

PACLIC 34 (2020)

**Proceedings of the 34th Pacific Asia  
Conference on Language, Information  
and Computation**

24–26 October, 2020

University of Science, Vietnam National University  
Hanoi, Vietnam

**©2020 The PACLIC 34 Organizing Committee and PACLIC Steering Committee**

All rights reserved. Except as otherwise expressly permitted under copyright law, no part of this publication may be reproduced, digitized, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, Internet or otherwise, without the prior permission of the publisher.

Copyright of contributed papers reserved by respective authors.

ISSN 2619-7782

### **Acknowledgments**

PACLIC 34 is hosted by University of Science, Vietnam National University, Hanoi in conjunction with The Association for Vietnamese Language and Speech Processing.

## Foreword

The 34th Pacific Asia Conference on Language, Information and Computation (PACLIC 34) is organized by the VNU University of Science, October 24–26, 2020. This edition of the PACLIC series of conferences, as its long tradition, also emphasizes the synergy of theoretical analysis and processing of natural language, aiming to enhance the interaction between researchers working in different fields of language study in the Asia-Pacific region as well as around the world.

For the first time in the history of PACLIC series organization, the conference is organized totally online due to the COVID-19 pandemic. We received 112 submissions, out of which 40 were accepted for oral presentations and 22 for poster presentations. The acceptance rate for oral presentations and poster presentations are 36% and 19% respectively. In addition to oral and poster presentations, the conference highlights four keynote talks and one satellite workshop. We are grateful to Alexander Waibel, Harald Baayen, Yunyao Li, Valia Kordoni for accepting to give a keynote talk. We also thank Jong-Bok Kim, Valia Kordoni and Thi Minh Huyen Nguyen for organizing the Workshop on Multi-word Expression in Asian Languages during the conference. Six papers have been accepted to present at this workshop.

PACLIC 34 would not be made possible without the support from many people, especially in the worldwide pandemic situation. We would like to express our sincere gratitude toward program committee members and sub-reviewers whose professional reviews allowed us to maintain the high quality standard of PACLIC. A special thank goes to Giang Son Tran from University of Science and Technology of Hanoi for maintaining the conference website. We are deeply indebted to the local organizing committee Phuong Le-Hong, The Quyen Ngo and My Linh Ha, as well as student staff members from VNU University of Science. We would also like to thank The Association for Vietnamese Language and Speech Processing (VLSP) for their financial and scientific support for the conference.

Le Minh Nguyen

Chi Mai Luong

Sanghoun Song

PACLIC 34 Program Committee Chairs

## **Organizers**

### **Steering Committee Standing Members**

Chu-Ren Huang, The Hong Kong Polytechnic University, Hong Kong

Jong-Bok Kim, Kyung Hee University, Seoul

Ryo Ootoguro, Waseda University, Tokyo

Rachel Edita O. Roxas, National University, Manila

Maosong Sun, Tsinghua University, Beijing

Benjamin T'sou, City University of Hong Kong, Hong Kong

Min Zhang, Soochow University, Suzhou

### **Organizing Committee**

NGUYEN Thi Minh Huyen, VNU University of Science (Chair)

Ryo Ootoguro, Waseda University (Co-chair)

PHAN Xuan Hieu, VNU University of Engineering and Technology (Co-chair)

Yasuhiro Katagiri, Future University Hakodate (Honorary chair)

PHAM Bao Son, Vietnam National University, Hanoi (Honorary chair)

### **Local Organizing Committee**

LE Hong Phuong, VNU University of Science (Chair)

TRAN Giang Son, University of Science and Technology of Hanoi (Co-chair)

HA My Linh, VNU University of Science

NGO The Quyen, VNU University of Science

TRAN Mai Vu, VNU University of Engineering and Technology

TRAN Thi Oanh, VNU International School

### **Program Committee Chairs**

LUONG Chi Mai, IOIT, Vietnam Academy of Science and Technology

NGUYEN Le Minh, Japan Advanced Institute of Science and Technology

Sanghoun Song, Korea University



## Reviewers

Wirote Aroonmanakun	Dongsik Lim	Rodolfo Jr Raga
Nguyen Bach	Te-Hsin Liu	Lavinia Salicchi
Philippe Blache	Wei Lu	Masashi Saraki
Thanh Hung Bui	Chi Mai Luong	Shu-Ing Shyu
Jasper Kyle Catapang	Erlyn Manguilimotan	Melanie Siegel
Alvin Cheng-Hsien Chen	Yuji Matsumoto	Pornsiri Singhapreecha
Emmanuele Chersoni	James Myers	Sanghoun Song
Sung-Kwon Choi	Ponrudee Netisopakul	Zhiyang Teng
Anh-Hien Dao	Xuan Bach Ngo	Oanh Tran
Dien Dinh	Le Minh Nguyen	Vu Tran
Alex Chengyu Fang	Minh Thuan Nguyen	Hong Viet Tran
Helena Gao	Minh-Tien Nguyen	Yuen-Hsien Tseng
Thanh-Le Ha	Thi Minh Huyen Nguyen	Benjamin Tsou
Yasunari Harada	Thi Thu Trang Nguyen	Yasushi Tsubota
Jeffrey J. Holliday	Vinh Van Nguyen	Sinh Vu
Munpyo Hong	Hoang Ky Nguyen	Hiroko Wakamatsu
Miao-Ling Hsieh	Tien Ha Nguyen	Xinyu Wang
Shu-Kai Hsieh	Tien Huy Nguyen	Tak-Sum Wong
Chu-Ren Huang	Jian-Yun Nie	Jiun-Shiung Wu
Shin'Ichiro Ishikawa	Nathaniel Oco	Rong Xiang
Jong-Bok Kim	Ethel Ong	Cheng-Zen Yang
Valia Kordoni	Chutamanee Onsuwan	Daisuke Yokomori
Pei-Jung Kuo	Jong C. Park	Satoru Yokoyama
Oi Yee Kwong	Hien Pham	Liang-Chih Yu
Huei-Ling Lai	Luan Pham	Zhang Yu
Huong Thanh Le	Quang Nhat Minh Pham	Niina Ning Zhang
Yong-Hun Lee	Anh Phan	Yuxiang Zhou
Phuong Le-Hong	Nattama Pongpairoj	

**Invited Speakers:**

Alexander Waibel, Carnegie Mellon University, Karlsruhe Institute of Technology

Harald Baayen, University of Tübingen

Yun Yao Li, IBM Almaden Research Center

Valia Kordoni, Humboldt-Universität zu Berlin

## Invited Talks

### **Alexander Waibel: Organic Machine Learning for “Intelligent” Language Interfaces**

There is good news and bad news in Speech and Language Processing: The good news is: Performance rates have dramatically improved and reach human parity (at least on matched test conditions), and Speech, Dialog, and Translation systems have gone mainstream and have become features of modern Tech Interfaces. The bad news, however: they are still barely usable and certainly not “intelligent”. What explains this discrepancy? Intelligence is the ability to respond to change and new situations. Rather than batch learning in static conditions on aggregated data and testing in matched conditions, human intelligence excels by learning and adapting continuously, incrementally and interactively, from mismatched new testing data. They must exploit multimodal information and advance with very little or no data. Learning must be a life-long process with local, personal data. We call this “Organic Machine Learning”.

In this talk, I share observations on where the technology is and where it isn’t and discuss some early research results with OML. We develop architectures for OML learning and apply them to AI language tasks such as Speech Translation and Speech Dialogs with Humanoid Robots.

### **Harald Baayen: How long you make your words crucially depends on their meanings**

Traditional approaches to human lexical processing assume that words have static form and meaning representations in the lexicon. Measures such as word frequency, number of neighbors, and word length are typically used to probe how word forms are processed. Measures such as number of synonyms or number of synonym sets in WordNet have been found to be useful for gauging semantic effects on lexical processing. Effectively, in research on the mental lexicon, measures of word form play a dominant role. For instance, the Chinese Lexical Database (Sun et al., 2018) makes available more than 200 measures of word form, but no measures of words’ meanings. Thus, the role of meaning in lexical processing is still not well understood.

A radically different approach to the mental lexicon is developed within the framework of the "Discriminative Lexicon" (Baayen et al., 2019). Central to this framework are simple fully connected two-layer networks (without hidden layers) that define mappings between high-dimensional numeric representations of word forms and high-dimensional numeric representations of word meanings (using semantic vectors aka word embeddings). These simple networks, formally equivalent to the mathematics underlying multivariate multiple regression, turn out to be surprisingly effective for predicting a wide range of lexical phenomena. In this presentation, the focus will be on predicting the acoustic durations with which words’ are realized in speech production. Evidence from English, Vietnamese, and Mandarin Chinese will be presented clarifying that how well a word’s form can be learned and predicted from its meaning is the crucial factor shaping its acoustic duration. Since learnability measures substantially out-perform measures such as word frequency as predictors of acoustic duration, the theory of the Discriminative Lexicon appears to provide a useful and productive new framework for understanding human lexical processing.

### **Yunyao Li: Towards Universal Natural Language Understanding**

Understanding the semantics of the natural language is a fundamental task in artificial intelligence. English semantic understanding has reached a mature state and successfully deployed in multiple IBM AI products and services, such as Watson Natural Language Understanding and Watson Compare and Comply. However, scaling existing products/services to support additional languages remain an open challenge. In this talk, we will discuss the open challenges in supporting universal natural language under-

standing. We will share our work in addressing these challenges in the past few years to provide the same unified semantic representation across languages. We will also showcase how such universal semantic understanding of natural languages can enable cross-lingual information extraction in concrete domains (e.g. insurance and compliance) and show promise towards seamless scaling existing NLP capabilities across languages with minimal efforts.

### **Valia Kordoni: Figurative Language in Big Data**

This talk focuses on metaphor analysis in big data, mainly in the area of education, that is, in multi-genre and heterogeneous course material, varying from video lectures, assignments, tutorial text to social web text posted on MOOC blogs and fora. While metaphor has been tackled in Natural Language Processing before, the focus of that research has never simultaneously been on the analysis of multilingual, multi-genre and heterogeneous texts for applications like Machine Translation. The work we will be presenting in this talk has been mainly carried out in TraMOOC (Translation for Massive Open Online Courses), an EU-funded Horizon 2020 collaborative project which has developed reliable Neural Machine Translation for Massive Open Online Courses (MOOCs).

## Table of Contents

### Regular Papers

Contextual Characters with Segmentation Representation for Named Entity Recognition in Chinese <i>Baptiste Blouin and Pierre Magistry</i>	2
Improving Sequence Tagging for Vietnamese Text using Transformer-based Neural Models . . . . <i>The Viet Bui, Thi Oanh Tran and Phuong Le-Hong</i>	13
A new look at Pattani Malay Initial Geminates: a statistical and machine learning approach . . . . <i>Francesco Burroni, Sireemas Maspong, Pittayawat Pittayaporn and Pimthip Kochaiyaphum</i>	21
Sketching the English Translations of Kumārajīva’s <i>The Diamond Sutra</i> : A Comparison of Individual Translators and Translation Teams . . . . . <i>Xi Chen, Vincent Xian Wang and Chu-Ren Huang</i>	30
Exploiting weak-supervision for classifying Non-Sentential Utterances in Mandarin Conversations <i>Xin-Yi Chen and Laurent Prévot</i>	42
Pay Attention to Categories: Syntax-Based Sentence Modeling with Metadata Projection Matrix . <i>Won Ik Cho and Nam Soo Kim</i>	51
Metaphoricity Rating of Chinese KIND Metaphor Expressions . . . . . <i>Siaw-Fong Chung, Meng-Hsien Shih, Yu-Hsiang Shen and Wei-Ting Tseng</i>	61
Latent Topic Refinement based on Distance Metric Learning and Semantics-assisted Non-negative Matrix Factorization . . . . . <i>Tran-Binh Dang, Ha-Thanh Nguyen and Le-Minh Nguyen</i>	70
TDP –A Hybrid Diacritic Restoration with Transformer Decoder . . . . . <i>Trung Duc Anh Dang and Thi Thu Trang Nguyen</i>	76
Construction of a VerbNet style lexicon for Vietnamese . . . . . <i>Ha My Linh, Le Van Cuong and Nguyen Thi Minh Huyen</i>	84
Utilizing Bert for Question Retrieval on Vietnamese E-commerce Sites . . . . . <i>Thi-Thanh Ha, Van-Nha Nguyen, Kiem-Hieu Nguyen, Kim-Anh Nguyen and Tien-Thanh Nguyen</i>	92
Language change in Report on the Work of the Government by Premiers of the People’s Republic of China . . . . . <i>Renkui Hou, Chu-Ren Huang and Kathleen Ahrens</i>	100
From Sense to Action: A Word-Action Disambiguation Task in NLP . . . . . <i>Shu-Kai Hsieh, Yu-Hsiang Tseng, Chiung-Yu Chiang, Richard Lian, Yong-fu Liao, Mao-Chang Ku and Ching-Fang Shih</i>	107
On the syntax of negative wh-constructions in Korean . . . . . <i>Okgi Kim</i>	113
Generation and Evaluation of Concept Embeddings Via Fine-Tuning Using Automatically Tagged Corpus . . . . . <i>Kanako Komiya, Daiki Yaginuma, Masayuki Asahara and Hiroyuki Shinnou</i>	122
Towards a Linguistically Motivated Segmentation for a Simultaneous Interpretation System . . . . <i>Youngeun Koo, Jiyoun Kim, Jungpyo Hong, Munpyo Hong and Sung-Kwon Choi</i>	129

<b>Towards Computational Linguistics in Minangkabau Language: Studies on Sentiment Analysis and Machine Translation</b> . . . . .	138
<i>Fajri Koto and Ikhwan Koto</i>	
<b>Vowel Effects on L2 Perception of English Consonants by Advanced Learners of English</b> . . . . .	149
<i>Yizhou Lan</i>	
<b>Predicting gender and age categories in English conversations using lexical, non-lexical, and turn-taking features</b> . . . . .	157
<i>Andreas Liesenfeld, Gábor Parti, Yuyin Hsu and Chu-Ren Huang</i>	
<b>Simple is Better! Lightweight Data Augmentation for Low Resource Slot Filling and Intent Classification</b> . . . . .	167
<i>Samuel Louvan and Bernardo Magnini</i>	
<b>Dialog policy optimization for low resource setting using Self-play and Reward based Sampling</b> .	178
<i>Tharindu Madusanka, Durashi Langappuli, Thisara Welmilla, Uthayasanker Thayasivam and Sanath Jayasena</i>	
<b>Learning to Describe Editing Activities in Collaborative Environments: A Case Study on GitHub and Wikipedia</b> . . . . .	188
<i>Edison Marrese-Taylor, Pablo Loyola, Jorge A. Balazs and Yutaka Matsuo</i>	
<b>A Multilingual Linguistic Domain Ontology</b> . . . . .	199
<i>Mariem Neji, Fatma Ghorbel, Bilel Gargouri, Nada Mimouni and Elisabeth Métais</i>	
<b>Iterative Multilingual Neural Machine Translation for Less-Common and Zero-Resource Language Pairs</b> . . . . .	207
<i>Minh Thuan Nguyen, Phuong Thai Nguyen, Van Vinh Nguyen and Minh Cong Nguyen Hoang</i>	
<b>Enhancing Quality of Corpus Annotation: Construction of the Multi-Layer Corpus Annotation and Simplified Validation of the Corpus Annotation</b> . . . . .	216
<i>Youngbin Noh, Kuntae Kim, Minho Lee, Cheolhun Heo, Yongbin Jeong, Yoosung Jeong, Younggyun Hahm, Taehwan Oh, Hyonsu Choe, Seokwon Park, Jin-Dong Kim and Key-Sun Choi</i>	
<b>Syntactic similarity of the sentences in a multi-lingual parallel corpus based on the Euclidean distance of their dependency trees</b> . . . . .	225
<i>Masanori Oya</i>	
<b>Plausibility and Well-formedness Acceptability Test on Deep Neural Nativeness Classification</b> . .	234
<i>Kwonsik Park and Sanghoun Song</i>	
<b>A Simple Disaster-Related Knowledge Base for Intelligent Agents</b> . . . . .	243
<i>Clark Emmanuel Paulo, Arvin Ken Ramirez, David Clarence Reducindo, Rannie Mark Mateo and Joseph Marvin Imperial</i>	
<b>Effective Approach to Develop a Sentiment Annotator For Legal Domain in a Low Resource Setting</b>	252
<i>Gathika Ratnayaka, Nisansa de Silva, Amal Shehan Perera and Ramesh Pathirana</i>	
<b>Deriving confirmation and justification — an expectative, compositional analysis of Japanese 'yo-ne'</b>	261
<i>Lukas Rieser</i>	
<b>Combining Thai EDUs: Principle and Implementation</b> . . . . .	270
<i>Chanatip Saetia, Supawat Taerungruang and Tawunrat Chalothorn</i>	
<b>Evaluation of Pretrained BERT Model by Using Sentence Clustering</b> . . . . .	279

*Naoki Shibayama, Rui Cao, Jing Bai, Wen Ma and Hiroyuki Shinnou*

**Music and speech are distinct in lexical tone normalization processing . . . . . 286**

*Ran Tao and Gang Peng*

**Construction of Associative Vocabulary Learning System for Japanese Learners . . . . . 294**

*Takehiro Teraoka and Tetsuo Yamashita*

**A corpus-based comparative study of light verbs in three Chinese speech communities . . . . . 302**

*Benjamin K Tsou and Ka-Fai Yip*

**Sensorimotor Enhanced Neural Network for Metaphor Detection . . . . . 312**

*Mingyu Wan, Baixi Xing, Qi Su, Pengyuan Liu and Chu-Ren Huang*

**A Parallel Corpus-driven Approach to Bilingual Oenology Term Banks: How Culture Differences  
Influence Wine Tasting Terms . . . . . 318**

*Vincent Xian Wang, Xi Chen, Songnan Quan and Chu-Ren Huang*

**Corpus-based Comparison of Verbs of Separation “Qie” and “Ge” . . . . . 329**

*Nga-In Wu, Chu-Ren Huang and Lap-Kei Lee*

**Association between declarative memory and language ability in older Chinese by education level 337**

*Chenwei Xie, Yun Feng and William Shi-Yuan Wang*

**A corpus-based analysis of Chinese relative clauses produced by Japanese and Thai learners . . . . 348**

*Yike Yang*

## **Poster Papers**

**Aspect-based Sentiment Analysis on Indonesia’s Tourism Destinations Based on Google Maps  
User Code-Mixed Reviews (Study Case: Borobudur and Prambanan Temples) . . . . . 359**

*Dian Arianto and Indra Budi*

**Imbalanced Chinese Multi-label Text Classification Based on Alternating Attention . . . . . 368**

*Hongliang Bi, Han Hu and Pengyuan Liu*

**How State-Of-The-Art Models Can Deal With Long-Form Question Answering . . . . . 375**

*Minh-Quan Bui, Vu Tran, Ha-Thanh Nguyen and Le-Minh Nguyen*

**Prosody Features of Collaborative Construction in Mandarin Conversation . . . . . 383**

*Yue Guan*

**ILP-based Opinion Sentence Extraction from User Reviews for Question DB Construction . . . . . 395**

*Masakatsu Hamashita, Takashi Inui, Koji Murakami and Keiji Shinzato*

**Composing Word Vectors for Japanese Compound Words Using Bilingual Word Embeddings . . . . 404**

*Teruo Hirabayashi, Kanako Komiya, Masayuki Asahara and Hiroyuki Shinnou*

**Exploring Discourse on Same-sex Marriage in Taiwan: A Case Study of Near-Synonym of HO-  
MOSEXUAL in Opposing Stances . . . . . 411**

*Han-Tang Hung and Shu-Kai Hsieh*

**A Simple and Efficient Ensemble Classifier Combining Multiple Neural Network Models on Social  
Media Datasets in Vietnamese . . . . . 420**

*Huy Duc Huynh, Hang Thi-Thuy Do, Kiet Van Nguyen and Ngan Thuy-Luu Nguyen*

<b>Text Mining of Evidence on Infants’ Developmental Stages for Developmental Order Acquisition from Picture Book Reviews</b> . . . . .	430
<i>Miho Kasamatsu, Takehito Utsuro, Yu Saito and Yumiko Ishikawa</i>	
<b>Expressing the Opposite: Acoustic Cues of Thai Verbal Irony</b> . . . . .	439
<i>Nimit Kumwapee and Sujinat Jitwiriyant</i>	
<b>Identifying Authors Based on Stylometric measures of Vietnamese texts</b> . . . . .	447
<i>Ho Ngoc Lam, Vo Diep Nhu, Dinh Dien and Nguyen Tuyet Nhung</i>	
<b>Marking Trustworthiness with Near Synonyms: A Corpus-based Study of “Renwei” and “Yiwei” in Chinese</b> . . . . .	453
<i>Bei Li, Chu-Ren Huang and Si Chen</i>	
<b>Empirical Study of Text Augmentation on Social Media Text in Vietnamese</b> . . . . .	462
<i>Son Luu, Kiet Nguyen and Ngan Nguyen</i>	
<b>Attention-based Domain Adaption Using Transfer Learning for Part-of-Speech Tagging: An Experiment on the Hindi Language</b> . . . . .	471
<i>Rajesh Kumar Mundotiya, Vikrant Kumar, Arpit Mehta and Anil Kumar Singh</i>	
<b>Understanding Transformers for Information Extraction with Limited Data</b> . . . . .	478
<i>Minh-Tien Nguyen, Dung Tien Le, Nguyen Hong Son, Bui Cong Minh, Do Hoang Thai Duong and Le Thai Linh</i>	
<b>A Study on Seq2seq for Sentence Compression in Vietnamese</b> . . . . .	488
<i>Thi-Trang Nguyen, Huu-Hoang Nguyen and Kiem-Hieu Nguyen</i>	
<b>Indirectly Determined Comparison and Difference: The Case of Japanese</b> . . . . .	496
<i>Toshiko Oda</i>	
<b>Extraction of Novel Character Information from Synopses of Fantasy Novels in Japanese using Sequence Labeling</b> . . . . .	505
<i>Yuji Oka and Kazuaki Ando</i>	
<b>Redefining Verbal Nouns in Japanese: From the Perspective of Polycategoriality</b> . . . . .	514
<i>David Y. Oshima and Midori Hayashi</i>	
<b>Speech Recognition for Endangered and Extinct Samoyedic languages</b> . . . . .	523
<i>Niko Partanen, Mika Hämäläinen and Tiina Klooster</i>	
<b>Neural Machine Translation from Historical Japanese to Contemporary Japanese Using Diachronically Domain-Adapted Word Embeddings</b> . . . . .	534
<i>Masashi Takaku, Toshio Hirasawa, Mamoru Komachi and Kanako Komiya</i>	
<b>Improving Semantic Similarity Calculation of Japanese Text for MT Evaluation</b> . . . . .	542
<i>Yuki Tanahashi, Kyoko Kanzaki, Eiko Yamamoto and Hitoshi Isahara</i>	
<b>Workshop on Multiword Expressions in Asian languages</b>	
<b>Predicative multi-word expressions in Persian</b> . . . . .	552
<i>Jens Fleischhauer</i>	
<b>Forms and Meanings of Lexical Reduplications in Cantonese: a corpus study</b> . . . . .	562
<i>Charles Lam</i>	



Abstract Meaning Representation for MWE: A study of the mapping of aspectuality based on Mandarin light verb <i>jiayi</i> . . . . .	568
<i>Lu Lu, Nianwen Xue and Chu-Ren Huang</i>	
Formulatic Language of Vietnamese Children with Autism Spectrum Disorders: A Corpus Lin- guistic Analysis . . . . .	575
<i>Hien Pham and Giang Nguyen Thi</i>	
The Framework of Multiword Expression in Indonesian Language . . . . .	582
<i>Totok Suhardijanto, Rahmad Mahendra, Zahroh Nuriah and Adi Budiwiyanto</i>	
Bilingual Multi-word Expressions, Multiple-correspondence, and their cultivation from parallel patents: The Chinese-English case . . . . .	589
<i>Benjamin K. Tsou, Ka Po Chow, John Lee, Ka-Fai Yip, Yaxuan Ji and Kevin Wu</i>	