks

Session I: Segmentation, Input, and Parsing

08:35 *Detecting Segmentation Errors in Chinese Annotated Corpus*
Chengjie Sun, Chang-Ning Huang, Xiaolong Wang, and Mu Li

08:55 *Using Word-Pair Identifier to Improve Chinese Input System*
Jia-Lin Tsai

09:15 *Chinese Deterministic Dependency Analyzer:*
Examining Effects of Global Features and Root Node Finder
Yuchang Cheng, Masayuki Asahara, and Yuji Matsumoto

09:35 *Break*

Session II: Part-of-Speech and Named Entity Recognition

09:45 *Chinese Classifier Assignment Using SVMs*
Hui Guo and Huayan Zhang

10:05 *Morphological features help POS tagging of unknown words across language varieties*
Huihsin Tseng, Daniel Jurafsky, and Christopher Manning

10:25 *Product Named Entity Recognition Based on Hierarchical Hidden Markov Model*
Feifan Liu, Jun Zhao, Bibo Lv, Bo Xu and Hao Yu

10:45 *Break*

Session III: Semantics and Dialogue

11:00 *Chinese Sketch Engine and the Extraction of Grammatical Collocations*
Chu-Ren Huang, Adam Kilgarriff, Yiching Wu, Chih-Ming Chiu,
Simon Smith, Pavel Rychly, Ming-Hong Bai and Keh-Jiann Chen

11:20 *Word Meaning Inducing via Character Ontology:*
A Survey on the Semantic Prediction of Chinese Two-Character Words
Shu-Kai Hsieh

11:40 *Domain Specific Word Extraction from Hierarchical Web Documents:*
A First Step Toward Building Lexicon Trees from Web Corpora
Jing-Shin Chang

12:00 *Turn-Taking in Mandarin Dialogue: Interactions of Tone and Intonation*
Gina-Anne Levow

12:20 *Lunch*

Session IV: General Session Posters

14:00 *Poster Briefings* (5 minutes each)

Learning a Log-Linear Model with Bilingual Phrase-Pair Features for Statistical Machine Translation
Bing Zhao and Alex Waibel

Integrating Collocation Features in Chinese Word Sense Disambiguation
Wanyin Li, Qin Lu, and Wenjie Li

NIL Is Not Nothing: Recognition of Chinese Network Informal Language Expressions
Yunqing Xia, Kam-Fai Wong and Wei Gao

The Robustness of Domain Lexico-Taxonomy: Expanding Domain Lexicon with CiLin
Chu-Ren Huang, Xiang-Bing Li and Jia-Fei Hong

Some Studies on Chinese Domain Knowledge Dictionary and Its Application to Text Classification
Jingbo Zhu and Wenliang Chen

Resolving Pronominal References in Chinese with the Hobbs Algorithm
Susan Converse

14:30 *Poster Viewing*

15:40 *Break*

Session V: Bakeoff Overview and Presentations

16:00 *The Second International Chinese Word Segmentation Bakeoff*
Thomas Emerson

16:20 *Combination of Machine Learning Methods for Optimum Chinese Word Segmentation*
Masayuki Asahara, Kenta Fukuoka, Ai Azuma, Chooi-Ling Goh, Yotaro Watanabe, Yuji Matsumoto, and Takashi Tsuzuki

16:35 *Unigram Language Model for Chinese Word Segmentation*
Aitao Chen, Yiping Zhou, Anne Zhang, and Gordon Sun

16:50 *Report To BMM-based Chinese Word Segmentor with Context-based Unknown Word Identifier for the Second International Chinese word Segmentation Bakeoff*
Jia-Lin Tsai

October 15

Session VI: Bakeoff Presentations

09:30 *An Example-Based Chinese Word Segmentation System for CWSB-2*
Chunyu Kit and Xiaoyue Liu

09:45 *Chinese Word Segmentation in FTRD Beijing*
Heng Li, Yuan Dong, Xinnian Mao, Haila Wang, and Wu Liu

10:00 *Perceptron Learning for Chinese Word Segmentation*
Yaoyong Li, Chuanjiang Miao, Kalina Bontcheva, and Hamish Cunningham

10:15 *Data-driven Language Independent Word Segmentation Using Character-Level Information*
Dong-Hee Lim and Seung-Shik Kang

10:35 *Break*

Session VII: Bakeoff Presentations

11:00 *A Maximum Entropy Approach to Chinese Word Segmentation*
Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo

11:15 *Description of the HKU Chinese Word Segmentation System for Sighan Bakeoff 2005*
Guohong Fu, Kang-Kwong Luke, and Percy Ping-Wai Wong

11:30 *A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005*
Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning

11:45 *Chinese Word Segmentation with Multiple Postprocessors in HIT-IRLab*
Huipeng Zhang, Ting Liu, Jinshan Ma, and Xiantao Liao

12:00 *SIGHAN Business Meeting*

Table of Contents