3rd Workshop on South and Southeast Asian Natural Language Processing 2012

(SANLP)

Held at the 24th International Conference on Computational Linguistics (COLING 2012)

Mumbai, India
8 December 2012

Editors:

Virach Sornlertamvanich
Abbas Malik

ISBN: 978-1-62748-351-3

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© (2012) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2013)

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

Computational evidence that Hindi and Urdu share a grammar but not the lexicon
K.V.S Prasad and Shafqat Mumtaz Virk 1

Semantic Relation Extraction from a Cultural Database
Canasai Kruengkrai, Virach Sornlertlamvanich, Watchira Buranasing and Thatsanee Charoenporn 15

Bengali Question Classification: Towards Developing QA System
Somnath Banerjee and Sivaji Bandyopadhyay 25

Morphological Analyzer for Kokborok
Khumbar Debbarma, Braja Gopal Patra, Dipankar Das and Sivaji Bandyopadhyay 41

Comparing Different Criteria for Vietnamese Word Segmentation
Quy T. Nguyen, Ngan L.T. Nguyen and Yusuke Miyao 53

A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language
Sajjad Ahmad Khan, Waqas Anwar, Usama Ijaz Bajwa and Xuan Wang 69

Morpheme Segmentation for Kannada Standing on the Shoulder of Giants
Suma Bhat 79

Manipuri Morpheme Identification
Kishorjit Nongmeikapam, Vidya Raj RK, Nirmal Y and Sivaji B 95

Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach
Nidhi Krail and Vishal Gupta 109

Using English Acoustic Models for Hindi Automatic Speech Recognition
Anik Dey, Li Ying and Pascale Fung 123

Tagger Voting for Urdu
Bushra Jawaid and Ondřej Bojar 135

BIS Annotation Standards With Reference to Konkani Language
Dr. Madhavi Sardesai, Jyoti Pawar, Shantaram Walawalikar and Edna Vaz 145

Automatic Extraction of Compound Verbs from Bangla Corpora
Sibanshu Mukhopadhayay, Tirthankar Dasgupta, Manjira Sinha and Anupam Basu . 153

Influences of particles on Vietnamese tonal Co-articulation
Thi Lan Nguyen and Do Dat Tran 163

Toward an amazigh language processing
Fatima Zahra Nejme, Siham Boulaknadel and Driss Aboutajdine 173

Bidirectional Bengali Script and Meetei Mayek Transliteration of Web Based Manipuri News Corpus
Thoudam Doren Singh 181

Rule-based Machine Translation between Indonesian and Malaysian
Raymond Hendy Susanto, Septina Dian Larasati and Francis M. Tyers 191

Building Multilingual Search Index using open source framework
Arjun Atreya, Swapnil Chaudhari, Pushpak Bhattacharyya and Ganesh Ramakrishnan 201

Automatic Searching for English-Vietnamese Documents on the Internet
Quoc Hung Ngo 211

Error tracking in search engine development
Swapnil Chaudhari, Arjun Atreya V, Pushpak Bhattacharyya
and Ganesh Ramakrishnan 221

An Efficient Database Design for IndoWordNet Development Using Hybrid Approach
Venkatesh Prabhu, Shilpa Desai, Hanumant Redkar, Neha Prabhugaonkar,
Apurva Nagvenkar and Ramdas Karmali 229

IndoWordNet Application Programming Interfaces
Neha Prabhugaonkar, Apurva Nagvenkar and Ramdas Karmali 237