Final Project Report
Detailed Form

I. Title of Proposed Project: ASEAN Language Speech Translation thru' U-STAR

II. Project Leader:

Full name: Prof Li Haizhou / Ms Aw Ai Ti
Institution: Institute for Infocomm Research
Address: 1 Fusionopolis Way, #21-01 Connexis, Singapore 138632
Phone: 6564082774
E-mail: haizhou.li@nus.edu.sg aaiti@i2r.a-star.edu.sg

III. Project Members:

Name	Position/ Degree	Department, Institution, Country	Email Address
Dr Rapid Sun	Director	Director of Research and Development Center, NIPTICT, Cambodia	sun.rapid@niptict.edu.kh
Dr. Hammam Riza	Deputy Chairman	Deputy Chairman IT, Energy and Material, BPPT, Indonesia	hammam.riza@bppt.go.id
Prof. Sevia M. Idrus	Deputy Dean	Faculty Of Engineering, UTM, Malaysia	sevia@fke.utm.my
Prof. Khin Mar Soe	Professor	Professor, NLP Lab, UCSY, Myanmar	khinmarsoe@ucsy.edu.mm
Dr. Chai Wutiwiwatchai	Executive Director	Executive Director, National Electronics and Computer Technology Center (NECTEC), Thailand	chai.wutiwiwatchai@nectec.or.th
Prof. NGUYEN Thi Thu Trang	Assistant Professor	Assistant Professor in Department of Software Engineering, School of Information and Communication Technology, HUST, Vietnam	trangntt@soict.hust.edu.vn
Prof. Luong Chi Mai	Assoc. Prof	Assoc. Prof, Speech and Language Processing, IOIT, Vietnam	lcmai@ioit.ac.vn

IV. Project Report

i) Introduction

Speech-to-speech translation system, which allows a user's speech to be translated into designated languages with synthesized voices, is a challenging and ongoing worldwide research carried out in many research institutes. Such system is expected to be the nexus of services delivery and applications for consumers and businesses in the future to bridge the language gaps.

Leveraging on U-STAR infrastructure, this project developed a mobile application to provide speech-to-speech translation. The project also developed data resources to facilitate the research & development of ASEAN speech translation technologies.

The project started on 1st July 2016 and completed on 30th June 2019.

ii) Project Activities

(1) Development of uniTRANS and localization of uniTRANS to ASEAN languages

To facilitate the communication among ASEAN community, a speech-to-speech mobile application – uniTRANS was developed. To further encourage the use of this application among the different language speaking users, the interface was localized to all native languages of the project team members. It was achieved through the contributions of all project team members on localizing the software resource files and images to the different local native languages.

(2) Leveraged Resources and Participants

As data are the most critical resources in speech translation research & development, the project team developed the guidelines for speech translation data collection. The guidelines outline the following specifications:

- Data: Content of utterances, Targeted speakers and Speaking style and environment
- Recording Device: Device and mechanism used for data collection
- Transcription Specifications: Transcribing Numbers, Transcribing Acronyms and Transcribing Foreign words and names
- Translation Specifications: Punctuation Insertion, Translating Numbers, Translation or Transliteration
- Speech Data Files and Naming Conventions

The following shows the statistics of the data collected by the project team.

Organization	Country	Language	Data Statistics
NIPTICT	Cambodia	Khmer	• 10K utterances collected and translated • 4K utterances selected to record as voice data.
BPPT	Indonesia	Bahasa Indonesia	• 5000 utterances collected, translated and recorded
UTM	Malaysia	Bahasa Melayu	• 5000 utterances have been collected and translated.
UCSY	Myanmar	Myanmar	• 4000 utterances collected and translated.
I2R	Singapore	Chinese	• 5000 utterances collected, translated and recorded
HUST	Vietnam	Vietnamese	• 6,500 Vietnamese text utterances collected • 3,000 Parallel text utterances • 1,200 recorded utterances (Vietnamese)
IOIT	Vietnam	Vietnamese	• 2000 utterances collected and recorded
NECTEC	Thailand	Thai	• 6000 utterances have been collected, translated, and NE annotated • 4000 utterances recorded

(3) Research Activities

I2R also collaborated with UCSY on two projects through student attachment. The first project is on Myanmar-to-English translation using syllable-based neural machine translation technique (Yi Mon Shwe Sin, *Khin Mar Soe,* UCSY; *Wu Kui, Aw Ai Ti*, I2R). As Myanmar language has rich morphology and word-based neural machine translation cannot model rare words effectively, a syllable-based NMT was proposed for this task.

The second project was on Myanmar word segmentation (*Hsu Myat Mo*, *Khin Mar Soe,* UCSY; *Zhou Nina, Aw Ai Ti,* I2R). As Myanmar scripts are written continuously as a sequence of characters without any delimiter between words, the project formulated the task as a

sequential labelling task on Myanmar character using Bi-LSTM.

(4) Knowledge Sharing and Exchange Activities

Two workshops were conducted to facilitate the sharing of knowledge among the team members. The first workshop was conducted on 4th October 2016 while the second workshop was held on 6th December 2017.

Workshop on 4th October 2016	
0900 to 0920	Comparison of Grapheme-to-Phoneme Conversion Methods on a Myanmar Pronunciation Dictionary *Ye Kyaw Thu and Win Pa Pa* Language and Speech Science Research Lab, Waseda University, Tokyo, Japan, Natural Language Processing Lab, University of Computer Studies, Yangon, Myanmar
0920 to 0940	The effect of dialect on the syllable accuracy of Vietnamese continuous speech recognition system *NGUYEN Hong Quang, TRINH Van Loan* Hanoi University of Science and Technology, Vietnam
0940 to 1000	Vietnamese LVCSR Development and Improvement *Van Huy Nguyen, Quoc Bao Nguyen, Chi Mai Luong, Tat Thang Vu* Thai Nguyen University of Technology, Vietnam Thai Nguyen University of Information and Communication Technology Institute of Information Technology (IOIT), Vietnam Academy of Science and Technology, Vietnam
1000 to 1020	Towards Indonesian Speech-to-speech Translation System *Agung Santosa, Hammam Riza, M. Teduh Ulinansyah, Gunarso, Made Gunawan, Elvira Nurfadhilah, Lyla R Aini, Harnum Annisa, Fara Ayuningtyas* Center for ICT – BPPT, Jakarta, Indonesia
1020 to 1040	Break
1040 to 1100	Network-based Speech Translation Services *[Zhongwei Li, Ai Ti Aw, Sharifah Mahani Aljunied, Haizhou Li] , [Rapid Sun, Vichet Chea] , Hammam Riza , [Sevia M. Idrus, Rubita Sudirman, Faizah Mohamad Nor] , [Khin Mar Soe, Win Pa Pa] , [Chai Wutiwiwatchai, Thepchai Supnithi] , [NGUYEN Hong Quang, NGUYEN Thi Thu Trang] , [Luong Chi Mai, Vu Tat Thang]* "ASEAN Language Speech Translation thru' U-STAR" Project Team
1100 to 1120	Context-dependent Bilingual Word Embedding with Sentence Similarity Constraint for Machine Translation *Kui Wu, Xuancong Wang, Ai Ti Aw* Institute for Infocomm Research, Singapore
1120 to	Extracting Parallel Sentences from Movie Subtitles

1140	*Boon Hong Yeo, Ai Ti Aw, Xuancong Wang* Institute for Infocomm Research, Singapore
1140 to 1200	An approach for Vietnamese-Japanese Statistical Machine Translation (SMT) *NGUYEN Thi Thu Trang, LE Thanh Huong* Hanoi University of Science and Technology, Vietnam
1200 to 1220	Natural Language Processing Development Trends in Malaysia and the Way Forward *Sevia Mahdaliza Idrus, Rubita Sudirman, Faizah Mohamad Nor* Universiti Teknologi Malaysia, Malaysia
1220 to 1400	Lunch
1400 to 1430	Opening Address & Project Introduction *Haizhou Li & Ai Ti Aw* Institute for Infocomm Research
1430 to 1500	uniTRANS Development and Localization *Zhongwei Li & Sharifah Mahani Aljunied* Institute for Infocomm Research
1500 to 1530	Activities update at NIPTICT *Vichet Chea, Rapid Sun* Deputy Director of Research and Innovation Center National Institute of Post Telecommunication Information Communication Technology (NIPTICT) Ministry of Posts and Telecommunications
1530 to 1600	Break
1600 to 1700	Project Discussion
1700	End of workshop

Workshop on 6th December 2017	
0900 to 0910	Registration
0910 to 0925	Opening Speech *Haizhou Li* I²R, Singapore
0925 to 0930	Logistics Announcement *Paul Yaozhu Chan, Ridong Jiang* I²R, Singapore
0930 to	Harnessing Data for the ASEAN-IVO

0945	*Rubita Sudirman, Nor Faizah* Mohammed UTM, Malaysia
0945 to 1000	Dynamic Semantic Boundary Detection for Speech Translation *Nina Zhou* I²R, Singapore
1000 to 1030	Tea Break
1030 to 1045	A MSD-HMM approach to Vietnamese LVCASR *Luong Chi Mai, Nguyen Van Huy* IOIT, Vietnam
1045 to 1100	How Myanmar ASR and TTS is going *Win Pa Pa, Thazin Myint Oo* UCSY, Myanmar
1100 to 1115	Khmer ASR System based on DNN Model - Development & Progress *Vichet Chea, Soky Kak* NIPTICT, Cambodia
1115 to 1130	Speech Translation Activities in Thailand *Chai Wutiwiwatchai, Sertsi Phuttapong* NECTEC, Thailand
1130 to 1145	E-I speech translation system *Hammam Riza, Gunarso* BPPT, Indonesia
1145 to 1200	Final Notes *Ai Ti Aw* I²R, Singapore
1200 to 1330	Lunch
1330 to 1715	Project Discussion
1715	End of Workshop

(5) Findings and Outcomes

A Myanmar-to-English translation prototype was developed. The prototype was trained on 228,767 parallel sentences using Syllable-based word segmentation and LSTM encoder-decoder. The prototype achieved the best performance when compared to word-based NMT and character-based NMT.

(6) Broader Impact

Most of the ASEAN languages are low resource languages. Data collected through this

project can motivate researchers working on low-resources NLP or MT to work on ASEAN languages and promote the research and development on ASEAN languages.

(7) Future Developments

Data resources are the fundamental building blocks of data-driven machine learning approach. The project team can continue to leverage on U-STAR infrastructure to collaborate on speech translation technologies and develop innovative applications using U-STAR services to benefit the community in human communication.

iii) Social Contribution

The project supported Dr. Nguyen Thi Thu Trang, Hanoi University of Science and Technology, to present their paper on "A Hybrid Method for Vietnamese Text Normalization" in the International Conference on Natural Language Processing and Information Retrieval (http://www.nlpir.net), June 28-30, 2019.