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

24–26 October, 2020
University of Science, Vietnam National University
Hanoi, Vietnam
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 Otoguro, 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 Otoguro, 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

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

Alexander Waibel, Carnegie Mellon University, Karlsruhe Institute of Technology
Harald Baayen, University of Tübingen
Yunyao 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  
*Baptiste Blouin and Pierre Magistry*  
2

Improving Sequence Tagging for Vietnamese Text using Transformer-based Neural Models  
*The Viet Bui, Thi Oanh Tran and Phuong Le-Hong*  
13

A new look at Pattani Malay Initial Geminates: a statistical and machine learning approach  
*Francesco Burroni, Sireemas Maspang, Pittayawat Pittayaporn and Pimthip Kochaiyaphum*  
21

Sketching the English Translations of Kumārajiva’s *The Diamond Sutra*: A Comparison of Individual Translators and Translation Teams  
*Xi Chen, Vincent Xian Wang and Chu-Ren Huang*  
30

Exploiting weak-supervision for classifying Non-Sentential Utterances in Mandarin Conversations  
*Xin-Yi Chen and Laurent Prévot*  
42

Pay Attention to Categories: Syntax-Based Sentence Modeling with Metadata Projection Matrix  
*Won Ik Cho and Nam Soo Kim*  
51

Metaphoricity Rating of Chinese KIND Metaphor Expressions  
*Siaw-Fong Chung, Meng-Hsien Shih, Yu-Hsiang Shen and Wei-Ting Tseng*  
61

Latent Topic Refinement based on Distance Metric Learning and Semantics-assisted Non-negative Matrix Factorization  
*Tran-Binh Dang, Ha-Thanh Nguyen and Le-Minh Nguyen*  
70

TDP –A Hybrid Diacritic Restoration with Transformer Decoder  
*Trung Duc Anh Dang and Thi Thu Trang Nguyen*  
76

Construction of a VerbNet style lexicon for Vietnamese  
*Ha My Linh, Le Van Cuong and Nguyen Thi Minh Huyen*  
84

Utilizing Bert for Question Retrieval on Vietnameses E-commerce Sites  
*Thi-Thanh Ha, Van-Nha Nguyen, Kiem-Hieu Nguyen, Kim-Anh Nguyen and Tien-Thanh Nguyen*  
92

Language change in Report on the Work of the Government by Premiers of the People’s Republic of China  
*Renkui Hou, Chu-Ren Huang and Kathleen Ahrens*  
100

From Sense to Action: A Word-Action Disambiguation Task in NLP  
*Shu-Kai Hsieh, Yu-Hsiang Tseng, Chiung-Yu Chiang, Richard Lian, Yong-fu Liao, Mao-Chang Ku and Ching-Fang Shih*  
107

On the syntax of negative wh-constructions in Korean  
*Okgi Kim*  
113

Generation and Evaluation of Concept Embeddings Via Fine-Tuning Using Automatically Tagged Corpus  
*Kanako Komiya, Daiki Yaginuma, Masayuki Asahara and Hiroyuki Shinnou*  
122

Towards a Linguistically Motivated Segmentation for a Simultaneous Interpretation System  
*Youngeun Koo, Jiyoun Kim, Jungpyo Hong, Munpyo Hong and Sung-Kwon Choi*  
129
Towards Computational Linguistics in Minangkabau Language: Studies on Sentiment Analysis and Machine Translation .......................................................... 138
Fajri Koto and Ikhwan Koto

Vowel Effects on L2 Perception of English Consonants by Advanced Learners of English .... 149
Yizhou Lan

Predicting gender and age categories in English conversations using lexical, non-lexical, and turn-taking features ................................................................. 157
Andreas Liesenfeld, Gábor Parti, Yuyin Hsu and Chu-Ren Huang

Simple is Better! Lightweight Data Augmentation for Low Resource Slot Filling and Intent Classification ................................................................. 167
Samuel Louvan and Bernardo Magnini

Dialog policy optimization for low resource setting using Self-play and Reward based Sampling . 178
Tharindu Madusanka, Durashi Langappuli, Thisara Welmilla, Uthayasanker Thayasivam and Sanath Jayasena

Learning to Describe Editing Activities in Collaborative Environments: A Case Study on GitHub and Wikipedia ................................................................. 188
Edison Marrese-Taylor, Pablo Loyola, Jorge A. Balazs and Yutaka Matsuo

A Multilingual Linguistic Domain Ontology ................................................................. 199
Mariem Neji, Fatma Ghorbel, Bilel Gargouri, Nada Mimouni and Elisabeth Métais

Iterative Multilingual Neural Machine Translation for Less-Common and Zero-Resource Language Pairs ................................................................. 207
Minh Thuan Nguyen, Phuong Thai Nguyen, Van Vinh Nguyen and Minh Cong Nguyen Hoang

Enhancing Quality of Corpus Annotation: Construction of the Multi-Layer Corpus Annotation and Simplified Validation of the Corpus Annotation ................................................................. 216
Youngbin Noh, Kuntae Kim, Minho Lee, Cheolhun Heo, Yongbin Jeong, Yoonsung Jeong, Younggyun Hahn, Taehwan Oh, Hyonsu Choe, Seokwon Park, Jin-Dong Kim and Key-Sun Choi

Syntactic similarity of the sentences in a multi-lingual parallel corpus based on the Euclidean distance of their dependency trees ................................................................. 225
Masanori Oya

Plausibility and Well-formedness Acceptability Test on Deep Neural Nativeness Classification .. 234
Kwonsik Park and Sanghoun Song

A Simple Disaster-Related Knowledge Base for Intelligent Agents ................................................................. 243
Clark Emmanuel Paulo, Arvin Ken Ramirez, David Clarence Reducindo, Rannie Mark Mateo and Joseph Marvin Imperial

Effective Approach to Develop a Sentiment Annotator For Legal Domain in a Low Resource Setting 252
Gathika Ratnayake, Nisansa de Silva, Amal Shehan Perera and Ramesh Pathirana

Deriving confirmation and justification — an expectative, compositional analysis of Japanese ‘yo-ne’ 261
Lukas Rieser

Combining Thai EDUs: Principle and Implementation ................................................................. 270
Chanatip Saetia, Supawat Tuerungruang and Tawunrat Chalothorn

Evaluation of Pretrained BERT Model by Using Sentence Clustering ................................................................. 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 HOMOSEXUAL 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
Abstract Meaning Representation for MWE: A study of the mapping of aspectuality based on Mandarin light verb jiayi .......................... 568
Lu Lu, Nianwen Xue and Chu-Ren Huang

Formulatic Language of Vietnamese Children with Autism Spectrum Disorders: A Corpus Linguistic Analysis ................................................... 575
Hien Pham and Giang Nguyen Thi

The Framework of Multiword Expression in Indonesian Language ......................... 582
Totok Suhardijanto, Rahmad Mahendra, Zahroh Nuriah and Adi Budiwiyanto

Bilingual Multi-word Expressions, Multiple-correspondence, and their cultivation from parallel patents: The Chinese-English case ....................... 589
Benjamin K. Tsou, Ka Po Chow, John Lee, Ka-Fai Yip, Yaxuan Ji and Kevin Wu