# Preface

CICLing 2014 was the 15th Annual Conference on Intelligent Text Processing and Computational Linguistics. The CICLing conferences provide a wide-scope forum for discussion of the art and craft of natural language processing research as well as the best practices in its applications.

This set of two books contains four invited papers and a selection of regular papers accepted for presentation at the conference. Since 2001, the proceedings of the CICLing conferences have been published in Springer's *Lecture Notes in Computer Science* series as volume numbers 2004, 2276, 2588, 2945, 3406, 3878, 4394, 4919, 5449, 6008, 6608, 6609, 7181, 7182, 7816, and 7817.

The set has been structured into 17 sections representative of the current trends in research and applications of Natural Language Processing:

- Lexical Resources
- Document Representation
- Morphology, POS-tagging, and Named Entity Recognition
- Syntax and Parsing
- Anaphora resolution
- Recognizing Textual Entailment
- Semantics and Discourse
- Natural Language Generation
- Sentiment Analysis and Emotion Recognition
- Opinion Mining and Social Networks
- Machine Translation and Multilingualism
- Information Retrieval
- Text Classification and Clustering
- Plagiarism Detection
- Style and Spelling Checking
- Speech Processing
- Applications

The 2014 event received submissions from 57 countries, a record high number in the 15-year history of the CICLing series. Exactly 300 papers (third highest number in the history of CICLing) by 635 authors were submitted for evaluation by the International Program Committee; see Figure 1 and Tables 1 and 2. This two-volume set contains revised versions of 85 regular papers selected for presentation; thus the acceptance rate for this set was 28.3%.

In addition to regular papers, the books feature invited papers by

- Jerry Hobbs, ISI, USA
- Bing Liu, University of Illinois, USA
- Suresh Manandhar, University of York, UK
- Johanna D. Moore, University of Edinburgh, UK

**Table 1.** Number of submissions and accepted papers by topic[1]

| Accepted | Submitted | % accepted | Topic |
|---|---|---|---|
| 19 | 45 | 42 | Semantics, pragmatics, discourse |
| 14 | 43 | 33 | Lexical resources |
| 12 | 31 | 39 | Machine translation & multilingualism |
| 12 | 33 | 36 | Practical applications |
| 12 | 35 | 34 | Emotions, sentiment analysis, opinion mining |
| 12 | 39 | 31 | Clustering and categorization |
| 12 | 56 | 21 | Text mining |
| 11 | 47 | 23 | Information retrieval |
| 10 | 29 | 34 | Underresourced languages |
| 8 | 26 | 31 | Syntax and chunking |
| 7 | 44 | 16 | Information extraction |
| 6 | 18 | 33 | Social networks and microblogging |
| 5 | 16 | 31 | Natural language generation |
| 4 | 11 | 36 | Noisy text processing and cleaning |
| 4 | 16 | 25 | Summarization |
| 3 | 4 | 75 | Spelling and grammar checking |
| 3 | 9 | 33 | Plagiarism detection |
| 3 | 12 | 25 | Word sense disambiguation |
| 3 | 16 | 19 | POS tagging |
| 2 | 5 | 40 | Coreference resolution |
| 2 | 7 | 29 | Computational terminology |
| 2 | 7 | 29 | Other |
| 2 | 9 | 22 | Textual entailment |
| 2 | 13 | 15 | Formalisms and knowledge representation |
| 2 | 17 | 12 | Named entity recognition |
| 2 | 19 | 11 | Morphology |
| 1 | 6 | 17 | Speech processing |
| 1 | 10 | 10 | Natural language interfaces |
| 1 | 11 | 9 | Question answering |
| 0 | 3 | 0 | Computational humor |

[1] As indicated by the authors. A paper may belong to several topics.

who presented excellent keynote lectures at the conference. Publication of full-text invited papers in the proceedings is a distinctive feature of the CICLing conferences. Furthermore, in addition to presentation of their invited papers, the keynote speakers organized separate vivid informal events; this is also a distinctive feature of this conference series. In addition, Professor Jens Allwood of the University of Gothenburg was a special guest of the conference.

With this event we continued with our policy of giving preference to papers with verifiable and reproducible results: in addition to the verbal description of their findings given in the paper, we encouraged the authors to provide a proof of their claims in electronic form. If the paper claimed experimental results, we asked the authors to make available to the community all the input data necessary to verify and reproduce these results; if it claimed to introduce an

**Table 2.** Number of submitted and accepted papers by country or region

| Country or region | Authors | Papers[2] | | Country or region | Authors | Papers[2] | |
|---|---|---|---|---|---|---|---|
| | Subm. | Subm. | Accp. | | Subm. | Subm. | Accp. |
| Afghanistan | 1 | 1 | – | Japan | 22 | 8.33 | 3 |
| Algeria | 2 | 0.67 | – | Jordan | 12 | 3.33 | – |
| Australia | 8 | 3 | 1 | Kazakhstan | 6 | 1.67 | 1.67 |
| Bangladesh | 9 | 3 | – | Korea (South) | 12 | 3.5 | 0.50 |
| Belgium | 3 | 2 | – | Latvia | 6 | 2 | 1 |
| Brazil | 18 | 6.17 | 2.17 | Malaysia | 4 | 1.67 | – |
| Bulgaria | 1 | 1 | – | Mexico | 19 | 12.42 | 2.67 |
| Canada | 13 | 7 | 4 | Mongolia | 1 | 0.5 | 0.5 |
| China | 57 | 21.1 | 7.35 | Morocco | 4 | 2 | – |
| Christmas Isl. | 1 | 0.2 | 0.2 | Nepal | 11 | 5 | 2 |
| Colombia | 3 | 1 | 1 | Norway | 1 | 0.2 | – |
| Croatia | 1 | 0.33 | 0.33 | Pakistan | 4 | 1.83 | – |
| Czech Rep. | 20 | 11.4 | 3 | Poland | 2 | 2 | – |
| Denmark | 3 | 0.38 | – | Portugal | 5 | 2.5 | 1 |
| Egypt | 12 | 7 | 1 | Romania | 10 | 5.67 | – |
| Ethiopia | 5 | 4 | 2 | Russia | 9 | 5.17 | – |
| Finland | 5 | 2 | 2 | Singapore | 9 | 2.78 | 1.78 |
| France | 29 | 12.42 | 9.67 | Slovenia | 2 | 0.67 | 0.67 |
| Germany | 19 | 7.33 | 4.33 | Spain | 13 | 3.7 | 0.67 |
| Greece | 1 | 0.33 | 0.33 | Sweden | 5 | 4 | 1 |
| Hong Kong | 4 | 2 | 1 | Switzerland | 6 | 5 | 2 |
| Hungary | 3 | 1 | – | Taiwan | 5 | 1 | – |
| India | 134 | 73.1 | 10.33 | Thailand | 2 | 1 | – |
| Indonesia | 3 | 1 | – | Tunisia | 20 | 8.83 | 1.83 |
| Iran | 4 | 2 | – | Turkey | 3 | 2.83 | 1.5 |
| Iraq | 0 | 1 | – | UK | 10 | 3.83 | 3.33 |
| Ireland | 0 | 0.5 | – | USA | 48 | 21.48 | 7.17 |
| Israel | 14 | 7 | 2 | ~~Viet Nam~~ | 5 | 1.67 | – |
| Italy | 6 | 2.5 | 1 | *Total:* | 635 | 296 | 85 |

[2] By the number of authors: e.g., a paper by two authors from the USA and one from UK is counted as 0.67 for the USA and 0.33 for UK.

algorithm, we encourage the authors to make the algorithm itself, in a programming language, available to the public. This additional electronic material will be permanently stored on the CICLing's server, www.CICLing.org, and will be available to the readers of the corresponding paper for download under a license that permits its free use for research purposes.

In a long run we expect that computational linguistics will have verifiability and clarity standards similar to those of mathematics: in mathematics, each claim is accompanied by a complete and verifiable proof (usually much longer than the claim itself); each theorem's complete and precise proof—and not just a vague description of its general idea—is made available to the reader. Electronic media allow computational linguists to provide material analogous to the proofs

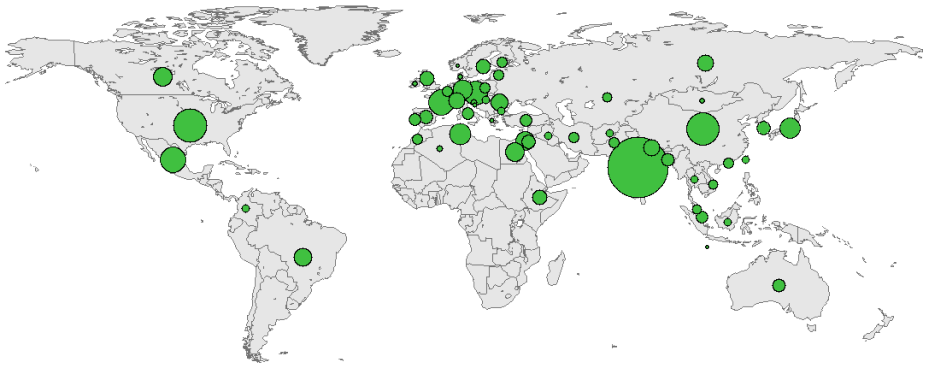

**Fig. 1.** Submissions by country or region. The area of a circle represents the number of submitted papers.

and formulas in mathematic in full length—which can amount to megabytes or gigabytes of data—separately from a 12-page description published in the book. More information can be found on www.CICLing.org/why_verify.htm.

To encourage providing algorithms and data along with the published papers, we selected a winner of our Verifiability, Reproducibility, and Working Description Award. The main factors in choosing the awarded submission were technical correctness and completeness, readability of the code and documentation, simplicity of installation and use, and exact correspondence to the claims of the paper. Unnecessary sophistication of the user interface was discouraged; novelty and usefulness of the results were not evaluated—instead, they were evaluated for the paper itself and not for the data.

The following papers received the Best Paper Awards, the Best Student Paper Award,[1] as well as the Verifiability, Reproducibility, and Working Description Award, ~~correspondingly~~:

1st Place: *A graph-based automatic plagiarism detection technique to handle artificial word reordering and paraphrasing*, by Niraj Kumar, India;

2nd Place: *Dealing with function words in unsupervised dependency parsing*, by David Mareček, Zdeněk Žabokrtský, Czech Republic;

3rd Place: *Extended CFG formalism for grammar checker and parser development*, by Daiga Deksne, Inguna Skadiņa, Raivis Skadiņš, Latvia;

and *How preprocessing affects unsupervised keyphrase extraction*, by Rui Wang, Wei Liu, Chris McDonald, Australia;

Student: *Iterative bilingual lexicon extraction from comparable corpora with topical and contextual knowledge*, by Chenhui Chu, Toshiaki Nakazawa, Sadao Kurohashi, Japan;

Verifiability: How document properties affect document relatedness measures, by Jessica Perrie, Aminul Islam, Evangelos Milios, Canada.

[1] The best student paper was selected among 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 (except for the Verifiability award) 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 its high scientific level, one of the success factors of CICLing conferences is their excellent cultural program. The attendees of the conference had a chance to visit wonderful historical and cultural attraction of ~~so little known country as~~ Nepal—the birthplace of the Buddha and the place where pagodas were invented before ~~they~~ spread to China and Japan to become an iconic image of East Asia. Of world's ~~10~~ highest mountains, ~~8~~ are in Nepal, including the ~~Earth's~~ highest ~~mountain,~~ Everest; the participants had a chance to see Everest during a~~n~~ tour ~~by~~ Himalayas on a small airplane. They also ~~saw~~ the Seto MachindraNath Chariot festival and visited three historical Durbar squares of the Kathmandu valley, UNESCO world cultural heritage. But probably the best of Nepal, after Himalayas, are its buddhist and hindu temples and monasteries, of which the participants ~~have~~ visited quite a few. Even the Organizing Committee Secretary and author of one of the best evaluated papers published in this ~~book~~ set was the hereditary Supreme Priest of an ancient Buddhist temple!

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 that constitute 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~~. I thank all those who served on the Program Committee, Software Reviewing Committee, Award Selection Committee, as well as additional reviewers, for their hard and very professional work. Special thanks go to Pushpak Bhattacharyya, Samhaa El-Beltagy, Aminul Islam, Cerstin Mahlow, Dunja Mladenic, Constantin Orasan, and Grigori Sidorov for their invaluable support in the reviewing process.

I would like to thank the conference staff, volunteers, and the members of the local ~~organization committee~~ headed by Professor Madhav Prasad Pokharel and advised by Professor Jai Raj Awasthi. In particular, I am very grateful to Mr. Sagun Dhakhwa, the secretary of the Organizing Committee, for his great effort in planning all the aspects of the conference. I want to thank Ms. Sahara Mishra for administrative support and Mr. Sushan Shrestha for ~~web page~~ design and technical support. I am deeply grateful to the administration of the Centre ~~F~~for Communication ~~&~~ Development Studies (CECODES) for their helpful support, warm hospitality, and in general for providing this wonderful opportunity ~~of holding~~ CICLing in Nepal. I acknowledge support from the project CONACYT Mexico--DST India 122030 "Answer Validation through Textual Entailment" and SIP-IPN grant 20144534.

The entire submission and reviewing process was supported for free by the EasyChair system (www.EasyChair.org). Last but not least, I deeply appreciate the ~~Springer staff's~~ patience and help in editing these volumes and getting them printed in very short time—it is always a great pleasure to work with Springer.

February 2014 Alexander Gelbukh

# Organization

CICLing 2014 was hosted by the Centre For Communication & Development Studies (CECODES), Nepal and was organized by the CICLing 2014 Organizing Committee in conjunction with the CECODES, the Natural Language and Text Processing Laboratory of the CIC (Centro de Investigación en Computación) of the IPN (Instituto Politécnico Nacional), Mexico, and Mexican Society of Artificial Intelligence (SMIA).

## Organizing Chair

Professor Madhav Prasad Pokharel

## Organizing Committee

Professor Madhav Prasad Pokharel (chair), Tribhuban University, Kathmandu
Professor Jai Raj Awasthi (adviser), CECODES, Lalitpur
Mr. Sagun Dhakhwa (secretary), CECODES, Lalitpur
Mr. Bhim Narayan Regmi, CECODES, Lalitpur
Mr. Krishna Prasad Parajuli, CECODES, Lalitpur
Mr. Sandeep Khatri, CECODES, Lalitpur
Mr. Kamal Poudel, CECODES, Lalitpur
Mr. Bhim Lal Gautam, CECODES, Lalitpur
Mr. Krishna Prasad Chalise, CECODES, Lalitpur
Mr. Dipesh Joshi, CECODES, Lalitpur
Dr. Prajol Shrestha, NLP Engineer, Vision Objects

## Program Chair

Alexander Gelbukh

## Program Committee

Ajith Abraham
Rania Al-Sabbagh
Sophia Ananiadou
Marianna Apidianaki
Alexandra Balahur
Kalika Bali
Leslie Barrett
Roberto Basili
Pushpak Bhattacharyya
Nicoletta Calzolari
Nick Campbell
Sandra Carberry
Michael Carl
Hsin-Hsi Chen
Dan Cristea
Bruce Croft
Mike Dillinger
Samhaa El-Beltagy

Tomaž Erjavec
Anna Feldman
Alexander Gelbukh
Dafydd Gibbon
Gregory Grefenstette
Eva Hajicova
Sanda Harabagiu
Yasunari Harada
Karin Harbusch
Ales Horak
Veronique Hoste
Nancy Ide
Diana Inkpen
Hitoshi Isahara
Aminul Islam
Guillaume Jacquet
Sylvain Kahane
Alma Kharrat
Adam Kilgarriff
Valia Kordoni
Leila Kosseim
Mathieu Lafourcade
Krister Lindén
Bing Liu
Elena Lloret
Bernardo Magnini
Cerstin Mahlow
Suresh Manandhar
Diana Mccarthy
Alexander Mehler
Rada Mihalcea
Evangelos Milios
Dunja Mladenic
Marie-Francine Moens
Masaki Murata
Preslav Nakov
Costanza Navarretta
Roberto Navigli
Vincent Ng
Joakim Nivre
Attila Novák
Kjetil Nørvåg
Kemal Oflazer
Constantin Orasan
Ekaterina Ovchinnikova
Ivandre Paraboni
Saint-Dizier Patrick
Maria Teresa Pazienza
Ted Pedersen
Viktor Pekar
Anselmo Peñas
Octavian Popescu
Marta R. Costa-Jussà
German Rigau
Horacio Rodriguez
Paolo Rosso
Vasile Rus
Kepa Sarasola
Roser Sauri
Hassan Sawaf
Satoshi Sekine
Serge Sharoff
Grigori Sidorov
Kiril Simov
Vivek Kumar Singh
Vaclav Snasel
Thamar Solorio
Efstathios Stamatatos
Carlo Strapparava
Tomek Strzalkowski
Maosong Sun
Stan Szpakowicz
Mike Thelwall
Jörg Tiedemann
Christoph Tillmann
George Tsatsaronis
Dan Tufis
Olga Uryupina
Karin Verspoor
Manuel Vilares Ferro
Aline Villavicencio
Piotr W. Fuglewicz
Savas Yildirim

## Software Reviewing Committee

Ted Pedersen
Florian Holz
Miloš Jakubíček
Sergio Jiménez Vargas
Miikka Silfverberg
Ronald Winnemöller

## Award Committee

Alexander Gelbukh
Eduard Hovy
Rada Mihalcea
Ted Pedersen
Yorick Wiks

## Additional ~~Referees~~

Mahmoud Abunasser
Naveed Afzal
Iñaki Alegria
Hanna Bechara
Houda Bouamor
Janez Brank
Chen Chen
Víctor Darriba
Owen Davison
Ismaïl El Maarouf
Mahmoud El-Haj
Milagros Fernández-Gavilanes
Daniel Fernández-González
Corina Forascu
Kata Gabor
Mercedes García Martínez
Diman Ghazi
Rohit Gupta
Francisco Javier Guzman
Kazi Saidul Hasan
Radu Ion
Zahurul Islam
Milos Jakubicek
Antonio Jimeno
Olga Kolesnikova
Mohammed Korayem
Tobias Kuhn
Majid Laali
Yulia Ledeneva
Andy Lücking
Tokunbo Makanju
Raheleh Makki
Akshay Minocha
Abidalrahman Moh'D
Zuzana Neverilova
An Ngoc Vo
Mohamed Outahajala
Michael Piotrowski
Soujanya Poria
Francisco Rangel
Amir Hossein Razavi
Francisco Ribadas-Pena
Alvaro Rodrigo
Armin Sajadi
Paulo Schreiner
Djamé Seddah
Karan Singla
Vit Suchomel
Aniruddha Tammewar
Yasushi Tsubota
Francisco José Valverde Albacete
Kassius Vargas Prestes
Tim Vor der Brück
Tadej Štajner

## Website and Contact

The webpage of the CICLing conference series is www.CICLing.org. It contains information about past CICLing conferences and their satellite events, includ-

ing links to published papers (many of them in open access) or their abstracts, photos, and video recordings of keynote talks. In addition, it contains data, algorithms, and open-source software accompanying accepted papers, ~~in~~ according to the CICLing verifiability, reproducibility, and working description policy. It also contains information about the forthcoming CICLing events, as well as contact options.