



**Proceedings of the
Second Chinese Language
Processing Workshop**

**Held in conjunction with
The 38th Annual Meeting of the
Association for Computational Linguistics**

**Edited by
Martha Palmer
Mitch Marcus
Aravind Joshi
Fei Xia**

**8 October 2000
Hong Kong University of Science and Technology (HKUST)
Hong Kong**

**Proceedings of the
Second Chinese Language
Processing Workshop**

**Held in conjunction with
The 38th Annual Meeting of the
Association for Computational Linguistics**

Edited by

Martha Palmer

Mitch Marcus

Aravind Joshi

Fei Xia

8 October 2000

**Hong Kong University of Science and Technology (HKUST)
Hong Kong**

©2000 The Association for Computational Linguistics

Order copies of this and other ACL workshop proceedings from:

Association for Computational Linguistics (ACL)
75 Paterson Street, Suite 9
New Brunswick, NJ 08901
USA
Tel: +1-732-342-9100
Fax: +1-732-342-9339
acl@aclweb.org

SPONSORS:

SIGDAT
 SIGLEX
 SIGPARSE

INDUSTRY SPONSOR:

Intel China Research Center

INVITED SPEAKER:

Chin-Chuan Cheng (City University of Hong Kong)

ORGANIZERS:

Martha Palmer
 Mitch Marcus
 Aravind Joshi
 Fei Xia

PROGRAM COMMITTEE:

Srinivas Bangalore	(AT&T Research Labs)	Keh-Yih Su	(Behavior Design Corporation)
Keh-Jiann Chen	(Academia Sinica)	Maosong Sun	(Tsinghua Univ.)
Zhengdong Dong	(Hownet)	Chew Lim Tan	(National Univ. of Singapore)
Shengli Feng	(Univ. of Kansas)	Benjamin K Tsou	(City Univ. of Hong Kong)
Kok Wee Gan	(HKUST)	Amy Weinberg	(Univ. of Maryland)
Laurie Gerber	(Univ. of Southern CA/ISI)	Ralph Weischedel	(BBN)
Changning Huang	(Microsoft Research China)	Andi Wu	(Microsoft)
Chu-Ren Huang	(Academia Sinica)	Dekai Wu	(HKUST)
Wanying Jin	(New Mexican State Univ.)	Nianwen Xue	(Univ. of Delaware)
Kim-Teng Lua	(National Univ. of Singapore)	Jin Yang	(Systran)
John Kovarik	(Department of Defense)	Shiwen Yu	(Peking Univ.)
K.L. Kwok	(Queens College)	Chunfa Yuan	(Tsinghua Univ.)
Mary Ellen Okurowski	(Department of Defense)	Joe Zhou	(Intel China Research Center)
Fuji Ren	(Hiroshima City Univ.)	Qiang Zhou	(Tsinghua Univ.)
Richard Sproat	(AT&T Research Lab)		

FURTHER INFORMATION:

Martha Palmer, Mitch Marcus, Aravind Joshi, and Fei Xia
 Department of Computer and Information Science
 University of Pennsylvania
 Philadelphia, PA 19104, USA
 Email: {mpalmer,mitch,joshi,fxia}@linc.cis.upenn.edu

ACKNOWLEDGMENT:

Special thanks to Yen-Lin Yin at the Univ. of Pennsylvania for helping us to organize this workshop.

WORKSHOP PROGRAM

Sunday, 8 October 2000

- 8:45-9:00 Welcome
- 9:00-9:50 *Invited Talk: Zero Anaphors in Chinese Discourse Processing*
Chin-Chuan Cheng
- 9:50-10:10 *Sense-Tagging Chinese Corpus*
Hsin-Hsi Chen and Chi-Ching Lin
- 10:10-10:30 *Enhancement of a Chinese Discourse Marker Tagger with C4.5*
Benjamin K. T'sou, Tom B.Y. Lai, Samuel W.K. Chan, Weijun Gao and Xuegang Zhan
- 10:30-11:00 *Break*
- 11:00-11:20 *Two Statistical Parsing Models Applied to the Chinese Treebank*
Daniel M. Bikel and David Chiang
- 11:20-11:40 *Using Co-occurrence Statistics as an Information Source for Partial Parsing of Chinese*
Elliott Franco Drabek and Qiang Zhou
- 11:40-12:00 *Statistics Based Hybrid Approach to Chinese Base Phrase Identification*
Tie-jun Zhao, Mu-yun Yang, Fang Liu, Jian-min Yao and Hao Yu
- 12:00-12:20 *A Block-Based Robust Dependency Parser for Unrestricted Chinese Text*
Ming Zhou
- 12:20-1:30 *Lunch*
- 1:30-2:30 *Poster session*
- 2:30-2:50 *Knowledge Extraction for Identification of Chinese Organization Names*
Keh-Jiann Chen and Chao-jan Chen
- 2:50-3:10 *Statistically-Enhanced New Word Identification in a Rule-Based Chinese System*
Andi Wu and Zixin Jiang
- 3:10-3:30 *A Trainable Method for Extracting Chinese Entity Names and Their Relations*
Yimin Zhang and Joe F Zhou
- 3:30-4:00 *Break*
- 4:00-4:20 *Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface*
Chu-Ren Huang, Feng-Yi Chen, Keh-Jiann Chen, Zhao-ming Gao and Kuang-Yu Chen
- 4:20-4:40 *Comparing Lexicalized Treebank Grammars Extracted from Chinese, Korean, and English*
Fei Xia, Chunghye Han, Martha Palmer and Aravind Joshi
- 4:40-5:00 *The Research of Word Sense Disambiguation Method Based on Co-occurrence Frequency of Hownet*
Erhong Yang, Guoqing Zhang and Yongkui Zhang
- 5:00-6:00 *Panel: Prioritizing Chinese Language Processing Resources*
Keh-Jiann Chen, Chu-Ren Huang, Bing Swen, Benjamin K. T'sou, Joe Zhou and Ming Zhou

Table of Contents

PRESENTATIONS

<i>Two Statistical Parsing Models Applied to the Chinese Treebank</i> Daniel M. Bikel and David Chiang	1
<i>Sense-Tagging Chinese Corpus</i> Hsin-Hsi Chen and Chi-Ching Lin	7
<i>Knowledge Extraction for Identification of Chinese Organization Names</i> Keh-Jiann Chen and Chao-jan Chen	15
<i>Using Co-occurrence Statistics as an Information Source for Partial Parsing of Chinese</i> Elliott Franco Drabek and Qiang Zhou	22
<i>Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface</i> Chu-Ren Huang, Feng-Yi Chen, Keh-Jiann Chen, Zhao-ming Gao and Kuang-Yu Chen	29
<i>Enhancement of a Chinese Discourse Marker Tagger with C4.5</i> Benjamin K. T'sou, Tom B.Y. Lai, Samuel W.K. Chan, Weijun Gao and Xuegang Zhan	38
<i>Statistically-Enhanced New Word Identification in a Rule-Based Chinese System</i> Andi Wu and Zixin Jiang	46
<i>Comparing Lexicalized Treebank Grammars Extracted from Chinese, Korean, and English Corpora</i> Fei Xia, Chungnye Han, Martha Palmer and Aravind Joshi	52
<i>The Research of Word Sense Disambiguation Method Based on Co-occurrence Frequency of Hownet</i> Erhong Yang, Guoqing Zhang, and Yongkui Zhang	60
<i>A Trainable Method for Extracting Chinese Entity Names and Their Relations</i> Yimin Zhang and Joe F Zhou	66
<i>Statistics Based Hybrid Approach to Chinese Base Phrase Identification</i> Tie-jun Zhao, Mu-yun Yang, Fang Liu, Jian-min Yao and Hao Yu	73
<i>A Block-Based Robust Dependency Parser for Unrestricted Chinese Text</i> Ming Zhou	78

POSTERS

<i>Annotating Information Structures in Chinese Texts Using HowNet</i> Kok Wee Gan and Ping Wai Wong	85
<i>Machine Learning Methods for Chinese Web Page Categorization</i> Ji He, Ah-Hwee Tan and Chew-Lim Tan	93

<i>Semantic Annotation of Chinese Phrases Using Recursive Graph</i> Donghong Ji.....	101
<i>Text Meaning Representation for Chinese</i> Wanying Jin.....	109
<i>How Should a Large Corpus Be Built?—A Comparative Study of Closure in Annotated Newspaper Corpora from Two Chinese Sources, Towards Building a Larger Representative Corpus Merged from Representative Sublanguage Collections</i> John J. Kovarik.....	116
<i>A Clustering Algorithm for Chinese Adjectives and Nouns</i> Yang Wen, Chunfa Yuan and Changning Huang.....	124
<i>Extraction of Chinese Compound Words - An Experimental Study on a Very Large Corpus</i> Jian Zhang, Jianfeng Gao and Ming Zhou.....	132
<i>An Algorithm for Situation Classification of Chinese Verbs</i> Xiaodan Zhu, Chunfa Yuan, K.F. Wong and Wenjie Li.....	140
 INVITED TALK	
<i>Zero Anaphors in Chinese Discourse Processing</i> Chin-Chuan Cheng.....	146