LANGUAGE TECHNOLOGIES FOR INDIAN SOCIAL MEDIA (सOCIAL-ईNDIA)

DURATION: FULL DAY

RATIONALE

The evolution of social media texts – such as blogs, micro-blogs (e.g., Twitter), and chats (e.g., Facebook messages) – has created many new opportunities for information access and language technology, but also many new challenges, making it one of the prime present-day research areas. Automatic processing of these types of texts warrants new strategies, in particular since they often are very 'noisy', that is, they are characterized by having a high percentage of spelling errors and containing creative spellings (*gr8* for 'great'), word play (*goooood* for 'good'), abbreviations (*OMG* for 'Oh my God!'), Meta tags (URLs, Hashtags), and so forth. So far, most of the research on social media texts has concentrated on English, whereas most of these texts now are in non-English languages. In social media, non-English speakers do not always use Unicode to write in their own language, they use phonetic typing, frequently insert English elements (through code-mixing and Anglicisms. See the following example 1), and often mix multiple languages to express their thoughts, making automatic language processing of social media texts a very challenging task. Thus it is clear that even though English still is the principal language for web communication, there is a growing need to develop technologies for other languages. Here we will concentrate on social media text in Indian languages, a nation with more than 20 official languages. ICON is a well-established gathering for the industrial and academic research communities both internationally and in India. Therefore, we believe that it is the best place to bring research attention towards developing language technologies for Indian social media text. The workshop will hold an embedded tutorial on code-mixing in social media. The three primary goals of the proposed workshop are:

1. To focus community awareness on language technologies for Indian social media.

2. Sharing of corpora and resources to promote future research.

3. Exchange of ideas and experiences amongst researchers.

Example 1. *ICON* **isbar** *Goa* **mein ho raha hai**! Great chance to visit *Goa*!

EMBEDDED TUTORIAL ON CODE-MIXING IN SOCIAL MEDIA

Abstract: Code-mixing, or mixing of more than one language in a single conversation or utterance is a common phenomenon in any multilingual society. Extreme multilinguality of India makes code-mixing extremely common on social media content posted in Indian languages and by Indian users. In this tutorial, we will talk about why code-mixing is, on one hand a computational challenge that must be solved to effectively process IL content, and on the other hand, a wonderful linguistic resource for studying several allied phenomena. The tutorial will also introduce some basic NLP techniques for code-mixed data.

DURATION: Half Day

TUTORIAL COORDINATORS

Monojit Choudhury
Microsoft Research Lab India
Website: http://research.microsoft.com/en-us/people/monojitc/

Kalika Bali
Microsoft Research Lab India
Website: http://research.microsoft.com/en-us/people/kalikab/

LIST OF TOPICS

We welcome original and unpublished submissions on all aspects of language technologies for Indian languages in the social media context. Topics of interest include, but are not limited to:

- Part of Speech (POS) Tagging
- Language Detection
- Morphological Analysis
- Name Entity Recognition (NER)
- Dependency Parsing
- Lexical Resources
- Annotated corpora
- Transliteration
- Sentiment Analysis

WORKSHOP ORGANIZERS

Amitava Das
University of North Texas, USA
Website: http://amitavadas.com/

BIO: Dr. Amitava Das is a researcher scientist at University of North Texas, USA. He is actively involved in research, teaching and organizational activities in the areas of Natural Language Processing. He has more than 40 research publications in reputed journals and conferences. He served as the PC member of several conferences like ACL, COLING, EMNLP, CICLING and has experience in conducting several workshops and conferences in the area of NLP, including Workshop on South and South East Asian NLP (WSSANLP) series, Workshop on Sentiment Analysis where AI meets Psychology (SAAIP) series, Conference on Mining Intelligence and Knowledge Exploration (MIKE) series and etc.

Björn Gambäck
Norwegian University of Science and Technology, Trondheim, Norway
Website: http://www.ntnu.edu/employees/gamback

BIO: Björn Gambäck is Professor of Language Technology at the Department of Computer and Information Science at NTNU, Norwegian University of Science and Technology, Trondheim, Norway, as well as Senior Research Expert at SICS, Swedish ICT AB, Stockholm, Sweden. He has previously worked at the University of the Saarland, Saarbrücken, Germany; Helsinki University, Finland; the Royal Institute of Technology, Stockholm, Sweden; and Addis Ababa University, Ethiopia. Prof. Gambäck has been the Coordinator or Principal Investigator of a dozen national and international projects, and has published over 100 scientific papers on subjects such as conversational agents, spoken dialogue translation, system evaluation, and machine learning applied to NLP, in particular focusing on developing processing tools and resources for under-resourced languages.

Dipankar Das
Jadavpur University
Website: http://www.dasdipankar.com/

BIO: Dipankar Das, is an Assistant Professor in the Department of Computer Science and Engineering, Jadavpur University. He worked as an Assistant Professor in the Department of Computer Science and Engineering, National Institute of Technology (NIT), Meghalaya, Govt. of India from 2012 to 2014. During his doctoral study he has been employing in the India-Japan collaborative project entitled “Sentiment Analysis where AI meets Psychology”. His research interests are in the area of Natural Language Processing, Emotion and Sentiment Analysis, Affect Computing, Information Extraction and Language Generation. He has had more than 50 publications in top conferences and journals and has served as an author over 15 Book Chapters and reviewer of several Books, Journals and Research Projects. He served as the PC member of several conferences (CICLING, MIKE etc) and has experience in conducting several workshops in the area of NLP (SAAIP series). He is a member of the IEEE, ACL, HUMAINE groups.