
IJCNLP 2008

**Sixth SIGHAN Workshop
on
Chinese Language Processing**

Proceedings of the Workshop

11–12 January 2008
Hyderabad, India

Preface

Welcome to Hyderabad for the Sixth SIGHAN Workshop on Chinese Language Processing! As the workshop co-chairs, we are indeed honored to pen the first few words in the SIGHAN-6 proceedings. Over the last few years, exciting research endeavors in Chinese language processing have been pursued vigorously all over the world. We feel privileged to be able to chair the Sixth SIGHAN, particularly at this juncture in our history when the Chinese language is receiving worldwide attention which places us at an important period of growth and change.

SIGHAN-6 received 14 regular paper submissions this year, 9 of which are accepted for presentation, contributing to yet another stimulating and inspiring scientific program. We would like to thank all the authors who submitted their research work, and all the reviewers who have put an immense amount of timely and quality work into the paper review. We would also like to express our gratitude and appreciation to the chair of SIGHAN, Prof. Benjamin Tsou, who has led SIGHAN to what it is today, especially for his invaluable advice at various stages of the development of SIGHAN-6.

Our workshop also uniquely features the Fourth International Chinese Language Processing Bakeoff, which was jointly held with the First CIPS Chinese Language Processing Evaluation over the summer of 2007. In addition to the classic Chinese word segmentation and named entity recognition tasks, there is a new track on Chinese POS-tagging. The bakeoff was co-organized by SIGHAN, ChineseLDC, and the Verifying Center on Chinese Language and Character Standards of the State Language Commission of PRC, and coordinated by Dr. Guangjin Jin of the Institute of Applied Linguistics, MOE, China. Special thanks to Dr. Jin and the bakeoff organizing committee for organizing another successful bakeoff. The generous support from the benchmarking corpora providers (Academia Sinica; City University of Hong Kong; Institute of Applied Linguistics, MOE, China; Microsoft Research Asia; Peking University; Shanxi University; and University of Colorado) is greatly appreciated. Certainly the bakeoff cannot stand without the support from the 28 participating teams. We have collected 23 bakeoff system reports in this volume.

SIGHAN-6 marks its second time to co-locate with IJCNLP, the flagship conference of the Asian Federation of Natural Language Processing. The organization of our workshop will not be so smooth without all the logistic support from the IJCNLP-08 committee, particularly the Workshop Committee, the Publication Chair Dr. Jing-Shin Chang, as well as everyone in the Local Organizing Committee. To them we would like to express our heartiest gratitude.

Last but not least, thank you for your active participation. Enjoy our program, and we wish you an educational and entertainment-filled experience in Hyderabad.

Olivia Oi Yee Kwong and Haizhou Li
November 2007

Organizers

Workshop Co-Chairs:

Olivia Oi Yee Kwong, City University of Hong Kong
Haizhou Li, Institute for Infocomm Research

Bakeoff Organizing Committee:

Guangjin Jin, Institute of Applied Linguistics, MOE, PRC (Co-ordinator)
Xiao Chen, City University of Hong Kong
Yongsheng Guo, Institute of Applied Linguistics, MOE, PRC
Changning Huang, Microsoft Research Asia
Qun Liu, Chinese Academy of Sciences

Program Committee:

Keh-Jiann Chen, Academia Sinica
Minghui Dong, Institute for Infocomm Research
Jianfeng Gao, Microsoft
Chu-Ren Huang, Academia Sinica
Xuanjing Huang, Fudan University
Donghong Ji, Wuhan University
Daniel Jurafsky, Stanford University
Chunyu Kit, City University of Hong Kong
Kui-Lam Kwok, Queens College
Gina-Anne Levow, University of Chicago
Dekang Lin, Google
Qun Liu, Chinese Academy of Sciences
Qin Lu, Hong Kong Polytechnic University
Qing Ma, Ryukoku University
Jianyun Nie, University of Montreal
Hwee Tou Ng, National University of Singapore
Martha Palmer, University of Colorado
Scott Piao, University of Manchester
Richard Sproat, University of Illinois at Urbana-Champaign
Keh-Yih Su, Behavior Design Corporation
Maosong Sun, Tsinghua University
Bing Swen, Peking University
Benjamin Tsou, City University of Hong Kong
Haifeng Wang, Toshiba (China) R&D Center
Kam-Fai Wong, Chinese University of Hong Kong
Dekai Wu, Hong Kong University of Science and Technology
Yujie Zhang, National Institute of Information and Communications Technology of Japan
Jun Zhao, Chinese Academy of Sciences

Tiejun Zhao, Harbin Institute of Technology
Ming Zhou, Microsoft Research Asia
Jingbo Zhu, Northeastern University

Additional Reviewers:

Xiangyu Duan, Chinese Academy of Sciences
Wei Gao, Chinese University of Hong Kong
Shiqi Li, Harbin Institute of Technology
Nianwen Xue, University of Colorado
Yongzheng Xue, Harbin Institute of Technology

Workshop Program

Friday, 11 January 2008

09:00–09:10 Opening Remarks

Session 1: Translation and Transliteration

09:10–09:35 *An Example-based Decoder for Spoken Language Machine Translation*
Zhou-Jun Li, Wen-Han Chao and Yue-Xin Chen

09:35–10:00 *Automatic Extraction of English-Chinese Transliteration Pairs using Dynamic Window and Tokenizer*
Chengguo Jin, Seung-Hoon Na, Dong-Il Kim and Jong-Hyeok Lee

10:00–10:25 *Mining Transliterations from Web Query Results: An Incremental Approach*
Jin-Shea Kuo, Haizhou Li and Chih-Lung Lin

10:25–11:00 Break

Session 2: Information Extraction and Word Sense Disambiguation

11:00–11:25 *An Effective Hybrid Machine Learning Approach for Coreference Resolution*
Feiliang Ren and Jingbo Zhu

11:25–11:50 *Use of Event Types for Temporal Relation Identification in Chinese Text*
Yuchang Cheng, Masayuki Asahara and Yuji Matsumoto

11:50–12:15 *Chinese Word Sense Disambiguation with PageRank and HowNet*
Jinghua Wang, Jianyi Liu and Ping Zhang

12:15–14:15 Lunch

Friday, 11 January 2008 (continued)

Session 3: Word Segmentation and Parsing

- 14:15–14:40 *Stochastic Dependency Parsing Based on A* Admissible Search*
Bor-shen Lin
- 14:40–15:05 *Analyzing Chinese Synthetic Words with Tree-based Information and a Survey on Chinese Morphologically Derived Words*
Jia Lu, Masayuki Asahara and Yuji Matsumoto
- 15:05–15:30 *Which Performs Better on In-Vocabulary Word Segmentation: Based on Word or Character?*
Zhenxing Wang, Changning Huang and Jingbo Zhu

15:30–16:00 Break

Session 4: Bakeoff Overview and Presentations

- 16:00–16:20 *The Fourth International Chinese Language Processing Bakeoff: Chinese Word Segmentation, Named Entity Recognition and Chinese POS Tagging*
Guangjin Jin and Xiao Chen
- 16:20–16:40 *A Two-Stage Approach to Chinese Part-of-Speech Tagging*
Aitao Chen, Ya Zhang and Gordon Sun
- 16:40–17:00 *NOKIA Research Center Beijing Chinese Word Segmentation System for the SIGHAN Bakeoff 2007*
Jiang Li, Rile Hu, Guohua Zhang, Yuezhong Tang, Zhanjiang Song and Xia Wang
- 17:00–17:20 *Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields*
Xinnian Mao, Yuan Dong, Saike He, Sencheng Bao and Haila Wang

Saturday, 12 January 2008

Session 5: Bakeoff Presentations

- 09:10–09:30 *BUPT Systems in the SIGHAN Bakeoff 2007*
Ying Qin, Caixia Yuan, Jiashen Sun and Xiaojie Wang
- 09:30–09:50 *The Character-based CRF Segmenter of MSRA&NEU for the 4th Bakeoff*
Zhenxing Wang, Changning Huang and Jingbo Zhu
- 09:50–10:10 *Chinese NER Using CRFs and Logic for the Fourth SIGHAN Bakeoff*
Xiaofeng Yu, Wai Lam, Shing-Kit Chan, Yiu Kei Wu and Bo Chen
- 10:10–10:30 *Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition*
Hai Zhao and Chunyu Kit
- 10:30–11:00 Break
- 11:00–12:30 SIGHAN Business Meeting

Table of Contents

Preface	iii
Organizers	v
Workshop Program	vii
<i>An Example-based Decoder for Spoken Language Machine Translation</i> Zhou-Jun Li, Wen-Han Chao and Yue-Xin Chen	1
<i>Automatic Extraction of English-Chinese Transliteration Pairs using Dynamic Window and Tokenizer</i> Chengguo Jin, Seung-Hoon Na, Dong-Il Kim and Jong-Hyeok Lee	9
<i>Mining Transliterations from Web Query Results: An Incremental Approach</i> Jin-Shea Kuo, Haizhou Li and Chih-Lung Lin	16
<i>An Effective Hybrid Machine Learning Approach for Coreference Resolution</i> Feiliang Ren and Jingbo Zhu	24
<i>Use of Event Types for Temporal Relation Identification in Chinese Text</i> Yuchang Cheng, Masayuki Asahara and Yuji Matsumoto	31
<i>Chinese Word Sense Disambiguation with PageRank and HowNet</i> Jinghua Wang, Jianyi Liu and Ping Zhang	39
<i>Stochastic Dependency Parsing Based on A* Admissible Search</i> Bor-shen Lin	45
<i>Analyzing Chinese Synthetic Words with Tree-based Information and a Survey on Chinese Morphologically Derived Words</i> Jia Lu, Masayuki Asahara and Yuji Matsumoto	53
<i>Which Performs Better on In-Vocabulary Word Segmentation: Based on Word or Character?</i> Zhenxing Wang, Changning Huang and Jingbo Zhu	61
<i>The Fourth International Chinese Language Processing Bakeoff: Chinese Word Segmentation, Named Entity Recognition and Chinese POS Tagging</i> Guangjin Jin and Xiao Chen	69
<i>A Two-Stage Approach to Chinese Part-of-Speech Tagging</i> Aitao Chen, Ya Zhang and Gordon Sun	82
<i>NOKIA Research Center Beijing Chinese Word Segmentation System for the SIGHAN Bakeoff 2007</i> Jiang Li, Rile Hu, Guohua Zhang, Yuezhong Tang, Zhanjiang Song and Xia Wang	86

<i>Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields</i> Xinnian Mao, Yuan Dong, Saike He, Sencheng Bao and Haila Wang	90
<i>BUPT Systems in the SIGHAN Bakeoff 2007</i> Ying Qin, Caixia Yuan, Jiashen Sun and Xiaojie Wang	94
<i>The Character-based CRF Segmenter of MSRA&NEU for the 4th Bakeoff</i> Zhenxing Wang, Changning Huang and Jingbo Zhu	98
<i>Chinese NER Using CRFs and Logic for the Fourth SIGHAN Bakeoff</i> Xiaofeng Yu, Wai Lam, Shing-Kit Chan, Yiu Kei Wu and Bo Chen	102
<i>Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition</i> Hai Zhao and Chunyu Kit	106
<i>An Agent-Based Approach to Chinese Word Segmentation</i> Samuel W.K. Chan and Mickey W.C. Chong	112
<i>Nanjing Normal University Segmenter for the Fourth SIGHAN Bakeoff</i> Xiaohe Chen, Bin Li, Junzhi Lu, Hongdong Nian and Xuri Tang	115
<i>Two Step Chinese Named Entity Recognition Based on Conditional Random Fields Models</i> Yuanyong Feng, Ruihong Huang and Le Sun	120
<i>A Morpheme-based Part-of-Speech Tagger for Chinese</i> Guohong Fu and Jonathan J. Webster	124
<i>Chinese Named Entity Recognition and Word Segmentation Based on Character</i> Jingzhou He and Houfeng Wang	128
<i>HMM and CRF Based Hybrid Model for Chinese Lexical Analysis</i> Degen Huang, Xiao Sun, Shidou Jiao, Lishuang Li, Zhuoye Ding and Ru Wan	133
<i>Chinese Tagging Based on Maximum Entropy Model</i> Ka Seng Leong, Fai Wong, Yiping Li and Ming Chui Dong	138
<i>Training a Perceptron with Global and Local Features for Chinese Word Segmentation</i> Dong Song and Anoop Sarkar	143
<i>A Study of Chinese Lexical Analysis Based on Discriminative Models</i> Guang-Lu Sun, Cheng-Jie Sun, Ke Sun and Xiao-Long Wang	147
<i>Word Boundary Token Model for the SIGHAN Bakeoff 2007</i> Jia-Lin Tsai	151
<i>An Improved CRF based Chinese Language Processing System for SIGHAN Bakeoff 2007</i> Xihong Wu, Xiaojun Lin, Xinhao Wang, Chunyao Wu, Yaozhong Zhang and Dianhai Yu	155

<i>Description of the NCU Chinese Word Segmentation and Part-of-Speech Tagging for SIGHAN Bakeoff 2007</i>	
Yu-Chieh Wu, Jie-Chi Yang and Yue-Shi Lee	161
<i>CRF-based Hybrid Model for Word Segmentation, NER and even POS Tagging</i>	
Zhiting Xu, Xian Qian, Yuejie Zhang and Yaqian Zhou	167
<i>CRFs-Based Named Entity Recognition Incorporated with Heuristic Entity List Searching</i>	
Fan Yang, Jun Zhao and Bo Zou	171
<i>A Chinese Word Segmentation System Based on Cascade Model</i>	
Jianfeng Zhang, Jiaheng Zheng, Hu Zhang and Hongye Tan	175
<i>Achilles: NiCT/ATR Chinese Morphological Analyzer for the Fourth Sighan Bakeoff</i>	
Ruiqiang Zhang and Eiichiro Sumita	178
Author Index	183

