ACL-IJCNLP 2015

# The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing

# Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing

July 30-31, 2015
Beijing, China

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 978-1-941643-57-0

# Preface

Welcome to the Eighth SIGHAN Workshop on Chinese Language Processing! Sponsored by the Association for Computational Linguistics (ACL) Special Interest Group on Chinese Language Processing (SIGHAN), this year's SIGHAN-8 workshop is being held in Beijing, China, on July 30-31, 2015, and is co-located with ACL-IJCNLP 2015. The workshop program includes three keynote speeches, research paper presentations and two Bake-offs. We hope that these events will bring together researchers and practitioners to share ideas and developments in various aspects of Chinese language processing.

We have received 17 valid submissions, each of which has been assigned to three reviewers. After a rigorous review process, we have accepted 5 papers for oral presentations (30% acceptance rate) and 6 papers for poster presentations, representing a global acceptance rate of 65%.

We are honored to welcome our distinguished speakers: Dr. Min Zhang (Distinguished Professor, Soochow University, China) and Rou Song (Professor, Beijing Language and Culture University, China) will give the first keynote speech "Discourse and Machine Translation." Yanxiong Lu and Lianqiang Zhou (WeChat Pattern Recognition Center at Tencent) will speak on "Intelligent Q&A System and NLP Open Platform." Finally, Dr. Lun-Wei Ku (Assistant Research Fellow, Academia Sinica, Taiwan) will speak on "From Lexical to Compositional Chinese Sentiment Analysis."

We would also like to thank the Bake-off organizers. The first task Chinese Spelling Check task was organized by Dr. Yuen-Hsien Tseng (National Taiwan Normal University), Dr. Lung-Hao Lee (National Taiwan Normal University), Dr. Li-Ping Chang (National Taiwan Normal University), and Dr. Hsin-Hsi Chen (National Taiwan University). The second Topic-Based Chinese Message Polarity Classification task is organized by Dr. Xiangwen Liao (Fuzhou University, China), Dr. Ruifeng Xu (Harbin Institute of Technology, China), Dr. Li Binyang (University of International Relation, China), and Dr. Liheng Xu (Institute of Automation, Chinese Academy of Sciences, China). A total of sixteen teams participated in these two tasks and have achieved good results.

Finally, we would like to thank all authors for their submissions. We appreciate your active participation and support to ensure a smooth and successful conference. The publication of these papers represents the joint effort of many researchers, and we are grateful to the efforts of the review committee for their work, and to the SIGHAN committee for their continuing support. We wish all a rewarding and eye-opening time at the workshop.

SIGHAN-8 Workshop Co-organizers
Liang-Chih Yu, Yuan Ze University
Zhifang Sui, Peking University
Yue Zhang, Singapore University of Technology and Design
Vincent Ng, University of Texas at Dallas

# Organizing Committee

**Organizers:**

Liang-Chih Yu, Yuan Ze University
Zhifang Sui, Peking University
Yue Zhang, Singapore University of Technology and Design
Vincent Ng, University of Texas at Dalles

**SIGHAN Committee:**

Chengqing Zong, Chinese Academy of Science
Min Zhang, Soochow University
Gina-Anne Levow, University of Washington
Nianwen Xue, Brandeis University

**Program Committee:**

Chia-Hui Chang, National Central University
Li-Ping Chang, National Taiwan Normal University
Wangxiang Che, Harbin Institute of Technology
Hsin-Hsi Chen, National Taiwan University
Kuan-hua Chen, National Taiwan University
Xiangyu Duan, Soochow University
Xianpei Han, Chinese Academy of Science
Xungjing Huang, Fudan University
Jing Jiang, Singapore Management University
Chunyu Kit, City University of Hong Kong
Wai Lam, Chinese University of Hong Kong
Chao-Hong Liu, Dublin City University
Lung-Hao Lee, National Taiwan University
Haizhou Li, Institute of Infocomm Research
Jyun-Jie Lin, Yuan Ze University
Yang Liu, Tsinghua University
Xiangwen Liao, Fuzhou University
Jianyun Nie, University of Montreal
Likun Qiu, Ludong University
Fuji Ren, The University of Tokoshima
Weiwei Sun, City University of Hong Kong
Yuen-Hsien Tseng, National Taiwan Normal University
Hsin-Min Wang, Academia Sinica
Kun Wang, Chinese Academy of Science
Derek F. Wong, University of Macau
Chung-Hsien Wu, National Chen Kung University
Ruifeng Xu, Harbin Institute of Technology
Chin-Sheng Yang, Yuan Ze University
Jui-Feng Yeh, National Chiayi University
Guodong Zhou, Soochow University
Qiang Zhou, TsingHua University
Jingbo Zhu, Northeastern University

# Invited Talk: Discourse and Machine Translation

**ZHANG Min, Soochow University, China**
**SONG Rou, Beijing Language and Culture University, China**

## Abstract

Discourse in linguistics refers to a unit of language longer than a single sentence. It has not been well studied in the research community of computational linguistics, but it has attracted more and more attention in very recent years. This talk consists of two parts, i.e., discourse and machine translation. We will first give an overview about discourse and review the research state-of-the-art of discourse from both linguistics and computational viewpoints, and then discuss how machine translation can benefit from discourse-level information. Finally, we conclude the talk with some future direction discussions.

## Biography

**ZHANG Min:** a distinguished professor and vice dean of the school of computer science and technology, director of the research Institute for Human Language Technology at Soochow University (China), received his Ph.D. degree in computer science from Harbin Institute of Technology (China) in 1997. He has studied and worked oversea in industry and academy at South Korea and Singapore since 1997 to 2013. His current research interests include machine translation and natural language processing. He has co-authored 2 Springer books and more than 130 papers in leading journals and conferences, and co-edited 13 books published by Springer and IEEE. He is an associate editor of IEEE T-ASLP (2015-2017).

**SONG Rou:** a professor and Ph.D. supervisor at Applied Linguistics and Computer Application in Beijing Language and Culture University, received his Bachelor degree in mathematics and mechanics from Beijing University in 1968 and his mater Master degree in computer science from Beijing University in 1981. He has been working on Chinese Information Processing study for tens of years as the PIs of more than 10 national-level projects with the research focuses on discourse analysis, Chinese word segmentation, Computer-aided proofreading, Chinese word attribute, Chinese Orthographic Computing and Chinese POS and so on. He has published more than 100 papers at leading journals and conferences in computer science and linguistics. He has developed and commercialized several softwares with two patents. He has received several awards from Beijing City and MOE, China. He has been appointed as guest professors in a few domestic and oversea universities and research institutes.

# Invited Talk: Intelligent Q&A System and NLP Open Platform

**LU Yanxiong and ZHOU Lianqiang**
**WeChat Pattern Recognition Center, Tencent**

## Abstract

Building a general Q&A system that can handle any subject is a very challenging AI task. Internet social platforms accumulate large amount of active users and UGC (User Generate Content) data, which become valuable crowdsourcing resources. In this talk, we will discuss the opportunity of using WeChat crowdsourcing resources to build an intelligent Q&A systems as well as some open questions and challenges under this topic.

Tencent Open Platform "Wen Zhi" provides comprehensive natural language processing APIs, including the modules of Lexical, Syntax, Semantics and Paragraph. It also provides the web crawling, data extraction and transcoding services. In this talk we will give an overview of Tencent NLP open platform as well as the techniques behind.

## Biography

**LU Yanxiong** is the senior researcher of WeChat Pattern Recognition Center, Tencent. He has been working on search query analysis, Q&A system and NLP related projects in Tencent. His current work focus on WeChat semantic analysis. His research interests include search engine, machine learning, NLP and big data analysis. Before joining in Tencent, Yanxiong worked in Baidu and graduated from Xidian University with master degree.

**ZHOU Lianqiang** has been working in the field of NLP and machine learning in Tencent, such as search query re-write, user interests mining, word segmentation, etc. He is now the senior researcher and team leader of NLP research group in Tencent Intelligent Computing and Search Lab. Before joining Tencent Lianqiang worked in several Internet companies and got his master degree from Harbin Institute of Technology.

# Invited Talk: From Lexical to Compositional Chinese Sentiment Analysis

**KU Lun-Wei**
**Academia Sinica, Taiwan**

## Abstract

Sentiment analysis determines the polarities and strength of sentiment-bearing expressions, and it has been an important and attractive research area due to its close affinity to applications. In the past research, sentiment analysis depended highly on lexical semantics. However, sentiment analysis is eager for the understanding of the context, and shallow features such as bag of words cannot fulfill this need. As a result, compositional semantics, which concerns the construction of meaning based on syntax, has been applied to sentiment analysis through different approaches. In the Chinese language, as morphological structures may represent the compositional semantics inside Chinese words, the compositional sentiment analysis can even start from determining the sentiment of morphemes, which will be touched in this talk.

This talk will begin from some background knowledge of sentiment analysis, such as how sentiment are categorized, where to find available corpora and which models are commonly applied, especially for the Chinese language. I will describe our work on compositional Chinese sentiment analysis from words to sentences. All our involved and recently developed related resources, including Chinese Morphological Dataset, Augmented NTU Sentiment Dictionary (aug-NTUSD), E-hownet with sentiment information, and Chinese Opinion Treebank, will also be introduced in this talk. I'll end by describing how we have begun to test our compositional model with word embeddings.

## Biography

**KU Lun-Wei** received her Ph.D. degree in Computer Science and Information Engineering from National Taiwan University. Then she joined the Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology (Yuntech), Taiwan, as an assistant professor. Since Aug. 2012, she joined the Institute of Information Science, Academia Sinica as an assistant research fellow. Previously, she was a postdoctoral researcher at the Department of Computer Science and Information Engineering, National Taiwan University, working on the project "Machine learning methods for ranking problems in multilingual information retrieval". She was a project researcher in Acer Product Value Lab, Taiwan, between Apr. 2003 and May 2004. At that time, she joined the project in speech recognition services for home media center. She was a software engineer/project manager in NaturalTel, a platform service provider of carriers, where she joined the development of speech entertainment service platform for Fareastone (Fetnet), Taiwan. Her international recognition includes CyberLink Technical Elite Fellowship in 2007, IBM Ph.D. Fellowship in 2008, ROCLING Doctorial Dissertation Distinction Award in 2009, and Good Design Award selected in 2012. Her research interests include natural language processing, information retrieval, sentiment analysis, and computational linguistics. She has been working on Chinese sentiment analysis since year 2005 and was the co-organizer of NTCIR MOAT Task (Multilingual Opinion Analysis Task, traditional Chinese side) from year 2006 to 2010. She is also one of the organizers of the SocialNLP workshop, which has been held jointly in IJCNLP 2013, Coling 2014, WWW 2015 and NAACL 2015. This year, she serves as the area chair of the sentiment analysis and opinion mining track in The 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015), as well as in The 2015 Conference on Empirical Methods on Natural Language Processing (EMNLP 2015). Other professional international activities she involved include The Publication Co-Chair, The 6th International Joint Conference on Natural Language Processing (IJCNLP-2013), Publicity Chair, The Twenty-fourth Conference on Computational Linguistics and Speech Processing (Rocling 2012), and Finance Chair, The Sixth Asia Information Retrieval Societies Conference (AIRS 2010).

# Table of Contents

# Workshop Program

**Thursday, July 30, 2015**

**09:00–09:10 Opening Session**

**09:10–10:30 Invited Talk**

*Discourse and Machine Translation*
Min Zhang and Rou Song

**10:30–10:50 Coffee Break**

**10:50–12:30 Workshop Session**

10:50–11:10 *Sequential Annotation and Chunking of Chinese Discourse Structure*
Frances Yung, Kevin Duh and Yuji Matsumoto

11:10–11:30 *Create a Manual Chinese Word Segmentation Dataset Using Crowdsourcing Method*
Shichang Wang, Chu-Ren Huang, Yao Yao and Angel Chan

11:30–11:50 *Chinese Named Entity Recognition with Graph-based Semi-supervised Learning Model*
Aaron Li-Feng Han, Xiaodong Zeng, Derek F. Wong and Lidia S. Chao

11:50–12:10 *Sentence selection for automatic scoring of Mandarin proficiency*
Jiahong Yuan, Xiaoying Xu, Wei Lai, Weiping Ye, Xinru Zhao and Mark Liberman

12:10–12:30 *ACBiMA: Advanced Chinese Bi-Character Word Morphological Analyzer*
Ting-Hao Huang, Yun-Nung Chen and Lingpeng Kong

**Thursday, July 30, 2015 (continued)**

**12:30–14:30 Lunch**

**14:30–15:30 Invited Talk**

*From Lexical to Compositional Chinese Sentiment Analysis*
Lun-Wei Ku

**15:30–16:00 Coffee Break**

**16:00–17:20 Bake-off Task 1: Chinese Spelling Check**

16:00–16:20 *Introduction to SIGHAN 2015 Bake-off for Chinese Spelling Check*
Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang and Hsin-Hsi Chen

16:20–16:40 *HANSpeller++: A Unified Framework for Chinese Spelling Correction*
Shuiyuan Zhang, Jinhua Xiong, Jianpeng Hou, Qiao Zhang and Xueqi Cheng

16:40–17:00 *Word Vector/Conditional Random Field-based Chinese Spelling Error Detection for SIGHAN-2015 Evaluation*
Yih-Ru Wang and Yuan-Fu Liao

17:00–17:20 *Introduction to a Proofreading Tool for Chinese Spelling Check Task of SIGHAN-8*
Tao-Hsing Chang, Hsueh-Chih Chen and Cheng-Han Yang

**Friday, July 31, 2015**

**09:00–10:30 Invited Talk**

*Intelligent Q&A System and NLP Open Platform*
Yanxiong Lu and Lianqiang Zhou

**10:30–11:00 Coffee Break**

**11:00–12:20 Bake-off Task 2: Topic-Based Chinese Message Polarity Classification**

11:00–11:20 *Overview of Topic-based Chinese Message Polarity Classification in SIGHAN 2015*
Xiangwen Liao, Binyang Li and Liheng Xu

11:20–11:40 *A Joint Model for Chinese Microblog Sentiment Analysis*
Yuhui Cao, Zhao Chen, Ruifeng Xu, Tao Chen and Lin Gui

11:40–12:00 *Learning Salient Samples and Distributed Representations for Topic-Based Chinese Message Polarity Classification*
Xin Kang, Yunong Wu and Zhifei Zhang

12:00–12:20 *An combined sentiment classification system for SIGHAN-8*
Qiuchi Li, Qiyu Zhi and Miao Li

**Friday, July 31, 2015 (continued)**

**12:20–14:00 Lunch**

**14:00–15:20 Poster Session**

*Linguistic Knowledge-driven Approach to Chinese Comparative Elements Extraction*
MinJun Park and Yulin Yuan

*A CRF Method of Identifying Prepositional Phrases in Chinese Patent Texts*
Hongzheng Li and Yaohong Jin

*Emotion in Code-switching Texts: Corpus Construction and Analysis*
Sophia Lee and Zhongqing Wang

*Chinese in the Grammatical Framework: Grammar, Translation, and Other Applications Anonymous*
Aarne Ranta, Tian Yan and Haiyan Qiao

*KWB: An Automated Quick News System for Chinese Readers*
Yiqi Bai, Wenjing Yang, Hao Zhang, Jingwen Wang, Ming Jia, Roland Tong and Jie Wang

*Chinese Semantic Role Labeling using High-quality Syntactic Knowledge*
Gongye Jin, Daisuke Kawahara and Sadao Kurohashi

*Chinese Spelling Check System Based on N-gram Model*
Weijian Xie, Peijie Huang, Xinrui Zhang, Kaiduo Hong, Qiang Huang, Bingzhou Chen and Lei Huang

*NTOU Chinese Spelling Check System in Sighan-8 Bake-off*
Wei-Cheng Chu and Chuan-Jie Lin

*Topic-Based Chinese Message Sentiment Analysis: A Multilayered Analysis System*
hongjie li, zhongqian sun and wei yang

*Rule-Based Weibo Messages Sentiment Polarity Classification towards Given Topics*
Hongzhao Zhou, Yonglin Teng, Min Hou, Wei He, Hongtao Zhu, Xiaolin Zhu and Yanfei Mu

*Topic-Based Chinese Message Polarity Classification System at SIGHAN8-Task2*
Chun Liao, Chong Feng, Sen Yang and Heyan Huang

**Friday, July 31, 2015 (continued)**

*CT-SPA: Text sentiment polarity prediction model using semi-automatically expanded sentiment lexicon*
Tao-Hsing Chang, Ming-Jhih Lin, Chun-Hsien Chen and Shao-Yu Wang

*Chinese Microblogs Sentiment Classification using Maximum Entropy*
Dashu Ye, Peijie Huang, Kaiduo Hong, Zhuoying Tang, Weijian Xie and Guilong Zhou

*NDMSCS: A Topic-Based Chinese Microblog Polarity Classification System*
Yang Wang, Yaqi Wang, Shi Feng, Daling Wang and Yifei Zhang

*NEUDM: A System for Topic-Based Message Polarity Classification*
Yaqi Wang, Shi Feng, Daling Wang and Yifei Zhang

**15:20–15:30 Closing Session**