ASEAN Language Speech Translation thru' U-STAR

Project Report

Project Leader: Dr. Li Haizhou, I^{2}R, Singapore

Speaker: Dr. Chai Wutiwiwatchai, NECTEC, Thailand

Project Members

| Name	Organization	Country
Rapid Sun	Director of Research and Development Center, NIPTICT	Cambodia
Hammam Riza	Deputy Chairman IT, Energy and Material, BPPT	Indonesia
Sevia M. Idrus	Professor, Communication Engineering Department, Faculty Of Electrical Engineering, UTM	Malaysia
Khin Mar Soe	Professor, NLP Lab, UCSY	Myanmar
Li Haizhou	Joint Appointment, Human Language Technology, I2R	Singapore
Chai Wutiwiwatchai	Director of Human Computer Communication Research Unit, NECTEC	Thailand
NGUYEN Thi Thu Trang	Assistant Professor in Department of Software Engineering, School of Information and Communication Technology, HUST	Vietnam
Luong Chi Mai	Assoc. Prof, Speech and Language Processing, IOIT	Vietnam

Project Introduction

To provide a single and unified platform for multilingual communications across the ASEAN nations through U-STAR

To collaborate and facilitate the development of speech translation technologies, language resources and translation app for ASEAN languages

Universal
Speech
Translation
Advanced
Research

http://www.ustar-consortium.com/members.html
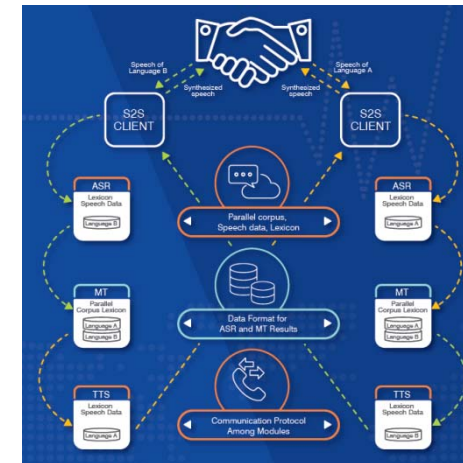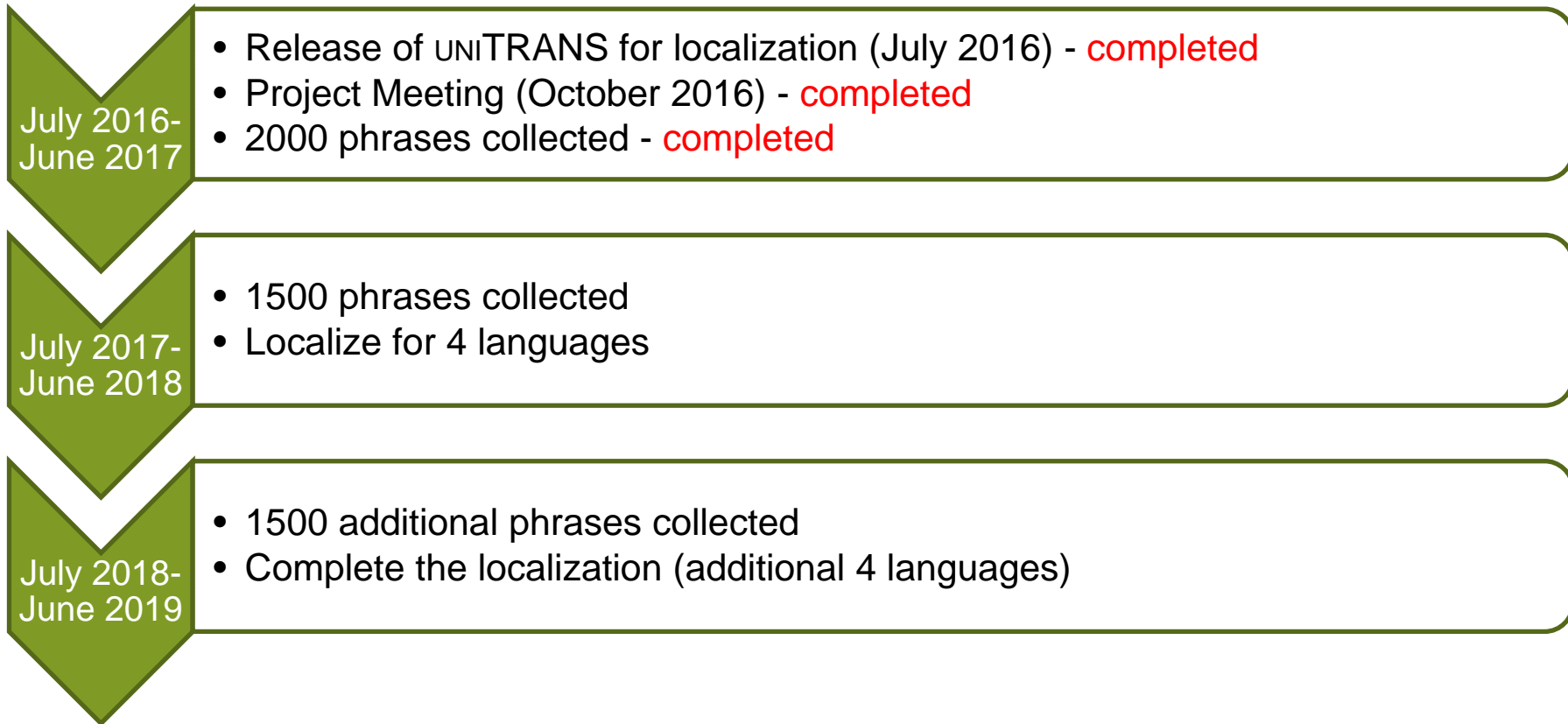
Duration
1st July 2016 to 30th June 2019

Target
➡ An integrated speech translation service for the ASEAN community.
➡ A common ASEAN speech translation evaluation language resource for ASEAN speech translation system evaluation.

Language Localization

Organization	Country	Language
NIPTICT	Cambodia	Khmer
BPPT	Indonesia	Bahasa Indonesia
UTM	Malaysia	Bahasa Melayu
UCSY	Myanmar	Myanmar
I^{2}R	Singapore	Chinese
NECTEC	Thailand	Thai
HUST	Vietnam	Vietnamese
IOIT	Vietnam	Vietnamese

Project Timeline

July 2016-June 2017	• Release of ᴜɴɪTRANS for localization (July 2016) - completed • Project Meeting (October 2016) - completed • 2000 phrases collected - completed
July 2017-June 2018	• 1500 phrases collected • Localize for 4 languages
July 2018-June 2019	• 1500 additional phrases collected • Complete the localization (additional 4 languages)

Speech Collection Specifications

Each of the 5000 phrases is recorded in a wave file with a close-talk microphone or a smart phone at 16KHz sample rate and 16 bit resolution at quiet environments. The quality by a close-talk microphone or a smart phone should be similar but you may want to select one of them.

ASEAN Speech Translation Data Collection

➡️ Completed the guideline for ASEAN Speech Data Collection in Dec 2016.

➡️ The guideline outlines 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

Current Status of ASEAN Speech Translation Data Collection

Organization	Country	Language	Status
NIPTICT	Cambodia	Khmer	• 10K utterances collected and translated • 4K utterances selected to record as a voice date. Recording in progress.
BPPT	Indonesia	Bahasa Indonesia	• 2000 utterances collected, translated and recorded
UTM	Malaysia	Bahasa Melayu	• 5000 utterances have been collected and translated. • Recording of 5000 utterances in progress (completed 8 subjects – 5000 utterances each)
UCSY	Myanmar	Myanmar	• 4000 utterances collected and translated. • Recording of 2000 utterances in progress.
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 • 30% recording progress

Audio Data Information Details

Organization	Device	Format	Speaker
UTM	• Samsung • iPhone	• 16KHz,16 bit rate, mono	• 4 male speakers • 5 female speakers
UCSY	• TASCAM DR-100MKIII	• 16KHz, 16 bit rate, mono	• 20 speakers
I^{2}R	• iPhone	• 16KHz,16 bit rate, mono	• 4 male speakers • 2 female speakers
HUST	• Recording tool (MySpeechRecord) • Samsung Galaxy A5 Apple iPhone 6s	• 16KHz,16 bit rate, mono	• 4 speakers
IOIT	• desktop computer	• 8KHz, 16bit, mono and pcm encoded • normal room environment	
NECTEC	• Apple iPhone via a specifically developed App	• 16 kHz, 16 bits, mono • PCM WAV • Quiet environment	• 12 male speakers • 12 female speakers

Upcoming Activities

➡ Project Meeting and Workshop
- 6th December in conjunction with IALP 2017
- Technical papers shared by team members and project update

➡ Continue with the data collection to develop a common ASEAN speech translation evaluation resource

Thank You!