

CALL FOR PAPERS



Workshop on Analytics for Noisy Unstructured Text Data

January 8, 2007, Hyderabad, India

<http://research.ihost.com/and2007>

Organizing Committee

Craig Knoblock, USC
Daniel Lopresti, Lehigh U.
Shourya Roy, IBM Research, India
L Venkata Subramaniam,
IBM Research, India

Program Committee

Eugene Agichtein, Emory U
Sophia Ananiadou, U of Manchester
Amit Bagga, Ask.com
Henry Baird, Lehigh U
Sreeram Balakrishnan, IBM Research
Pushpak Bhattacharyya, IIT Bombay
Gerald DeJong, UIUC
David Doermann, U of Maryland
Tim Finin, UMBC
Venkatesh Ganti, Microsoft Research
Lise Getoor, U of Maryland
Venu Govindaraju, SUNY Buffalo
Matthew Hurst, Nielsen BuzzMetrics
C V Jawahar, IIT Hyderabad
Nick Koudas, U of Toronto
Emiel Krahmer, Tilburg U
Raghu Krishnapuram, IBM Research
Shiau Hong Lim, UIUC
R Manmatha, U of Mass, Amherst
Yuji Matsumoto, NAIST
Prem Natarajan, BBN Technologies
Sunita Sarawagi, IIT Bombay
Sudeshna Sarkar, IIT Kharagpur
Klaus U Schulz, U of Munich
Kazem Taghva, UNLV

Contact

L. Venkata Subramaniam
lvsu@in.ibm.com
Shourya Roy
rshourya@in.ibm.com

IBM RESEARCH

Supported by
IBM Research



Endorsed by the
[International Association
for Pattern Recognition](http://www.iaprg.org)

Noisy unstructured text data is found in informal settings such as online chat, SMS, emails, message boards, newsgroups, blogs, wikis and web pages. Also, text produced by processing spontaneous speech, printed text, handwritten text contains processing noise. Text produced under such circumstances is typically highly noisy containing spelling errors, abbreviations, non-standard words, false starts, repetitions, missing punctuations, missing case information, pause filling words such as “um” and “uh.” Such text can be seen in large amounts in contact centers, on-line chat rooms, OCRed text documents, SMS corpus etc. The theme of the IJCAI 2007 Conference is "AI and its benefits to society." In keeping with this theme, this workshop proposes to look at text analytics of highly noisy text that is produced in such everyday applications in society.

The goal of the workshop is to focus on the problems encountered in analyzing noisy documents coming from various sources. The nature of the text warrants moving beyond traditional text analytics techniques. We hope that the workshop will allow researchers to present current research and development in addressing this challenge. We also believe that as a result of this workshop there will be sharing of real life noisy data sets and will result in their becoming available to a wider research community.

Topics of Interest (not limited to)

- NLP techniques for handling noisy unstructured data
- Characterization of the types of noise in documents
- Genre recognition based on the type of noise
- Robust parsing
- Characterizing, modeling and accounting for historical language change
- Methods for detecting and correcting spelling and grammatical errors in noisy text
- Information Extraction and Retrieval from noisy text
- Automatic classification and clustering of imprecise documents
- Noise-invariant document summarization techniques
- Issues in keyword search in presence of noise in unstructured data
- Machine Translation for noisy text
- Text analysis techniques for analysis and mining of call-logs, transcribed calls, web logs, chat logs, email exchanges
- Business Intelligence (BI) applications for contact centers that deal with noisy data
- Surveys on aspects of text analytics for noisy unstructured data

Submission Guidelines

We invite papers up to 8 pages in length in the style specified at
<http://research.ihost.com/and2007/submission.html>
Make submissions online at <http://www.easychair.org/AND2007>

Important Dates

Paper Submission: September 25th, 2006
Notification of Acceptance: October 23rd, 2006
Camera-Ready papers due: November 8th, 2006