Program for USTAR Workshop

4th October

Venue: Connexis Building, Fusionopolis

0830 to 0900 – Registration

Technical Session

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	**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
1040 to 1100	**Break**
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 1140	**Extracting Parallel Sentences from Movie Subtitles** *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

Project Meeting

2pm to 5:30pm – Project Discussion (with afternoon tea)

Dinner

6:00pm

Abstract

1. **Comparison of Grapheme-to-Phoneme Conversion Methods on a Myanmar Pronunciation Dictionary**
   Ye Kyaw Thu and Win Pa Pa

   Grapheme-to-Phoneme (G2P) conversion is the task of predicting the pronunciation of a word given its grapheme or written form. It is a highly important part of both automatic speech recognition (ASR) and text-to-speech (TTS) systems. In this paper, we evaluate six G2P conversion approaches: Adaptive Regularization of Weight Vectors (AROW) based structured learning, Conditional Random Field (CRF), Joint-sequence models (JSM), phrase-based statistical machine translation (PBSMT), Support Vector Machine (SVM) based pointwise classification, Weighted Finite-state Transducers (WFST) on a manually tagged Myanmar phoneme dictionary. The G2P bootstrapping experimental results measured with both automatic phoneme error rate (PER) calculation and also manual checking in terms of voiced/unvoiced, tone, consonant and vowel errors. The results show that CRF, PBSMT and WFST approaches are robust for G2P bootstrapping on Myanmar language.

2. **The effect of dialect on the syllable accuracy of Vietnamese continuous speech recognition system**
   NGUYEN Hong Quang, TRINH Van Loan

   This paper presents the effect of dialect on Vietnamese continuous speech recognition system's performance. In this paper, the speech recognition system was built by the Kaldi toolkit. First, experiments examined the choices of the parameters for better recognition results. Secondly, we evaluate recognition results of each Vietnamese dialect's model and the general model trained from training data of both three dialects. Finally, the influence of the gender is also analyzed. The experiments on Northern data containing only male voices show that the trigram language model archives 19.7% less error than bigram. With Mel-frequency cepstral coefficients (MFCC) and fundamental frequency F0, and tone information in lexicon, the word error rate is reduced by 53% from the baseline in which the system using only MFCC and no tone in lexicon. The Northern, Central, and Southern trained model can reduce the error rate by 20.7%, 13.8%, and 12.5% respectively compared to general model trained from both three dialects' data.

3. **Vietnamese LVCSR Development and Improvement**
   Van Huy Nguyen, Quoc Bao Nguyen, Chi Mai Luong, Tat Thang Vu

   This paper presents some IOIT's contributions over recent years. We focused on evaluating and improving the performance of Vietnamese LVCSR with tones. For the VoiceTra project, we developed a new and better Vietnamese ASR engine that is further integrated to VoiceTarU4 software. We also published some new results on developing the toneme set, tonal acoustic models, and optimizing feature for Vietnamese ASR.

4. **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

This paper describes our research for developing Indonesian speech-to-speech translation (SST) system. The research aims to advance the state-of-the-art Indonesian automatic speech recognition (ASR), statistical machine translation (SMT) and text-to-speech (TTS) techniques. The project combines Indonesian ASR system that developed earlier and two undergoing projects which are Indonesian SMT and TTS system. The Indonesian ASR system was developed using hidden Markov model (HMM) based engine, the Indonesian SMT system is developed using statistical method while the Indonesian TTS system is developed using HMM-based synthesis method. In developing those systems we use the following language resources: parallel corpus consists of around 250,000 sentences, 9 million sentences monolingual corpus, and more than 130 hours speech corpus. The SST system is designed as a web service and enables any ASR, MT and TTS other than the currently available ones, to be integrated so it can help the advancement of ASR, MT and TTS research in Bahasa Indonesia and any NLP related research in general.

5. **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]

Network-based Speech Translation system integrates automatic speech recognition, machine translation and speech synthesis through network to remove language barriers by translating speech from one language to other language. The U-STAR (Universal Speech Translation Advanced Research) Consortium is a global research collaboration group to support these activities. This paper reports the work done under the *ASEAN Language Speech Translation thru U-STAR* project supported by ASEAN-IVO to promote the use of U-STAR platform for translating ASEAN languages. We introduce the universal speech translation freeware, UNITRANS developed by Institute for Infocomm Research (I^2R) and project members. We also discuss the Web service APIs based on REST and JSON for developers to develop new applications and services leverage on the U-STAR platform.

6. **Context-dependent Bilingual Word Embedding with Sentence Similarity Constraint for Machine Translation**
Kui Wu, Xuancong Wang, Ai Ti Aw

In this work, we propose a context-based bilingual word embedding framework that leverages the information of large amount of parallel sentence pairs which share the same semantic meaning. Such information is abundantly available but has not been fully utilized in previous work of context-based bilingual word embedding models, which only exploit local contextual information through a short window sequence at the word level. To incorporate such information, we define a sentence similarity matching objective which is enforced as a constraint into the original bilingual word embedding objective. They are jointly optimized to better learn the bilingual word embedding. Experimental results show that the proposed model is superior to previous methods on machine translation quality.

7. **Extracting Parallel Sentences from Movie Subtitles**
   Boon Hong Yeo, Ai Ti Aw, Xuancong Wang

   Parallel corpus is a mandatory resource for developing machine learning based statistical translation engine. The size and coverage of parallel corpus available for training affects the translation accuracy of the engine directly. To have more training data available for developing the translation engine for conversational domain, we propose a method to extract parallel data from Movie Subtitles using dynamic time warping, cosine similarity and beam search algorithm. The proposed method extracts 30% of parallel sentences from a set of Indonesian-English movie subtitles with an accuracy of 98%.

8. **An approach for Vietnamese-Japanese Statistical Machine Translation (SMT)**
   NGUYEN Thi Thu Trang, LE Thanh Huong

   This work aims at spoken text statistical machine translation of Vietnamese-Japanese. Firstly, a SMT baseline system for this language pair was built using Moses and Giza+, and experimented with existing corpora from OPUS/Ted talks and different word segmentation tools. Vitk and MeCab were finally adopted for Vietnamese and Japanese word segmentation due to their quality. The *GNOME* corpus (250,000 sentence pairs) from OPUS received the best BLEU score of 57.5 for Japanese-Vietnamese and 64.2 for Vietnamese-Japanese. The one from Ted talks (53,000 sentence pairs) received the worse score of 7.0 and 6.2 responsively. The reasons were found in (i) the translation quality of these corpora, (ii) the word density of these corpora, and (iii) some language problems such as the difference in lexical boundary or in word re-ordering of the two languages. Since most of spoken texts are not syntactically corrected, the phrase-based approach is adopted for the re-ordering task. Reordering rules are extracted from inconsistent blocks, i.e. phrases that words inside the phrase are not all aligned to each other, but to words outside the phrase. Some on-going works are (i) Applying extracted re-ordering rules to accelerate the quality of the translation and (ii) Improving the quality of an existing parallel corpus.

9. **Natural Language Processing Development Trends in Malaysia and the Way Forward**
   Sevia Mahdaliza Idrus, Rubita Sudirman, Faizah Mohamad Nor

   In the field of speech translation, the Malay Language speech translation apps are not readily available. There have been studies done on Malay language processing and which are still on-going; however, but these are not done on a large scale and neither are they intended for public use. Due to this lack of development in the area of Malay speech translation, this study will fill in the void in this field. Therefore, this paper reports the need for the uniTRANS app localization for the Malay Language due to the limited readily available corpus in the language. The local apps is targeted to consist of 5000 common Malay phrases, that would be beneficial for mainly travelers for their ease of communication with the locals, due to its ability to translate from Speech-to-Speech (S2S) and Speech-to-Text (S2T), Text-to-Speech (T2S) and Text-to-Text (T2T).