2006

COLING • ACL

COLING•ACL 2006

Fifth SIGHAN Workshop on Chinese Language Processing

Proceedings of the Workshop

Chairs:
Hwee Tou Ng and Olivia O. Y. Kwong

22-23 July 2006
Sydney, Australia

Production and Manufacturing by
BPA Digital
11 Evans St
Burwood VIC 3125
AUSTRALIA

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 1-932432-70-1

Table of Contents

Preface

The Fifth SIGHAN Workshop on Chinese Language Processing will be held in Sydney, Australia, on July 22 – 23, 2006, co-located with COLING/ACL 2006. The annual SIGHAN workshop is an international forum for presenting the latest research on Chinese language processing. This year, the workshop attracted 24 submissions to the main session, out of which we accepted 8 as oral paper presentations and 6 as poster paper presentations.

The Third International Chinese Language Processing Bakeoff was also organized in conjunction with this workshop. In addition to the Chinese word segmentation task of the first two bakeoffs, this year's bakeoff also included the Chinese named entity recognition task. Altogether 29 teams participated in the bakeoff, organized by Gina-Anne Levow and Olivia Oi Yee Kwong. The increase in the number of participating teams compared to the last two bakeoffs is testimony to the healthy growth of research interest in Chinese language processing.

We would like to thank all authors who submitted papers to this workshop, and all program committee members who worked hard to review the submissions. Special thanks to Gina-Anne Levow who did a fantastic job organizing a successful bakeoff. We would also like to acknowledge the help of the following people who provided the corpora used in the bakeoff: Keh-Jiann Chen and Henning Chiu (Academia Sinica), Mu Li (Microsoft Research Asia), Martha Palmer and Nianwen Xue (University of Pennsylvania/University of Colorado), Stephanie Strassel (Linguistic Data Consortium), and Benjamin K. Tsou and Olivia Oi Yee Kwong (City University of Hong Kong). We also thank Benjamin K. Tsou, Martha Palmer, and Suzanne Stevenson for their guidance and advice in our organization of this workshop.

We hope that you will have a great time attending this workshop in Sydney!

Hwee Tou Ng and Olivia Oi Yee Kwong
June 2006

Organizers

Workshop Chair:

Hwee Tou Ng, National University of Singapore

Workshop Co-Chair:

Olivia Oi Yee Kwong, City University of Hong Kong

Bakeoff Coordinators:

Gina-Anne Levow, University of Chicago
Olivia Oi Yee Kwong, City University of Hong Kong

Program Committee:

Aitao Chen, Yahoo!
Keh-Jiann Chen, Academia Sinica
David Chiang, USC Information Sciences Institute
Pascale Fung, Hong Kong University of Science and Technology
Jianfeng Gao, Microsoft Research
Julia Hockenmaier, University of Pennsylvania
Xuanjing Huang, Fudan University
Daniel Jurafsky, Stanford University
Kui-Lam Kwok, Queens College, CUNY
Gina-Anne Levow, University of Chicago
Haizhou Li, Institute for Infocomm Research
Mu Li, Microsoft Research Asia
Qun Liu, Chinese Academy of Sciences
Xiaoqiang Luo, IBM
Qing Ma, Ryukoku University
Yuji Matsumoto, Nara Institute of Science and Technology
Martha Palmer, University of Colorado
Fuchun Peng, Yahoo!
Richard Sproat, University of Illinois at Urbana-Champaign
Maosong Sun, Tsinghua University
Haifeng Wang, Toshiba (China) R&D Centre
Kam-Fai Wong, Chinese University of Hong Kong
Fei Xia, University of Washington at Seattle
Nianwen Xue, University of Pennsylvania
Jun Zhao, Chinese Academy of Sciences
Tiejun Zhao, Harbin Institute of Technology
Guodong Zhou, Institute for Infocomm Research
Ming Zhou, Microsoft Research Asia
Jingbo Zhu, Northeastern University

Workshop Program

Saturday, 22 July 2006

09:00–09:10 Opening Remarks

Session 1: Lexicon Construction

09:10–09:35 *Improving Context Vector Models by Feature Clustering for Automatic Thesaurus Construction*
Jia-Ming You and Keh-Jiann Chen

09:35–10:00 *Regional Variation of Domain-Specific Lexical Items: Toward a Pan-Chinese Lexical Resource*
Oi Yee Kwong and Benjamin K. Tsou

10:00–10:25 *Mining Atomic Chinese Abbreviation Pairs: A Probabilistic Model for Single Character Word Recovery*
Jing-Shin Chang and Wei-Lun Teng

10:25–11:00 Break

Session 2: Part-of-Speech Tagging, Semantics, and Discourse

11:00–11:25 *Features, Bagging, and System Combination for the Chinese POS Tagging Task*
Fei Xia and Lap Cheung

11:25–11:50 *Semantic Analysis of Chinese Garden-Path Sentences*
Yaohong Jin

11:50–12:15 *A Clustering Approach for Unsupervised Chinese Coreference Resolution*
Chi-shing Wang and Grace Ngai

12:15–13:45 Lunch

Saturday, 22 July 2006 (continued)

Session 3: Translation and Retrieval

13:45–14:10 *Latent Features in Automatic Tense Translation between Chinese and English*
Yang Ye, Victoria Li Fossum and Steven Abney

14:10–14:35 *Cluster-Based Language Model for Sentence Retrieval in Chinese Question Answering*
Youzheng Wu, Jun Zhao and Bo Xu

Session 4: Posters

14:35–15:30 *The Role of Lexical Resources in CJK Natural Language Processing*
Jack Halpern

Hybrid Models for Chinese Named Entity Recognition
Lishuang Li, Tingting Mao, Degen Huang and Yuansheng Yang

Realization of the Chinese BA-construction in an English-Chinese Machine Translation System
Xiaohong Wu, Sylviane Cardey and Peter Greenfield

A Hybrid Approach to Chinese Base Noun Phrase Chunking
Fang Xu, Chengqing Zong and Jun Zhao

A SVM-Based Model for Chinese Functional Chunk Parsing
Yingze Zhao and Qiang Zhou

Broadcast Audio and Video Bimodal Corpus Exploitation and Application
Yu Zou, Min Hou, Yudong Chen, Fengguo Hu and Li Fu

15:30–16:00 Break

Saturday, 22 July 2006 (continued)

Session 5: Bakeoff Overview and Presentations

16:00–16:20 *The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition*
Gina-Anne Levow

16:20–16:35 *Chinese Named Entity Recognition with Conditional Random Fields*
Wenliang Chen, Yujie Zhang and Hitoshi Isahara

16:35–16:50 *France Telecom R&D Beijing Word Segmenter for Sighan Bakeoff 2006*
Wu Liu, Heng Li, Yuan Dong, Nan He, Haitao Luo and Haila Wang

16:50–17:05 *Voting between Dictionary-Based and Subword Tagging Models for Chinese Word Segmentation*
Dong Song and Anoop Sarkar

17:05–17:20 *BMM-Based Chinese Word Segmentor with Word Support Model for the SIGHAN Bakeoff 2006*
Jia-Lin Tsai

17:20–17:35 *On Closed Task of Chinese Word Segmentation: An Improved CRF Model Coupled with Character Clustering and Automatically Generated Template Matching*
Richard Tzong-Han Tsai, Hsieh-Chuan Hung, Cheng-Lung Sung, Hong-Jie Dai and Wen-Lian Hsu

17:35–17:50 *Chinese Word Segmentation with Maximum Entropy and N-gram Language Model*
Xinhao Wang, Xiaojun Lin, Dianhai Yu, Hao Tian and Xihong Wu

Sunday, 23 July 2006

Session 6: Bakeoff Presentations

09:00–09:15 *On Using Ensemble Methods for Chinese Named Entity Recognition*
Chia-Wei Wu, Shyh-Yi Jan, Richard Tzong-Han Tsai and Wen-Lian Hsu

09:15–09:30 *Chinese Word Segmentation and Named Entity Recognition by Character Tagging*
Kun Yu, Sadao Kurohashi, Hao Liu and Toshiaki Nakazawa

09:30–09:45 *Boosting for Chinese Named Entity Recognition*
Xiaofeng Yu, Marine Carpuat and Dekai Wu

09:45–10:00 *Chinese Word Segmentation and Named Entity Recognition Based on a Context-Dependent Mutual Information Independence Model*
Min Zhang, GuoDong Zhou, LingPeng Yang and DongHong Ji

10:00–10:15 *Word Segmentation and Named Entity Recognition for SIGHAN Bakeoff3*
Suxiang Zhang, Ying Qin, Juan Wen and Xiaojie Wang

10:15–10:30 *An Improved Chinese Word Segmentation System with Conditional Random Field*
Hai Zhao, Chang-Ning Huang and Mu Li

10:30–11:00 Break

11:00–12:00 SIGHAN Business Meeting