Lecture Notes in Computer Science

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
  TU Dortmund University, Germany
Demetri Terzopoulos
  University of California, Los Angeles, CA, USA
Doug Tygar
  University of California, Berkeley, CA, USA
Gerhard Weikum
  Max Planck Institute for Informatics, Saarbruecken, Germany
CICLing 2014 was the 15\textsuperscript{th} 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 \textit{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 639 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

<table>
<thead>
<tr>
<th>Accepted</th>
<th>Submitted</th>
<th>% Accepted</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>45</td>
<td>42</td>
<td>Semantics, pragmatics, discourse</td>
</tr>
<tr>
<td>14</td>
<td>43</td>
<td>33</td>
<td>Lexical resources</td>
</tr>
<tr>
<td>12</td>
<td>31</td>
<td>39</td>
<td>Machine translation and multilingualism</td>
</tr>
<tr>
<td>12</td>
<td>33</td>
<td>36</td>
<td>Practical applications</td>
</tr>
<tr>
<td>12</td>
<td>35</td>
<td>34</td>
<td>Emotions, sentiment analysis, opinion mining</td>
</tr>
<tr>
<td>12</td>
<td>40</td>
<td>30</td>
<td>Clustering and categorization</td>
</tr>
<tr>
<td>12</td>
<td>56</td>
<td>21</td>
<td>Text mining</td>
</tr>
<tr>
<td>11</td>
<td>48</td>
<td>23</td>
<td>Information retrieval</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
<td>34</td>
<td>Underresourced languages</td>
</tr>
<tr>
<td>8</td>
<td>26</td>
<td>31</td>
<td>Syntax and chunking</td>
</tr>
<tr>
<td>7</td>
<td>44</td>
<td>16</td>
<td>Information extraction</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>33</td>
<td>Social networks and microblogging</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>31</td>
<td>Natural language generation</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>36</td>
<td>Noisy text processing and cleaning</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>25</td>
<td>Summarization</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>75</td>
<td>Spelling and grammar checking</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>33</td>
<td>Plagiarism detection</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>25</td>
<td>Word sense disambiguation</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>19</td>
<td>POS tagging</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>40</td>
<td>Coreference resolution</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>29</td>
<td>Computational terminology</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>29</td>
<td>Other</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>22</td>
<td>Textual entailment</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>15</td>
<td>Formalisms and knowledge representation</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>12</td>
<td>Named entity recognition</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>10</td>
<td>Morphology</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>17</td>
<td>Speech processing</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>Natural language interfaces</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>9</td>
<td>Question answering</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>Computational humor</td>
</tr>
</tbody>
</table>

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

These speakers 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
### Table 2. Number of submitted and accepted papers by country or region

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>1</td>
<td>1</td>
<td>–</td>
<td>Japan</td>
<td>22</td>
<td>8.33</td>
</tr>
<tr>
<td>Algeria</td>
<td>2</td>
<td>0.67</td>
<td>–</td>
<td>Jordan</td>
<td>12</td>
<td>3.33</td>
</tr>
<tr>
<td>Australia</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>Kazakhstan</td>
<td>6</td>
<td>1.67</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>9</td>
<td>3</td>
<td>–</td>
<td>Korea (South)</td>
<td>12</td>
<td>3.5</td>
</tr>
<tr>
<td>Belgium</td>
<td>3</td>
<td>2</td>
<td>–</td>
<td>Latvia</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Brazil</td>
<td>18</td>
<td>6.17</td>
<td>2.17</td>
<td>Malaysia</td>
<td>4</td>
<td>1.67</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>1</td>
<td>1</td>
<td>–</td>
<td>Mexico</td>
<td>19</td>
<td>12.42</td>
</tr>
<tr>
<td>Canada</td>
<td>13</td>
<td>7</td>
<td>4</td>
<td>Mongolia</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>China</td>
<td>57</td>
<td>21.1</td>
<td>7.35</td>
<td>Morocco</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Christmas Isl.</td>
<td>1</td>
<td>0.2</td>
<td>0.2</td>
<td>Nepal</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Colombia</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>Norway</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>Croatia</td>
<td>1</td>
<td>0.33</td>
<td>0.33</td>
<td>Pakistan</td>
<td>4</td>
<td>1.83</td>
</tr>
<tr>
<td>Czech Rep.</td>
<td>20</td>
<td>11.4</td>
<td>3</td>
<td>Poland</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Denmark</td>
<td>3</td>
<td>0.38</td>
<td>–</td>
<td>Portugal</td>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>Egypt</td>
<td>12</td>
<td>7</td>
<td>1</td>
<td>Romania</td>
<td>10</td>
<td>5.67</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>Russia</td>
<td>9</td>
<td>5.17</td>
</tr>
<tr>
<td>Finland</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>Singapore</td>
<td>9</td>
<td>2.78</td>
</tr>
<tr>
<td>France</td>
<td>29</td>
<td>12.42</td>
<td>9.67</td>
<td>Slovenia</td>
<td>2</td>
<td>0.67</td>
</tr>
<tr>
<td>Germany</td>
<td>19</td>
<td>7.33</td>
<td>4.33</td>
<td>Spain</td>
<td>13</td>
<td>3.7</td>
</tr>
<tr>
<td>Greece</td>
<td>1</td>
<td>0.33</td>
<td>0.33</td>
<td>Sweden</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>Switzerland</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Hungary</td>
<td>3</td>
<td>1</td>
<td>–</td>
<td>Taiwan</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>India</td>
<td>136</td>
<td>75.1</td>
<td>10.33</td>
<td>Thailand</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Indonesia</td>
<td>3</td>
<td>1</td>
<td>–</td>
<td>Tunisia</td>
<td>20</td>
<td>8.83</td>
</tr>
<tr>
<td>Iran</td>
<td>4</td>
<td>2</td>
<td>–</td>
<td>Turkey</td>
<td>3</td>
<td>2.83</td>
</tr>
<tr>
<td>Iraq</td>
<td>0</td>
<td>1</td>
<td>–</td>
<td>UK</td>
<td>10</td>
<td>3.83</td>
</tr>
<tr>
<td>Ireland</td>
<td>0</td>
<td>0.5</td>
<td>–</td>
<td>USA</td>
<td>48</td>
<td>21.48</td>
</tr>
<tr>
<td>Israel</td>
<td>14</td>
<td>7</td>
<td>2</td>
<td>Vietnam</td>
<td>5</td>
<td>1.67</td>
</tr>
<tr>
<td>Italy</td>
<td>6</td>
<td>2.5</td>
<td>1</td>
<td><strong>Total:</strong></td>
<td>639</td>
<td>300</td>
</tr>
</tbody>
</table>

$^2$ 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.

an algorithm, we encouraged 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 the 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
Fig. 1. Submissions by country or region. The area of a circle represents the number of submitted papers.

media allow computational linguists to provide material analogous to the proofs 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, as well as the Verifiability, Reproducibility, and Working Description Award, respectively:

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

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

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 the wonderful historical and cultural attractions of the lesser-known country Nepal—the birthplace of the Buddha and the place where pagodas were invented before their spread to China and Japan to become an iconic image of East Asia. Of the world’s ten highest mountains, eight are in Nepal, including the highest one Everest; the participants had a chance to see Everest during a tour of the Himalayas on a small airplane. They also attended the Seto MachindraNath Chariot festival and visited three historical Durbar squares of the Kathmandu valley, a UNESCO world cultural heritage site. But probably the best of Nepal, after the Himalayas, are its buddhist and hindu temples and monasteries, of which the participants visited quite a few. Even the Organizing Committee secretary and author of one of the best evaluated papers published in this 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. Firstly, 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 the rest. I thank all those who served on the Program Committee, Software Reviewing Committee, Award Selection Committee, as well as the 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 Organizing 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 the website design and technical support. I am deeply grateful to the administration of the Centre for Communication and Development Studies (CECODES) for their helpful support, warm hospitality, and in general for providing this wonderful opportunity to hold 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 patience and help of Springer staff 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 and 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 the Mexican Society of Artificial Intelligence (SMIA).

Organizing Chair

Madhav Prasad Pokharel

Organizing Committee

Madhav Prasad Pokharel  
(Chair) Tribhuvan University, Kathmandu
Jai Raj Awasthi (Advisor) CECODES, Lalitpur
Sagun Dhakhwa (Secretary) CECODES, Lalitpur
Bhim Narayan Regmi CECODES, Lalitpur
Krishna Prasad Parajuli CECODES, Lalitpur
Sandeep Khatri CECODES, Lalitpur
Kamal Poudel CECODES, Lalitpur
Bhim Lal Gautam CECODES, Lalitpur
Krishna Prasad Chalise CECODES, Lalitpur
Dipesh Joshi CECODES, Lalitpur
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
Dunjna 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 Reviewers

Mahmoud Abunasser
Naveed Afzal
Inaki 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
Miros 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
Abirruthda 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, including 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, according to the CICLing verifiability, reproducibility, and working description policy. It also contains information about the forthcoming CICLing events, as well as contact options.
## Table of Contents – Part I

### Lexical Resources

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using Word Association Norms to Measure Corpus Representativeness</td>
<td>1</td>
</tr>
<tr>
<td>Reinhard Rapp</td>
<td></td>
</tr>
<tr>
<td>Optimality Theory as a Framework for Lexical Acquisition</td>
<td>14</td>
</tr>
<tr>
<td>Thierry Poibeau</td>
<td></td>
</tr>
<tr>
<td>Verb Clustering for Brazilian Portuguese</td>
<td>25</td>
</tr>
<tr>
<td>Carolina Scarton, Lin Sun, Karin Kipper-Schuler, Magali Sanches Duran,</td>
<td></td>
</tr>
<tr>
<td>Martha Palmer, and Anna Korhonen</td>
<td></td>
</tr>
<tr>
<td>Spreading Relation Annotations in a Lexical Semantic Network Applied</td>
<td>40</td>
</tr>
<tr>
<td>to Radiology</td>
<td></td>
</tr>
<tr>
<td>Lionel Ramadier, Manel Zarrouk, Mathieu Lafourcade, and Antoine Micheau</td>
<td></td>
</tr>
<tr>
<td>Issues in Encoding the Writing of Nepal’s Languages</td>
<td>52</td>
</tr>
<tr>
<td>Pat Hall, Bal Krishna Bal, Sagun Dhakhwa, and Bhim Narayan Regmi</td>
<td></td>
</tr>
<tr>
<td>Compound Terms and Their Multi-word Variants: Case of German and</td>
<td>68</td>
</tr>
<tr>
<td>Russian Languages</td>
<td></td>
</tr>
<tr>
<td>Elizaveta Clouet and Béatrice Daille</td>
<td></td>
</tr>
<tr>
<td>A Fully Automated Approach for Arabic Slang Lexicon Extraction</td>
<td>79</td>
</tr>
<tr>
<td>from Microblogs</td>
<td></td>
</tr>
<tr>
<td>Hady ElSahar and Samhaa R. El-Beltagy</td>
<td></td>
</tr>
<tr>
<td>Simple TF-IDF Is Not the Best You Can Get for Regionalism Classification</td>
<td>92</td>
</tr>
<tr>
<td>Hiram Calvo</td>
<td></td>
</tr>
<tr>
<td>Improved Text Extraction from PDF Documents for Large-Scale Natural</td>
<td>102</td>
</tr>
<tr>
<td>Language Processing</td>
<td></td>
</tr>
<tr>
<td>Jörg Tiedemann</td>
<td></td>
</tr>
</tbody>
</table>

### Document Representation

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependency-Based Semantic Parsing for Concept-Level Text Analysis</td>
<td>113</td>
</tr>
<tr>
<td>Soujanya Poria, Basant Agarwal, Alexander Gelbukh, Amir Hussain, and</td>
<td></td>
</tr>
<tr>
<td>Newton Howard</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Obtaining Better Word Representations via Language Transfer</td>
<td>128</td>
</tr>
<tr>
<td>Changliang Li, Bo Xu, Gaowei Wu, Xiuying Wang, Wendong Ge, and Yan Li</td>
<td></td>
</tr>
<tr>
<td>Exploring Applications of Representation Learning in Nepali</td>
<td>138</td>
</tr>
<tr>
<td>Anjan Nepal and Alexander Yates</td>
<td></td>
</tr>
<tr>
<td>Topic Models Incorporating Statistical Word Senses</td>
<td>151</td>
</tr>
<tr>
<td>Guoyu Tang, Yunqing Xia, Jun Sun, Min Zhang, and Thomas Fang Zheng</td>
<td></td>
</tr>
<tr>
<td>How Preprocessing Affects Unsupervised Keyphrase Extraction</td>
<td>163</td>
</tr>
<tr>
<td>(Best Paper Award, Third Place)</td>
<td></td>
</tr>
<tr>
<td>Rui Wang, Wei Liu, and Chris McDonald</td>
<td></td>
</tr>
<tr>
<td><strong>Morphology, POS-tagging, and Named Entity Recognition</strong></td>
<td></td>
</tr>
<tr>
<td>Methods and Algorithms for Unsupervised Learning of Morphology</td>
<td>177</td>
</tr>
<tr>
<td>(Invited Paper)</td>
<td></td>
</tr>
<tr>
<td>Burcu Can and Suresh Manandhar</td>
<td></td>
</tr>
<tr>
<td>Morphological Analysis of the Bishnupriya Manipuri Language Using</td>
<td>206</td>
</tr>
<tr>
<td>Finite State Transducers</td>
<td></td>
</tr>
<tr>
<td>Nayan Jyoti Kalita, Navanath Saharia, and Smriti Kumar Sinha</td>
<td></td>
</tr>
<tr>
<td>A Hybrid Approach to the Development of Part-of-Speech Tagger</td>
<td>214</td>
</tr>
<tr>
<td>for Kafi-noonoo Text</td>
<td></td>
</tr>
<tr>
<td>Zelalem Mekuria and Yaregal Assabie</td>
<td></td>
</tr>
<tr>
<td>Modified Differential Evolution for Biochemical Name Recognizer</td>
<td>225</td>
</tr>
<tr>
<td>Utpal Kumar Sikdar, Asif Ekbal, and Sriparna Saha</td>
<td></td>
</tr>
<tr>
<td><strong>Syntax and Parsing</strong></td>
<td>237</td>
</tr>
<tr>
<td>Extended CFG Formalism for Grammar Checker and Parser</td>
<td></td>
</tr>
<tr>
<td>Development (Best Paper Award, Third Place)</td>
<td></td>
</tr>
<tr>
<td>Daiga Deksne, Inguna Skadiņa, and Raivis Skadiņš</td>
<td></td>
</tr>
<tr>
<td>Dealing with Function Words in Unsupervised Dependency Parsing</td>
<td>250</td>
</tr>
<tr>
<td>(Best Paper Award, Second Place)</td>
<td></td>
</tr>
<tr>
<td>David Mareček and Zdeněk Žabokrtský</td>
<td></td>
</tr>
<tr>
<td>When Rules Meet Bigrams</td>
<td>262</td>
</tr>
<tr>
<td>Eric Wehrli and Luka Nerima</td>
<td></td>
</tr>
<tr>
<td>Methodology for Connecting Nouns to Their Modifying Adjectives</td>
<td>271</td>
</tr>
<tr>
<td>Nir Ofek, Lior Rokach, and Prasenjit Mitra</td>
<td></td>
</tr>
</tbody>
</table>
Constituency Parsing of Complex Noun Sequences in Hindi............ 285
Arpita Batra, Soma Paul, and Amba Kulkarni

Amharic Sentence Parsing Using Base Phrase Chunking ............ 297
Abeba Ibrahim and Yaregal Assabie

**Anaphora Resolution**

A Machine Learning Approach to Pronominal Anaphora Resolution
in Dialogue Based Intelligent Tutoring Systems ....................... 307
Nobal B. Niraula and Vasile Rus

A Maximum Entropy Based Honorificity Identification for Bengali
Pronominal Anaphora Resolution .................................... 319
Apurbalal Senapati and Utpal Garain

**Recognizing Textual Entailment**

Statistical Relational Learning to Recognise Textual Entailment ...... 330
Miguel Ríos, Lucia Specia, Alexander Gelbukh, and Ruslan Mitkov

Annotation Game for Textual Entailment Evaluation .................. 340
Zuzana Nevšílová

**Semantics and Discourse**

Axiomatizing Complex Concepts from Fundamentals (Invited Paper) ... 351
Jerry R. Hobbs and Andrew Gordon

A Semantics Oriented Grammar for Chinese Treebanking ............... 366
Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu

Unsupervised Interpretation of Eventive Propositions ................. 379
Anselmo Peñas, Bernardo Cabaleiro, and Mirella Lapata

Sense-Specific Implicative Commitments ............................ 391
Gerard de Melo and Valeria de Paiva

A Tiered Approach to the Recognition of Metaphor .................... 403
David B. Bracewell, Marc T. Tomlinson, Michael Mohler, and Bryan Rink

Knowledge Discovery with CRF-Based Clustering of Named Entities
without a Priori Classes .................................................. 415
Vincent Claveau and Abir Ncibi

Semi-supervised SRL System with Bayesian Inference............... 429
Alejandra Lorenzo and Christophe Cerisara
A Sentence Similarity Method Based on Chunking and Information Content .......................................................... 442  
Dan Ștefănescu, Rajendra Banjade, and Vasile Rus

An Investigation on the Influence of Genres and Textual Organisation on the Use of Discourse Relations .......................................................... 454  
Félix-Hervé Bachand, Elnaz Davoodi, and Leila Kosseim

Discourse Tagging for Indian Languages .................................................. 469  
Sobha Lalitha Devi, S. Lakshmi, and Sindhuja Gopalan

Natural Language Generation

Classification-Based Referring Expression Generation ............... 481  
Thiago Castro Ferreira and Ivandré Paraboni

Generating Relational Descriptions Involving Mutual Disambiguation ... 492  
Caio V.M. Teixeira, Ivandré Paraboni, Adriano S.R. da Silva, and Alan K. Yamasaki

Bayesian Inverse Reinforcement Learning for Modeling Conversational Agents in a Virtual Environment .................................................. 503  
Lina M. Rojas-Barahona and Christophe Cerisara

Learning to Summarize Time Series Data .............................. 515  
Pranay Kumar Venkata Sowdaboina, Sutanu Chakraborti, and Somayajulu Sripada

Author Index .................................................. 529
## Table of Contents – Part II

### Sentiment Analysis and Emotion Recognition

**Sentence-Level Sentiment Analysis in the Presence of Modalities**  
(Invited Paper) ................................................................. 1  
*Yang Liu, Xiaohui Yu, Bing Liu, and Zhongshuai Chen*

**Word-Level Emotion Recognition Using High-Level Features**  
(Invited Paper) ................................................................. 17  
*Johanna D. Moore, Leimin Tian, and Catherine Lai*

**Constructing Context-Aware Sentiment Lexicons with an Asynchronous Game with a Purpose**  
*Marina Boia, Claudiu Cristian Musat, and Boi Faltings*  
*32*

**Acknowledging Discourse Function for Sentiment Analysis**  
*Phillip Smith and Mark Lee*  
*45*

**A Method of Polarity Computation of Chinese Sentiment Words Based on Gaussian Distribution**  
*Ruijing Li, Shumin Shi, Heyan Huang, Chao Su, and Tianhang Wang*  
*53*

**A Sentence Vector Based Over-Sampling Method for Imbalanced Emotion Classification**  
*Tao Chen, Ruifeng Xu, Qin Lu, Bin Liu, Jun Xu, Lin Yao, and Zhenyu He*  
*62*

**Reader Emotion Prediction Using Concept and Concept Sequence Features in News Headlines**  
*Yuanlin Yao, Ruifeng Xu, Qin Lu, Bin Liu, Jun Xu, Chengtian Zou, Li Yuan, Shuwei Wang, Lin Yao, and Zhenyu He*  
*73*

**Identifying the Targets of the Emotions Expressed in Health Forums**  
*Sandra Bringay, Eric Kergosien, Pierre Pompidor, and Pascal Poncelet*  
*85*

**Investigating the Role of Emotion-Based Features in Author Gender Classification of Text**  
*Calkin Suero Montero, Myriam Munezero, and Tuomo Kakkonen*  
*98*

### Opinion Mining and Social Networks

**A Review Corpus for Argumentation Analysis**  
*Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska*  
*115*
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Looking for Opinion in Land-Use Planning Corpora</td>
<td>128</td>
</tr>
<tr>
<td><em>Eric Kergosien, Cédric Lopez, Mathieu Roche, and Maguelonne Teisseire</em></td>
<td></td>
</tr>
<tr>
<td>Cross-Lingual Product Recommendation Using Collaborative Filtering</td>
<td>141</td>
</tr>
<tr>
<td>with Translation Pairs</td>
<td></td>
</tr>
<tr>
<td><em>Kanako Komiya, Shohei Shibata, and Yoshiyuki Kotani</em></td>
<td></td>
</tr>
<tr>
<td>Identifying a Demand Towards a Company in Consumer-Generated Media</td>
<td>153</td>
</tr>
<tr>
<td><em>Yuta Kikuchi, Hiroya Takamura, Manabu Okumura, and Satoshi Nakazawa</em></td>
<td></td>
</tr>
<tr>
<td>Standardizing Tweets with Character-Level Machine Translation</td>
<td>164</td>
</tr>
<tr>
<td><em>Nikola Ljubešić, Tomaz Erjavec, and Darja Fišer</em></td>
<td></td>
</tr>
<tr>
<td>#impressme: The Language of Motivation in User Generated Content</td>
<td>176</td>
</tr>
<tr>
<td><em>Marc T. Tomlinson, David B. Bracewell, Wayne Krug, and David Hinote</em></td>
<td></td>
</tr>
<tr>
<td>Mining the Personal Interests of Microbloggers via Exploiting Wikipedia Knowledge</td>
<td>188</td>
</tr>
<tr>
<td><em>Miao Fan, Qiang Zhou, and Thomas Fang Zheng</em></td>
<td></td>
</tr>
<tr>
<td>Website Community Mining from Query Logs with Two-Phase Clustering</td>
<td>201</td>
</tr>
<tr>
<td><em>Lidong Bing, Wai Lam, Shoaib Jameel, and Chunliang Lu</em></td>
<td></td>
</tr>
<tr>
<td>Extracting Social Events Based on Timeline and User Reliability</td>
<td>213</td>
</tr>
<tr>
<td>Analysis on Twitter</td>
<td></td>
</tr>
<tr>
<td><em>Bayar Tsolmon and Kyung-Soon Lee</em></td>
<td></td>
</tr>
</tbody>
</table>

**Machine Translation and Multilingualism**

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beam-Width Adaptation for Hierarchical Phrase-Based Translation</td>
<td>224</td>
</tr>
<tr>
<td><em>Fei Su, Gang Chen, Xinyan Xiao, and Kaile Su</em></td>
<td></td>
</tr>
<tr>
<td>Training Phrase-Based SMT without Explicit Word Alignment</td>
<td>233</td>
</tr>
<tr>
<td><em>Cyrine Nasri, Kamel Smaili, and Chiraz Latiri</em></td>
<td></td>
</tr>
<tr>
<td>Role of Paraphrases in PB-SMT</td>
<td>242</td>
</tr>
<tr>
<td><em>Santanu Pal, Pintu Lohar, and Sudip Kumar Naskar</em></td>
<td></td>
</tr>
<tr>
<td>Inferring Paraphrases for a Highly Inflected Language from a</td>
<td>254</td>
</tr>
<tr>
<td>Monolingual Corpus</td>
<td></td>
</tr>
<tr>
<td><em>Kfir Bar and Nachum Dershowitz</em></td>
<td></td>
</tr>
</tbody>
</table>
Improving Egyptian-to-English SMT by Mapping Egyptian into MSA .................................................... 271
    Nadir Durrani, Yaser Al-Onaizan, and Abraham Ittycheriah

Bilingually Learning Word Senses for Translation ................. 283
    João Casteleiro, Gabriel Pereira Lopes, and Joaquim Silva

Iterative Bilingual Lexicon Extraction from Comparable Corpora
    with Topical and Contextual Knowledge
    (Best Student Paper Award) ...................................... 296
        Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi

Improving Bilingual Lexicon Extraction from Comparable Corpora
    Using Window-Based and Syntax-Based Models ................. 310
        Amir Hazem and Emmanuel Morin

An IR-Based Strategy for Supporting Chinese-Portuguese Translation
    Services in Off-line Mode........................................ 324
        Jordi Centelles, Marta Ruiz Costa-jussá, Rafael E. Banchs, and
            Alexander Gelbukh

Cross Lingual Snippet Generation Using Snippet Translation System ... 331
        Pintu Lohar, Pinaki Bhaskar, Santanu Pal, and
            Sivaji Bandyopadhyay

A Novel Machine Translation Method for Learning Chinese as a Foreign
    Language ......................................................... 343
        Tiansi Dong and Armin B. Cremers

Information Retrieval

A New Relevance Feedback Algorithm Based on Vector Space Basis
    Change ..................................................................... 355
        Rabeb Mbarek, Mohamed Tmar, and Hawete Hattab

How Complementary Are Different Information Retrieval Techniques?
    A Study in Biomedicine Domain .................................. 367
        Xiangdong An and Nick Cercone

Performance of Turkish Information Retrieval: Evaluating the Impact
    of Linguistic Parameters and Compound Nouns .................. 381
        Hatem Haddad and Chedi Bechikh Ali

Text Classification and Clustering

How Document Properties Affect Document Relatedness Measures
    (Verifiability Award) ................................................ 392
        Jessica Perrie, Aminul Islam, and Evangelos Milios
Multi-attribute Classification of Text Documents as a Tool for Ranking and Categorization of Educational Innovation Projects ........ 404
  Alexey An, Bakytkan Dauletbakov, and Eugene Levner

Named Entities as New Features for Czech Document Classification .... 417
  Pavel Král

A Knowledge-Poor Approach to Turkish Text Categorization ........... 428
  Savaş Yıldırım

Credible or Incredible? Dissecting Urban Legends ...................... 441
  Marco Guerini and Carlo Strapparava

Intelligent Clustering Scheme for Log Data Streams .................... 454
  Basanta Joshi, Umanga Bista, and Manoj Ghimire

Text Summarization

Graph Ranking on Maximal Frequent Sequences for Single Extractive Text Summarization .......................................................... 466
  Yulia Ledeneva, René Arnulfo García-Hernández, and Alexander Gelbukh

Plagiarism Detection

A Graph Based Automatic Plagiarism Detection Technique to Handle Artificial Word Reordering and Paraphrasing (Best Paper Award) ...... 481
  Niraj Kumar

Identification of Plagiarism Using Syntactic and Semantic Filters ...... 495
  R. Vijay Sundar Ram, Efstathios Stamatatos, and Sobha Lalitha Devi

Style and Spelling Checking

Readability Classification of Bangla Texts .................................... 507
  Zahurul Islam, Md. Rasheddur Rahman, and Alexander Mehler

State-of-the-Art in Weighted Finite-State Spell-Checking ................ 519
  Tommi A. Pirinen and Krister Lindén

Spelling Correction for Kazakh ................................................ 533
  Aibek Makazhanov, Olzhas Makhambetov, Islam Sabyrgaliyev, and Zhandos Yessenbayev
## Speech Processing

A Preliminary Study on the VOT Patterns of the Assamese Language and Its Nalbaria Variety .......................... 542

*Sanghamitra Nath, Himangshu Sarma, and Utpal Sharma*

## Applications

Evaluation of Sentence Compression Techniques Against Human Performance ..................................................... 553

*Prasad Perera and Leila Kosseim*

Automatically Assessing Children’s Writing Skills Based on Age-Supervised Datasets ................................. 566

*Nelly Moreno, Sergio Jimenez, and Julia Baquero*

## Author Index

............................................................... 579