Issues in Vietnamese Language Processing

Hồ Tú Bảo

Vietnamese Academy of Science and Technology

Japan Advanced Institute of Science and Technology

More languages than you might have thought

6912 distinct languages (230 spoken in Europe, 2197 in Asia)

- We meet here today to talk about Vietnamese language and speech processing.
- Aujourd'hui nous nous réunissons ici pour discuter le traitement de langue et de parole vietnamienne.
- Сегодня мы встречаемся здесь, чтобы говорить о обработке вьетнамского языка и речи.
- 今日我々はここに集まりベトナム語処理について議論します.
- 오늘 우리는 여기에 모여서 베트남어와 발언처리에 대하여 의론하겠습니다.
- أننا نجتمع هنا اليوم لنتحدث عن اللغة الفيتنامية و لغة الخطاب
- Hôm nay chúng ta gặp nhau ở đây để nói về xử lý ngôn ngữ và tiếng nói tiếng Việt.

Everyone can understand foreign languages

Imagine a day in some near future you can read everything and talk to anyone in any language …

Translation and machine translation

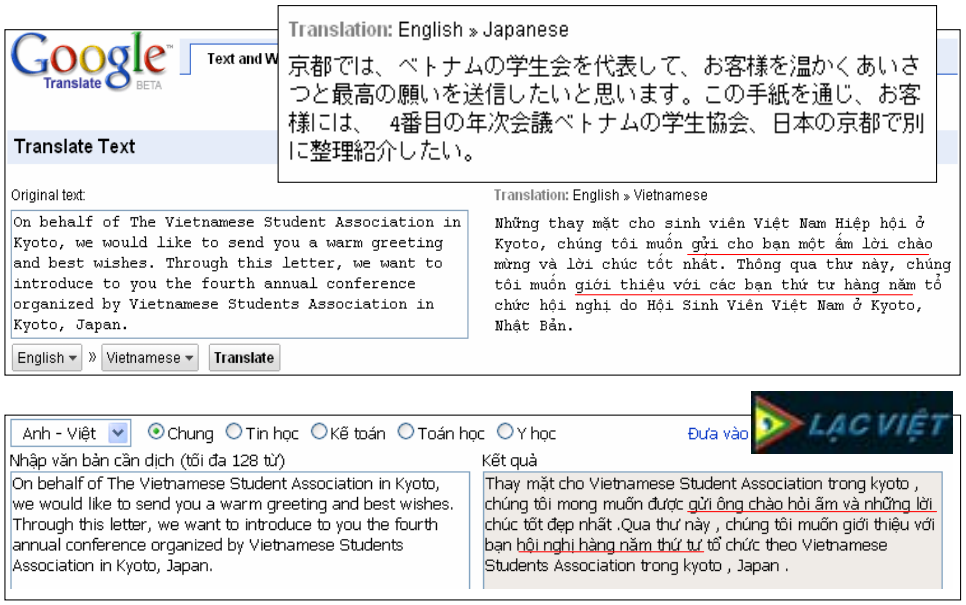
- Translate the following sentence into English

  "Ông già đi nhanh quá"?

- Many possible translations

1. [Ông già] [đi] [nhanh quá] → The old man walks too fast
   → My father walks too fast
2. [Ông già] [đi] [nhanh quá] → The old man died too fast
   → My father died too fast
3. [Ông] [già đi] [nhanh quá] → You get old too fast
   → Grandfather gets old too fast

Google: English-Vietnamese translation
26.9.08 (translate.google.com, 35 languages)



Two approaches to machine translation

Linguistic rule-based machine translation

- ➔ words are translated by using linguistic rules about the two languages the correspondence transfer between them (morphology, syntax, etc)
- ➔ Requires understanding of natural language

Statistical machine translation

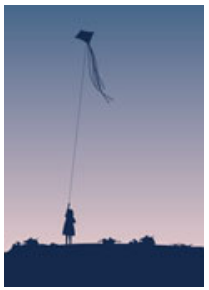
- ➔ generate translations using statistical learning methods based on bilingual text corpora (statistically similar)
- ➔ Requires large and qualified bilingual text corpora.

DOMINATING!

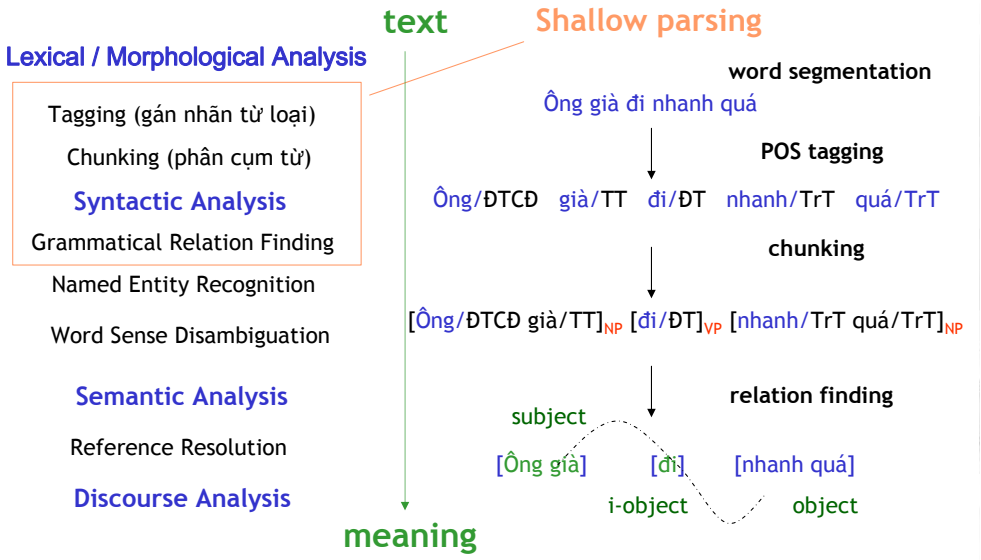
Natural language processing (NLP)

- Goal: automated language understanding 言語理解の自動化
  - ➔ this isn't possible 不可能
  - ➔ instead, go for sub-goals of text analysis, e.g., 下位目標として
    - ⇨ word sense disambiguation, phrase recognition, semantic associations
- Common current approach: statistical analyses over very large text collections 大規模テキスト集合を統計的に解析

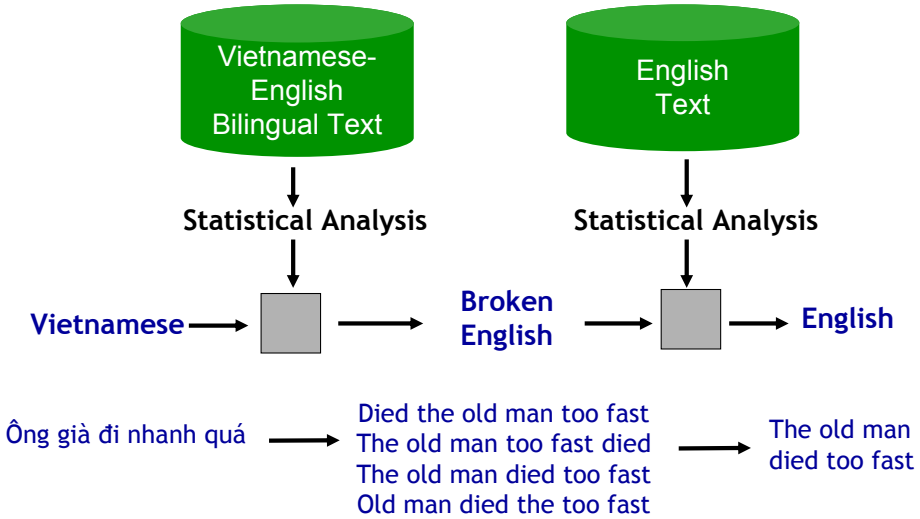
Consider a word like "string" or "rope." No computer today has any way to understand what those things mean. For example, you can pull something with a string, but you cannot push anything. You can tie a package with string, or fly a kite, but you cannot eat a string or make it into a balloon. In a few minutes, any young child could tell you a hundred ways to use a string – or not to use a string – but no computer knows any of this.



From text to the meaning
Natural Language Processing (NLP)

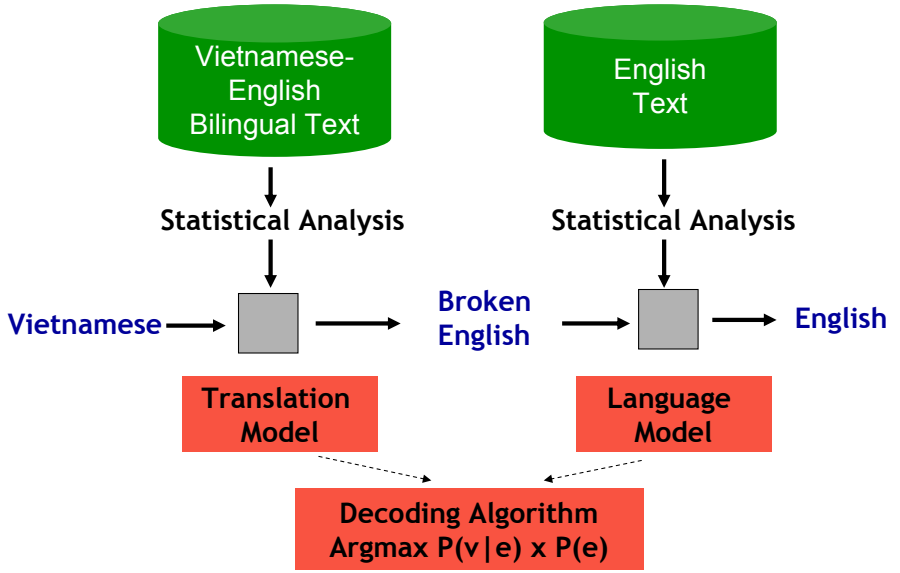


Statistical machine translation



(Slides 6-7 adapted from tutorial on SMT, K. Knight and P. Koehn)

Statistical machine translation

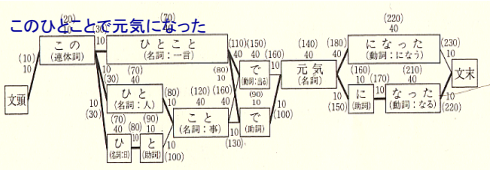


Vietnamese language

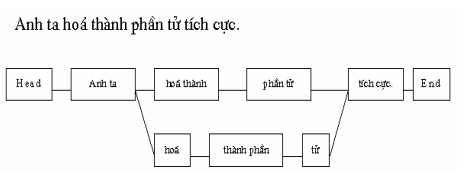
- Vietnamese is an analytic (words are composed of a single morpheme) language.
  - ➔ ngôn ngữ (analytic), lang-gua-ge (synthetic), 言語 (synthetic)
- Vietnamese does not use morphological marking of case, gender, number, and tense.
  - ➔ Trưa nay tôi ăn ba thằng tôm
- Syntax conforms to Subject Verb Object word order
  - ➔ Cái thằng chồng em nó chẳng ra gì.
    FOCUS CLASSIFIER husband I he not turn.out what
    "That husband of mine, he is good for nothing."
- The written language uses the Vietnamese alphabet ("national script"), based on the Latin alphabet.

About Vietnamese language processing

- Work on machine translation or in top layers but less basic work at lower layers
- Lack of common itinerary
- Work done in isolation, no inheritance → people have to do their work from the scratch without sharing and collaboration
- Almost no resources and tools for VLSP



Many tools such as ChaSen, Yamcha, …



No tool to do such a simple task

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.

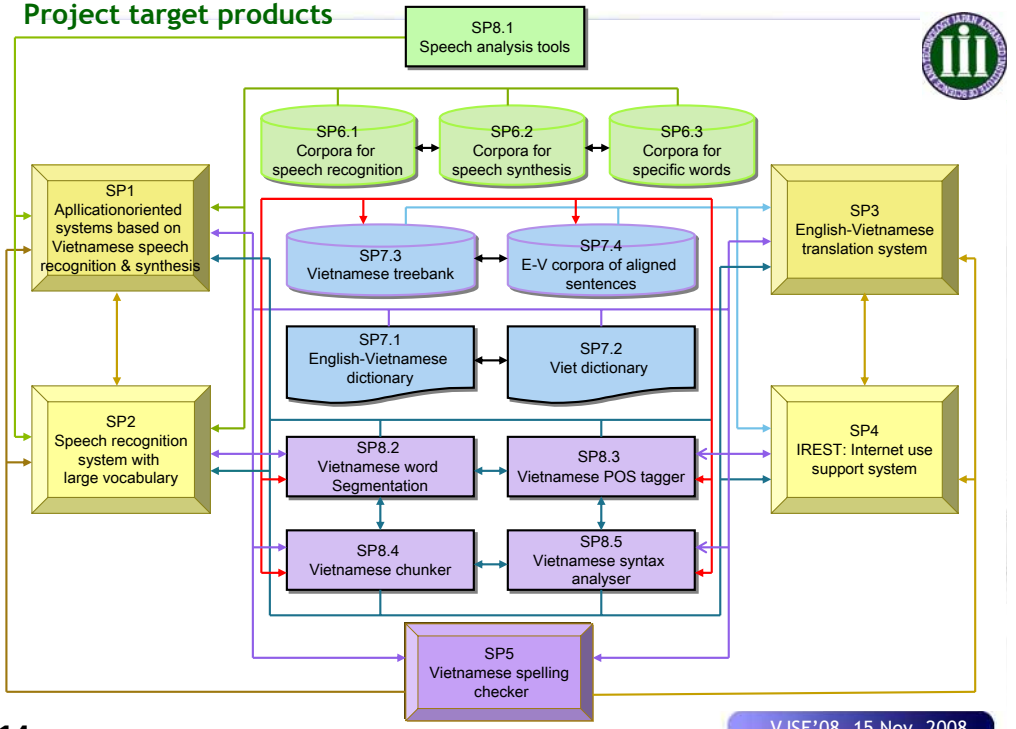
Building and developing several typical VLSP products for public end-users.

Pragmatics: Speech, text and Web data mining

Natural language processing methods

Tools, corpora, resources

Project target products

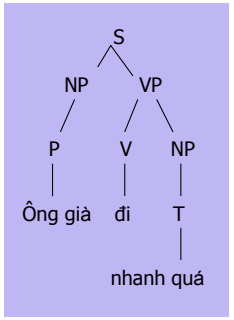


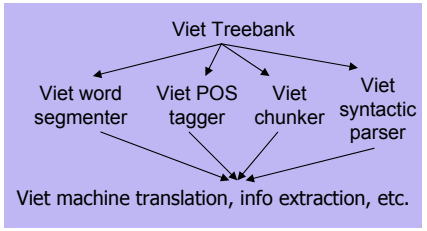
Setting up the "standards" for VLSP

- VLSP: Vietnamese Language and Speech Processing
- 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

Viet Treebank

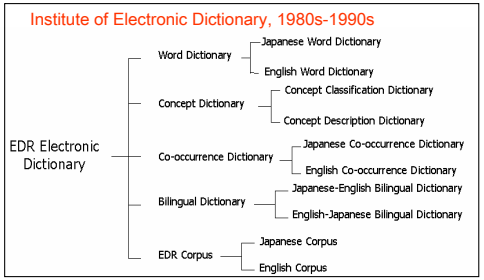
- A **Treebank** or **parsed corpus** is a text corpus in which each sentence has been parsed, i.e. annotated with syntactic structure.
  - → **English**: Penn Treebank (4.5M words) and many others;
  - → **Chinese**: Penn Chinese Treebank (507K words), Sinica Treebank (61,087 trees, 361K words);
  - → **Japanese**: ATR Dependency corpus, Kyoto Text Corpus, Verbmobil treebanks;
  - → **Korean**: Korean Treebank (5078 trees, 54K words)
- **Viet Treebank (7.2007-5.2009):**
  - → 10,000 trees
  - → 1,000,000 morphemes

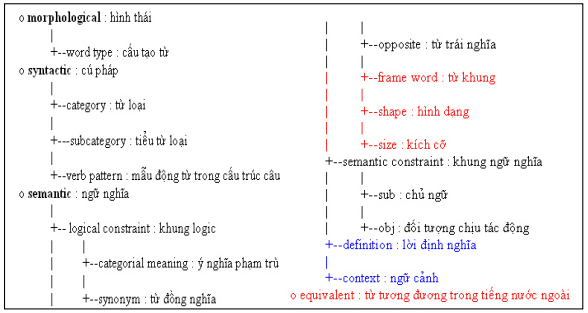




Vietnamese Machine Readable Dictionaries

- Build a model of VCL (Vietnamese Computational Lexicon) by learning from other language's MRDs.
- 35,000 Vietnamese common used words in modern Vietnamese
- Develop a tool for building VCL with XML representation.





English-Vietnamese parallel corpus

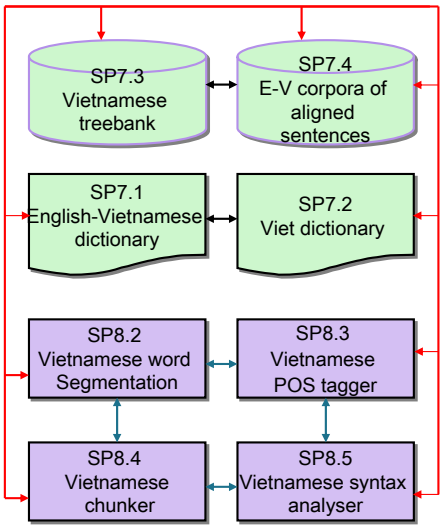
- Set of many pairs of corresponding sentences in English and Vietnamese
- Importance: Size and quality (LDC: English-French corpus of 2.8M sentences, source from Canadian Parliament)
- Our corpus in its first phase: 100,000 sentences pairs
  - ➔ Manual and semi-automatic collection of parallel text
  - ➔ Automatic alignment

Parallel Corpus (L1-L2)	Sentences	L1 Words	L2 Words
German-English	1,313,096	34,700,362	36,663,083
Greek-English	662,090	18,834,758	18,827,241
Spanish-English	1,304,116	37,870,751	36,429,274
Finnish-English	1,257,720	24,895,790	34,802,617
French-English	1,334,080	41,573,117	37,436,222
Italian-English	1,251,315	36,411,166	36,510,033
Dutch-English	1,326,412	36,784,168	36,690,392
Portuguese-English	1,287,757	37,342,426	36,355,907
Swedish-English	1,164,536	28,882,142	32,053,628

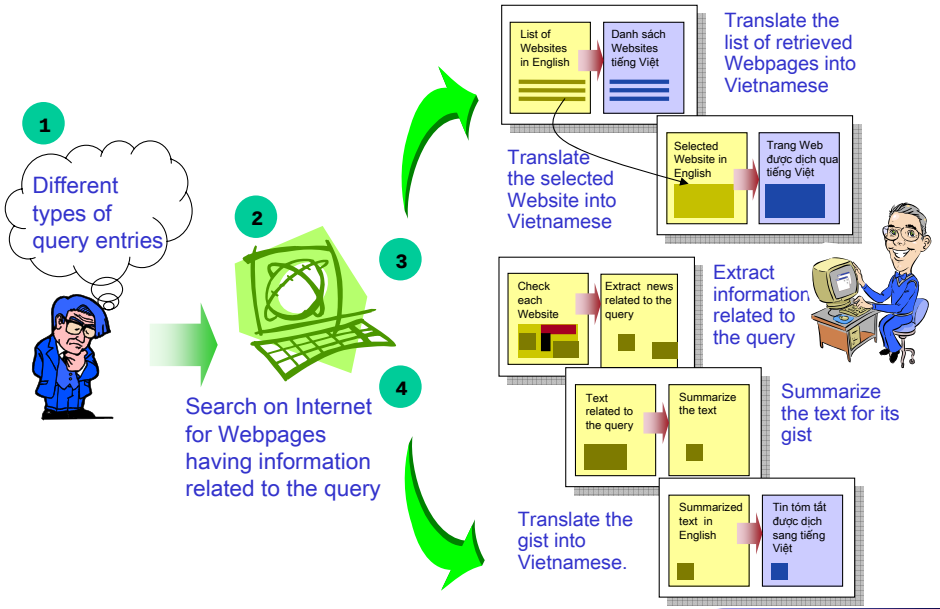
(http://www.euromatrix.net)

VLSP tools

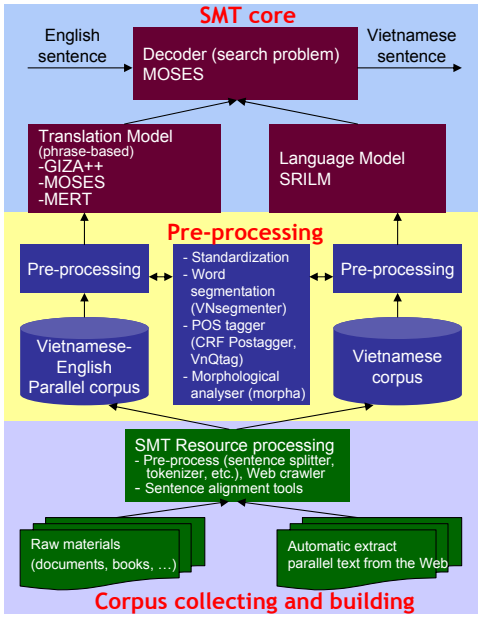
- All the tools are constructed based on the same view of words, label assignment, sentences, and resources.
- Using statistical and machine learning methods in building such tools.
- Tools and resources will be given to the public.



IREST: Support for exploiting the Internet
(Information Retrieval, Extraction, Summarization, Translation)



English-Vietnamese translation EVSMT1.0



Issues in Vietnamese SMT

- Corpus building
- Language Modeling
- Translation Model
- Decoder
- Others

Language	Tokens	Token types
English	342,035	8,128
Vietnamese	285,137	5805

Table 2: The parallel corpus

Models	BLEU score
Baseline phrase-based	0.5826
Words+POS	0.5964
Words+POS+Morphological	**0.6014**

Table 3: Comparison by the Bleu score

Sentences in Japanese-Vietnamese corpus

…

Ở bất cứ đâu, người xa xứ vẫn mong được trở về sum họp dưới mái ấm gia đình trong 3 ngày Tết.
家族みなで一つの屋根の下でテトを迎えることはみなの願いである。

Thị trường chứng khoán của Nhật tuần này có tăng?
今週の市場は上昇するか?

Sống trong đời sống cần có một tấm lòng, dù không để làm gì cả, dù chỉ để … gió cuốn đi
心無くして生きられない．例え風に吹かれるだけであっても
…

Toward Japanese-Vietnamese translation

- Dream of a translation system for Japanese-Vietnamese and Vietnamese-Japanese
- The most feasible way is statistical machine translation, but it requires a big parallel corpus of Japanese-Vietnam sentences.
- Hope Vietnamese students in Japan to contribute their collection of such sentence pairs. If each gives 50 pairs, 500 people give 25,000 pairs, and it allows us to apply some fund for the project.
- The sentences are encoded by UTF-8, written in a file, pair by pair with a blank line between pairs as in the previous page.
- Send to **jvcorpus@jaist.ac.jp**, Subject: JVcorpus.

Message from VLSP

- VLSP is a part of ICT in Vietnam, and plays a significant role in the development of the country.
- It is a long way requiring collaboration and contribution of many people, and can learn much from processing of other languages, in particular JLSP (itinerary, methods, etc.).
- If you give a hand to VLSP, we can hope that some day in future people in Vietnam and Japan can understand better each other thank to the translation system.

Acknowledgements

- The national project KC01.01.05/06-10
- Projects members: Luong Chi Mai, Ngo Cao Son, Ho Bao Quoc, Dinh Dien, Cao Hoang Tru, Nguyen Thi Minh Huyen, Vu Luong, Le Thanh Huong, Nguyen Phuong Thai, Nguyen Le Minh, Le Minh Hoang, Phan Xuan Hieu, Pham Ngoc Khanh, Ha Thanh Le, Nguyen Phuong Thao, Nguyen Viet Cuong, VLSP forum, among others.