COLING 2012

24th International Conference on Computational Linguistics

Proceedings of the Second Workshop on Advances in Text Input Methods (WTIM 2)

Workshop chairs:
Kalika Bali, Monojit Choudhury and Yoh Okuno

15 December 2012
Mumbai, India

Diamond sponsors

Tata Consultancy Services
Linguistic Data Consortium for Indian Languages (LDC-IL)

Gold Sponsors

Microsoft Research
Beijing Baidu Netcon Science Technology Co. Ltd.

Silver sponsors

IBM, India Private Limited
Crimson Interactive Pvt. Ltd.
Yahoo
Easy Transcription & Software Pvt. Ltd.

Proceedings of the Second Workshop on Advances in Text Input Methods (WTIM 2)
Kalika Bali, Monojit Choudhury and Yoh Okuno (eds.)
Revised preprint edition, 2012

Published by The COLING 2012 Organizing Committee
Indian Institute of Technology Bombay,
Powai,
Mumbai-400076
India
Phone: 91-22-25764729
Fax: 91-22-2572 0022
Email: pb@cse.iitb.ac.in

Also available online in the ACL Anthology at http://aclweb.org

Preface

It is our great pleasure to present the proceedings of the Second Workshop on Advances in Text Input Methods (WTIM-2) held in conjunction with Coling 2012, on 15th December 2012, in Mumbai, India. This workshop is a sequel to the first WTIM which was held in conjunction with IJCNLP 2011 in November 2011, Chiang Mai, Thailand. The aim of the current workshop remains the same as the previous one that is to bring together the researchers and developers of text input technologies around the world, and share their innovations, research findings and issues across different applications, devices, modes and languages.

The proceedings contain nine contributions, five as long papers and the rest as short papers or demonstration proposals. Together they cover research on various languages including Assamese, Arabic, Bangla, Chinese, Dzongkha and Japanese, as well as keyboard design aspects of many languages using Brahmi derived scripts. The workshop featured two invited talks by Paul Butcher, Chief Software Architect of SwiftKey and Ram Prakash H., Founder and CEO of Tachyon Technologies, both of whom gave insights into development and deployment of commercial text input systems SwiftKey and Quillpad respectively. We would like to thank both the speakers for taking the time to share their experiences. The volume also includes a paper by Ram Prakash H. based on his invited talk.

In order to facilitate more interaction between the participants and presenters, all papers in WTIM-2 were presented as posters during a long session, which was preceded by short elevator pitches. In line with the same objective of increased interaction, we organized two focused sessions namely an open discussion on data and resources and a panel discussion on research and community building for text input methods.

We would like to take this opportunity to thank all the panelists, participants, presenters and authors for making WTIM-2 an enriching experience. We would also like to thank our PC members who did a wonderful job of critically reviewing the submissions and providing constructive feedback to the authors. Thanks are also due to Coling 2012 organizers for giving us this opportunity and helping us with various phases of organization, and to Microsoft Research Lab India for sponsorship. Last but not the least we would like to extend our gratitude to Hisami Suzuki, Microsoft, the founding co-chair of WTIM series, who advised us on different aspects of organization of the workshop.

Kalika Bali, Monojit Choudhury, Yoh Okuno
Organizing Co-Chairs
WTIM 2012

Committees

Organizing Co-chairs

Kalika Bali, Microsoft Research Lab India
Monojit Choudhury, Microsoft Research Lab India
Yoh Okuno, SwiftKey

Program Committee

Achraf Chalabi, Microsoft ATLC, Egypt
Hiroshi Manabe
Hiroyuki Tokunaga, Preferred Infrastructure
Hisami Suzuki, Microsoft Research Lab Redmond
Jugal Kalita, University of Colorado, Colorado Springs
Jun Hatori, Apple
Pushpak Bhattacharyya, IIT Bombay
Richa, LDC-IL, Central Institute of Indian Languages Mysore
Samit Bhattacharya, IIT Guwahati
Sarvnaz Karimi, CSIRO, Sydney
Shinsuke Mori, Kyoto University
Sriganesh Madhvanath, HP Labs India
Taku Kudo, Google Japan
Tim Paek, Microsoft Research Lab Redmond
Vasudeva Varma, IIIT Hyderabad
Virach Sornlertlamvanich, NECTEC
Xianchao Wu, Baidu

Invited Talks

SwiftKey: Building a commercial success upon firm theoretical foundations
Speaker: *Paul Butcher, SwiftKey*

Abstract: At the heart of SwiftKey's success are well motivated Machine Learning and Natural Language Processing principles. But that foundation is only the start, it's also required relentless focus on User Experience, solving endless real world issues and building and connecting with the vibrant community of SwiftKey users worldwide. This talk will take you through the story of how we turned a great IME into the most successful paid Android application in the world.

Speaker's Bio: Paul is Chief Software Architect of NLP company SwiftKey, creators of the market-leading input method by the same name.

Quillpad multilingual predictive transliteration system
Speaker: *Ram Prakash H, Tachyon Technologies*

Abstract: Transliteration has been one of the common methods for multilingual text input. Many earlier methods employed transliteration schemes for defining one to one mapping of input alphabet combinations to output alphabet combinations. Though such well-defined mappings made it easier to write a transliteration program, the end user was burdened with learning the mappings. Further, though transliteration schemes try to map the alphabet combinations phonetically, it is unavoidable to introduce non intuitive combinations into the scheme. An alternative is to use predictive transliteration, where user could input a word, by intuitively combining the input alphabet phonetically and the predictive transliteration system should correctly convert it to the target language. In this talk, I will present the challenges that must be addressed by such a system, and describe how Quillpad can be trained for performing predictive transliteration between any two alphabets.

Speaker's Bio: Ram Prakash is the founder of Tachyon Technologies, and developer of Quillpad, the first online Indian language phonetic transliteration based input system. He was listed as one of the twenty MIT TR-35 2010 Young Innovators from India for Quillpad.

Table of Contents

Statistical Input Method based on a Phrase Class n-gram Model
Hirokuni Maeta and Shinsuke Mori 1

An Ensemble Model of Word-based and Character-based Models for Japanese and Chinese Input Method
Yoh Okuno and Shinsuke Mori 15

Multi-objective Optimization for Efficient Brahmic Keyboards
Albert Brouillette, Devraj Sarmah and Jugal Kalita 29

Using Collocations and K-means Clustering to Improve the N-pos Model for Japanese IME
Long Chen, Xianchao Wu and Jingzhou He 45

phloat : Integrated Writing Environment for ESL learners
Yuta Hayashibe, Masato Hagiwara and Satoshi Sekine 57

Bangla Phonetic Input Method with Foreign Words Handling
Khan Md. Anwarus Salam, Nishino Tetsuro and Setsuo Yamada 73

LuitPad: A fully Unicode compatible Assamese writing software
Navanath Saharia and Kishori M. Konwar 79

Romanized Arabic Transliteration
Achraf Chalabi and Hany Gerges 89

Forward Transliteration of Dzongkha Text to Braille
Tirthankar Dasgupta, Manjira Sinha and Anupam Basu 97

Quillpad Multilingual Predictive Transliteration System
Ram Prakash H 107

Second Workshop on Advances in Text Input Methods (WTIM 2)

Program

Saturday, 15 December 2012

0900-09:10	Welcome Address by the Organizing Co-chairs
09:10–10:00	**Invited Talk** *SwiftKey: Building a commercial success upon firm theoretical foundations* Paul Butcher, SwiftKey
	Poster Boasters
10:00–10:12	*Statistical Input Method based on a Phrase Class n-gram Model* Hirokuni Maeta and Shinsuke Mori
10:12–10:24	*An Ensemble Model of Word-based and Character-based Models for Japanese and Chinese Input Method* Yoh Okuno and Shinsuke Mori
10:24–10:36	*Multi-objective Optimization for Efficient Brahmic Keyboards* Albert Brouillette, Devraj Sarmah and Jugal Kalita
10:36–10:48	*Using Collocations and K-means Clustering to Improve the N-pos Model for Japanese IME* Long Chen, Xianchao Wu and Jingzhou He
10:48–11:00	*phloat : Integrated Writing Environment for ESL learners* Yuta Hayashibe, Masato Hagiwara and Satoshi Sekine
11:00–11:05	*Bangla Phonetic Input Method with Foreign Words Handling* Khan Md. Anwarus Salam, Nishino Tetsuro and Setsuo Yamada
11:05–11:10	*LuitPad: A fully Unicode compatible Assamese writing software* Navanath Saharia and Kishori M. Konwar
11:10–11:15	*Romanized Arabic Transliteration* Achraf Chalabi and Hany Gerges
11:15–11:30	*Forward Transliteration of Dzongkha Text to Braille* Tirthankar Dasgupta, Manjira Sinha and Anupam Basu
11:30–12:00	Coffee break
12:00–12:45	*Quillpad Multilingual Predictive Transliteration System* Ram Prakash H
	Open Discussion
12:45–13:30	Data and Resources for Research on Text Input Methods
13:30–14:30	Lunch

Saturday, 15 December 2012 (continued)

	Poster and Demo Session
14:30–16:00	All long, short and demo paper will be presented as posters for better interaction among the participants and presenters
16:00–16:30	Coffee break
	Panel Discussion
16:30–17:45	Future of Text Input Systems: Research directions and community building
17:45	Vote of Thanks and Closing