Preface

CICLing-2004 is the fifth Annual Conference on Intelligent Text Processing and Computational Linguistics, see www.CICLing.org. CICLing conferences are intended to provide a balanced view of the cutting edge developments in both theoretical foundations of computational linguistics and practice of natural language text processing with its numerous applications. A feature of CICLing conferences is their wide scope that covers nearly all areas of computational linguistics and all aspects of natural language processing applications. These conferences are a forum for dialogue between the specialists working in the two areas.

This year we were honored by the presence of our invited speakers ***Martin Kay*** of Stanford University, ***Philip Resnik*** of the University of Maryland, ***Ricardo Baeza-Yates***, of the University of Chile, and ***Nick Campbell*** of the ATR Spoken Language Translation Research Laboratories. They delivered excellent extended lectures and organized vivid discussions.

Of 129 submissions received (74 full papers and 44 short papers), after careful international reviewing 74 papers have been selected for presentation (40 full papers and 35 short papers), by 176 authors from 21 countries: Korea (37), Spain (34), Japan (22), Mexico (15), China (11), Germany (10), Ireland (10), UK (10), Singapore (6), Canada (3), Czech Rep. (3), France (3), Brazil (2), Sweden (2), Taiwan (2), Turkey (2), USA (2), Chile (1), Romania (1), Thailand (1), The Netherlands (1); the figures in parentheses stand for the number of authors from the corresponding country.

In addition to high scientific level, one of the success factors of CICLing conferences is their excellent cultural program. CICLing-2004 was held in Korea, the beautiful and wonderful Country of the Morning Calm, as Korean people call their land. The participants enjoyed three full-day excursions to the most important natural and historical attractions around Seoul city; see photos at www.CICLing.org. Full-day excursions allow for friendly personal interaction between participants and give them a chance to make friends with the most famous experts in the field, who are not easily accessible at larger conferences.

A conference is the result of the work of many people. First of all I would like to thank the members of the Program Committee for the time and effort they devoted to the reviewing of the submitted articles and to the selection process. Especially helpful were Manuel Vilares, John Tait, Alma Kharrat, Karin Verspoor, Viktor Pekar, and many others—a complete list would be too long.

Obviously I thank the authors for their patience in the preparation of the papers, not to mention the very development of their scientific results that form this book. I also express my most cordial thanks to the members of the local Organizing Committee for their considerable contribution to making this conference become a reality. Last but not least, I thank our host—the ITRI of the Chung-Ang University. would like to also thank RITOS-2 of CYTED for their support of the CICLing conferences.

December 2003 Alexander Gelbukh

Conference Chair

Alexander Gelbukh (CIC-IPN, Mexico / Chung-Ang U., Korea)

Program Committee

1. Boitet, Christian (CLIPS-IMAG, France),
2. Bolshakov, Igor (CIC-IPN, Mexico),
3. Bontcheva, Kalina (U. Sheffield, UK),
4. Calzolari, Nicoletta (ILC-CNR, Italy),
5. Carroll, John, (U. Sussex, UK)
6. Cristea, Dan (U. Iasi, Romania),
7. Gelbukh, Alexander (**Chair**, CIC-IPN, Mexico / Chung-Ang U., Korea)
8. Hallett, Cătălina (U. Brighton, UK),
9. Han, SanYong (Chung-Ang U., Korea)
10. Harada, Yasunari (Waseda U, Japan),
11. Hasida, Kôiti (Electrotechnical Laboratory-AIST, Japan),
12. Hirst, Graeme (U. Toronto, Canada),
13. Hovy, Eduard (ISI of U. Southern Carolina, USA),
14. Johnson, Frances (Manchester Metropolitan U., UK),
15. Kharrat, Alma (Microsoft Research, USA),
16. Kilgarriff, Adam (U. Brighton, UK),
17. Kittredge, Richard (CoGenTex Inc., USA / Canada),
18. Kübler, Sandra (U. Tübingen, Germany),
19. López López, Aurelio (INAOE, Mexico),
20. Loukanova, Roussanka (Indiana U, USA / Bulgaria),
21. Lüdeling, Anke (U. Stuttgart, Germany),
22. Maegard, Bente (Centre for Language Technology, Denmark),
23. Martín-Vide, Carlos (U. Rovira i Virgili, Spain),
24. Mel'čuk, Igor (U. Montreal, Canada),
25. Metais, Elisabeth (U. Versailles, France),
26. Mihalcea, Rada (U. North Texas, USA),
27. Mitkov, Ruslan (U. Wolverhampton, UK),
28. Murata, Masaki (KARC-CRL, Japan),
29. Narin'yani, Alexander (Russian Institute of Artificial Intelligence, Russia),
30. Nirenburg, Sergei (New Mexico U, USA),
31. Palomar, Manuel (U. Alicante, USA / Spain),
32. Pedersen, Ted (U. Minnesota Duluth, USA),
33. Pekar, Viktor (U. Wolverhampton, UK)
34. Pineda Cortes, Luis Alberto (UNAM, Mexico),
35. Piperidis, Stelios (Institute for Language and Speech Processing, Greece),
36. Pustejovsky, James (Brandeis U., USA)
37. Ren, Fuji (U. Tokushima, Japan),
38. Riloff, Ellen (U. Utah, USA)
39. Sag, Ivan (Standford U, USA),
40. Sharoff, Serge (U. Leeds, UK),

41. Sidorov, Grigori (CIC-IPN, Mexico),
42. Sun Maosong (Tsinghua U, China),
43. Tait, John (U. Sunderland, UK),
44. Trujillo, Arturo (UMIST, UK),
45. T'sou Ka-yin, Benjamin (City U. Hong Kong, Hong Kong),
46. Van Guilder, Linda (MITRE Corp., USA),
47. Verspoor, Karin (Intelligenesis Corp., USA / The Netherlands),
48. Vilares Ferro, Manuel (U. La Coruña, Spain),
49. Wilks, Yorick (U. Sheffield, UK).

Additional Reviewers

1. Babych, Bogdan (U. Leeds, Centre for Translation, UK)
2. Campbell, Nick (ATR Human Information Science Labs, Japan)
3. Liang Shao-Fen (U. Sunderland; Taiwan)
4. Llopis, Fernando (U. Alicante, Spain)
5. Martínez-Barco, Patricio (U. Alicante, Spain)
6. Montoyo, Andrés (U. Alicante, Spain)
7. Oakes, Michael (U. Sunderland; Taiwan)
8. Saiz Noeda, Maximiliano (U. Alicante, Spain)
9. Stokoe, Christopher (U. Sunderland, UK)
10. Vicedo, José L. (U. Alicante, Spain)

Organizing Committee

1. Han, SanYong (**chair**)
2. Gelbukh, Alexander (**co-chair**)
3. Chang, Tae Gyu
4. Shim, Duk Sun
5. Kim, Jun Sung
6. Park, Ho Hyun
7. Kim, Hyung Suk
8. Moon, Suk Whan
9. Choi, Ok Kyung
10. Kang, Nam Oh
11. Shin, Kwang Chul

Organization. Website and Contact

The conference was organized by the Electronic Commerce and Internet Applications Laboratory of the Chung-Ang University, Seoul, Korea, and the Natural Language Laboratory of the Center for Computing Research of the National Polytechnic

Institute, Mexico City, Mexico. The conference was hosted by the ITRI of the Chung-Ang University.

The website of the CICLing conferences is www.CICLing.org (mirrored at www.cic.ipn.mx/cicling). Contact: CICLing.org; see
also www.gelbukh.com.

Table of Contents

Computational Linguistics

Computational Linguistics Formalisms

Towards an LFG Syntax-Semantics Interface for Frame Semantics Annotation 1
Anette Frank, Katrin Erk

Projections from Morphology to Syntax in the Korean Resource Grammar: Implementing Typed Feature Structures 13
Jong-Bok Kim, Jaehyung Yang

A Systemic-Functional Approach to Japanese Text Understanding 25
Noriko Ito, Toru Sugimoto, Michio Sugeno

Building and Using a Russian Resource Grammar in GF 37
Janna Khegai, Aarne Ranta

An Application of a Semantic Framework for the Analysis of Chinese Sentences 41
Li Tang, Donghong Ji, Yu Nie, Lingpeng Yang

Semantics and Dialogue

A Modal Logic Framework for Human-Computer Spoken Interaction 45
Luis Villaseñor-Pineda, Manuel Montes-y-Gómez, Jean Caelen

Agents Interpreting Imperative Sentences 54
Miguel Pérez-Ramírez, Chris Fox

Intention Retrieval with a Socially-supported Belief System 66
Naoko Matsumoto, Akifumi Tokosum

Extracting Domain Knowledge for Dialogue Model Adaptation 70
Kuei-Kuang Lin, Hsin-Hsi Chen

Syntax and Parsing

A Probabilistic Chart Parser implemented with an Evolutionary Algorithm 79
Lourdes Araujo

Probabilistic Shift-Reduce Parsing Model Using Rich Contextual Information 91
Yong-Jae Kwak, So-Young Park, Joon-Ho Lim, Hae-Chang Rim

Evaluation of Feature Combination for Effective Structural Disambiguation 95
So-Young Park, Yong-Jae Kwak, Joon-Ho Lim, Hae-Chang Rim

Parsing incomplete sentences revisited 99
Manuel Vilares, Victor M. Darriba, Jesús Vilares

Unlexicalized Dependency Parser for Variable Word Order Languages based on Local Contextual Pattern 109
Hoojung Chung, Hae-Chang Rim

A cascaded syntactic analyser for Basque 121
Itziar Aduriz, Maxux J. Aranzabe, Jose Mari Arriola, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, Larraitz Uria

Lexical Analysis

An Analysis of Sentence Boundary Detection Systems for English and Portuguese Documents 132
Carlos N. Silla Jr., Celso A. A. Kaestner

Towards Language-Independent Sentence Boundary Detection 139
Do-Gil Lee, Hae-Chang Rim

Korean Compound Noun Decomposition Using Syllabic Information Only 143
Seong-Bae Park, Jeong-Ho Chang, Byoung-Tak Zhang

Named Entity Recognition

Learning Named Entity Classifiers using Support Vector Machines 155
Thamar Solorio, Aurelio López López

An Internet-based method for Verification of Extracted Proper Names.................. 165
Angelo Dalli

Boundary Correction of Protein Names Adapting Heuristic Rules 169
Tomohiro Mitsumori, Sevrani Fation, Masaki Murata, Kouichi Doi, Hirohumi Doi

Word Sense Disambiguation

Word Sense Disambiguation Based on Weight Distribution Model with Multiword Expression 173
Hee-Cheol Seo, Young-Sook Hwang, Hae-Chang Rim

Combining EWN and sense-untagged corpus for WSD 185
Iulia Nica, Mª. Antònia Martí, Andrés Montoyo, Sonia Vázquez

Feature Selection for Chinese Character Sense Discrimination 198
Zheng-Yu Niu, Dong-Hong Ji

The Role of Temporal Expressions in Word Sense Disambiguation....................... 206
Sonia Vázquez, Estela Saquete, Andrés Montoyo, Patricio Martínez-Barco, Rafael Muñoz

Anaphora Resolution

An Empirical Study on Pronoun Resolution in Chinese 210
Wang Houfeng, Mei Zheng

Lexicon and Corpus

Language-independent Methods for Compiling Monolingual Lexical Data............ 214
Christian Biemann, Stefan Bordag, Gerhard Heyer, Uwe Quasthoff, Christian Wolff

Getting One's First Million… Collocations 226
Igor A. Bolshakov

Automatic Syntactic Analysis for Detection of Word Combinations 240
Alexander Gelbukh, Grigori Sidorov, Sang-Yong Han, Erika Hernández-Rubio

A Small System Storing Spanish Collocations 245
Igor A. Bolshakov, Sabino Miranda-Jiménez

A Semi-Automatic Tree Annotating Workbench for Building a Korean Treebank 250
Joon-Ho Lim, So-Young Park, Yong-Jae Kwak, Hae-Chang Rim

Extracting Semantic Categories of Nouns for Syntactic Disambiguation from Human-Oriented Explanatory Dictionaries 254
Hiram Calvo, Alexander Gelbukh

Hierarchies Measuring Qualitative Variables 258
Serguei Levachkine, Adolfo Guzmán-Arenas

Bilingual Resources

Invited talk:
Substring Alignment using Suffix Trees 271
Martin Kay

Invited talk:
Exploiting Hidden Meanings: Using Bilingual Text for Monolingual Annotation 280
Philip Resnik

Acquisition of Word Translations Using Local Focus-based Learning in Ainu-Japanese Parallel Corpora 297
Hiroshi Echizen-ya, Kenji Araki, Yoshio Momouchi, Koji Tochinai

Sentence Alignment for Spanish-Basque Bitexts: Word Correspondences vs. Markup similarity 302
Arantza Casillas, Idoia Fernández, Raquel Martínez

Two-level Alignment by Words and Phrases Based on Syntactic Information 306
Seonho Kim, Juntae Yoon, Dong-Yul Ra

Machine Translation

Exploiting a Mono-Bilingual Dictionary for English-Korean Translation Selection and Sense Disambiguation 318
Hyun Ah Lee, Juntae Yoon, Gil Chang Kim

Source Language Effect on Translating Korean Honorifics 330
Kyonghee Paik, Kiyonori Ohtake, Francis Bond, Kazuhide Yamamoto

An Algorithm for Determining DingYu Structural Particle using Grammar Knowledge and Statistical Information 334
Fuji Ren

Natural Language Generation

Generating natural word orders in a semi-free word order language: Treebank-based linearization preferences for German 346
Gerard Kempen, Karin Harbusch

Guideline for developing a software life cycle process in natural language generation projects 351
M^a del Socorro Bernardos

Human-Computer Interaction Applications

A plug and play spoken dialogue interface for smart environments 355
Germán Montoro, Xavier Alamán, Pablo A. Haya

Evaluation of Japanese Dialogue Processing Method Based on Similarity Measure Using tf · AoI 366
Yasutomo Kimura, Kenji Araki, Koji Tochinai

Towards Programming in Everyday Language: A Case for Email Management 378
Toru Sugimoto, Noriko Ito, Shino Iwashita, Michio Sugeno

Speech Recognition and Synthesis

Invited talk:
Specifying Affect and Emotion for Expressive Speech Synthesis 390
Nick Campbell

Overcoming the Sparseness Problem of Spoken Language Corpora using Other Large corpora of distinct characteristics 402
Sehyeong Cho, SangHun Kim, Jun Park, YoungJik Lee

A Syllabification Algorithm for Spanish 406
Heriberto Cuayáhuitl

Experiments on the Construction of a Phonetically Balanced Corpus from the Web 410
Luis Villaseñor-Pineda, Manuel Montes-y-Gómez, Dominique Vaufreydaz, Jean-François Serignat

Intelligent Text Processing

Indexing

Head/Modifier Frames for Information Retrieval 414
Cornelis H.A. Koster

Performance Analysis of Semantic Indexing in Text Retrieval 427
Bo-Yeong Kang, Hae-Jung Kim, Sang-Jo Lee

A Model for Extracting Keywords of Document Using Term Frequency and Distribution 431
Jae-Woo Lee, Doo-Kwon Baik

A Combining Approach to Automatic Keyphrases Indexing for Chinese News Documents 435
Wang Houfeng, Li Sujian, Yu Shiwen, Kang Byeong Kwu

Information Retrieval

Invited talk:
Challenges in the Interaction of Information Retrieval and Natural Language Processing 439
Ricardo Baeza-Yates

The Challenge of Creative Information Retrieval 451
Tony Veale

Using T-Ret System to Improve Incident Report Retrieval 463
Joe Carthy, David C. Wilson, Ruichao Wang, John Dunnion, Anne Drummond

Question Answering and Sentence Retrieval

Spanish Question Answering Evaluation 467
Anselmo Peñas, Felisa Verdejo, Jesús Herrera

Comparative Analysis of Term Distributions in a Sentence and in a Document for Sentence Retrieval 479
Kyoung-Soo Han, Hae-Chang Rim

Browsing

Contextual Exploration of Text Collections 483
Manuel Montes-y-Gómez, Manuel Pérez-Coutiño, Luis Villaseñor-Pineda, Aurelio López-López

Automatic Classification and Skimming of Articles in a News Video Using Korean Closed-caption 492
Jung-Won Cho, Seung-Do Jeong, Byung-Uk Choi

Filtering

A Framework for Evaluation of Information Filtering Techniques in an Adaptive Recommender System 496
John O'Donovan, John Dunnion

Lexical Chains versus Keywords for Topic Tracking 500
Joe Carthy

Filtering Very Similar Text Documents: A Case Study 504
Jiří Hroza, Jan Žižka, Aleš Bourek

Information Extraction

Using Information Extraction to Build a Directory of Conference Announcements 515
Karl-Michael Schneider

Unsupervised Event Extraction from Biomedical Text based on Event and Pattern Information 527
Hong-woo Chun, Young-sook Hwang, Hae-Chang Rim

Thai Syllable-Based Information Extraction Using Hidden Markov Models 531
Lalita Narupiyakul, Calvin Thomas, Nick Cercone, Booncharoen Sirinaovakul

The impact of enriched linguistic annotation on the performance of extracting relation triples 541
Sanghee Kim, Paul Lewis, Kirk Martinez

Text Categorization

An kNN Model-based Approach and its Application in Text Categorization........ 553
Gongde Guo, Hui Wang, David Bell, Yaxin Bi, Kieran Greer

Automatic Learning Features Using Bootstrapping for Text Categorization........ 565
Chen Wenliang, Zhu Jingbo, Wu Honglin, Yao Tianshun

Recomputation of Class Relevance Scores for Improving Text Classification........ 575
Sang-Bum Kim, Hae-Chang Rim

Raising High-Degree Overlapped Character Bigrams into Trigrams for Dimensionality Reduction in Chinese Text Categorization........ 579
Dejun Xue, Maosong Sun

Information Retrieval and Text Categorization with Semantic Indexing 591
Paolo Rosso, Antonio Molina, Ferran Pla, Daniel Jiménez, Vicent Vidal

Document Clustering

Sampling and Feature Selection in a Genetic Algorithm for Document Clustering 596
Arantza Casillas, Mayte T. González de Lena, Raquel Martínez

A New Efficient Clustering Algorithm for Organizing Dynamic Data Collection........ 608
Kwangcheol Shin, Sangyong Han

Domain-informed Topic Detection 612
Cormac Flynn, John Dunnion

Summarization

Assessing the Impact of Lexical Chain Scoring Methods and Sentence Extraction Schemes on Summarization........ 622
William Doran, Nicola Stokes, Joe Carthy, John Dunnion

A Term Weighting Method based on Lexical Chain for Automatic Summarization 631
Young-In Song, Kyoung-Soo Han, Hae-Chang Rim

Language Identification

Centroid-Based Language Identification Using Letter Feature Set 635
Hidayet Takcı, İbrahim Soğukpınar

Author Index 645