Preface

Welcome to the proceedings of the 12th China National Conference on Computational Linguistics (12th CCL) and the First International Symposium on Natural Language Processing Based on Naturally Annotated Big Data (1st NLP-NABD). The conference was hosted by Soochow University.

CCL is a bi-annual conference that started in 1991. It is the flagship conference of the Chinese Information Processing Society (CIPS), which is the largest NLP scholar and expert community in China. CCL is a premier nation-wide forum for disseminating new scholarly and technological work in computational linguistics, with a major emphasis on computer processing of the languages in China such as Mandarin, Tibetan, Mongolian, and Uyghur.

Affiliated with the 12th CCL, The First International Symposium on Natural Language Processing Based on Naturally Annotated Big Data (NLP-NABD) covered all the NLP topics, with particular focus on methodologies and techniques relating to naturally annotated big data. In contrast to manually annotated data such as treebanks that are constructed for specific NLP tasks, naturally annotated data come into existence through users’ normal activities, such as writing, conversation, and interactions on the Web. Although the original purposes of these data typically were unrelated to NLP, they can nonetheless be purposefully exploited by computational linguists to acquire linguistic knowledge. For example, punctuation marks in Chinese text can help word boundaries identification, social tags in social media can provide signals for keyword extraction, categories listed in Wikipedia can benefit text classification. The natural annotation can be explicit, as in the above examples, or implicit, as in Hearst patterns (e.g., “Beijing and other cities” implies “Beijing is a city”). This symposium focuses on numerous research challenges ranging from very-large-scale unsupervised/semi-supervised machine leaning (deep learning, for instance) of naturally annotated big data to integration of the learned resources and models with existing handcrafted “core” resources and “core” language computing models.

The Program Committee selected 127 papers (95 Chinese papers and 32 English papers) out of 252 submissions from China, Hong Kong (region), Japan, and Poland for publication. The English papers cover the following topics:

- Word segmentation (7)
- Sentiment analysis, opinion mining and text classification (4)
- Text mining, open-domain information extraction and machine reading of the Web (3)
- Statistical and machine learning methods in NLP (3)
- Machine translation (3)
- Tagging and Chunking (2)
- Language resources and annotation (2)
The final program for the 12th CCL and the First NLP-NABD was the result of a great deal of work by many dedicated colleagues. We want to thank, first of all, the authors who submitted their papers, and thus contributed to the creation of the high-quality program that allowed us to look forward to an exciting joint conference. We are deeply indebted to all the Program Committee members for providing high-quality and insightful reviews under a tight schedule. We are extremely grateful to the sponsors of the conference. Finally, we extend a special word of thanks to all the colleagues of the Organizing Committee and Secretariat for their hard work in organizing the conference, and to Springer for their assistance in publishing the proceedings in due time.

On behalf of the program and organizing committees, we hope all we have done will make the conference successful, and make it interesting for all the participants. We also believe that your visit to Suzhou, a famous and beautiful historical and cultural city in China, will be a really valuable memory.

Maosong Sun (CCL Program Committee Chair)
Ting Liu, Le Sun, and Min Zhang (CCL Program Committee Co-Chairs)
Maosong Sun, Dekang Lin, and Haifeng Wang (NLP-NABD Program Committee Chairs)
Organization

General Chairs
Bo Zhang, Tsinghua University, China
Haoming Zhang, Ministry of Education, China
Zhendong Dong, Hownet, China

Program Committee
12th CCL Program Chair
Maosong Sun, Tsinghua University, China

12th CCL Program Co-chairs
Ting Liu, Harbin Institute of Technology, China
Le Sun, Institute of Software, Chinese Academy of Sciences, China
Min Zhang, Soochow University, China

12th CCL Program Committee
Dongfeng Cai, Shenyang Aerospace University, China
Baobao Chang, Peking University, China
Qunxiu Chen, Tsinghua University, China
Xiaohe Chen, Nanjing Normal University, China
Xueqi Cheng, Institute of Computing Technology, Chinese Academy of Sciences, China
Key-Sun Choi, KAIST, Korea
Li Deng, Microsoft Research, USA
Alexander Gelbukh, National Polytechnic Institute, Mexico
Josef van Genabith, Dublin City University, Ireland
Randy Goebel, University of Alberta, Canada
Tingting He, Huazhong Normal University, China
Isahara Hitoshi, Toyohashi University of Technology, Japan
Heyan Huang, Beijing Polytechnic University, China
Xuanjing Huang, Fudan University, China
Donghong Ji, Wuhan University, China
Turgen Ibrahim, Xinjiang University, China
Shiyong Kang, Ludong University, China
Sadao Kurohashi, Kyoto University, Japan
Kiong Lee, ISO TC37, Korea
Hang Li, Huawei, Hong Kong, SAR China
Ru Li, Shanxi University, China
VIII Organization

Dekang Lin  
Google, USA

Qun Liu  
Institute of Computing Technology, Chinese Academy of Sciences, China

Shaoming Liu  
Fuji Xerox, Japan

Qin Lu  
Polytechnic University of Hong Kong, Hong Kong, SAR China

Wolfgang Menzel  
University of Hamburg, Germany

Jian-Yun Nie  
University of Montreal, Canada

Yanqiu Shao  
Beijing Language and Culture University, China

Xiaodong Shi  
Xiamen University, China

Rou Song  
Beijing Language and Culture University, China

Jian Su  
Institute for Infocomm Research, Singapore

Benjamin Ka Yin Tsou  
The Hong Kong Institute of Education, Hong Kong, SAR China

Haifeng Wang  
Baidu, China

Fei Xia  
University of Washington, USA

Feiyu Xu  
DFKI, Germany

Nianwen Xue  
Brandeis University, USA

Ping Xue  
Research & Technology, The Boeing Company

Erhong Yang  
Beijing Language and Culture University, China

Tianfang Yao  
Shanghai Jiaotong University, China

Shiwen Yu  
Peking University, China

Quan Zhang  
Institute of Acoustics, Chinese Academy of Sciences, China

Jun Zhao  
Institute of Automation, Chinese Academy of Sciences, China

Guodong Zhou  
Soochow University, China

Ming Zhou  
Microsoft Research Asia, China

Jingbo Zhu  
Northeast University, China

First NLP-NABD Program Chairs

Maosong Sun  
Tsinghua University, China

Dekang Lin  
Google, USA

Haifeng Wang  
Baidu, China

First NLP-NABD Program Committee

Key-Sun Choi  
KAIST, Korea

Li Deng  
Microsoft Research, USA

Alexander Gelbukh  
National Polytechnic Institute, Mexico

Josef van Genabith  
Dublin City University, Ireland

Randy Goebel  
University of Alberta, Canada

Isahara Hitoshi  
Toyohashi University of Technology, Japan
Xuanjing Huang  
Fudan University, China

Donghong Ji  
Wuhan University, China

Sadao Kurohashi  
Kyoto University, Japan

Kiong Lee  
ISO TC37, Korea

Hang Li  
Huawei, Hong Kong

Hongfei Lin  
Dalian Polytechnic University, China

Qun Liu  
Institute of Computing, Chinese Academy of Sciences, China

Shaoming Liu  
Fuji Xerox, Japan

Ting Liu  
Harbin Institute of Technology, China

Yang Liu  
Tsinghua University, China

Qin Lu  
Polytechnic University of Hong Kong, Hong Kong, SAR China

Wolfgang Menzel  
University of Hamburg, Germany

Hwee Tou Ng  
National University of Singapore, Singapore

Jian-Yun Nie  
University of Montreal, Canada

Jian Su  
Institute for Infocomm Research, Singapore

Zhifang Sui  
Peking University, China

Le Sun  
Institute of Software, Chinese Academy of Sciences, China

Benjamin Ka Yin Tsou  
The Hong Kong Institute of Education, Hong Kong, SAR China

Fei Xia  
University of Washington, USA

Feiyu Xu  
DFKI, Germany

Nianwen Xue  
Brandeis University, USA

Ping Xue  
Research & Technology, The Boeing Company

Jun Zhao  
Institute of Automation, Chinese Academy of Sciences, China

Guodong Zhou  
Soochow University, China

Ming Zhou  
Microsoft Research Asia, China

Organizing Committee

Organizing Committee Chair

Qiaoming Zhu  
Soochow University, China

Organizing Committee Co-chair

Yang Liu  
Tsinghua University, China

Longhua Qian  
Soochow University, China
# Table of Contents

## Word Segmentation

Improving Chinese Word Segmentation Using Partially Annotated Sentences ................................. 1  
\textit{Kaixu Zhang, Jinsong Su, and Changle Zhou}

Chinese Natural Chunk Research Based on Natural Annotations in Massive Scale Corpora: Exploring Work on Natural Chunk Recognition Using Explicit Boundary Indicators ................................................................. 13  
\textit{Zhi-e Huang, En-dong Xun, Gao-qi Rao, and Dong Yu}

A Kalman Filter Based Human-Computer Interactive Word Segmentation System for Ancient Chinese Texts ......................... 25  
\textit{Tongfei Chen, Weimeng Zhu, Xueqiang Lv, and Junfeng Hu}

Chinese Word Segmentation with Character Abstraction ................................. 36  
\textit{Le Tian, Xipeng Qiu, and Xuanjing Huang}

A Refined HDP-Based Model for Unsupervised Chinese Word Segmentation ................................. 44  
\textit{Wenzhe Pei, Dongxu Han, and Baobao Chang}

Enhancing Chinese Word Segmentation with Character Clustering ................................. 52  
\textit{Yijia Liu, Wanxiang Che, and Ting Liu}

Integrating Multi-source Bilingual Information for Chinese Word Segmentation in Statistical Machine Translation ................................. 61  
\textit{Wei Chen, Wei Wei, Zhenbiao Chen, and Bo Xu}

## Open-Domain Q&A

Interactive Question Answering Based on FAQ. ........................................ 73  
\textit{Song Liu, Yi-Xin Zhong, and Fu-Ji Ren}

## Discourse, Coreference and Pragmatics

Document Oriented Gap Filling of Definite Null Instantiation in FrameNet ................................. 85  
\textit{Ning Wang, Ru Li, Zhangzhang Lei, Zhiqiang Wang, and Jingpan Jin}

Interesting Linguistic Features in Coreference Annotation of an Inflectional Language ................................. 97  
\textit{Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska}
Statistical and Machine Learning Methods in NLP

Semi-supervised Learning with Transfer Learning ................................. 109
   Huiwei Zhou, Yan Zhang, Degen Huang, and Lishuang Li

Online Distributed Passive-Aggressive Algorithm for Structured Learning ................................................................. 120
   Jiayi Zhao, Xipeng Qiu, Zhao Liu, and Xuanjing Huang

Power Law for Text Categorization .................................................. 131
   Wuying Liu, Lin Wang, and Mianzhu Yi

Semantics

Natural Language Understanding for Grading Essay Questions in Persian Language ......................................................... 144
   Iman Mokhtari-Fard

Text Mining, Open-Domain Information Extraction and Machine Reading of the Web

Learning to Extract Attribute Values from a Search Engine with Few Examples ................................................................. 154
   Xingxing Zhang, Tao Ge, and Zhifang Sui

User-Characteristics Topic Model ...................................................... 166
   Wenfeng Li, Xiaojie Wang, and Shaowei Jiang

Mining User Preferences for Recommendation: A Competition Perspective ........................................................................ 179
   Shaowei Jiang, Xiaojie Wang, Caixia Yuan, and Wenfeng Li

Sentiment Analysis, Opinion Mining and Text Classification

A Classification-Based Approach for Implicit Feature Identification ................................................................. 190
   Lingwei Zeng and Fang Li

Role of Emoticons in Sentence-Level Sentiment Classification ................................................................. 203
   Martin Min, Tanya Lee, and Ray Hsu

   Aijun Li, Qiang Fang, Yuan Jia, and Jianwu Dang

Pests Hidden in Your Fans: An Effective Approach for Opinion Leader Discovery ................................................................. 227
   Binyang Li, Kam-fai Wong, Lanjun Zhou, Zhongyu Wei, and Jun Xu
**Lexical semantics and Ontologies**

Exploiting Lexicalized Statistical Patterns in Chinese Linguistic Analysis ........................................................ 238
  *Yu Zhao and Maosong Sun*

**Language Resources and Annotation**

Development of Traditional Mongolian Dependency Treebank .......... 247
  *Xiangdong Su, Guanglai Gao, and Xueliang Yan*

Chinese Sentence Compression: Corpus and Evaluation ................. 257
  *Chunliang Zhang, Minghan Hu, Tong Xiao, Xue Jiang, Lixin Shi, and Jingbo Zhu*

**Machine Translation**

Graphic Language Model for Agglutinative Languages: Uyghur as Study Case ............................................................ 268
  *Miliwan Xuehelaiti, Kai Liu, Wenbin Jiang, and Tuergen Yibulayin*

iCPE: A Hybrid Data Selection Model for SMT Domain Adaptation ... 280
  *Longyue Wang, Derek F. Wong, Lidia S. Chao, Yi Lu, and Junwen Xing*

Multi-classifier Combination for Translation Error Detection .......... 291
  *Jinhua Du, Junbo Guo, Sha Wang, and Xiyuan Zhang*

**Speech Recognition and Synthesis**

Automatic Discrimination of Pronunciations of Chinese Retroflex and Dental Affricates ................................................. 303
  *Akemi Hoshino and Akio Yasuda*

A New Word Language Model Evaluation Metric for Character Based Languages ......................................................... 315
  *Peilu Wang, Ruihua Sun, Hai Zhao, and Kai Yu*

**Tagging and Chunking**

Bidirectional Sequence Labeling via Dual Decomposition ............... 325
  *Zhiguo Wang, Chengqing Zong, and Nianwen Xue*

Semantic Analysis of Chinese Prepositional Phrases for Patent Machine Translation .................................................. 333
  *Renfen Hu, Yun Zhu, and Yaohong Jin*
Large-scale Knowledge Acquisition and Reasoning

Massive Scientific Paper Mining: Modeling, Design and Implementation ................................................. 343
    Yang Zhou, Shufan Ji, and Ke Xu

Author Index ................................................................. 353