ASEAN Language Speech Translation thru' U-STAR

27-28 November 2018
Indonesia

Dr. LI Haizhou, I²R, Singapore
Ms. AW Ai Ti, I²R, Singapore

Speaker: Dr. Emre Yilmaz, NUS, Singapore

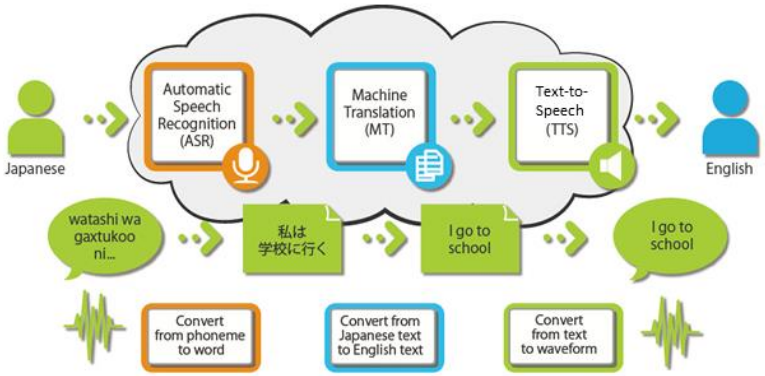
Project Members

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	Executive Director, National Electronics and Computer Technology Center (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



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

Universal Speech Translation Advanced Research

http://www.ustar-

Language Localization

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

ASEAN Speech Translation Data Collection Guidelines

Data Specifications

- Content of utterances: should be common and appropriate for general use and reflect a variety of practical situations

- Target Speakers: as many speakers as possible and well distributed across the ages 15-75 years (minimum of 5 male speakers and 5 female speakers)

- Speaking style and environment: fluent expressions, and these should be read naturally and fluently by the speakers (no fillers, hesitations, and long pauses are recorded)

Transcription & Translation Specifications

- Transcription of Numbers, Acronyms, Foreign words and names
- Punctuation Insertion & Translation of Number: as natural as possible (e..g phone number:12345678)
- Translation or Transliteration: always translate unless no English equivalent.

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	• 5000 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.
I^2R	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 • 70% 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^2R	• 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	• 4 male speakers • 4 female speakers
NECTEC	• Apple iPhone via a specifically developed App	• 16 kHz, 16 bits, mono • PCM WAV • Quiet environment	• 12 male speakers • 12 female speakers

Syllable-based Recurrent Neural Network for Myanmar Word Segmentation

Hsu Myat Mo, UCSY
Zhou Nina, Aw Ai Ti, I2R

Background

- Myanmar scripts are written continuously as a sequence of characters without any delimiter between words.
- Word is formed by characters and syllables with certain rules.

Method

- Existing methods use dictionary, rule, traditional machine learning approaches e.g., HMM or CRF.
- Investigate syllable-based segmentation scheme and use BLSTM for sequential labelling

Unsegmented sentence: အတိတ်မှသင်ခန်းစာသည်အနာဂတ်ရဲ့အားအင်။

Syllable-based segmented sentence: |အ/Bတိတ်/Iမှ/B|သင်/Bခန်း/Iစာ/I|သည်/B|အ/Bနာ/Iဂတ်/I|ရဲ့/B|အား/Bအင်/I။/B|

Data

#	Sent.	Word	Syllable	Char
Train	47K	1.0M	1.6M	4.785M
Dev.	2K	46,500	59,296	151,447
Test	2K	38,000	60,004	226,583

CRF Method

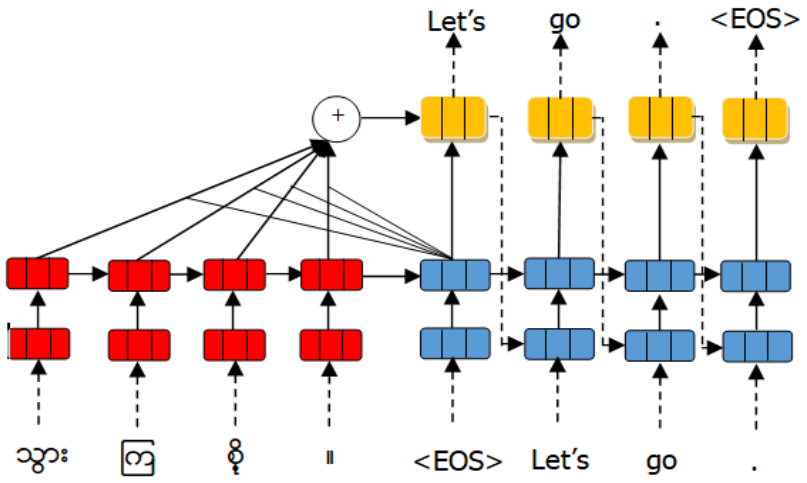
CRF (Dev.)	Prec	Recall	F1-score
Char-based	94.03	93.93	93.98
Syllable-based	93.90	94.08	93.99

BLSTM Method

Models		Precision / Recall / F1-score	
		Dev.	Test
BiLSTM (SGD)	Char	89.40/ 88.48 / 88.94	81.67/ 84.13 / 82.88
	Syllable	93.67 / 94.0 / 93.83	92.73/ 92.40 / 92.56
BiLSTM-CRF (SGD)	Char	88.23/ 87.35 / 87.79	82.32/ 83.04 / 82.68
	Syllable	92.03/ 93.59 / 92.80	91.20/ 91.90 / 91.55

Myanmar-to-English Syllable-based Neural Machine Translation System

Yi Mon Shwe Sin, UCSY
Wu Kui, Aw Ai Ti, I2R



- Word level neural machine translation (NMT) cannot model rare words in translating languages with rich morphology. Therefore, a syllable-based NMT is presented by segmenting the source sentences at syllable level using a Myanmar syllable segmenter
- Syllable-based NMT cannot translate unknown words well and a Name Entity database is incorporated to improve name entity translation

System	BLEU
Word-based NMT	21.88
Character-based NMT	20.71
Syllable-based NMT	26.50
Syllable-based NMT+ NE-database	27.94

Myanmar-English	Parallel Sentences
Original Corpus	228,767
NE-database	27,024
Total	255,791


Thank
You!