Current status of *Vietnamese Treebank*
usefulness of collaboration with *Asian Language Treebank*

VU TAT THANG

Dept. of Multimedia Human-Machine Language Technology,

Institute of Information Technology,
Vietnam Academy of Science and Technology.
Content

- IOIT and International Collaborations
- Vietnamese Language
- VLSP Standard
- Current status of Vietnamese Processing
- Propose idea
Content

- IOIT and International Collaborations
- Vietnamese Language
- VLSP Standard
- Current status of Vietnamese Processing
- Propose idea
IOIT – a member of ASEAN MT

- 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
IOIT – a member of A-STAR (U-STAR)

- A-STAR (Asian Speech Translation Advanced Research), 2008-2010
- U-STAR (Universal Speech Translation Advanced Research), 2010 – till now
Mechanism of S2s system

Vietnamese

I go to school

Tôi đi đến trường

Statistical Speech Recognition

Large Scale Vietnamese Speech Corpora

Corpus-based Speech Synthesis

Large Scale English Speech Corpora

Multilingual Speech Recognition

Spoken Language Translation

English

I go to school

Tôi đi đến trường

Statistical Machine Translation + Multi-engine approach

Large Scale Parallel Corpora of Vietnamese and English

Large Scale English Text Corpora

Large Scale Vietnamese Text Corpora
Content

- IOIT and International Collaborations
- Vietnamese Language
- VLSP Standard
- Current status of Vietnamese Processing
- Propose idea
Vietnamese Language

- Spoken as mother tongue by
  - 86% of Vietnam’s population
  - ~3 million overseas Vietnamese – most live in US
- It is part of the Austro-asiatic language family (168 languages)
- Many vocabulary has been borrowed from Chinese
- Writing system:
  - Formerly, Chinese writing system
  - Today: Latin alphabet, with additional diacritics for tones and certain letters
- Dialects: Northern, Central, Southern
Vietnamese language

- Vietnamese language was established a long time ago
- Chinese characters were 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

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

- IOIT and International Collaborations
- Vietnamese Language
- VLSP Standard
- Current status of Vietnamese Processing
- Propose idea
Setting up the VLSP “standards” for the public

- Importance of “standards” in VLSP: choose an unified view from various schools on Vietnamese language
- Guide for words recognition and description: morphological, syntactic, semantic criteria
- Guide for constituent labeling: noun phrase, verb phrase, clause, etc.
- Guide for sentence split
- Others
VLSP national project

National project with eleven active research groups on VLSP (Vietnamese Language and Speech Processing)

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

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

26/11/2015
ML and statistical methods in NLP

1992 ACL: 24% (8/34) some ML/Stat, 76% (26/34) no ML/Stat

1994 ACL: 35% (14/40) some ML/Stat, 65% (26/40) no ML/Stat

1996 ACL: 39% (16/41) some ML/Stat, 61% (25/41) no ML/Stat

1999 ACL: 60% (41/69) some ML/Stat, 40% (28/69) no ML/Stat

2001 NAACL: 87% (27/31) some ML/Stat, 13% (4/31) no ML/Stat

2005 ACL: 96% (74/77) some ML/Stat

Pages 11-12 from Marie Claire, ECML/PKDD 2005
Considering words "nhà cửa", "sắc đẹp", "hiệu sách". They are words in the following sentences:
a. Nhà cửa bè bộn quá
b. Cô ấy giữ gìn sắc đẹp.
c. Ngoài hiệu sách có bán cuốn này

And they are not words in:
a. Ở nhà cửa ngõ chẳng đóng gì cả.
b. Bức này màu sắc đẹp hơn.
c. Ngoài cửa hiệu sách báo bày la liệt.

Many tools such as ChaSen, Yamcha, …
Example: Guideline for POS tagging

- 36 word labels in English, from Penn Treebank (1989)
- 30 word labels in Chinese, from Chinese TreeBank (1998)
- 47 word labels in Thai, from Orchid corpus (1997)
- How many for Vietnamese?

<table>
<thead>
<tr>
<th></th>
<th>Chinese Treebank</th>
<th>Penn Treebank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tổng số thể</td>
<td>33</td>
<td>36</td>
</tr>
<tr>
<td>Danh từ (nouns)</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Danh từ chỉ thời gian (temporal nouns)</td>
<td>NT</td>
<td>NN, NNS</td>
</tr>
<tr>
<td>Danh động từ (verbal nouns)</td>
<td>NN</td>
<td>NN, NNS</td>
</tr>
<tr>
<td>Danh từ riêng (proper nouns)</td>
<td>NR</td>
<td>NNP, NNPS</td>
</tr>
<tr>
<td>Các danh từ khác</td>
<td>NN</td>
<td>NN, NNS</td>
</tr>
<tr>
<td>Localizer</td>
<td>1 (LC)</td>
<td>0</td>
</tr>
<tr>
<td>Đề từ (pronouns)</td>
<td>1 (PN)</td>
<td>4 (PRP, PRP$, WP, WP$)</td>
</tr>
<tr>
<td>Động từ</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Động từ khuyết thiếu (modal)</td>
<td>ВV</td>
<td>MD</td>
</tr>
<tr>
<td>Các động từ khác</td>
<td>ВV, ВА, ВС, VE</td>
<td>ВB, ВBD, ВBG, ВBN, ВBP, ВBZ</td>
</tr>
<tr>
<td>Trạng từ</td>
<td>1 (AD)</td>
<td>4 (RB, RBR, RBS, WRB)</td>
</tr>
<tr>
<td>Giới từ</td>
<td>1 (P)</td>
<td>1 (IN)</td>
</tr>
<tr>
<td>DP-related</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Định từ (Determiner)</td>
<td>DT</td>
<td>DT, WDT, PDT</td>
</tr>
<tr>
<td>Số</td>
<td>CD, OD</td>
<td>CD</td>
</tr>
<tr>
<td>Measure word</td>
<td>M</td>
<td>-</td>
</tr>
<tr>
<td>Liên từ</td>
<td>2 (CC, CS)</td>
<td>2 (CC, IN)</td>
</tr>
<tr>
<td>Tiếu từ (particles)</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Khác</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>Thần tử</td>
<td>IJ</td>
<td>UH</td>
</tr>
<tr>
<td>Sound word (từ tương thanh??)</td>
<td>ON</td>
<td>-</td>
</tr>
<tr>
<td>Punctuation</td>
<td>PU</td>
<td>-</td>
</tr>
</tbody>
</table>
Project target products

SP6.1 Corpora for speech recognition
SP6.2 Corpora for speech synthesis
SP6.3 Corpora for specific words

SP7.1 English-Vietnamese dictionary
SP7.2 Viet dictionary
SP7.3 Vietnamese treebank
SP7.4 E-V corpora of aligned sentences

SP8.1 Speech analysis tools
SP8.2 Vietnamese word Segmentation
SP8.3 Vietnamese POS tagger
SP8.4 Vietnamese chunker
SP8.5 Vietnamese syntax analyser

SP1 Application-oriented systems based on Vietnamese speech recognition & synthesis
SP2 Speech recognition system with large vocabulary
SP3 English-Vietnamese translation system
SP4 IREST: Internet use support system

SP5 Vietnamese spelling checker

To be standard for long term development
Content

- IOIT and International Collaborations
- Vietnamese Language
- VLSP Standard
- Current status of Vietnamese Processing
- Propose idea
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.
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
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
SP7.3: Viet Treebank

- A **Treebank** or **parsed corpus** is a text corpus in which each sentence has been parsed, i.e. annotated with syntactic structure.
  - **English**: Penn Treebank (4.5M words) and many others;
  - **Chinese**: Penn Chinese Treebank (507K words), Sinica Treebank (61,087 trees, 361K words);
  - **Japanese**: ATR Dependency corpus, Kyoto Text Corpus, Verbmobil treebanks;
  - **Korean**: Korean Treebank (5078 trees, 54K words)
- **Viet Treebank (2012):**
  - 10,000 trees
  - 1,000,000 morphemes

![Diagram](image-url)
SP7.3: Viet Treebank

- Study various existing treebanks, modern theories for syntax and Vietnamese language

- Build guidelines for word segmentation, POS, and syntax
  - “Nhà cửa bè borderline quá” and “Ô nhà cửa ngoại chẳng động gì cả” (“the house is in jumble” and “at home the door is not closed”)
  - “Cô ấy giữ giữ sắc đẹp” and “Bức này màu sắc đẹp hơn” (She keeps her beauty” and “this painting has better color”)

- Build the tools

- Labeling
  
  Agreement between labelers (95%)
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 sentences with POS from VieTreebank 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 on > 9,000 sentences with POS as in VieTreebank
  - 94%

26/11/2015
Content

- IOIT and International Collaborations
- Vietnamese Language
- VLSP Standard
- Current status of Vietnamese Processing
- Propose idea
We need **Asian Language Treebank**

- ALT is the key resources of most of Asian languages.
  - Word segmenter
  - POS tagger
  - Chunker
  - Syntactic parser
  - ... Search engines,
  - Information retrieval
  - machine translation
  - QA system
  - ...

- Can constructs from **multi-lingual corpora** among all **Asian languages** with
  - The same standard of infrastructure
  - The same kind of tool
  - ...
- Accelerates research of NLP for Asian languages
  - We have Treebank for English, Japanese, Vietnamese
  - How about Indonesian, Thai, Khmer, Laos, Malay, Myammar, Philippine..