Fourth Workshop on South and Southeast Asian Natural Language Processing

(WSSANLP 2013)

Held at the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)

Nagoya, Japan
14 – 18 October 2013

ISBN: 978-1-62993-547-8

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2013) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2014)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2634
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

Fast Bootstrapping of Grapheme to Phoneme System for Under-resourced Languages - Application to the Iban Language
Sarah Samson Juan and Laurent Besacier 1

LexToPlus: A Thai Lexeme Tokenization and Normalization Tool
Choochart Haruechaiyasak and Alisa Kongthon 9

A Three-Layer Architecture for Automatic Post Editing System Using Rule-Based Paradigm
Mahsa Mohaghegh, Abdolhossein Sarrafzadeh and Mehdi Mohammadi 17

Statistical Stemming for Kannada
Suma Bhat 25

On Application of Conditional Random Field in Stemming of Bengali Natural Language Text
Sandipan Sarkar and Sivaji Bandyopadhyay 34

Urdu Hindi Machine Transliteration using SMT
M. G. Abbas Malik, Christian Boitet, Laurent Besacier and Pushpak Bhattcharyya 43

Urdu Spell Checking: Reverse Edit Distance Approach
Saadat Iqbal, Muhammad Waqas Anwar, Usama Ijaz Bajwa and Zobia Rehman 58

Information Mining from Islamic Scriptures
Abdul Rauf Saeed and Syed Waqar Jaffry 66

English to Urdu Hierarchical Phrase-based Statistical Machine Translation
Nadeem Khan, Muhammad Waqas Anwar, Usama Ijaz Bajwa and Nadir Durrani 72

Cliticization and Endoclitics Generation of Pashto Language
Azizud Din 77

Malayalam Clause Boundary Identifier: Annotation and Evaluation
Sobha Lalitha Devi and Lakshmi S 83