CLP 2012

# The Second CIPS-SIGHAN Joint Conference on Chinese Language Processing

20-21 December 2012
Tianjin University
Tianjin, China

Production and Manufacturing by
*Chinese Information Processing Society of China*

*No.4 Zhongguancun South 4th Street*
*Haidian District, Beijing, China*

To order hard copies of this proceedings, please contact:

Mail Order Division, Chinese Information Processing Society of China
No.4 Zhongguancun South 4th Street
Haidian District, Beijing, China
Tel: +86-010-62562961
cips@iscas.ac.cn

# Preface

In the big data age, Chinese language data online is expanding rapidly, and the application of natural language processing technology is drawing growing interest from the research community across the globe to harness Chinese language content. The rise of China as a global power with increasing influence on the world stage is only fanning this interest. The Chinese language also has a number of characteristics that make Chinese language processing particularly challenging and intellectually rewarding.

To meet the challenge, the CIPS-SIGHAN Joint Conference on Chinese Language Processing(CLP) is organized under the auspices of CIPS (Chinese Information Processing Society of China) and SIGHAN, a Special Interest Group of the ACL. CLP-2012 is the second conference jointly organized by the Chinese Language Processing Society of China (CIPS) and the ACL Special Interest Group on Chinese Language Processing (SIGHAN). The first conference, CLP-2010, was held on Aug 28-29, 2010 in Beijing, China, in conjunction with COLING 2010.

The goal of CLP2012 is to provide a platform for researchers around the world to present their research, share ideas, explore new research directions, and advance the state-of-the-art in Chinese language processing. The conference will also feature an international bakeoff on four tracks: word segmentation on Chinese Mirco-blog data, Chinese personal name disambiguation, simplified Chinese parsing, and traditional Chinese parsing.

The four bakeoff tasks have attracted 31 groups to submit their results.The proceedings also includes 4 overview papers that introduce the bakeoff tasks as well as the 32 bakeoff papers.

We would like to thank CIPS and SIGHAN for their continuing support of the conference, as well as the Asian Information Retrieval Society for allowing us to be a co-event of their Eighth Asian Information Retrieval Societies Conference (AIRS-2012). Especially we would like to thank professors, Zhifang Sui, Houfeng Wang, Qiang Zhou, Liang-Chih Yu, and Yuexian Hou, for initiating and proposing to hold this conference, and we are deeply indebted to all the reviewers for their tireless and generous work. Besides, we really appreciate Prof. Chunliang Zhang and Doctor Huizhen Wang for their dedication with all the publicity and publication issues. Most of all, we are grateful that the two keynote speakers, Prof. Xiaoyan Zhu and Prof. Guodong Zhou, share their inspiration in NLP research. Finally, we would like to thank all the authors for submitting their papers and reports to the conference.

We wish you all an enjoyable and thought-provoking conference.

| | |
|---|---|
| Le Sun, Hsin-His Chen | *CLP2012 General Co-Chairs* |
| Jingbo Zhu, Fei Xia, Houfeng Wang | *CLP2012 Program Co-Chairs* |

# Organizers

**General Chairs:**

Le Sun,*Chinese Information Processing Society of China*
*Hsin-His Chen,* SIGHAN & National Taiwan University

**Program Chairs:**

Jingbo Zhu, *Northeastern University*
*Fei Xia,* University of Washington
Houfeng Wang, *Peking University*

**Bakeoff Chairs:**

**Chinese Micro blog Word Segmentation:*

*Huiming Duan,* Peking University
Zhifang Sui, *Peking University*

**Simplified Chinese Parsing:*

*Qiang Zhou,* Tsinghua University

*Traditional Chinese Parsing:
Yuen-Hsien Tseng, *National Taiwan Normal University*

**Chinese Personal Name disambiguation:*
*Houfeng Wang,* Peking University
Sujian Li, *Peking University*

**Publications Chair:**

*Huizhen Wang,* Northeastern University

**Publicity Chair:**

Chunliang Zhang, *Northeastern University*

**Local Arrangements Chair:**

*Yuexian Hou,* Tianjin University

**Reviewers:**

| | | |
|---|---|---|
| Wanxiang Che | Jiajun Chen | Jinying Chen |
| Keh-Jiann Chen | Huiming Duan | Xuanjing Huang |
| Donghong Ji | Heng Ji | Olivia Kwong |
| Juanzi Li | Mu Li | Sujian Li |
| Chin-Yew Lin | Hongfei Lin | Yang Liu |
| Qin Lu | Yajuan Lv | Shaoping Ma |
| Jianyun Nie | Xiaodong Shi | Keh-Yih Su |
| Zhifang Sui | Maosong Sun | Yuen-Hsien Tseng |
| Xiaojun Wan | Bin Wang | Houfeng Wang |
| Xiaojie Wang | Kam-Fai Wong | Hua Wu |
| Yunfang Wu | Yunqing Xia | Deyi Xiong |
| Jinan Xu | Nianwen Xue | Muyun Yang |
| Kun Yu | Weidong Zhan | Jiajun Zhang |
| Min Zhang | Hai Zhao | Jun Zhao |
| Guodong Zhou | Ming Zhou | Qiang Zhou |

# Table of Contents

## Keynote Speaks:

*QA: from Turing Test to Intelligent Information Service*
Xiaoyan Zhu . . . . . 1

*Linguistic foundation for NLP*
Guodong Zhou . . . . . 2

## Research Papers:

*A Language Modeling Approach to Identifying Code-Switched Sentences and Words*
Liang-Chih Yu, Wei-Cheng He and Wei-Nan Chien . . . . . 3

*Semi-automatic Annotation of Chinese Word Structure*
Jianqiang Ma, Chunyu Kit and Dale Gerdemann . . . . . 9

*Building a Chinese Lexical Taxonomy*
Xiaopeng Bai and Nianwen Xue . . . . . 18

*Extending and Scaling up the Chinese Treebank Annotation*
Xiuhong Zhang and Nianwen Xue . . . . . 27

### Task 1: Micro-blog word segmentation

*The CIPS-SIGHAN CLP 2012 ChineseWord Segmentation onMicroBlog Corpora Bakeoff*
Huiming Duan, Zhifang Sui, Ye Tian and Wenjie Li . . . . . 35

*Word Segmentation on Chinese Mirco-Blog Data with a Linear-Time Incremental Model*
Kaixu Zhang, Maosong Sun and Changle Zhou . . . . . 41

*Soochow University Word Segmenter for SIGHAN 2012 Bakeoff*
Yan Fang, Zhongqing Wang, Shoushan Li, Zhongguo Li, Richen Xu and Leixin Cai . . . . . 47

*CRFs-Based Chinese Word Segmentation for Micro-Blog with Small-Scale Data*
Longyue Wang, Derek F. Wong, Lidia S. Chao and Junwen Xing . . . . . 51

*A Cascaded Approach for CIPS-SIGHAN Micro-Blog Word Segmentation Bakeoff 2012*
Bei Shi, Xianpei Han and Le Sun . . . . . 58

*Adapting Conventional Chinese Word Segmenter for Segmenting Micro-blog Text: Combining Rule-based and Statistic-based Approaches*
Ning Xi, Bin Li, Guangchao Tang, Shujian Huang, Yinggong Zhao, Hao Zhou, Xinyu Dai and Jiajun Chen . . . . . 63

*Cascaded Chinese Weibo Segmentation Based on CRFs*
keli Zhong, xue Zhou, hangyu Li and caixia Yuan . . . . . 69

*Rules-based Chinese Word Segmentation on MicroBlog for CIPS-SIGHAN on CLP2012*
Jing Zhang, Degen Huang, Xia Han and Wei Wang . . . . . 74

*Semi-supervised Chinese Word Segmentation for CLP2012*
Saike HE, Nan HE, Songxiang CEN and Jun LU . . . . . 79

*Micro blogs Oriented Word Segmentation System*
Liu Yijia, Zhang Meishan, Che Wanxiang, Liu Ting and Deng Yihe . . . . . 85

*Rules Design in Word Segmentation of Chinese Micro-Blog*
Hao Zong, Derek F. Wong and Lidia S. Chao . . . . . 90

*A Comparison of Chinese Word Segmentation on News and Microblog Corpora with a Lexicon Based Method*
Yuxiang Jia, Hongying Zan, Ming Fan and Zhimin Wang . . . . . 95

*A MMSM-based Hybrid Method for Chinese MicroBlog Word Segmentation*
Xiao Sun, Chengcheng Li, Chenyi Tang and Jiaqi Ye . . . . . 99

*Chinese Tweets Segmentation based on Morphemes*
Chaoyue Wang and Guohong Fu . . . . . 106

**Task 2: Chinese personal name disambiguation**

*The Task 2 of CIPS-SIGHAN 2012 Named Entity Recognition and Disambiguation in Chinese Bakeoff*
Zhengyan He, Houfeng Wang and Sujian Li . . . . . 108

*SIR-NERD: A Chinese Named Entity Recognition and Disambiguation System using a Two-StageMethod*
Zehuan Peng, Le Sun and Xianpei Han . . . . . 115

*A Template Based Hybrid Model for Chinese Personal Name Disambiguation*
Hao Zong, Derek F. Wong and Lidia S. Chao . . . . . 121

*Attribute based Chinese Named Entity Recognition and Disambiguation*
Han Wei, Liu Guang, Mao Yuzhao and Huang Zhenni . . . . . 127

*Chinese Name Disambiguation Based on Adaptive Clustering with the Attribute Features*
Wei Tian, Xiao Pan, Zhengtao Yu, yantuan Xian and xiuzhen Yang . . . . . 132

*Explore Chinese Encyclopedic Knowledge to Disambiguate Person Names*
Jie Liu, Ruifeng Xu, Qin Lu and Jian Xu . . . . . 138

*A Joint Chinese Named Entity Recognition and Disambiguation System*
Longyue Wang, Shuo Li, Derek F. Wong and Lidia S. Chao . . . . . 146

*Chinese Personal Name Disambiguation Based on Vector Space Model*
Qing-hu FAN, Hong-ying ZAN, Yu-mei CHAI, Yu-xiang JIA and Gui-ling NIU . . . . . 152

**Task 3: Simplified Chinese parsing**

*Evaluation Reportof the third Chinese Parsing Evaluation: CIPS-SIGHAN-ParsEval-2012*
Qiang Zhou . . . . . 159

*Multiple TreeBanks Integration for Chinese Phrase Structure Grammar Parsing Using Bagging*
Meishan Zhang, Wanxiang Che and Ting Liu . . . . . 168

*Parsing TCT with Split Conjunction Categories*
Dongchen Li and Xihong Wu . . . . . 174

*Parsing Simplified Chinese and Traditional Chinese with Sentence Structure Grammar*
Xiangli Wang . . . . . 179

*A Simplified Chinese Parser with Factored Model*
Qiuping Huang, Liangye He, Derek F. Wong and Lidia S. Chao . . . . . 188

*Parsing TCT with a Coarse-to-fine Approach*
Dongchen Li and Xihong Wu . . . . . 194

**Task 4: Traditional Chinese parsing**

*Traditional Chinese Parsing Evaluation at SIGHAN Bake-offs 2012*
Yuen-Hsien Tseng, Lung-Hao Lee and Liang-Chih Yu . . . . . 199

*NEU Systems in SIGHAN Bakeoff 2012*
Ji Ma, LongFei Bai, Zhuo Liu, Ao Zhang and Jingbo Zhu . . . . . 206

*Adapting Multilingual Parsing Models to Sinica Treebank*
Liangye He, Derek F. Wong and Lidia S. Chao . . . . . 211

*Improving PCFG Chinese Parsing with Context-Dependent Probability Re-estimation*
Yu-Ming Hsieh, Ming-Hong Bai, Jason S. Chang and Keh-Jiann Chen . . . . . 216

*Sentence Parsing with Double Sequential Labeling in Traditional Chinese Parsing Task*
Shih-Hung Wu, Hsien-You Hsieh and Liang-Pu Chen . . . . . 222

*A Conditional Random Field-based Traditional Chinese Base Phrase Parser for SIGHAN Bake-off 2012 Evaluation*
Yih-Ru Wang and Yuan-Fu Liao . . . . . 231

*Hierarchical Maximum Pattern Matching with Rule Induction Approach for Sentence Parsing*
Yi-Syun Tan, Yuan-Cheng Chu and Jui-Feng Yeh . . . . . 237

# Conference Program

**Thursday, December 20, 2012**

8:30–8:50 Opening

8:50–9:50 Keynote Speech: Xiaoyan Zhu, *QA: from Turing Test to Intelligent Information Service*

9:50–10:10 Coffee Break

**Session 1: Overview of All tasks**

10:10–10:25 *The CIPS-SIGHAN CLP 2012 ChineseWord Segmentation onMicroBlog Corpora Bakeoff*
Huiming Duan, Zhifang Sui, Ye Tian and Wenjie Li

10:25–10:40 *The Task 2 of CIPS-SIGHAN 2012 Named Entity Recognition and Disambiguation in Chinese Bakeoff*
Zhengyan He, Houfeng Wang and Sujian Li

10:40–10:55 *Evaluation Reportof the third Chinese Parsing Evaluation: CIPS-SIGHAN-ParsEval-2012*
Qiang Zhou

10:55–11:10 *Traditional Chinese Parsing Evaluation at SIGHAN Bake-offs 2012*
Yuen-Hsien Tseng, Lung-Hao Lee and Liang-Chih Yu

**Session 2: Research Paper**

11:10–11:25 *A Language Modeling Approach to Identifying Code-Switched Sentences and Words*
Liang-Chih Yu, Wei-Cheng He and Wei-Nan Chien

11:25–11:40 *Semi-automatic Annotation of Chinese Word Structure*
Jianqiang Ma, Chunyu Kit and Dale Gerdemann

11:40–11:55 *Building a Chinese Lexical Taxonomy*
Xiaopeng Bai and Nianwen Xue

11:55–12:10 *Extending and Scaling up the Chinese Treebank Annotation*
Xiuhong Zhang and Nianwen Xue

**Thursday, December 20, 2012 (continued)**

12:10–13:30 Lunch

**Session 3: Bakeoff 1 Micro-blog word segmentation**

13:30–13:45 *Word Segmentation on Chinese Mirco-Blog Data with a Linear-Time Incremental Model*
Kaixu Zhang, Maosong Sun and Changle Zhou

13:45–14:00 *Soochow University Word Segmenter for SIGHAN 2012 Bakeoff*
Yan Fang, Zhongqing Wang, Shoushan Li, Zhongguo Li, Richen Xu and Leixin Cai

14:00–14:15 *CRFs-Based Chinese Word Segmentation for Micro-Blog with Small-Scale Data*
Longyue Wang, Derek F. Wong, Lidia S. Chao and Junwen Xing

14:15–14:30 *A Cascaded Approach for CIPS-SIGHAN Micro-Blog Word Segmentation Bakeoff 2012*
Bei Shi, Xianpei Han and Le Sun

14:30–15:00 Coffee Break

**Session 4: Bakeoff 2 Chinese personal name disambiguation**

15:00–15:15 *SIR-NERD: A Chinese Named Entity Recognition and Disambiguation System using a Two-StageMethod*
Zehuan Peng, Le Sun and Xianpei Han

15:15–15:30 *A Template Based Hybrid Model for Chinese Personal Name Disambiguation*
Hao Zong, Derek F. Wong and Lidia S. Chao

15:30–15:45 *Attribute based Chinese Named Entity Recognition and Disambiguation*
Han Wei, Liu Guang, Mao Yuzhao and Huang Zhenni

15:45–16:00 *Chinese Name Disambiguation Based on Adaptive Clustering with the Attribute Features*
Wei Tian, Xiao Pan, Zhengtao Yu, yantuan Xian and xiuzhen Yang

**Thursday, December 20, 2012 (continued)**

**Session 5: Bakeoff 4 Traditional Chinese parsing**

16:00–16:15 *NEU Systems in SIGHAN Bakeoff 2012*
Ji Ma, LongFei Bai, Zhuo Liu, Ao Zhang and Jingbo Zhu

16:15–16:30 *Adapting Multilingual Parsing Models to Sinica Treebank*
Liangye He, Derek F. Wong and Lidia S. Chao

16:30–16:45 *Improving PCFG Chinese Parsing with Context-Dependent Probability Re-estimation*
Yu-Ming Hsieh, Ming-Hong Bai, Jason S. Chang and Keh-Jiann Chen

16:45–17:00 *Sentence Parsing with Double Sequential Labeling in Traditional Chinese Parsing Task*
Shih-Hung Wu, Hsien-You Hsieh and Liang-Pu Chen

**Friday, December 21, 2012**

8:30–9:30 Keynote Speech: Guodong Zhou, *Linguistic foundation for NLP*

9:30–9:50 Coffee Break

**Session 1: Bakeoff 3 Simplified Chinese parsing**

9:50–10:05 *Multiple TreeBanks Integration for Chinese Phrase Structure Grammar Parsing Using Bagging*
Meishan Zhang, Wanxiang Che and Ting Liu

10:05–10:20 *Parsing TCT with Split Conjunction Categories*
Dongchen Li and Xihong Wu

10:20–10:35 *Parsing Simplified Chinese and Traditional Chinese with Sentence Structure Grammar*
Xiangli Wang

**Friday, December 21, 2012 (continued)**

**Session 2: Bakeoff Posters**

10:35–11:35 Bakeoff Posters

11:35–11:45 Closing

**Bakeoff Poster List**

*1. Adapting Conventional Chinese Word Segmenter for Segmenting Micro-blog Text: Combining Rule-based and Statistic-based Approaches*
Ning Xi, Bin Li, Guangchao Tang, Shujian Huang, Yinggong Zhao, Hao Zhou, Xinyu Dai and Jiajun Chen

*2. Cascaded Chinese Weibo Segmentation Based on CRFs*
keli Zhong, xue Zhou, hangyu Li and caixia Yuan

*3. Rules-based Chinese Word Segmentation on MicroBlog for CIPS-SIGHAN on CLP2012*
Jing Zhang, Degen Huang, Xia Han and Wei Wang

*4. Semi-supervised Chinese Word Segmentation for CLP2012*
Saike HE, Nan HE, Songxiang CEN and Jun LU

*5. Micro blogs Oriented Word Segmentation System*
Liu Yijia, Zhang Meishan, Che Wanxiang, Liu Ting and Deng Yihe

*6. Rules Design in Word Segmentation of Chinese Micro-Blog*
Hao Zong, Derek F. Wong and Lidia S. Chao

*7. A Comparison of Chinese Word Segmentation on News and Microblog Corpora with a Lexicon Based Method*
Yuxiang Jia, Hongying Zan, Ming Fan and Zhimin Wang

*8. A MMSM-based Hybrid Method for Chinese MicroBlog Word Segmentation*
Xiao Sun, Chengcheng Li, Chenyi Tang and Jiaqi Ye

*9. Chinese Tweets Segmentation based on Morphemes*
Chaoyue Wang and Guohong Fu

*10. Explore Chinese Encyclopedic Knowledge to Disambiguate Person Names*
Jie Liu, Ruifeng Xu, Qin Lu and Jian Xu

**Friday, December 21, 2012 (continued)**

*11. A Joint Chinese Named Entity Recognition and Disambiguation System*
Longyue Wang, Shuo Li, Derek F. Wong and Lidia S. Chao

*12. Chinese Personal Name Disambiguation Based on Vector Space Model*
Qing-hu FAN, Hong-ying ZAN, Yu-mei CHAI, Yu-xiang JIA and Gui-ling NIU

*13. A Simplified Chinese Parser with Factored Model*
Qiuping Huang, Liangye He, Derek F. Wong and Lidia S. Chao

*14. Parsing TCT with a Coarse-to-fine Approach*
Dongchen Li and Xihong Wu

*15. A Conditional Random Field-based Traditional Chinese Base Phrase Parser for SIGHAN Bake-off 2012 Evaluation*
Yih-Ru Wang and Yuan-Fu Liao

*16. Hierarchical Maximum Pattern Matching with Rule Induction Approach for Sentence Parsing*
Yi-Syun Tan, Yuan-Cheng Chu and Jui-Feng Yeh