Activities on Speech and Machine Translation in Vietnam

Luong Chi Mai
lcmai@ioit.ac.vn
Institute of Information Technology
Vietnam Academy of Science and Technology

Outline

- Situation on development of Vietnamese Language and Speech Processing
- International Collaboration
- VLSP community's activities
- Conclusion and Suggestion

Vietnamese language

- Vietnamese language was established a long time ago
- Chinese characters was used for a long time
- Unique writing system of Vietnam called Chu Nom (字喃) in the 10th century
- Romanced script to represent the Quốc Ngữ since the beginning of the 20th century

Nam quốc sơn hà Nam đế cư
南 国 山 河 南 帝 居

Over Mountains and Rivers of the South, Reigns the Emperor of the South

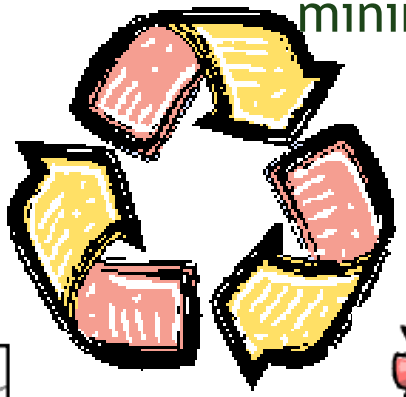
Vietnam Language and Seech Processing(VLSP) National Project (KC01.01.05/06-10)

National project with eleven active research VLSP groups Hanoi - Ho Chi Minh City with two objectives:

Building VLSP infrastructure, especially indispensable resources and tools for the VLSP development.

Building and developing several typical VLSP products for public end-users.

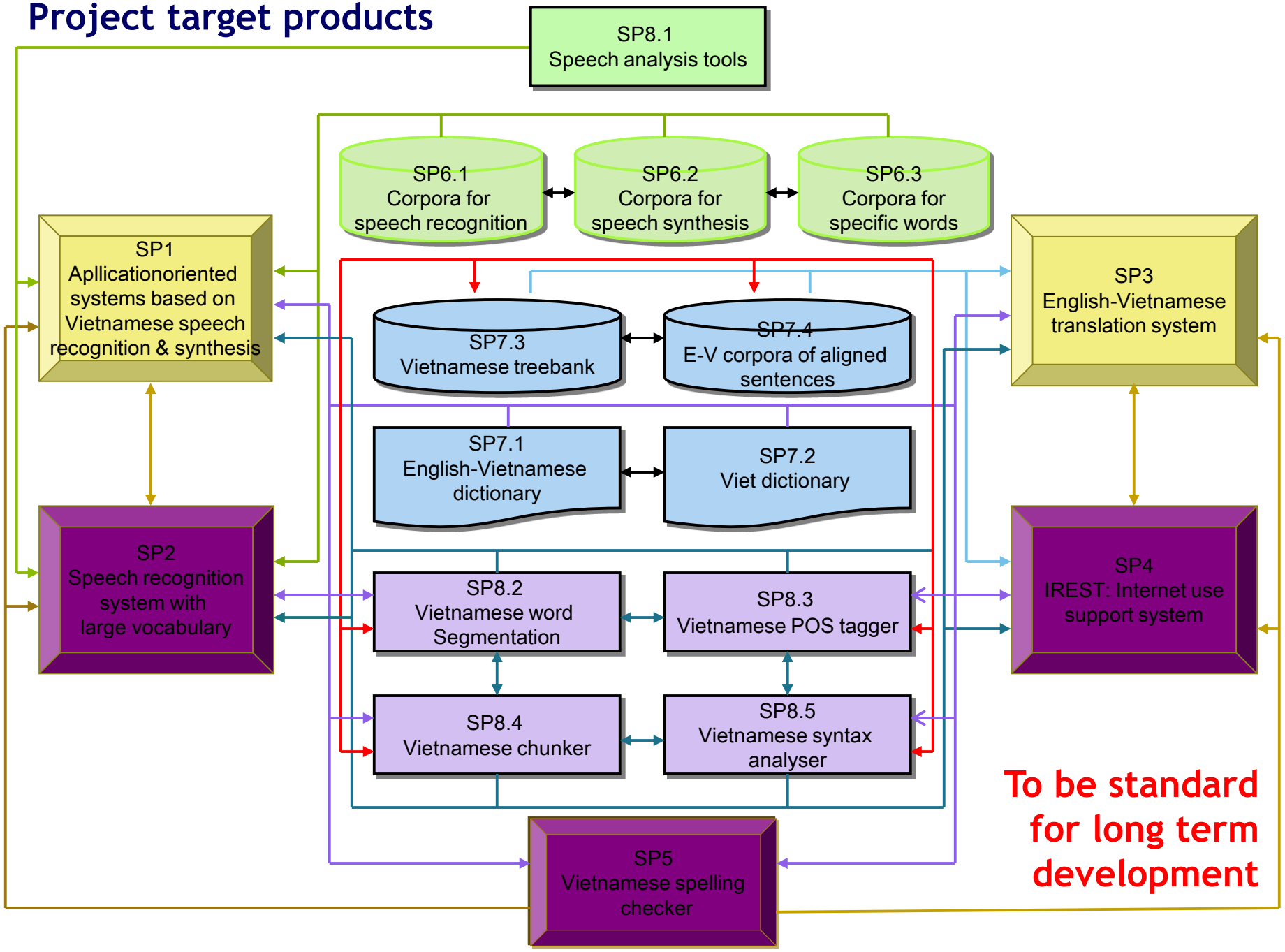
Pragmatics: Speech, text and Web data mining



Natural language processing methods

Tools, corpora, resources

Project target products



VLSP website: open to the public

http://vlsp.vietlp.org:8080/demo/?page=resources

NLP tools + resources

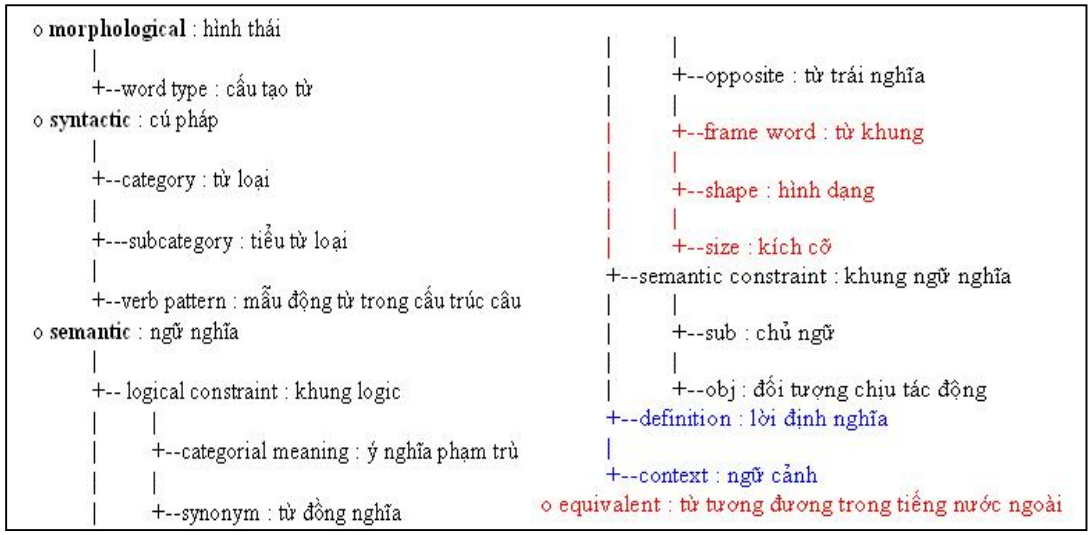
- All the tools: Word segmentation, POS tagging, Chunking, Syntax analysis are constructed based on the same view of words, label assignment, sentences, Viet dictionary and Viet Treebank.
- Using statistical and machine learning methods in building such tools.
- All the tools and resources is given to the R&D community.

NLP Tools

- Word segmentation
  - Methods: n-gram + dictionary + regular expression
  - 97,1% based on VieTreebank with annotated 220.000 vietnamese words
  - 98,2% based on 100 sentences not included in VieTreebank
- POS tagger
  - Methods: MEMs, CRFs
  - Training: 20.000 VN sentences with POS taken from Viet Treebank and VN dictionary
  - 90%
- Syntactic parser 1
  - Method: HPSG grammar
  - P = 82%, R = 74%, F-score = 78% tested on 100 sentences in VieTreebank
- Syntactic parser 2
  - Method: LPCFG, Bikel's implementation
  - F-score = 78% tested on 9600 sentences in VieTreebank
- Chunker
  - CRF, online learning based on more than 9.000 sentences with POS
  - 94%

NLP Resources

- VietTreebank
  - 10,000 trees; 1,000,000 morphemes
  - Tools: text graphical edit, log and history view, agreement check, search by words, syntactic patterns
- Vietnamese Machine Readable Dictionary
  - Model of VCL (Vietnamese Computational Lexicon) by learning from other language's MRDs with morphological, syntactic and semantic information.
  - 35,000 Vietnamese common used words in modern Vietnamese
  - Develop a tool for building VCL with XML representation



National Project for buiding Vietnamese WordNet 2012-2015

- Developing Vietnamese WordNet with the following features:
  - Vietnamese WordNet with 50.000 words (30.000 popular words and 20.000 domain-based)
  - 30.000 synset
  - Accuracy: 95% for terms in the same synset, 90% in the relationship between different synsets
  - Develop API for WordNet users
  - Develop a tool to access, verify and update
  - Propose guideline for long term WordNet development

Speech Resource

- **IOIT2013 - Daily News – Reading Corpus**
  - Phonetically-balanced 9200 sentences chosen from 5M sentences of daily news websites
  - Recorded in clean speech, microphone Sennheiser HMD410-6, *48* kHz sampling rate, 16 bit /sample.
  - Dialect : Northern
  - #Speakers : around 500 speakers (1/2 F, 1/2 M) ≅ 200 hours
  - #Utterances: ≅ 500 sent./spk
- **VOV – Broadcasting Corpus**
  - Clean speech
  - Dialect : Northern
  - #Sentences: ≅ 24.000 sentences
  - #Speakers : 30 Broadcasters (More F than M) ≅ 20 hours
  - Syllable Alignment, text included
- **VoiceTra - Small telephone speech corpus**
  - collecting data from app VoiceTra4U (outsite) in Apple Store (users can download it)
  - 2.6 hours, 841 speakers, 3K sentences (questions, answers) → **VoiceTra-dev**: only 35.9 minutes has meaning, 18.7 minutes, 803 utterances
  - (we are a partner in U-STAR consortium for speech-to-speech translation. You can find info of the Consortium form U-STAR website)

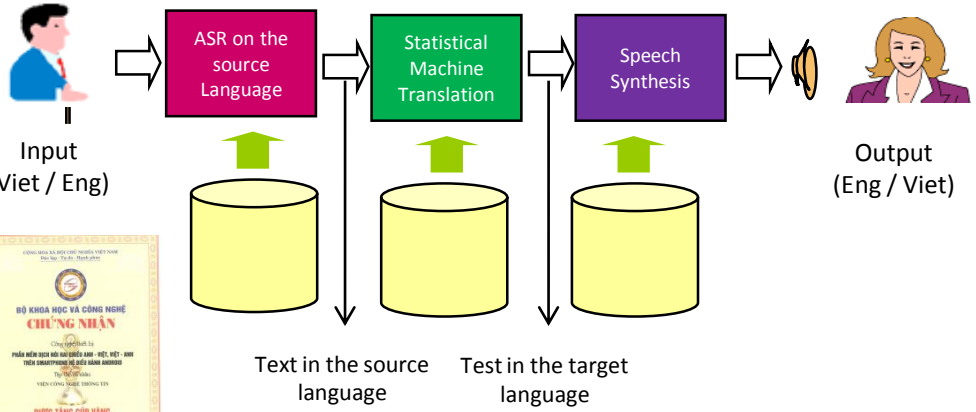
Vietnamese ASR

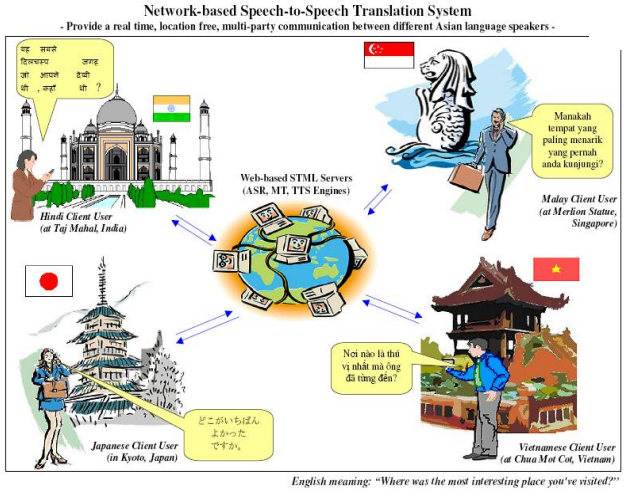
- **Vietnamese ASR - AM & LM models:**
  - Analysis the significant sets of lexicon units for Vietnamese Speech recognition: 45 units: 19 vowels (include diphthongs and semivowels), 25 consonants, 1 SIL
  - Feature Parameter: 16 kHz Sampling Frequency, 20 ms frame length and 10 ms frame shift
  - Training on VOV and part of IOIT2013 (only 1/3 of voices)
  - Test on BTEC test
  - 94,69% accuracy

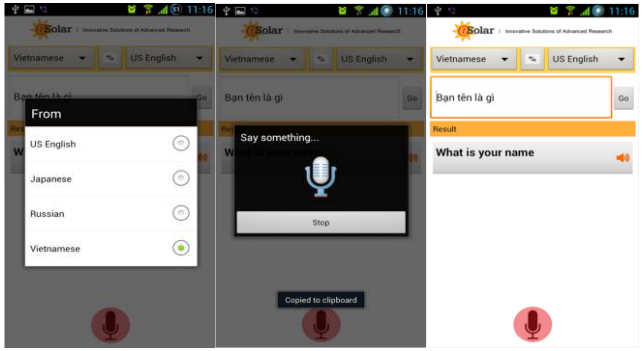
National Project KC01.03/11-15: "Research and Development of Speech to Speech Translation Technology for bidirectional Vietnamese-English in communication of travelling domain", 1/2011-12/2013

CONTRIBUTION

- **iSolarSpeech** – S2S bidirectinal Vietnamese-English system on Android smartphone.
- Posibility to extend to other pair of languages.







Specification of iSolarSpeech

- Communication on both North and South dialect of Vietnam
- Not very much noisy environment
- Travelling domain

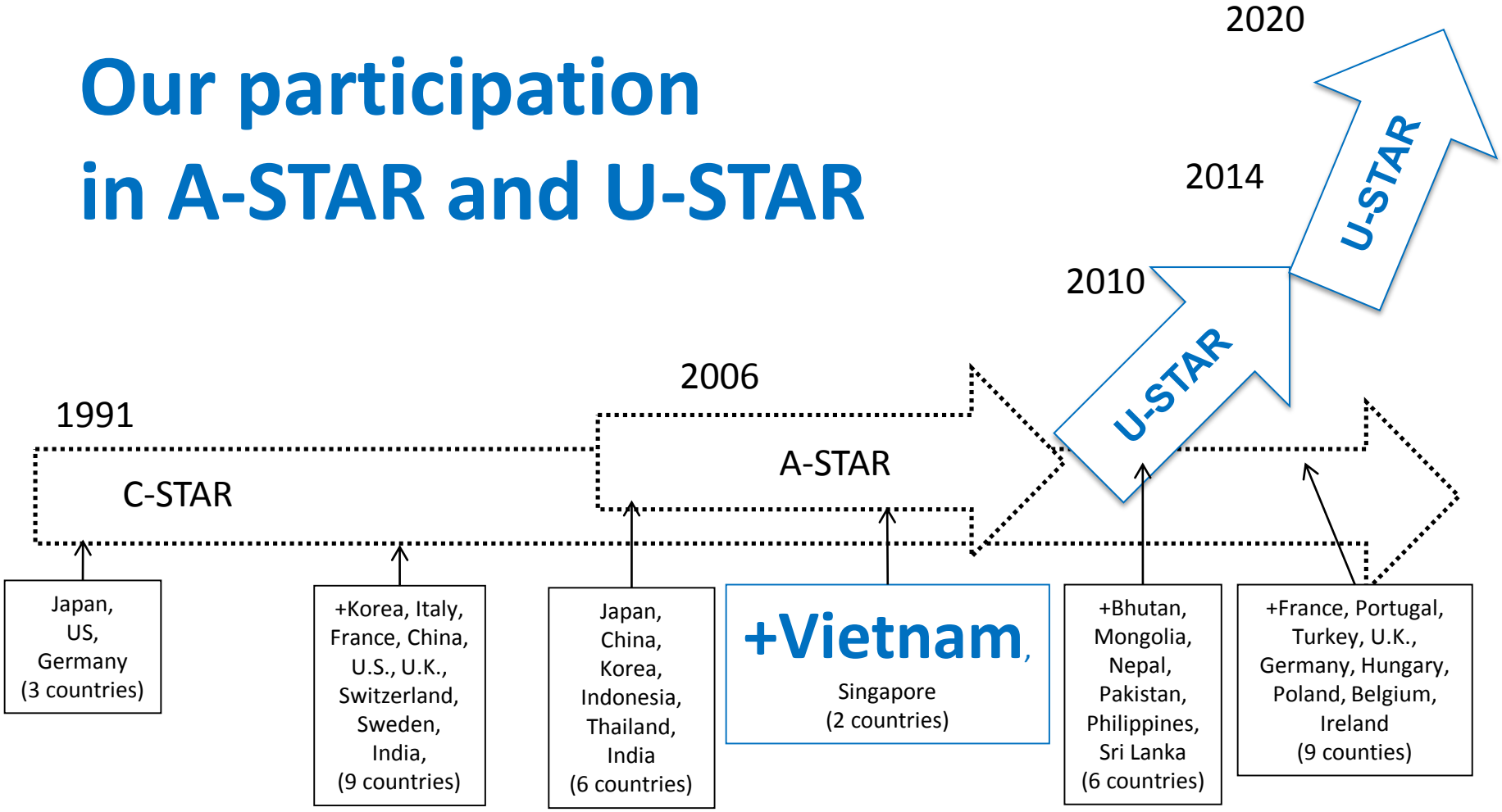
Outline

- Situation on development of Vietnamese Language and Speech Processing
- International Collaboration
- VLSP community's activities
- Conclusion and Suggestion

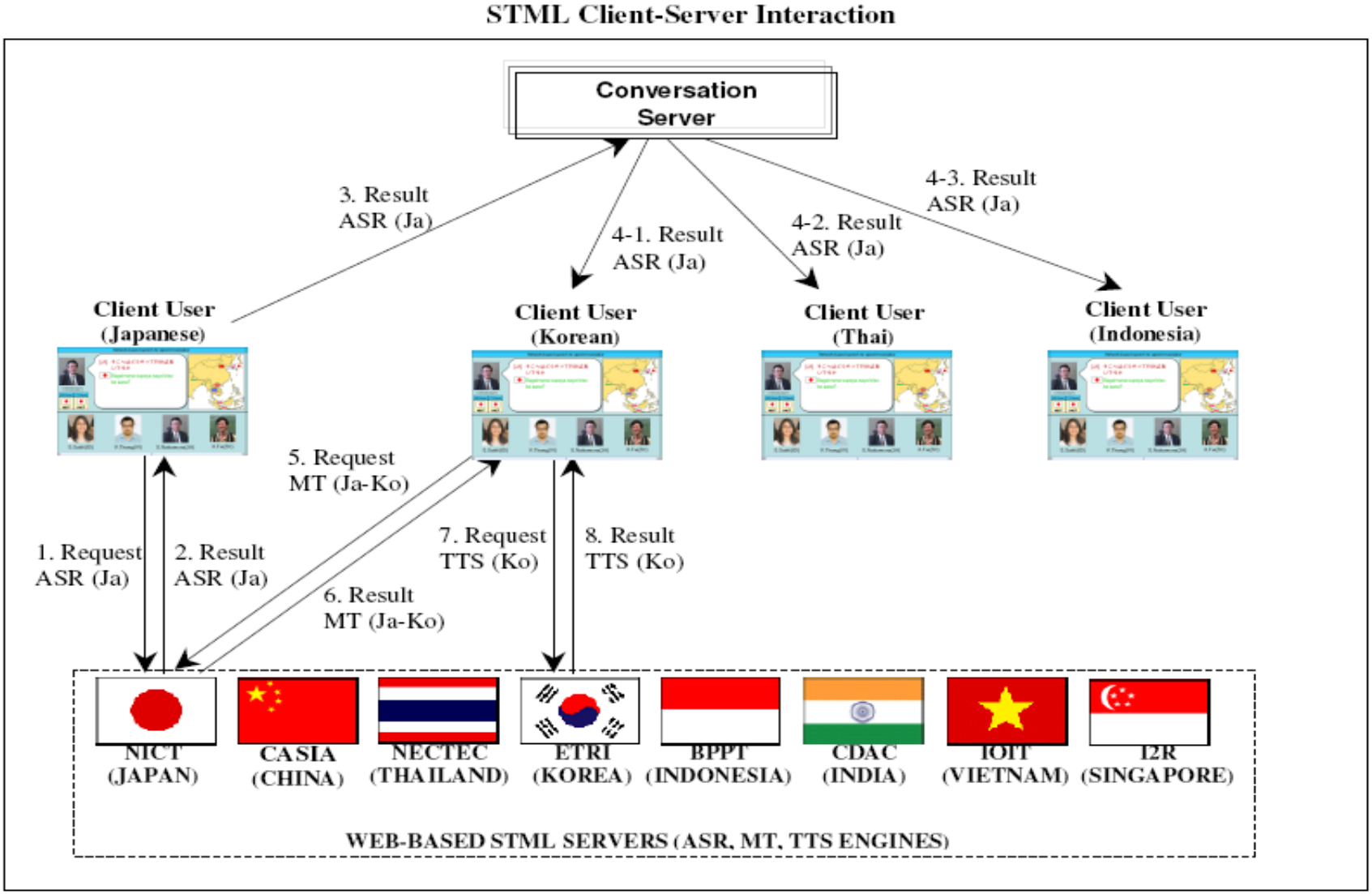
International Collaboration

- A-STAR (Asian Speech Translation Advanced Research), 2008-2010
  U-STAR (Universal Speech Translation Advanced Research), 2010 – till now
  Lead by NICT - Japan
- Member of “Network-based ASEAN Languages Translation Public Service Project”, 2012- 2015. Lead by NECTEC – Thailand
  - The communication among people in the ASEAN region has increased gradually and will become extreme especially after 2015 when the ASEAN Community begins. The automatic machine translation (MT) system has become more and more important to facilitate the cross-language communication, but has been limited for ASEAN countries.
  - Sharing language data
  - Develop platform
  - Integration of translation system

Our participation in A-STAR and U-STAR



Participation in A-STAR



Recognition Results

- **Recognition Accuracy Rates (%) on BTEC test set) A-STAR:**

Language	Phn	Dialect	Speaker (m, f)	Utterances	Total time	WA
Japanese	26	No accent	4200 (1600, 2600)	172,674	270.9	94.87%
English	44	3 (US, BRT, AUS)	532 (266, 266)	207,724	202.0	92.29%
Chinese	85	4 (BJ, SH, CT, TW)	536 (268, 268)	207,257	249.2	90.65%
Indonesia	33	4 (JV, SN, BT, ST)	400 (200, 200)	84,000	79.5	92.47%
Vietnamese	45	Northern	30	23,424	29.8	89.75%

- **Discussions:**
  - **Experimental results of Vietnamese SR is still below 90%, may because of:**

    → **Quality of training data**
  - **In real environment (demo system), It seems to work better for female-speakers**

    → **Data from female are most parts in data set.**
- **Improvement of accuracy (now): 94.69%**

Vietnamese ASR in U-STAR

- **VoiceTra4U – Baseline system**
  - AM: Kaldi gmm-bmmi
  - Speech data: IOIT2013
  - LM: 3-gram, using BTEC1 text
  - Dict: 9k words, G2P based on our Dictionary
  - Feature: MFCC
  - WER:
    - 61.01% on VoiceTra-dev
    - 16.16% on BTEC-dev
  - Development is going on, using DNN-based bottleneck features (BNF), more training data , retrain 3-gram LM on more text data, ... → improvement
    - On VioceTra-dev: IOIT2013+VOV+GlobalPhone+VoiceTra+BTEC1, 8745 of Vocab, BNF → WER: 27.73
    - On BTEC-dev: IOIT2013+VOV+GlobalPhone+VoiceTra+BTEC1+Webtxt, 9099 of Vocab, BNF: WER: 9.07

Outline

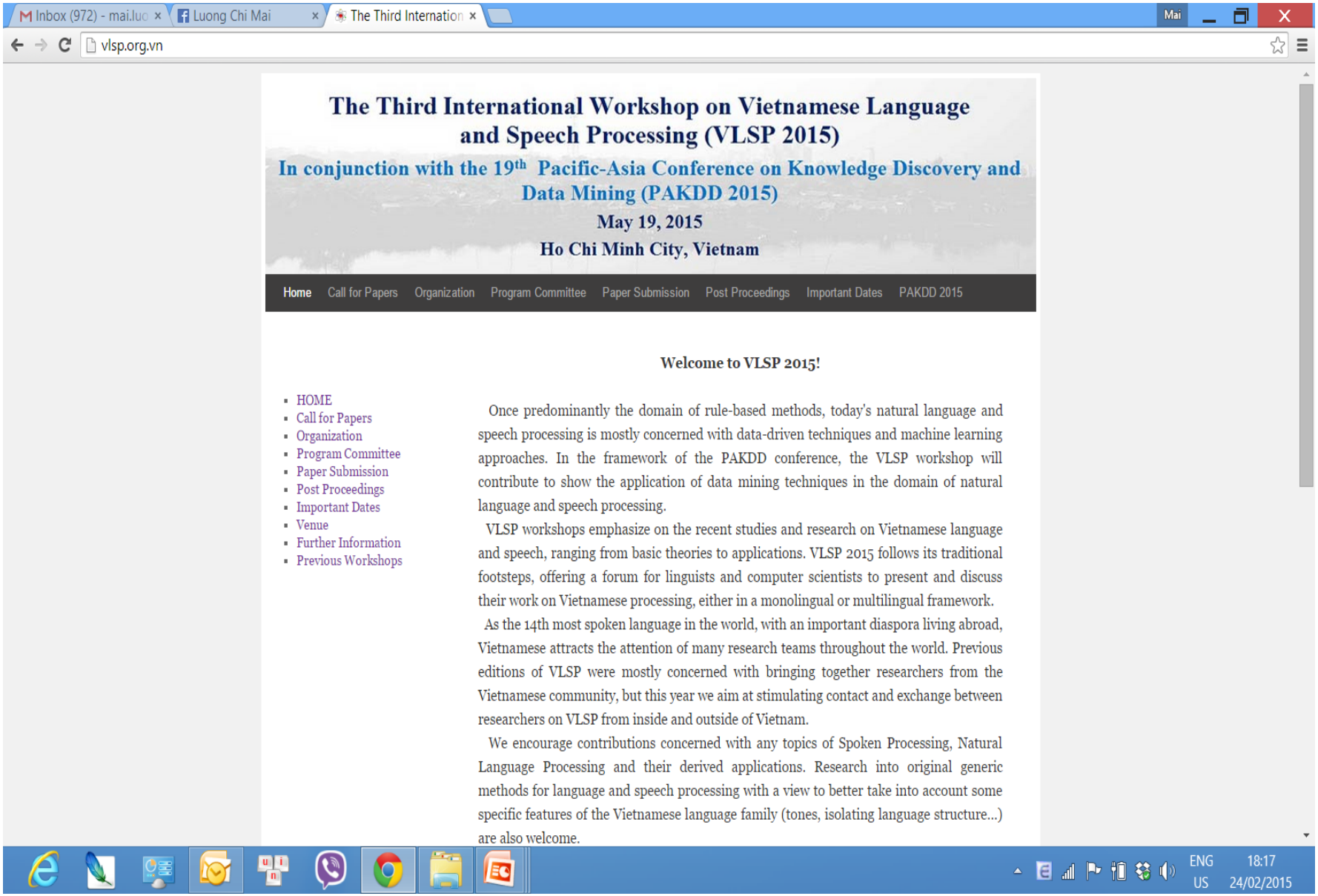
- Situation on development of Vietnamese Language and Speech Processing
- International Collaboration
- VLSP community's activities
- Conclusion and Suggestion

VLSP Activities

- Workshops on VLSP community (4 times) as a satellite events for IEEE-RIVF 2010, 2011, 2013 and upcoming PAKDD 2015 in HoChiMinh City

  http://vlsp.org.vn/
- Participation to a campaign IWSLT 2013 (Heidelberg, Gemany) and IWSLT 2014 (Lake Tahoe, CA, USA)
  - ASR for English
  - SMT for English-French
- Host IWSLT 2015 in Hanoi on Dec. 3-4, 2015

  Please join us: include other Asian Languages for a campaign (Chinese, Thai, Vietnamese,…)


Inbox (972) - mai.luo
Luong Chi Mai
The Third Internation
Mai
vlsp.org.vn
The Third International Workshop on Vietnamese Language and Speech Processing (VLSP 2015)
In conjunction with the 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2015)
May 19, 2015
Ho Chi Minh City, Vietnam
Home
Call for Papers
Organization
Program Committee
Paper Submission
Post Proceedings
Important Dates
PAKDD 2015
Welcome to VLSP 2015!
HOME
Call for Papers
Organization
Program Committee
Paper Submission
Post Proceedings
Important Dates
Venue
Further Information
Previous Workshops
Once predominantly the domain of rule-based methods, today's natural language and speech processing is mostly concerned with data-driven techniques and machine learning approaches. In the framework of the PAKDD conference, the VLSP workshop will contribute to show the application of data mining techniques in the domain of natural language and speech processing.
VLSP workshops emphasize on the recent studies and research on Vietnamese language and speech, ranging from basic theories to applications. VLSP 2015 follows its traditional footsteps, offering a forum for linguists and computer scientists to present and discuss their work on Vietnamese processing, either in a monolingual or multilingual framework.
As the 14th most spoken language in the world, with an important diaspora living abroad, Vietnamese attracts the attention of many research teams throughout the world. Previous editions of VLSP were mostly concerned with bringing together researchers from the Vietnamese community, but this year we aim at stimulating contact and exchange between researchers on VLSP from inside and outside of Vietnam.
We encourage contributions concerned with any topics of Spoken Processing, Natural Language Processing and their derived applications. Research into original generic methods for language and speech processing with a view to better take into account some specific features of the Vietnamese language family (tones, isolating language structure...) are also welcome.
ENG
US
18:17
24/02/2015

IWSLT 2014 Track
Welcome to Hanoi 3-4 Dec 2015

Conclusion and Suggestion

- Some activities and contribution to national and international level from Vietnam for multilingual translation.
- For further collaboration: mutual benefit-based sharing
  - Language resources (parallel corpus, Building corpus with POS-tagged, alignment, and treebank, …)
  - Basic tools and API
  - Tutorial courses
  - Shared task evaluation
  - Joint project in regional/global levels
  - Setting progress milestones