

**Proceedings of the**  
**Third SIGHAN Workshop**  
**on Chinese Language Learning**

**Held in cooperation with ACL-2004**

**25 July 2004**  
**Barcelona, Spain**

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)  
73 Landmark Center  
East Stroudsburg, PA 18301  
USA  
Tel: +1-570-476-8006  
Fax: +1-570-476-0860  
[acl@aclweb.org](mailto:acl@aclweb.org)



## **ORGANIZERS:**

Chair: Qin Lu & Oliver Streiter

Proceedings: Qin Lu & Oliver Streiter

## **PROGRAM COMMITTEE:**

Andi Wu - Microsoft, USA

Changning Huang - Microsoft, China

Chu-ren Huang - Academia Sinica, Taiwan

Joyce Chai - Michigan State Univ, USA

Keh-Jian Chen - Academia Sinica, Taiwan

Li Wenjie - the Hong Kong Polytechnic University, Hong Kong

Martha Palmer - Univ. of Pennsylvania, USA

Nianwen Xue - Univ. of Pennsylvania, USA

Oliver Streiter - EURAC, Italy

Qiang Zhou - Tsinghua University, China

Qing Ma - Ryukoku University, Japan

Qin Lu - The Hong Kong Polytechnic University

Sui Zhifang - Peking University, China

Sun Maosong - Tsinghua University, China

Tom Emerson - Basis Technology Corp, U.S.A.

## **FURTHER INFORMATION:**

Dr. Qin Lu

Department of Computing,

The Hong Kong Polytechnic University,

Hung Hom, Kowloon,

Hong Kong

# Table of Contents

<i>Segmentation of Chinese Long Sentences Using Commas</i> Meixun Jin, Mi-Young Kim, Dongil Kim and Jong-Hyeok Lee .....	1
<i>A Preliminary Study on Probabilistic Models for Chinese Abbreviations</i> Jing-Shin Chang and Yu-Tso Lai .....	9
<i>Document Re-ranking based on Global and Local Terms</i> Lingpeng Yang, DongHong Ji and Li Tang .....	17
<i>Adaptive Compression-based Approach for Chinese Pinyin Input</i> JinHu Huang and David Powers .....	24
<i>An Enhanced Model for Chinese Word Segmentation and Part-of-Speech Tagging</i> Feng Jiang, Hui Liu, Yuquan Chen and Ruzhan Lu .....	28
<i>Character-Sense Association and Compounding Template Similarity: Automatic Semantic Classification of Chinese Compounds</i> Chao-Jan Chen .....	33
<i>Chinese Chunking with Another Type of Spec</i> Hongqiao Li, Changning Huang, Jianfeng Gao and Xiaozhong Fan .....	41
<i>Combining Neural Networks and Statistics for Chinese Word Sense Disambiguation</i> Zhimao Lu, Ting Liu and Sheng Li .....	49
<i>Chinese Word Segmentation by Classification of Characters</i> Chooi-Ling GOH, Masayuki Asahara and Yuji Matsumoto .....	57
<i>Automated Alignment and Extraction of Bilingual Domain Ontology for Medical Domain Web Search</i> Jui-Feng Yeh, Chung-Hsiwn Wu, Ming-Jun Chen and Liang-chih Yu .....	65
<i>A Statistical Model for Hangeul-Hanja Conversion in Terminology Domain</i> Jin-Xia Huang, Sun-Mee Bae and Key-sun Choi .....	72
<i>Chinese Term Extraction from Web Pages Based on Compound Term Productivity</i> Hiroshi Nakagawa, Hiroyuki Kojima and Akira Maeda .....	79
<i>Using Synonym Relations in Chinese Collocation Extraction</i> Wanyin Li, Qin Lu and Ruifeng Xu .....	86
<i>The Construction of A Chinese Shallow Treebank</i> Ruifeng Xu, Qin Lu, Yin Li and Wanyin Li .....	94
<i>Combining Prosodic and Text Features for Segmentation of Mandarin Broadcast News</i> Gina-Anne Levow .....	102
<i>Automatic Semantic Role Assignment for a Tree Structure</i> Jia-Ming You and Keh-Jiann Chen .....	109
<i>A Large-Scale Semantic Structure for Chinese Sentences</i> Li Tang, Donghong Ji and Lingpeng Yang .....	116
<i>Do We Need Chinese Word Segmentation for Statistical Machine Translation?</i> Jia Xu, Richard Zens and Hermann Ney .....	122
<i>A Semi-Supervised Approach to Build Annotated Corpus for Chinese Named Entity Recognition</i> Xiaoshan Fang, Jianfeng Gao and Huanye Sheng .....	129

<i>A New Chinese Natural Language Understanding Architecture Based on Multilayer Search Mechanism</i> Wanxiang Che, Ting Liu and Sheng Li.....	134
<i>Aligning Bilingual Corpora Using Sentences Location Information</i> Weigang Li, Ting Liu, Zhen Wang and Sheng Li .....	141
<i>An Integrated Method for Chinese Unknown Word Extraction</i> Zhiyong Luo and Rou Song.....	148

## Technical Program Schedule

### Sunday, July 25

- 8:45-8:50            Welcome
- 8:50-9:10            *Segmentation of Chinese Long Sentences Using Commas*  
Meixun Jin, Mi-Young Kim, Dongil Kim and Jong-Hyeok Lee
- 9:15-9:35            *A Preliminary Study on Probabilistic Models for Chinese Abbreviations*  
Jing-Shin Chang and Yu-Tso Lai
- 9:40-10:00          *Document Re-ranking based on Global and Local Terms*  
Lingpeng Yang, DongHong Ji and Li Tang

### Coffee Break

- 10:30-10:50         *Chinese Chunking with Another Type of Spec*  
Hongqiao Li, Changning Huang, Jianfeng Gao and Xiaozhong Fan
- 10:55-11:15         *Chinese Word Segmentation by Classification of Characters*  
Chooi-Ling GOH, Masayuki Asahara and Yuji Matsumoto
- 11:20-11:40         *Automated Alignment and Extraction of Bilingual Domain Ontology for  
Medical Domain Web Search*  
Jui-Feng Yeh, Chung-Hsiwn Wu, Ming-Jun Chen and Liang-chih Yu
- 11:45-12:05         *Using Synonym Relations in Chinese Collocation Extraction*  
Wanyin Li, Qin Lu and Ruifeng Xu

### Lunch Break

- 13:50-14:10         *Combining Prosodic and Text Features for Segmentation of Mandarin  
Broadcast News*  
Gina-Anne Levow
- 14:15-14:35         *Automatic Semantic Role Assignment for a Tree Structure*  
Jia-Ming You and Keh-Jiann Chen
- 14:40-15:00         *A Large-Scale Semantic Structure for Chinese Sentences*  
Li Tang, Donghong Ji and Lingpeng Yang
- 15:05-15:25         *Aligning Bilingual Corpora Using Sentences Location Information*  
Weigang Li, Ting Liu, Zhen Wang and Sheng Li

### Coffee Break

## **Poster Session**

- 15:50-17:10 *An Integrated Method for Chinese Unknown Word Extraction*  
Zhiyong Luo and Rou Song
- 15:50-17:10 *Adaptive Compression-based Approach for Chinese Pinyin Input*  
JinHu Huang and David Powers
- 15:50-17:10 *Character-Sense Association and Compounding Template Similarity:  
Automatic Semantic Classification* Chao-Jan Chen
- 15:50-17:10 *Combining Neural Networks and Statistics for Chinese Word Sense  
Disambiguation*  
Zhimao Lu, Ting Liu and Sheng Li
- 15:50-17:10 *A Statistical Model for Hangeul-Hanja Conversion in Terminology Domain*  
Jin-Xia Huang, Sun-Mee Bae and Key-sun Choi
- 15:50-17:10 *Chinese Term Extraction from Web Pages Based on Compound Term  
Productivity*  
Hiroschi Nakagawa, Hiroyuki Kojima and Akira Maeda
- 15:50-17:10 *The Construction of A Chinese Shallow Treebank*  
Ruifeng Xu, Qin Lu, Yin Li and Wanyin Li
- 15:50-17:10 *Do We Need Chinese Word Segmentation for Statistical Machine  
Translation?*  
Jia Xu, Richard Zens and Hermann Ney
- 15:50-17:10 *A New Chinese Natural Language Understanding Architecture Based on  
Multilayer Search Mechanism*  
Wanxiang Che, Ting Liu and Sheng Li
- 15:50-17:10 *A Semi-Supervised Approach to Build Annotated Corpus for Chinese  
Named Entity Recognition*  
Xiaoshan Fang, Jianfeng Gao and Huanye Sheng
- 15:50-17:10 *An Enhanced Model for Chinese Word Segmentation and Part-of-Speech  
Tagging*  
Feng Jiang, Hui Liu, Yuquan Chen and Ruzhan Lu

## **SIGHAN**

### **Meeting**

- 17:20-18:20 *Organizational Meeting*  
*Discussion of Future Bakeoff*



THIS IS A BLANK PAGE

PLEASE IGNORE

## Author Index

Asahara, Masayuki .....	57	Yu, Liang-chih .....	65
Bae, Sun-Mee .....	72	Zens, Richard .....	122
Chang, Jing-Shin .....	9		
Che, Wanxiang .....	134		
Chen, Chao-Jan .....	33		
Chen, Keh-Jiann .....	109		
Chen, Ming-Jun .....	65		
Chen, Yuquan .....	28		
Choi, Key-sun .....	72		
Fan, Xiaozhong .....	41		
Fang, Xiaoshan .....	129		
Gao, Jianfeng .....	41, 129		
GOH, Chooi-Ling .....	57		
Huang, Changning .....	41		
Huang, JinHu .....	24		
Huang, Jin-Xia .....	72		
Ji, DongHong .....	17		
Ji, Donghong .....	116		
Jiang, Feng .....	28		
Jin, Meixun .....	1		
Kim, Dongil .....	1		
Kim, Mi-Young .....	1		
Kojima, Hiroyuki .....	79		
Lai, Yu-Tso .....	9		
Lee, Jong-Hyeok .....	1		
Levow, Gina-Anne .....	102		
Li, Hongqiao .....	41		
Li, Sheng .....	49, 134, 141		
Li, Wanyin .....	86, 94		
Li, Weigang .....	141		
Li, Yin .....	94		
Liu, Hui .....	28		
Liu, Ting .....	49, 134, 141		
Lu, Qin .....	86, 94		
Lu, Ruzhan .....	28		
Lu, Zhimao .....	49		
Luo, Zhiyong .....	148		
Maeda, Akira .....	79		
Matsumoto, Yuji .....	57		
Nakagawa, Hiroshi .....	79		
Ney, Hermann .....	122		
Powers, David .....	24		
Sheng, Huanye .....	129		
Song, Rou .....	148		
Tang, Li .....	17, 116		
Wang, Zhen .....	141		
Wu, Chung-Hsiwn .....	65		
Xu, Jia .....	122		
Xu, Ruifeng .....	86, 94		
Yang, Lingpeng .....	17, 116		
Yeh, Jui-Feng .....	65		
You, Jia-Ming .....	109		