

PREFACE

VLSP stands for Vietnamese Language and Speech Processing Consortium. It is an initiative which came from Institute of Information Technology - Vietnam Academy of Science and Technology, to establish a community working on speech and text processing for Vietnamese language. The first national project KC01.03/06-10 from 2007-2009 which received strong support from Ministry of Science and Technology, gathered eight active research groups from universities and institutes in Vietnam and overseas. A main goal of the first project is to set up long term strategy on Vietnamese language processing in order to provide and to involve community to enrich shared language resources and tools for R&D purpose. From the years 2012 to 2018, the Consortium has organized five VLSP workshops in conjunction with large international conferences organized in Vietnam. The sixth international workshop on Vietnamese Language and Speech Processing will be organized in connection with the 2019 Conference of the Pacific Association for Computational Linguistics (PACLING 2019), which will be held in Hanoi in October 2019. VLSP workshops are leading by Assoc. Prof. Luong Chi Mai from IOIT-VAST for a branch of speech processing, and Dr. Nguyen Thi Minh Huyen from HUS-VNU Hanoi - for text processing.

VLSP 2012 was the first and kick-off international workshop on Vietnamese Language and Speech Processing, organized in conjunction with IEEE-RIVF 2012. The first VLSP evaluation campaign happened in the second VLSP workshop, associated with the IEEE-RIVF 2013 Conference. This first campaign dealt with two different tasks. One concerns the very essential tools for Vietnamese language processing, word segmentation and part-of-speech (POS) tagging. The other concerns one of the most important Natural Language Processing (NLP) applications: the machine translation task. The third VLSP workshop (2015) was organized for the first time as a satellite event of PAKDD 2015, with a call for participation of only scientific and technical papers in NLP domain.

VLSP 2016, the fourth international workshop, was once again organized in conjunction with the IEEE-RIVF conference. From this event onwards, the organizers put the workshop priority on organizing shared tasks in Vietnamese processing. These shared tasks are meant to promote the most efficient methods for the important NLP tools. The organization of this campaign with sponsor from academia and industry will permit to build and offer to the VLSP community gold datasets for Vietnamese text and speech processing. In 2016, two popular tasks were considered: named-entity recognition (NER) and sentiment analysis (SA). The NER evaluation dealt with three types of entity: person, organization and location. The SA task consisted of classifying reviews of technology products as negative, positive or neutral.

This fifth international workshop on Vietnamese Language and Speech Processing VLSP 2018 was the largest in terms of number of shared tasks and number of participating teams. Like VLSP 2016, the workshop attracted not only academic research institutions but also R&D teams from the industry sector. The workshop was carefully prepared as one important event for CICLing 2018 the 19th International Conference on Computational Linguistics and Intelligent Text Processing. Four shared tasks were organized for text processing as well as

for speech processing. For text processing, the campaign dealt with two tasks as in 2016:

- Named Entity Recognition (NER). Two sources of data were considered: online newspapers and social media texts. As previously, the task dealt with three entity types: person, organization and location.
- Sentiment Analysis (SA). The task in this campaign focused on aspect-based sentiment analysis of reviews and comments on hotels and restaurants.

For speech processing, two tasks have been organized for the first time:

- Automatic Speech Recognition (ASR): Speech recognition with Northern, Central, and Southern dialects.
- Text-to-Speech (TTS): Speech synthesis for one or more regions (Northern, Central, and Southern).

In this special issue of the Journal of Computer Science and Cybernetics we introduce technical papers of the best systems for each task in VLSP 2018. Each paper presents the main technology used to develop the system, analyzes the data used for training and testing, and discusses obtained results. For the Vietnamese language, although there existed some research projects and publications on NER and SA tasks before 2016, no systematic comparison of the performance of NER systems had been done. At VLSP 2018, the NER and SA shared task has been organized for the second time, providing for NER a bigger data set containing texts from various domains, but without morpho-syntactic annotation, and for SA two datasets containing reviews in the hospitality domain. In order to offer an overview of the NER and SA tasks, we describe in this special issue the datasets as well as the evaluation results of the systems participating in these campaigns. The two speech processing tasks focus on Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) for Vietnamese, two fields where the significant effort of research teams has produced systems whose performance is up to the international standard for under-resourced languages. All papers in this special issue are under standard review process of the Journal with the emphasis on content to technology questions.

Luong Chi Mai
Section Editor
Journal of Computer Science and Cybernetics