

*Commenced Publication in 1973*

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

## Editorial Board

David Hutchison

*Lancaster University, UK*

Takeo Kanade

*Carnegie Mellon University, Pittsburgh, PA, USA*

Josef Kittler

*University of Surrey, Guildford, UK*

Jon M. Kleinberg

*Cornell University, Ithaca, NY, USA*

Alfred Kobsa

*University of California, Irvine, CA, USA*

Friedemann Mattern

*ETH Zurich, Switzerland*

John C. Mitchell

*Stanford University, CA, USA*

Moni Naor

*Weizmann Institute of Science, Rehovot, Israel*

Oscar Nierstrasz

*University of Bern, Switzerland*

C. Pandu Rangan

*Indian Institute of Technology, Madras, India*

Bernhard Steffen

*University of Dortmund, Germany*

Madhu Sudan

*Massachusetts Institute of Technology, MA, USA*

Demetri Terzopoulos

*University of California, Los Angeles, CA, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Gerhard Weikum

*Max-Planck Institute of Computer Science, Saarbruecken, Germany*

Alexander Gelbukh (Ed.)

# Computational Linguistics and Intelligent Text Processing

10th International Conference, CICLing 2009  
Mexico City, Mexico, March 1-7, 2009  
Proceedings

Volume Editor

Alexander Gelbukh  
Center for Computing Research  
National Polytechnic Institute  
Mexico City, 07738, Mexico  
E-mail: [gelbukh@gelbukh.com](mailto:gelbukh@gelbukh.com)

Library of Congress Control Number: 2009921294

CR Subject Classification (1998): H.3, I.2.7, I.7, I.2, F.4.3

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN	0302-9743
ISBN-10	3-642-00381-8 Springer Berlin Heidelberg New York
ISBN-13	978-3-642-00381-3 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

[springer.com](http://springer.com)

© Springer-Verlag Berlin Heidelberg 2009  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper      SPIN: 12625819      06/3180      5 4 3 2 1 0

# Preface

CICLing 2009 marked the 10<sup>th</sup> anniversary of the Annual Conference on Intelligent Text Processing and Computational Linguistics. The CICLing conferences provide a wide-scope forum for the discussion of the art and craft of natural language processing research as well as the best practices in its applications.

This volume contains five invited papers and the regular papers accepted for oral presentation at the conference. The papers accepted for poster presentation were published in a special issue of another journal (see the website for more information). Since 2001, the proceedings of CICLing conferences have been published in Springer's Lecture Notes in Computer Science series, as volumes 2004, 2276, 2588, 2945, 3406, 3878, 4394, and 4919.

This volume has been structured into 12 sections:

- Trends and Opportunities
- Linguistic Knowledge Representation Formalisms
- Corpus Analysis and Lexical Resources
- Extraction of Lexical Knowledge
- Morphology and Parsing
- Semantics
- Word Sense Disambiguation
- Machine Translation and Multilingualism
- Information Extraction and Text Mining
- Information Retrieval and Text Comparison
- Text Summarization
- Applications to the Humanities

A total of 167 papers by 392 authors from 40 countries were submitted for evaluation by the International Program Committee, see Tables 1 and 2. This volume contains revised versions of 44 papers, by 120 authors, selected for oral presentation; the acceptance rate was 26.3%. It also features invited papers by

- Jill Burstein, Educational Testing Service, USA
- Ken Church, Microsoft, USA
- Dekang Lin, Google, USA
- Bernardo Magnini, Fondazione Bruno Kessler, Italy

who presented excellent keynote lectures at the conference. Publication of extended full-text invited papers in the proceedings is a distinctive feature of the CICLing conferences. What is more, in addition to presenting their invited papers, the keynote speakers organized separate vivid informal events; this is also a distinctive feature of this conference series.

The 2009 event was accompanied by a five-day pre-conference Lexicom Americas Workshop 2009, organized by Lexicography MasterClass and led by Adam

**Table 1.** Statistics of submissions and accepted papers by country or region

Country or region	Authors		Papers <sup>1</sup>		Country or region	Authors		Papers <sup>1</sup>	
	Subm	Subm	Accp			Subm	Subm	Accp	
Algeria	4	1.33	–		Italy	14	5.93	2.43	
Argentina	3	1.33	1		Japan	12	4.5	1	
Brazil	14	4.8	–		Jordan	1	1	–	
Canada	6	2.58	1.25		Korea	12	5	–	
China	17	7.25	–		Lithuania	2	2	–	
Colombia	3	0.75	0.75		Macao	4	1	–	
Croatia	5	1	1		Mexico	44	15.38	3.88	
Czech Rep.	9	5	3		Myanmar	2	1	–	
Egypt	2	1	1		Norway	2	1	1	
Estonia	1	1	1		Portugal	4	2	–	
Finland	1	1	1		Romania	8	4	–	
France	11	4.37	–		Russia	2	2	–	
Germany	9	6	2		Spain	60	19.95	6.87	
Greece	3	1.25	1		Sweden	3	3	1	
Hong Kong	7	2.5	–		Switzerland	13	5	1	
Hungary	2	1	–		Tunisia	9	4	–	
India	32	13	2		Turkey	10	4	1	
Iran	5	2	–		UK	9	3.32	1.32	
Ireland	3	1	1		USA	40	21.75	7.5	
Israel	2	1	1		Vietnam	2	2	–	
					Total:	392	167	44	

<sup>1</sup> Counted by authors. E.g., for a paper by 3 authors, 2 from Mexico and 1 from USA, we added  $\frac{2}{3}$  to Mexico and  $\frac{1}{3}$  to USA.

Kilgarriff, Lexicography MasterClass, UK, and Jan Pomikálek, Masaryk University, Czech Republic. The main conference program also included a discussion panel on the future of corpus design, organized by Adam Kilgarriff.

The following papers received the Best Paper Awards and the Best Student Paper Award, correspondingly:

- 1<sup>st</sup> Place: *Cross-Language Frame Semantics Transfer in Bilingual Corpora*, by Roberto Basili, Diego De Cao, Danilo Croce, Bonaventura Coppola, and Alessandro Moschitti (page 332);
- 2<sup>nd</sup> Place: *Detecting Protein-Protein Interactions in Biomedical Texts Using a Parser and Linguistic Resources*, by Gerold Schneider, Kaarel Kaljurand, and Fabio Rinaldi (page 406);
- 3<sup>rd</sup> Place: *Learning to Learn Biological Relations from a Small Training Set*, by Laura Alonso i Alemany and Santiago Bruno (page 418);
- Student: *Enriching Statistical Translation Models Using a Domain-Independent Multilingual Lexical Knowledge Base*, by Miguel García, Jesús Giménez, and Lluís Màrquez (page 306).

**Table 2.** Statistics of submissions and accepted papers by topic<sup>2</sup>

Accepted	Submitted	Topic
13	39	Information extraction
12	31	Text mining
11	31	Clustering and categorization
8	19	Syntax and chunking (linguistics)
8	20	Statistical methods (mathematics)
8	29	Lexical resources
7	22	Information retrieval
7	24	Semantics and discourse
7	29	Formalisms and knowledge representation
7	31	Other
6	18	Word sense disambiguation
5	29	Symbolic and linguistic methods
3	11	Natural language interfaces
2	3	Emotions and humor
2	7	Text generation
2	8	Machine translation
2	9	Morphology
2	10	Summarization
1	1	Textual entailment
1	2	Parsing algorithms (mathematics)
–	2	Spell checking
–	2	POS tagging
–	3	Speech processing
–	3	Anaphora resolution

<sup>2</sup> According to the topics indicated by the authors. A paper may be assigned to more than one topic.

The best student paper was selected from those papers of which the first author was a full-time student, excluding the papers that received a Best Paper Award. The authors of the awarded papers were given extended time for their presentations.

In addition, the Best Presentation Award and the Best Poster Award winners were selected by a ballot among the attendees of the conference.

Besides their high scientific level, one of the success factors of CICLing conferences is their excellent cultural program. CICLing 2009 was held in Mexico, a wonderful country, rich in culture, history, and nature. The participants of the conference had a chance to see the legendary 2000-years-old Teotihuacan pyramids, a monarch butterfly wintering site where the old pines are covered with millions of butterflies as if they were leaves, a great cave with 85-meter halls and a river flowing out of it, Aztec warriors dancing in the street in their colorful plumages, and the largest anthropological museum in the world; see photos at [www.CICLing.org](http://www.CICLing.org).

I would like to thank all those involved in the organization of this conference. In the first place these are the authors of the papers constituting this book: it is the

excellence of their research work that gives value to the book and sense to the work of all other people involved. I thank the Program Committee members for their hard and very professional work. Very special thanks go to Manuel Vilares and his group, Rada Mihalcea, and Ted Pedersen for their invaluable support in the reviewing process.

I express my most cordial thanks to the members of the Local Organizing Committee for their considerable contribution to making this conference become a reality. I thank the Mexican Government for providing financial support, the Mexican Society of Artificial Intelligence for valuable collaboration, and the Center for Computing Research (CIC) of the National Polytechnic Institute (IPN), Mexico, for hosting the conference.

The entire submission, reviewing, and selection process, as well as putting together the proceedings, was supported for free by the EasyChair system ([www.EasyChair.org](http://www.EasyChair.org)); I express my gratitude to its author Andrei Voronkov for his constant support and help. Last but not least, I deeply appreciate the Springer staff's patience and help in editing this volume – it is always a great pleasure to work with them.

January 2009

Alexander Gelbukh

# Organization

CICLing 2009 was organized by the Natural Language and Text Processing Laboratory ([nlp.cic.ipn.mx](http://nlp.cic.ipn.mx)) of the Center for Computing Research (CIC) of the National Polytechnic Institute (IPN), Mexico, in collaboration with the Mexican Society of Artificial Intelligence (SMIA, [www.smia.org.mx](http://www.smia.org.mx)).

## Program Chair

Alexander Gelbukh

## Program Committee

John Atkinson  
Christian Boitet  
Nicoletta Calzolari  
John Carroll  
Kenneth Church  
Dan Cristea  
Walter Daelemans  
Mona Talat Diab  
Oren Etzioni  
Alexander Gelbukh  
Gregory Grefenstette  
Yasunari Harada  
Eva Hajičová  
Graeme Hirst  
Eduard Hovy  
Nancy Ide  
Diana Inkpen  
Frederick Jelinek  
Alma Kharat  
Adam Kilgariff  
Alexander Koller  
Henry Lieberman

Dekang Lin  
Aurelio López  
Igor Mel'čuk  
Rada Mihalcea  
Masaki Murata  
Nicolas Nicolov  
Kemal Oflazer  
Constantin Orasan  
Ted Pedersen  
Viktor Pekar  
Stelios Piperidis  
Fuji Ren  
Fabio Rinaldi  
Horacio Rodríguez  
Vasile Rus  
Franco Salvetti  
Serge Sharoff  
Grigori Sidorov  
Thamar Solorio  
Maosong Sun  
Karin Verspoor  
Manuel Vilares Ferro

## Award Committee

Alexander Gelbukh  
Eduard Hovy  
Rada Mihalcea

Ted Pedersen  
Yorick Wiks



## Additional Referees

Mohamed Abdel Fattah	Joji Maeno
Mikhail Alexandrov	Nobuaki Minematsu
Muath Alzghool	Tomoko Ohkuma
Chris Biemann	Juan Otero Pombo
Maya Carrillo Ruiz	Ryo Otoguro
Victor Darriba	Feng Pan
Georgiana Dinu	Michael Piotrowski
Iustin Dornescu	Natalia Ponomareva
Milagros Fernández Gavilanes	Prokopis Prokopidis
Oana Frunza	Jonathon Read
René Arnulfo García Hernández	Francisco Jose Ribadas Pena
Byron Georgantopoulos	Gerold Schneider
Chikara Hashimoto	Petr Sgall
Laura Hasler	Kiyoaki Shirai
Laritz Hernández Rojas	Takeo Tatsumi
Kaarel Kaljurand	Lorenzo Thione
Fazel Keshtkar	Martin Volk
Rob Koeling	Christopher Walker
Marco Kuhlmann	Zdenek Zabokrtsky
Yulia Ledeneva	Carlos Mario Zapata Jaramillo

## Organizing Committee

Ignacio García Araoz, CIC	Raquel López Alamilla, CIC
Rosalía García, SMIA	Oralia del Carmen Pérez Orozco, CIC
Alexander Gelbukh, CIC & SMIA	Carlos Alberto Reyes García, SMIA
Olga Kolesnikova, CIC	Grigori Sidorov, CIC & SMIA
Yulia Ledeneva, UAEM	Sulema Torres Ramos, CIC & SMIA

## Website and Contact

The website of the CICLing conferences is [www.CICLing.org](http://www.CICLing.org). It contains information on the past CICLing conferences and satellite events, abstracts of all published papers, photos from past CICLing conferences, video recordings of most keynote talks, as well as the information on the forthcoming CICLing conferences. Contact options can be found on the website.

# Table of Contents

## Trends and Opportunities

Has Computational Linguistics Become More Applied? (Invited Paper) .....	1
<i>Kenneth Church</i>	
Opportunities for Natural Language Processing Research in Education (Invited Paper) .....	6
<i>Jill Burstein</i>	

## Linguistic Knowledge Representation Formalisms

Information Structure in a Formal Framework: Unification-Based Combinatory Categorical Grammar .....	28
<i>Maarika Traat</i>	
A Karaka Based Annotation Scheme for English .....	41
<i>Ashwini Vaidya, Samar Husain, Prashanth Mannem, and Dipti Misra Sharma</i>	

## Corpus Analysis and Lexical Resources

Substring Statistics (Invited Paper) .....	53
<i>Kyoji Umemura and Kenneth Church</i>	
Evaluation of the Syntactic Annotation in EPEC, the Reference Corpus for the Processing of Basque .....	72
<i>Larraitx Uria, Ainara Estarrona, Izaskun Aldezabal, Maria Jesús Aranzabe, Arantza Díaz de Ilarraza, and Mikel Iruskietta</i>	
Reducing Noise in Labels and Features for a Real World Dataset: Application of NLP Corpus Annotation Methods .....	86
<i>Rebecca J. Passonneau, Cynthia Rudin, Axinia Radeva, and Zhi An Liu</i>	
Unsupervised Classification of Verb Noun Multi-Word Expression Tokens .....	98
<i>Mona T. Diab and Madhav Krishna</i>	

# Extraction of Lexical Knowledge

Semantic Mapping for Related Term Identification . . . . .	111
<i>Rafael E. Banchs</i>	
An Improved Automatic Term Recognition Method for Spanish . . . . .	125
<i>Alberto Barrón-Cedeño, Gerardo Sierra, Patrick Drouin, and Sophia Ananiadou</i>	
Bootstrapping a Verb Lexicon for Biomedical Information Extraction . . .	137
<i>Giulia Venturi, Simonetta Montemagni, Simone Marchi, Yutaka Sasaki, Paul Thompson, John McNaught, and Sophia Ananiadou</i>	
<i>TermeX</i> : A Tool for Collocation Extraction . . . . .	149
<i>Davor Delač, Zoran Krleža, Jan Šnajder, Bojana Dalbelo Bašić, and Frane Šarić</i>	

# Morphology and Parsing

Guessers for Finite-State Transducer Lexicons . . . . .	158
<i>Krister Lindén</i>	
Combining Language Modeling and Discriminative Classification for Word Segmentation (Invited Paper) . . . . .	170
<i>Dekang Lin</i>	
Formal Grammar for Hispanic Named Entities Analysis . . . . .	183
<i>Gretel Barceló, Eduardo Cendejas, Grigori Sidorov, and Igor A. Bolshakov</i>	
Automatic Extraction of Clause Relationships from a Treebank . . . . .	195
<i>Oldřich Krůza and Vladislav Kuboň</i>	
A General Method for Transforming Standard Parsers into Error-Repair Parsers . . . . .	207
<i>Carlos Gómez-Rodríguez, Miguel A. Alonso, and Manuel Vilares</i>	

# Semantics

Topic-Focus Articulation from the Semantic Point of View . . . . .	220
<i>Marie Duží</i>	
The Value of Weights in Automatically Generated Text Structures . . . . .	233
<i>Dana Dannélls</i>	
AORTE for Recognizing Textual Entailment . . . . .	245
<i>Reda Siblini and Leila Kosseim</i>	

## Word Sense Disambiguation

Semi-supervised Word Sense Disambiguation Using the Web as Corpus .....	256
<i>Rafael Guzmán-Cabrera, Paolo Rosso, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, and David Pinto-Avendaño</i>	
Semi-supervised Clustering for Word Instances and Its Effect on Word Sense Disambiguation .....	266
<i>Kazunari Sugiyama and Manabu Okumura</i>	
Alleviating the Problem of Wrong Coreferences in Web Person Search (Invited Paper) .....	280
<i>Octavian Popescu and Bernardo Magnini</i>	
Improved Unsupervised Name Discrimination with Very Wide Bigrams and Automatic Cluster Stopping .....	294
<i>Ted Pedersen</i>	

## Machine Translation and Multilinguism

Enriching Statistical Translation Models Using a Domain-Independent Multilingual Lexical Knowledge Base (Best Student Paper Award) .....	306
<i>Miguel García, Jesús Giménez, and Lluís Màrquez</i>	
Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation .....	318
<i>John Tinsley, Mary Hearne, and Andy Way</i>	
Cross-Language Frame Semantics Transfer in Bilingual Corpora (Best Paper Award, 1 <sup>st</sup> Place) .....	332
<i>Roberto Basili, Diego De Cao, Danilo Croce, Bonaventura Coppola, and Alessandro Moschitti</i>	
A Parallel Corpus Labeled Using Open and Restricted Domain Ontologies .....	346
<i>Ester Boldrini, Sergio Ferrández, Ruben Izquierdo, David Tomás, and Jose Luis Vicedo</i>	
Language Identification on the Web: Extending the Dictionary Method .....	357
<i>Radim Řehůřek and Milan Kolkus</i>	

## Information Extraction and Text Mining

Business Specific Online Information Extraction from German Websites .....	369
<i>Yeong Su Lee and Michaela Geierhos</i>	

Low-Cost Supervision for Multiple-Source Attribute Extraction . . . . .	382
<i>Joseph Reisinger and Marius Paşca</i>	
An Integrated Architecture for Processing Business Documents in Turkish . . . . .	394
<i>Serif Adali, A. Coskun Sonmez, and Mehmet Gokturk</i>	
Detecting Protein-Protein Interactions in Biomedical Texts Using a Parser and Linguistic Resources (Best Paper Award, 2 <sup>nd</sup> Place) . . . . .	406
<i>Gerold Schneider, Kaarel Kaljurand, and Fabio Rinaldi</i>	
Learning to Learn Biological Relations from a Small Training Set (Best Paper Award, 3 <sup>rd</sup> Place) . . . . .	418
<i>Laura Alonso i Alemany and Santiago Bruno</i>	
Using a Bigram Event Model to Predict Causal Potential . . . . .	430
<i>Brandon Beamer and Roxana Girju</i>	
Semantic-Based Temporal Text-Rule Mining . . . . .	442
<i>Kjetil Nørvåg and Ole Kristian Fivelstad</i>	
Generating Executable Scenarios from Natural Language . . . . .	456
<i>Michal Gordon and David Harel</i>	
Determining the Polarity and Source of Opinions Expressed in Political Debates . . . . .	468
<i>Alexandra Balahur, Zornitsa Kozareva, and Andrés Montoyo</i>	

## Information Retrieval and Text Comparison

Query Translation and Expansion for Searching Normal and OCR-Degraded Arabic Text . . . . .	481
<i>Tarek Elghazaly and Aly Fahmy</i>	
NLP for Shallow Question Answering of Legal Documents Using Graphs . . . . .	498
<i>Alfredo Monroy, Hiram Calvo, and Alexander Gelbukh</i>	
Semantic Clustering for a Functional Text Classification Task . . . . .	509
<i>Thomas Lippincott and Rebecca Passonneau</i>	
Reducing the Plagiarism Detection Search Space on the Basis of the Kullback-Leibler Distance . . . . .	523
<i>Alberto Barrón-Cedeño, Paolo Rosso, and José-Miguel Benedí</i>	
Empirical Paraphrasing of Modern Greek Text in Two Phases: An Application to Steganography . . . . .	535
<i>Katia Lida Kermanidis and Emmanouil Magkos</i>	

BorderFlow: A Local Graph Clustering Algorithm for Natural Language Processing .....	547
<i>Axel-Cyrille Ngonga Ngomo and Frank Schumacher</i>	

Generalized Mongue-Elkan Method for Approximate Text String Comparison .....	559
<i>Sergio Jimenez, Claudia Becerra, Alexander Gelbukh, and Fabio Gonzalez</i>	

## Text Summarization

Estimating Risk of Picking a Sentence for Document Summarization ...	571
<i>Chandan Kumar, Prasad Pingali, and Vasudeva Varma</i>	
The Decomposition of Human-Written Book Summaries .....	582
<i>Hakan Ceylan and Rada Mihalcea</i>	

## Applications to the Humanities

Linguistic Ethnography: Identifying Dominant Word Classes in Text .....	594
<i>Rada Mihalcea and Stephen Pulman</i>	

Author Index .....	603
--------------------	-----