

Proceedings of
ROCLING X (1997)
International Conference
Research on Computational Linguistics

August 22-24, 1997
Academia Sinica, Taipei, Taiwan

Sponsored by:

Association for Computational Linguistics
and Chinese Language Processing
Institute of Information Science, Academia Sinica
Institute of Linguistics (Preparatory Office), Academia Sinica

Co-sponsored by:

Behavior Design Corporation
CCL Industrial Technology Research Institute
Global View Corporation
IBM Taiwan Corporation
Institute for Information Industry
Matsushita Electric Institute of Technology (Taipei) Co., LTD.
Microsoft Taiwan Corporation
National Science Council
Philips Taiwan LTD.
Telecommunication Laboratories ChungHwa Telecom Co., LTD.

Conference Chair's Message

It is our great pleasure to welcome you all to the city of Taipei and to the ROCLING X (1997) International Conference. After nine very successful years, this year's conference reflects the growing international participation in ROCLING by being the first international conference of its kind in Taiwan. We would like to invite all of you to experience this modern oriental metropolis with hospitality in the Chinese style. We are sure you will find this conference to be an excellent forum for innovation and technical discussions, and a very natural environment for extending friendship and fellowship here in Taipei. Please join us for this very special event.

The day when computers can process language as well as humans, and maybe even better and faster, is still just a dream, but a dream that is getting closer and closer to reality. Problems such as the recognition of compound words and proper names, disambiguation, speech recognition and synthesis, the representation of meaning, and fast and robust processing are just some of the problems that have prevented this dream from becoming a reality. However, as some of the papers in this conference will show, significant progress is being made towards solving these problems. Tutorials on the lexical semantics of verbs and on the use of finite-state methods, as well as a panel on the question of the need for psychologically real processing of language, serve not only to provide a forum for the presentation of knowledge and ideas, but also to enhance intellectual exchange and interaction.

Although Taiwan is an island off the eastern shore of China, it has been a very good heir to classical Chinese culture. Traditional Chinese flavor flourishes everywhere. Taipei is also the world capital for Chinese cuisine. Please join us at the reception to be held on August 22nd, and at the banquet featuring "real" Chinese food on August 23rd. On behalf of the Organizing Committee of ROCLING X, we would like to welcome you all here in Taipei in August 1997.

Lin-shan Lee

ROCLING X Conference Chair

ORGANIZING COMMITTEE

Honorary chairman: Yuan-Tseh Lee, Academia Sinica

Chairman: Lin-Shan Lee, Academia Sinica

PROGRAM COMMITTEE

Co-chairmen: Keh-Jiann Chen, Academia Sinica

Chu-Ren Huang, Academia Sinica

Richard Sproat, Bell Labs

Steve Abney, University of Tuebingen

Rens Bod, University of Amsterdam

Jyun-sheng Chang, Tsing Hua University

Hsin-hsi Chen, Taiwan University

Shin-Horng Chen, Chiao Tung University

Lee-feng Chien, Academia Sinica

Key-sun Choi, KAIST

Jerry Hobbs, SRI International

Changning Huang, Tsing Hua University

Geunbae Lee, Pohang University of Science and Technology

Hsi-Jian Lee, Chiao Tung University

Kim Teng Lua, National University of Singapore

Yuji Matsumoto, Nara Institute of Science and Technology

Masaaki Nagata, NTT

Martha Palmer, University of Penn.

Jerry Seligman, Chung Cheng University

Keh-yih Su, Tsing Hua University

Hozumi Tanaka, Tokyo Institute of Technology

Chiu-yu Tseng, Academia Sinica

Benjamin K. T'sou, City University of Hong Kong

Jhing-fa Wang, Cheng Kung University

Dekai Wu, Hong Kong University of Science and Technology

David Yarowsky, Johns Hopkins University

Victor Zue, MIT

Table of Contents

Keynote Speech

Setting Parameters: Triggering and Mis-triggering	1
<i>Janet D. Fodor</i>	

Invited Speech

Encoding Events across Languages	2
<i>Beth Levin</i>	
Opposing Effects of Word and Character Frequency in the Processing of Chinese text	3
<i>Ovid T.L. Tzeng, Chih Wei Hue, Daisy L. Hung</i>	

Session 1a: Lexical Semantics

Meaning Representation and Meaning Instantiation for Chinese Nominals	4
<i>Kathleen Ahrens, Lily Chang, Keh-Jiann Chen, Chu-Ren Huang</i>	
Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy	19
<i>Jay J. Jiang, David W. Conrath</i>	
Towards a Representation of Verbal Semantics - An Approach Based on Near-Synonyms	34
<i>Mei-Chih Tsai, Chu-Ren Huang, Keh-Jiann Chen, Kathleen Ahrens</i>	
*Word Sense Disambiguation Based on the Information Theory	49
<i>Ho Lee, Dae-Ho Baek, Hae-Chang Rim</i>	

Session 1b: Parsing

An Agreement Error Correction Method Based on a Multicriteria Approach: An Application to Arabic Language	59
<i>Lamia Belguith Hadrich, Abdelmajid Ben Hamadou, Chafik Aloulou</i>	
Incorporating Bigram Constraints into an LR Table	76
<i>Hiroki Imai, Hui Li, Hozumi Tanaka</i>	
A Level-Synchronous Approach to Ill-Formed Sentence Parsing	89
<i>Yi-Chung Lin, Keh-Yih Su</i>	
*The Application of the Similarities between the Morphemes of the English and Chinese Languages to Represent Chinese Characters Phonetically with English Letters to Facilitate Computer Applications Manually and by Voice with the Character-Based Languages Chinese, Japanese and Korean	109
<i>Stanley K. Chan</i>	

Session 2a: Lexical Analysis

A Multivariate Gaussian Mixture Model for Automatic Compound Word Extraction	123
<i>Jing-Shin Chang, Keh-Yih Su</i>	
Proper Name Extraction from Web Pages for Finding People in Internet	143
<i>Hsin-Hsi Chen, Guo-Wei Bian</i>	
Unknown Word Detection for Chinese by a Corpus-based Learning Method	159
<i>Keh-Jiann Chen, Ming-Hong Bai</i>	

Session 2b: Message Understanding

Analyzing the Complexity of A Domain with Respect to An Information Extraction Task	175
<i>Amit Bagga</i>	
Human Judgment as a Basis for Evaluation of Discourse-Connective-based Full-text Abstraction in Chinese	195
<i>Benjamin K T'sou, Hing-Lung Lin, Tom B.Y. Lai</i>	
An Assessment on Character-based Chinese News Filtering Using Latent Semantic Indexing	209
<i>Shih-Hung Wu, Pey-Ching Yang, Von-Wun Soo</i>	

Poster Session

The Role of Shared Attention in Human-Computer Conversation	224
<i>Hideki Kozima, Akira Ito</i>	
Chinese Word Segmentation and Part-of-Speech Tagging in One Step	229
<i>Tom B.Y. Lai, Maosong Sun, Benjamin K.T'sou, S. Caesar Lan</i>	
Corpus-Based Chinese Text Summarization System	237
<i>Jun-Jie Li, Key-Sun Choi</i>	
A Study on the Portability of a Grammatical Inference system	242
<i>Hsue-Hueh Shih, Steve Young</i>	
Fast Lexical Post-Processing on Cursive Script Recognition	247
<i>Marco A. Torres, Susumu Kuroyanagi, Akira Iwata</i>	
The Study of Recurrent Neural Networks for Language Modeling	252
<i>Wen-Jyun Wang, Jyun-Hsiao Lee, Ji-Yi Liu (in Chinese)</i>	
A Simple Heuristic Approach for Word Segmentation	257
<i>Wing-Kwong Wong, Chenming Hsu, Jie-Iao Chen, Jien-Chi Yu</i>	
Prosody Generation in a Chinese TTS System Based on a Hierarchical Word Prosody Template Tree	262
<i>Chung-Hsien Wu, Jau-Hung Chen</i>	
Ambiguity Resolution Using Lexical Association	267
<i>Juntae Yoon, Seonho Kim, Mansuk Song</i>	
The Description of the Intra-State Feature Space in Speech Recognition	272
<i>Fang Zheng, Mingxing Xu, Wenhui Wu</i>	
Similarity Comparison between Chinese Sentences	277
<i>Lina Zhou, James Liu</i>	
Attributive Clauses in Chinese: Theory and Implementation	282
<i>Xiaokang Zhou, Francis Y. Lin</i>	

Session 4: Speech Processing

Integrating Long-Distance Language Modeling to the Phoneme-to-Text Conversion Task	287
<i>Tai-Hsuan Ho, Kae-Cherng Yang, Juei-Sung Lin, Lin-Shan Lee</i>	
Automatic Speaker Identification Based on Fuzzy Theory and Neural Network Using Genetic Algorithm	300
<i>Ching-Tang Hsieh, Eugene Lai, You-Chuang Wang</i>	

*A General Public Application of Pedagogic and Linguistic Vocations of Speech Synthesis: Ordictee	316
<i>Marc Guyomard, Jacques Siroux, Dominique Pernici, Christophe Royer</i>	
*A Conversational Agent for Food-ordering Dialog Based on VenusDictate	325
<i>Hsien-Chang Wang, Jhing-Fa Wang, Yi-Nan Liu</i>	
*Truncation on Combined Word-Based and Class-Based Language Model Using Kullback-Leibler Distance Criterion	335
<i>Kae-Cherng Yang, Tai-Hsuan Ho, Juei-Sung Lin, Lin-Shan Lee</i>	
Session 5: Speech and Language Processing	
Recognizing Korean Unknown Proper Nouns by Using Automatically Extracted Lexical Clues	345
<i>Bong-Rae Park, Young-Sook Hwang, Hae-Chang Rim</i>	
Logical Operators and Quantifiers in Natural Language	357
<i>Shin-ichiro Kamei, Kazunori Muraki</i>	
Chinese Text Compression Using Chinese Language Information Processing	368
<i>Jun Gao, Xixian Chen (in Chinese)</i>	
*Combining Multiword Units into a Hidden Markov Model for Part-of-Speech Tagging	380
<i>Jae-Hoon Kim</i>	
*A Robust Keyword Spotting System for Mandarin Speech	390
<i>Chung-Hung Chien, Hsiao-Chuan Wang</i>	
*A First Study on Mandarin Prosodic State Detection	399
<i>Yuan-Fu Liao, Wern-Jun Wang, Shu-Ling Lee, Sin-Horng Chen</i>	
*Rejection in Speech Recognition Based on CDCPMs	412
<i>Mingxing Xu, Fang Zheng, Wenhui Wu</i>	
Authors Index	420

Setting Parameters: Triggering and Mis-Triggering

Janet Dean Fodor
CUNY Graduate Center

Abstract

The principles-and-parameters theory of language proposed by Chomsky (1981) greatly simplifies the task of language learning. However, recent research in learnability theory has made it clear that for natural languages there can be no instant "automatic" triggering of parameters. This is because the trigger properties in natural languages are often deep properties, not recognizable without parsing the input sentence.

Current approaches such as Gibson and Wexler (1994) therefore use the sentence parsing routines to identify triggers. Unfortunately, the proposed mechanism for doing so is very inefficient. I show that this is because it does not respect the Parametric Principle: it evaluates millions of particular grammars, rather than establishing the values of 20 or 30 parameters.

By tracing out why this is so, I have found a remedy for it. There is a way of using the parser that does implement the Parametric Principle, and permits efficient learning with no exponential complexity increase. But it calls for a new model of how sentence parsing contributes to learning, and a new conception of parameters and of their triggers: they are one and the same thing, and consist of features or small treelets, made available by UG and adoptable into individual grammars.

This conclusion is in accord with most current theories of syntactic parameterization, including the Minimalist program, HPSG and TAG theory.

References

- Chomsky, N., Lectures on Government and Binding, Foris Publications, Dordrecht, Holland, 1981.
- Fodor, J. D., "Unambiguous triggers," to appear in *Linguistic Inquiry*. (Manuscript supplied)
- Gibson, E. and Wexler, K., "Triggers," *Linguistic Inquiry* 25.3, 1994, pp. 407-454.

Encoding Events Across Languages

Beth Levin

Department of Linguistics

Northwestern University

Abstract

All languages must describe the same events in the world, yet the resources that languages have available for expressing certain types of events vary, giving rise to rather different ways of expressing such events across languages. The best known example of this type of cross-linguistic variation involves differences in the expression of motion events --- differences which have long been noted in manuals of translation and have more recently been introduced into the typological literature by Leonard Talmy. This talk will investigate systematic cross-linguistic similarities and divergences in the expression of events, primarily through a case study of the expression of motion events. Taking Talmy's observations as a starting point, I propose that differences among languages can be characterized in terms of the compositional resources available for expressing telic (i.e., bounded) events, rather than in terms of an inherent difference in language type (i.e., Talmy's path vs.~manner languages), as is more commonly done. This approach receives support from cross linguistic variation in the expressions of certain other event types, which can also be understood as involving compositional resources. In addition, I briefly discuss the implications of such systematic divergences across languages for natural language processing, in general, and machine translation, in particular.

Opposing Effects of Word and Character Frequency in the Processing of Chinese Text

Ovid J. L. Tzeng, Daisy L. Hung, Chih Wei Hue

Abstract

Six experiments were conducted to explore the nature of lexical access in the processing of Chinese text. Subjects were asked to perform a lexical decision task under various stimulus presentation conditions. The stimuli were always Chinese words of two characters which could be presented simultaneously or sequentially. Under the sequential conditions, three inter-character-intervals were manipulated to examine the time courses of various frequency effects due either to word or character. Results of these experiments showed a general positive effect of word frequency under all conditions, whereas negative effects of character frequency were consistently observed at ICI longer than 50 msec. These opposing effects of frequency depict a complicated pattern of word recognition which can be explained only by proposing a dual-route model of lexical access in the processing of Chinese text.