14th Workshop on Building and Using Comparable Corpora (BUCC 2021)

In Conjunction with the 13th International Conference on Recent Advances in Natural Language Processing (RANLP 2021)

Online
6 September 2021

Editors:

Richard Rapp
Serge Sharoff
Pierre Zweigenbaum

ISBN: 978-1-7138-4098-5

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.

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: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

Machine Translation in Low Resource Setting
Pushpak Bhattacharyya 1

Mining Bilingual Word Pairs from Comparable Corpus using Apache Spark Framework
Sanjanasri JP, Vijay Krishna Menon, Soman KP and Krzysztof Wolk 2

Effective Bitext Extraction From Comparable Corpora Using a Combination of Three Different Approaches
Steintor Steingrimsson, Pintu Lohar, Hrafn Loftsson and Andy Way 8

Syntax-aware Transformers for Neural Machine Translation: The Case of Text to Sign Gloss Translation
Santiago Egea Gomez, Euan McGill and Horacio Saggion 18

Employing Wikipedia as a resource for Named Entity Recognition in Morphologically complex under-resourced languages
Aravind Krishnan, Stefan Ziehe, Franziska Pannach and Caroline Sporleder 28

Semi-Automated Labeling of Requirement Datasets for Relation Extraction
Jeremias Bohn, Jannik Fischbach, Martin Schmitt, Hinrich Schuetze and Andreas Vogelsang . . . 40

Majority Voting with Bidirectional Pre-translation For Bitext Retrieval
Alexander Jones and Derry Tanti Wijaya 46

EM Corpus: a comparable corpus for a less-resourced language pair Manipuri-English
Rudali Huidrom, Yves Lepage and Khogendra Khomdram 60

On Pronunciations in Wiktionary: Extraction and Experiments on Multilingual Syllabification and Stress Prediction
Winston Wu and David Yarowsky 68

A Dutch Dataset for Cross-lingual Multilabel Toxicity Detection
Ben Burtenshaw and Mike Kestemont 75