

Foreword

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 and 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. Documents with historical language can also be considered noisy with respect to today’s knowledge about the language. Such text contains important historical, religious and ancient medical knowledge that is useful. 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 analytics of highly noisy text that is produced in 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. This workshop brings together a diverse group of researchers to present current research and development in addressing this challenge. As a result of this workshop some new real life noisy data sets have also become available to a wider research community.

We were fortunate to assemble a diverse group of researchers from the Natural Language Processing, Machine Learning and Knowledge Management communities to help us in organizing this workshop. The workshop call for papers had a very good response. We received 30 submissions spanning a diverse set of issues relevant to noisy text analytics. Each submission was reviewed by at least three members of the program committee.

To encourage discussion, the workshop program is structured into topic-oriented oral and poster sessions. In addition to the contributed papers, the program also contains a keynote address and a panel discussion – on the topic of whether noisy text analytics is at all possible, and if it is then how.

We would like to thank our organizing and program committees for their many invaluable inputs and thoughtful reviews. We would like to thank Monojit Choudhury, Matthew Hurst, Ted Pedersen and Sudeshna Sarkar for sharing noisy text datasets prepared by them. We would also like to thank the others who pointed us to many relevant noisy text datasets. We thank the International Association for Pattern Recognition for endorsing this workshop and instituting a best student paper award. We would like to thank Raghuram Krishnapuram for chairing the committee to decide the best student paper award. We also thank IBM Research for providing financial support for the workshop.

Craig Knoblock
Daniel Lopresti
Shourya Roy
L. Venkata Subramaniam
Workshop Co-Chairs