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

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 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

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

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

Poster Papers

Workshop on Multiword Expressions in Asian languages