Mining parallel corpora for multilingual machine translation system

Dr. NGUYEN Viet Son
Dr. DO Thi Ngoc Diep

ASEAN IVO, Kuala Lumpur, 26/11/2015

International Research Institute MICA
Multimedia, Information, Communication & Applications
UMI 2954

Hanoi University of Science and Technology
1 Dai Co Viet - Hanoi - Vietnam

Machine Translation Technology

- Rules based (expert method)
  - lexicons, morphological
  - syntactic, sematic

→ need **specialist knowledge**

costly in **time** & **human resources**

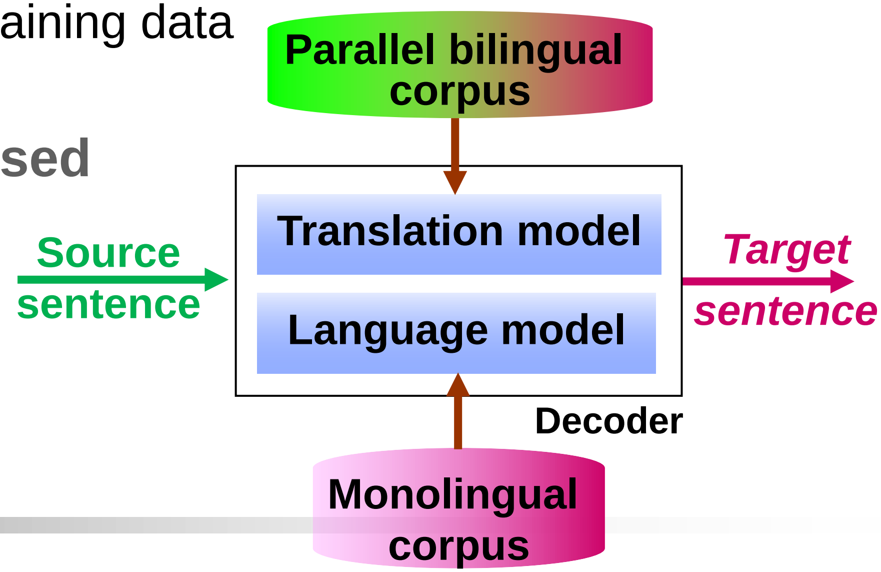
- Example-based
- **Statistical**

→ **quickly build** a translation system

based on **large parallel bilingual corpus**

  - Adapted with a lot of language pairs
  - Extract statistical information from learning databases
  - Depends: quantity & quality of training data

- Hybrid = Statistical + Rules based
- For common language pairs:
  - EN - FR
  - EN - JA
  - EN - CH
  - etc.



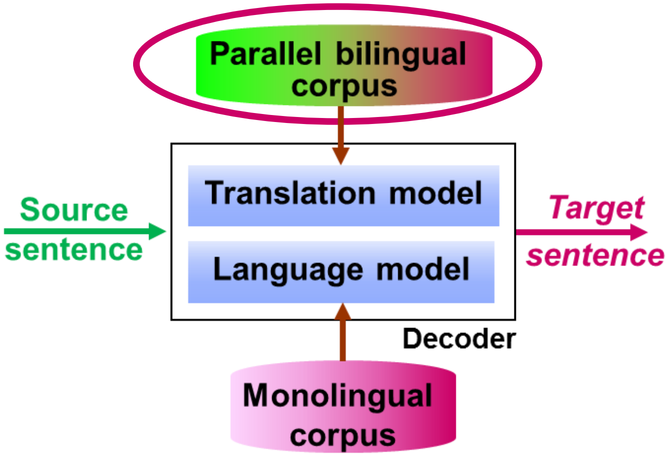
MT for ASEAN languages

- For **ASEAN (under-resourced) languages**
  - **DO NOT HAVE** any large parallel corpus for **ASEAN language pairs**
  - **FEW** available parallel corpus **even with common languages** & **LIMITTED** in some domains:
    - BTEC (tourist): **NOT OPEN** for community
    - OPENSubtitles (film subtitle): **SHORT** sentence, **DIALOG** format
    - TATOEBA (human translation): starting in **TECHNICAL** domain

Objectives

- **Methodology of retrieving a large parallel bilingual text corpus**
  - Firstly for ASEAN language pairs
  - Extracting parallel corpus from comparable corpora: Website, wiki, digital documents, etc.
  - Extendable for any under-resourced language pairs

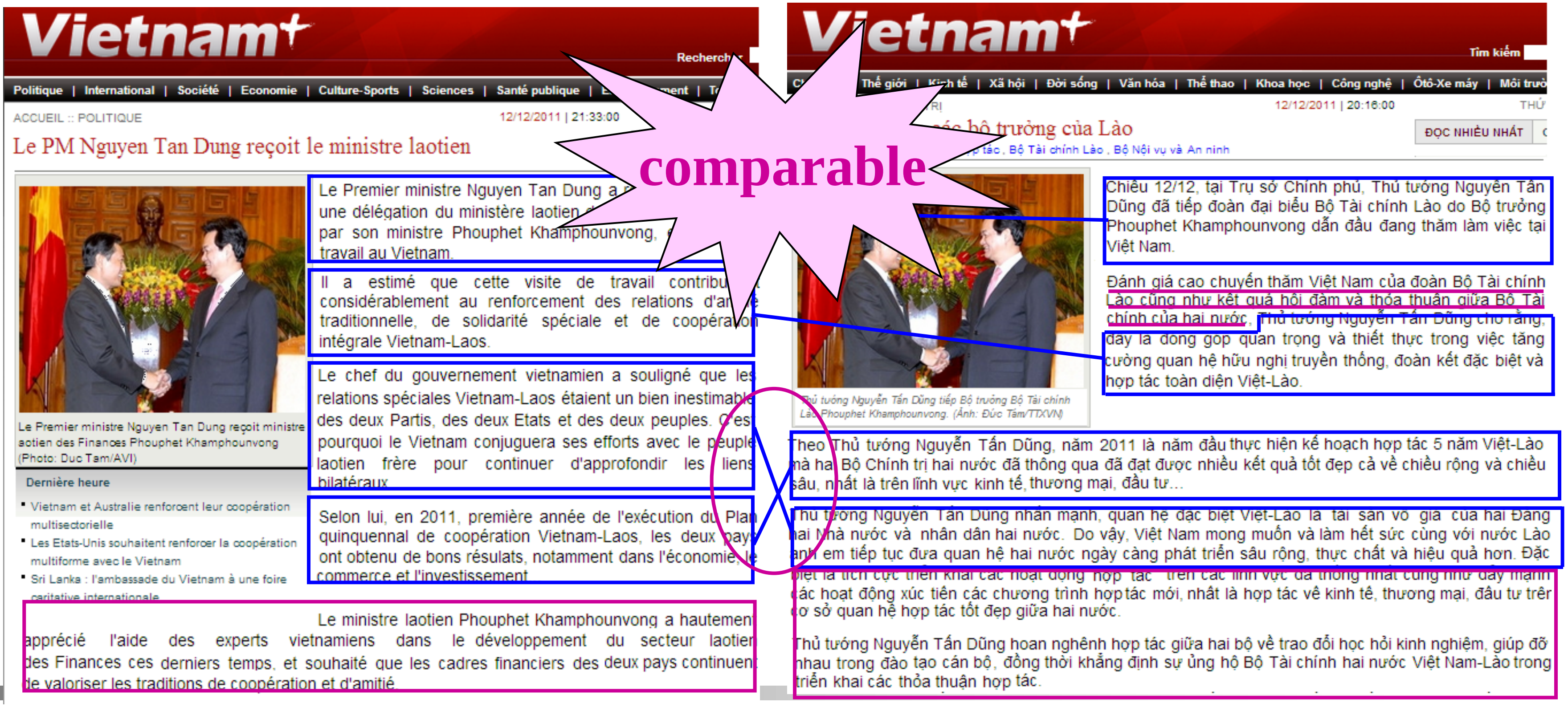
- **Building a machine translation system for ASEAN languages in using the statistical approach.**



Our recent researches

Extracting parallel corpus

- Collect parallel text from **multilingual websites**
- Hypotheses:
  - (One) multilingual website(s) of an under-resourced language
  - None of available parallel data
  - No supplementation data: lexical, morphological, dict. info.



comparable
Vietnam+
Recherche
Politique | International | Société | Economie | Culture-Sports | Sciences | Santé publique | E... ment | T...
Vietnam+
Tìm kiếm
Ch... | Thế giới | Kinh tế | Xã hội | Đời sống | Văn hóa | Thể thao | Khoa học | Công nghệ | Ôtô-Xe máy | Môi trườ
ĐỌC NHIỀU NHẤT

ACCUEIL :: POLITIQUE

12/12/2011 | 21:33:00

Le PM Nguyen Tan Dung reçoit le ministre laotien

Le Premier ministre Nguyen Tan Dung reçoit ministre laotien des Finances Phouphet Khamphounvong (Photo: Duc Tam/AVI)

Le Premier ministre Nguyen Tan Dung a une délégation du ministère laotien par son ministre Phouphet Khamphounvong, travail au Vietnam.

Il a estimé que cette visite de travail contribu considérablement au renforcement des relations d'ar traditionnelle, de solidarité spéciale et de coopéra intégrale Vietnam-Laos.

Le chef du gouvernement vietnamien a souligné que les relations spéciales Vietnam-Laos étaient un bien inestimable des deux Partis, des deux Etats et des deux peuples. C'est pourquoi le Vietnam conjuguera ses efforts avec le peuple laotien frère pour continuer d'approfondir les liens bilatéraux.

Selon lui, en 2011, première année de l'exécution du Plan quinquennal de coopération Vietnam-Laos, les deux pays ont obtenu de bons résulats, notamment dans l'économie, le commerce et l'investissement.

Le ministre laotien Phouphet Khamphounvong a hautement apprécié l'aide des experts vietnamiens dans le développement du secteur laotien des Finances ces derniers temps, et souhaité que les cadres financiers des deux pays continuent de valoriser les traditions de coopération et d'amitié.

Dernière heure

- Vietnam et Australie renforcent leur coopération multisectorielle
- Les Etats-Unis souhaitent renforcer la coopération multiforme avec le Vietnam
- Sri Lanka : l'ambassade du Vietnam à une foire caritative internationale

12/12/2011 | 20:16:00

...ác bộ trưởng của Lào

...p tác , Bộ Tài chính Lào , Bộ Nội vụ và An ninh

Thủ tướng Nguyễn Tấn Dũng tiếp Bộ trưởng Bộ Tài chính Lào Phouphet Khamphounvong. (Ảnh: Đức Tâm/TTXVN)

Chiều 12/12, tại Trụ sở Chính phủ, Thủ tướng Nguyễn Tấn Dũng đã tiếp đoàn đại biểu Bộ Tài chính Lào do Bộ trưởng Phouphet Khamphounvong dẫn đầu đang thăm làm việc tại Việt Nam.

Đánh giá cao chuyến thăm Việt Nam của đoàn Bộ Tài chính Lào cũng như kết quả hội đàm và thỏa thuận giữa Bộ Tài chính của hai nước, Thủ tướng Nguyễn Tấn Dũng cho rằng, đây là đóng góp quan trọng và thiết thực trong việc tăng cường quan hệ hữu nghị truyền thống, đoàn kết đặc biệt và hợp tác toàn diện Việt-Lào.

Theo Thủ tướng Nguyễn Tấn Dũng, năm 2011 là năm đầu thực hiện kế hoạch hợp tác 5 năm Việt-Lào mà hai Bộ Chính trị hai nước đã thông qua đã đạt được nhiều kết quả tốt đẹp cả về chiều rộng và chiều sâu, nhất là trên lĩnh vực kinh tế, thương mại, đầu tư…

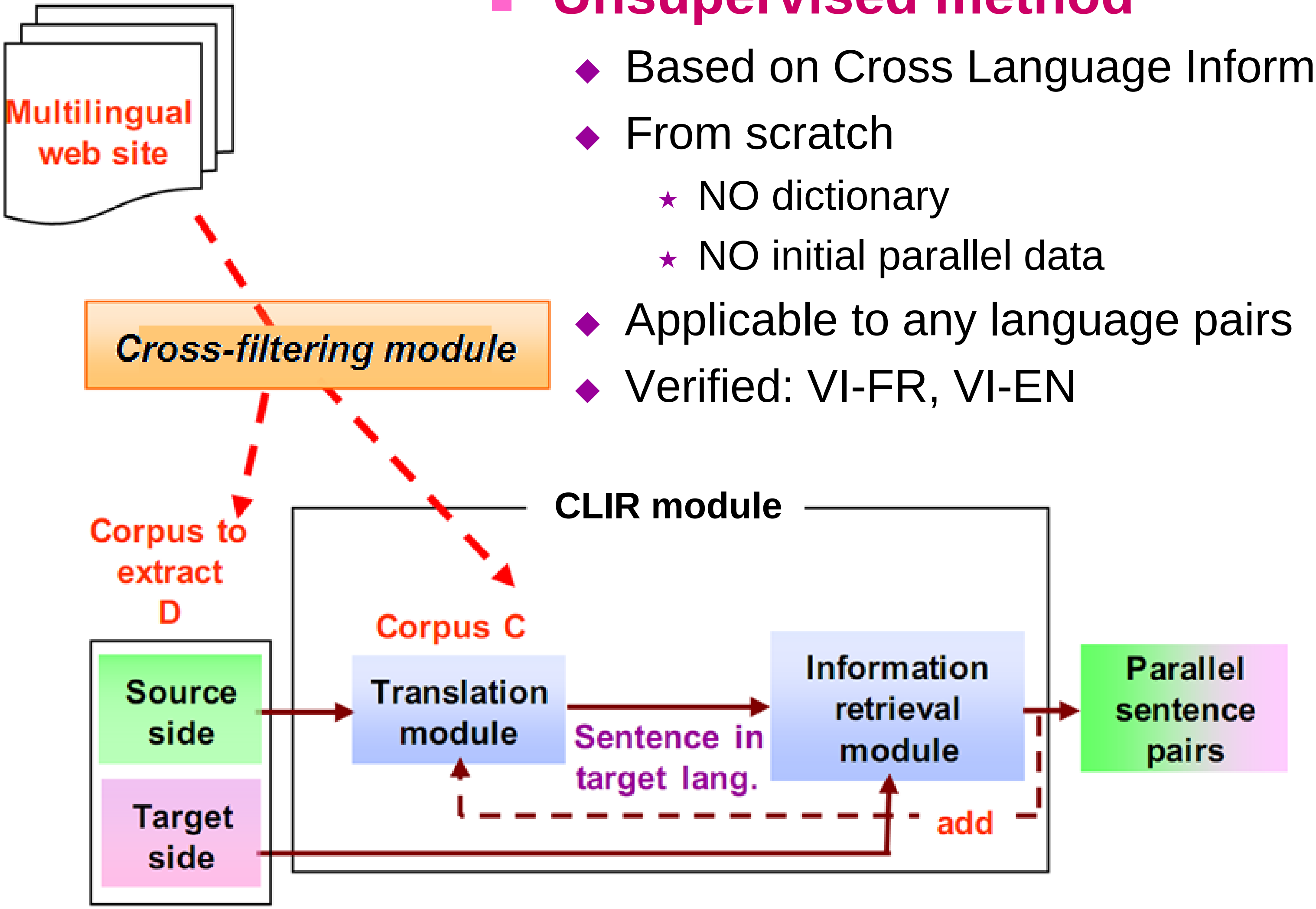
Thủ tướng Nguyễn Tấn Dũng nhấn mạnh, quan hệ đặc biệt Việt-Lào là tài sản vô giá của hai Đảng hai Nhà nước và nhân dân hai nước. Do vậy, Việt Nam mong muốn và làm hết sức cùng với nước Lào anh em tiếp tục đưa quan hệ hai nước ngày càng phát triển sâu rộng, thực chất và hiệu quả hơn. Đặc biệt là tích cực triển khai các hoạt động hợp tác trên các lĩnh vực đã thống nhất cũng như đẩy mạnh các hoạt động xúc tiến các chương trình hợp tác mới, nhất là hợp tác về kinh tế, thương mại, đầu tư trên cơ sở quan hệ hợp tác tốt đẹp giữa hai nước.

Thủ tướng Nguyễn Tấn Dũng hoan nghênh hợp tác giữa hai bộ về trao đổi học hỏi kinh nghiệm, giúp đỡ nhau trong đào tạo cán bộ, đồng thời khẳng định sự ủng hộ Bộ Tài chính hai nước Việt Nam-Lào trong triển khai các thỏa thuận hợp tác.

Our recent researches

Extracting parallel corpus

- **Unsupervised method**
  - Based on Cross Language Information Retrieval
  - From scratch
    - NO dictionary
    - NO initial parallel data
  - Applicable to any language pairs
  - Verified: VI-FR, VI-EN



Multilingual web site

Cross-filtering module

Corpus to extract D

Source side

Target side

CLIR module

Corpus C

Translation module

Sentence in target lang.

Information retrieval module

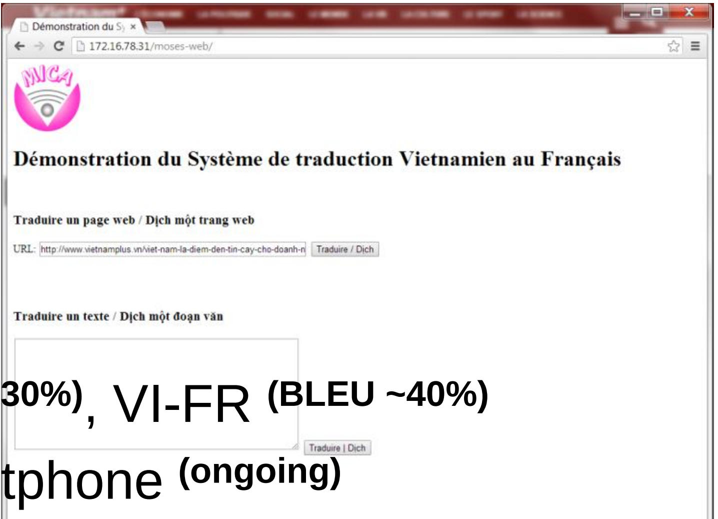
add

Parallel sentence pairs

Our recent researches

Extracting parallel corpus

- From website VnAgency
  - **50K VI-FR** parallel sentence pairs
  - **40K VI-EN** parallel sentence pairs
- Applications:
  - **Webpage MT system**: VI-EN **(BLEU ~30%)**, VI-FR **(BLEU ~40%)**
  - **FR-VI Speech Translation** on smartphone **(ongoing)**



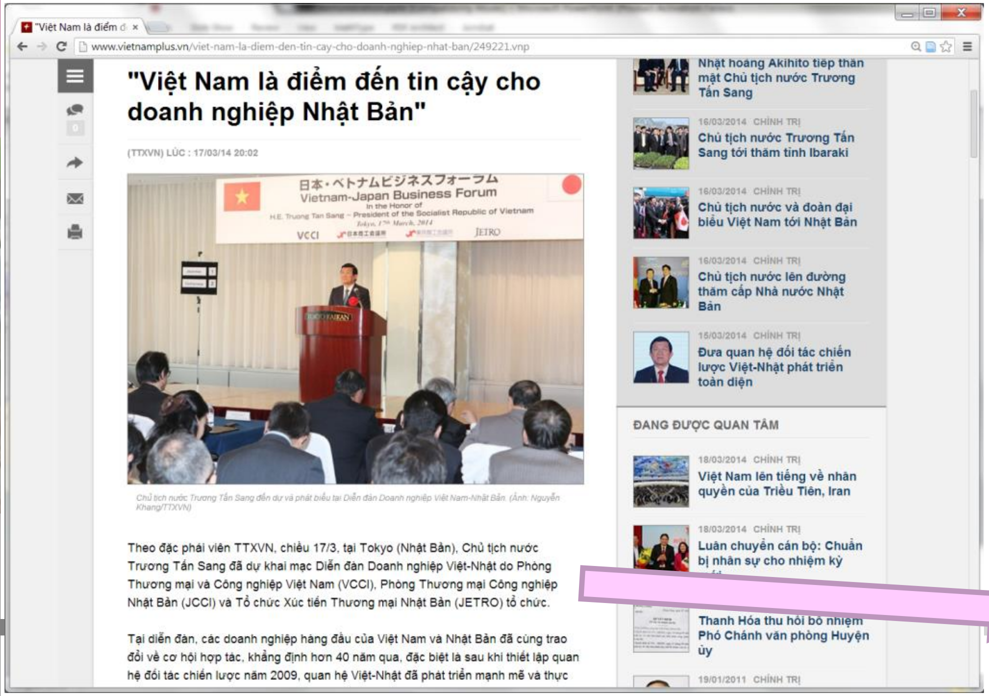
Démonstration du Système de traduction Vietnamien au Français

Traduire un page web / Dịch một trang web

URL: http://www.vietnamplus.vn/viet-nam-la-diem-den-tin-cay-cho-doanh-n Traduire / Dich

Traduire un texte / Dịch một đoạn văn

Traduire | Dich



"Việt Nam là điểm đ
www.vietnamplus.vn/viet-nam-la-diem-den-tin-cay-cho-doanh-nghiep-nhat-ban/249221.vnp
0
日本・ベトナムビジネスフォーラム
Vietnam-Japan Business Forum
In the Honor of
H.E. Truong Tan Sang – President of the Socialist Republic of Vietnam
Tokyo, 17th March, 2014
VCCI
JETRO
TOKYO KAIKAN

"Việt Nam là điểm đến tin cậy cho doanh nghiệp Nhật Bản"

(TTXVN) LÚC : 17/03/14 20:02

Chủ tịch nước Trương Tấn Sang đến dự và phát biểu tại Diễn đàn Doanh nghiệp Việt Nam-Nhật Bản. (Ảnh: Nguyễn Khang/TTXVN)

Theo đặc phái viên TTXVN, chiều 17/3, tại Tokyo (Nhật Bản), Chủ tịch nước Trương Tấn Sang đã dự khai mạc Diễn đàn Doanh nghiệp Việt-Nhật do Phòng Thương mại và Công nghiệp Việt Nam (VCCI), Phòng Thương mại Công nghiệp Nhật Bản (JCCI) và Tổ chức Xúc tiến Thương mại Nhật Bản (JETRO) tổ chức.

Tại diễn đàn, các doanh nghiệp hàng đầu của Việt Nam và Nhật Bản đã cùng trao đổi về cơ hội hợp tác, khẳng định hơn 40 năm qua, đặc biệt là sau khi thiết lập quan hệ đối tác chiến lược năm 2009, quan hệ Việt-Nhật đã phát triển mạnh mẽ và thực

Nhật hoàng Akihito tiếp thân mật Chủ tịch nước Trương Tấn Sang

16/03/2014 CHÍNH TRỊ
Chủ tịch nước Trương Tấn Sang tới thăm tỉnh Ibaraki

16/03/2014 CHÍNH TRỊ
Chủ tịch nước và đoàn đại biểu Việt Nam tới Nhật Bản

16/03/2014 CHÍNH TRỊ
Chủ tịch nước lên đường thăm cấp Nhà nước Nhật Bản

15/03/2014 CHÍNH TRỊ
Đưa quan hệ đối tác chiến lược Việt-Nhật phát triển toàn diện

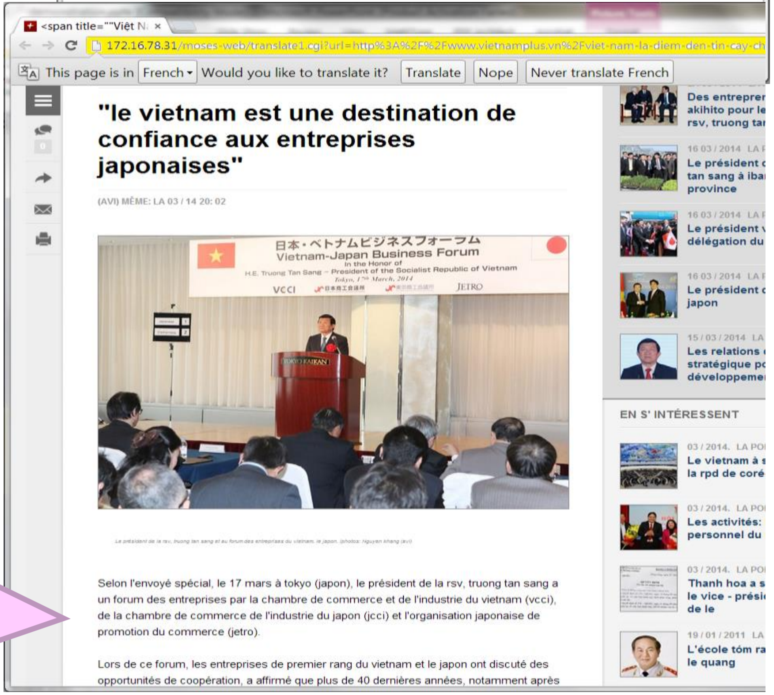
ĐANG ĐƯỢC QUAN TÂM

18/03/2014 CHÍNH TRỊ
Việt Nam lên tiếng về nhân quyền của Triều Tiên, Iran

18/03/2014 CHÍNH TRỊ
Luân chuyển cán bộ: Chuẩn bị nhân sự cho nhiệm kỳ

Thanh Hóa thu hồi bổ nhiệm Phó Chánh văn phòng Huyện ủy

19/01/2011 CHÍNH TRỊ



<span title=""Việt N
172.16.78.31/moses-web/translate1.cgi?url=http%3A%2F%2Fwww.vietnamplus.vn%2Fviet-nam-la-diem-den-tin-cay-ch
This page is in French ▾ Would you like to translate it?
Translate
Nope
Never translate French
0
日本・ベトナムビジネスフォーラム
Vietnam-Japan Business Forum
In the Honor of
H.E. Truong Tan Sang – President of the Socialist Republic of Vietnam
Tokyo, 17th March, 2014
VCCI
JETRO
TOKYO KAIKAN

"le vietnam est une destination de confiance aux entreprises japonaises"

(AVI) MÊME: LA 03 / 14 20: 02

Le président de la rsv, truong tan sang et au forum des entreprises du vietnam. le japon. (photos: Nguyen khang (avi)

Selon l'envoyé spécial, le 17 mars à tokyo (japon), le président de la rsv, truong tan sang a un forum des entreprises par la chambre de commerce et de l'industrie du vietnam (vcci), de la chambre de commerce de l'industrie du japon (jcci) et l'organisation japonaise de promotion du commerce (jetro).

Lors de ce forum, les entreprises de premier rang du vietnam et le japon ont discuté des opportunités de coopération, a affirmé que plus de 40 dernières années, notamment après

Des entrepren
akihito pour le
rsv, truong tan

16 03 / 2014 LA P
Le président d
tan sang à iba
province

16 03 / 2014 LA P
Le président v
délégation du

16 03 / 2014 LA P
Le président d
japon

15 / 03 / 2014 LA
Les relations
stratégique po
développemen

EN S' INTÉRESSENT

03 / 2014. LA POL
Le vietnam à s
la rpd de coré

03 / 2014. LA POL
Les activités:
personnel du

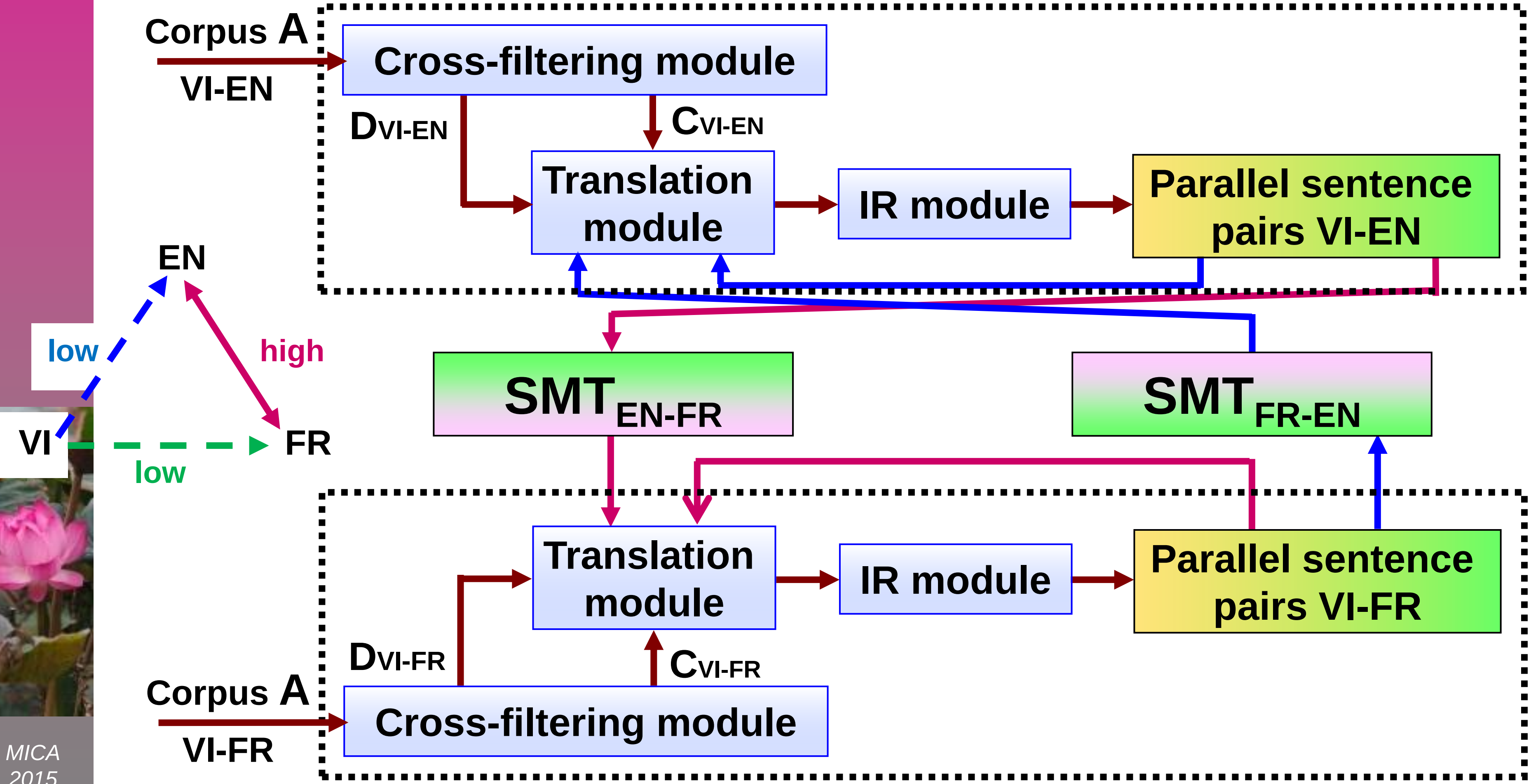
03 / 2014. LA POL
Thanh hoa a s
le vice - prési
de le

19 / 01 / 2011 LA
L'école tóm ra
le quang

Our recent researches

Extracting parallel corpus

- **Unsupervised method with triangulation**



Corpus A VI-EN

Cross-filtering module

$D_{VI\text{-}EN}$

$C_{VI\text{-}EN}$

Translation module

IR module

Parallel sentence pairs VI-EN

EN

low

high

VI

FR

low

$SMT_{EN\text{-}FR}$

$SMT_{FR\text{-}EN}$

Translation module

IR module

Parallel sentence pairs VI-FR

$D_{VI\text{-}FR}$

$C_{VI\text{-}FR}$

Corpus A VI-FR

Cross-filtering module

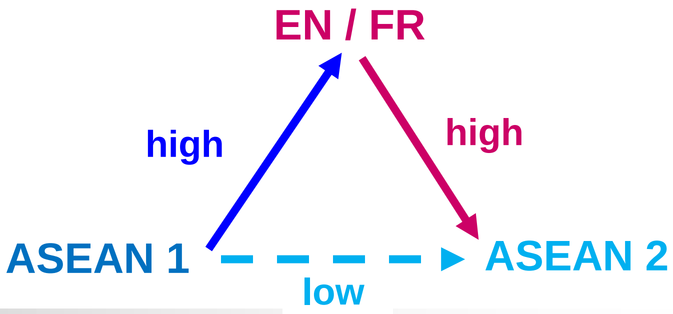
Our recent researches

Extracting parallel corpus

- **2014: KH-EN language pair**
  - Try to apply our extracting methods
  - **Difficulties in Natural Language Processing** for KH language problems
  - Need **cooperation with Cambodia research groups**
- **2015: JA-VI language pair**
  - In cooperation with Multilingual Translation lab. (NICT)
  - Use available parallel text resources:
    - TED talks data: 100K pairs
    - Wiki data: 76K pairs
  - Hard task due to **different language families** **(BLEU ~10%)**

Call for collaborations

- **Methodology of extracting parallel corpus:**
  - Multi-levels of extraction (sentence, phrase, fragment)
  - Independent language pairs: ASEAN, JA, EN, FR, etc.
  - From different resources: Website, Wiki, etc.
  - Multilingual extraction methods:
    - **Unsupervised**: ASEAN - common language (EN/FR) pairs
    - **Unsupervised with triangulation**: ASEAN languages pairs
- **Research on MT for different family language pairs**
- **Exchange researchers on MT**
- **Train students on MT**



References

[1] Thi-Ngoc-Diep Do, Viet-Bac Le, Brigitte Bigi, Laurent Besacier, Eric Castelli. *Mining a Comparable Text Corpus for a Vietnamese - French Statistical Machine Translation System*. 4th Workshop on Statistical Machine Translation, Athens – Greece, 2009.

[2] Thi-Ngoc-Diep Do, Laurent Besacier, Eric Castelli. *A Fully Unsupervised Approach for Mining Parallel Data from Comparable Corpora*. 14th Annual Conference of the European Association. for Machine Translation, Saint-Raphaël – France, 2010.

[3] Thi-Ngoc-Diep Do, Eric Castelli, Laurent Besacier. *Mining Parallel Data from Comparable Corpora via Triangulation*. International Conference on Asian Language Processing, Penang – Malaysia, 2011.

[4] Thi-Ngoc-Diep Do, Masao Utiyama, Eiichiro Sumita, *Machine Translation from Japanese and French to Vietnamese, the Difference among Language Families*, International Conference on Asian Language Processing, Oct 2015.

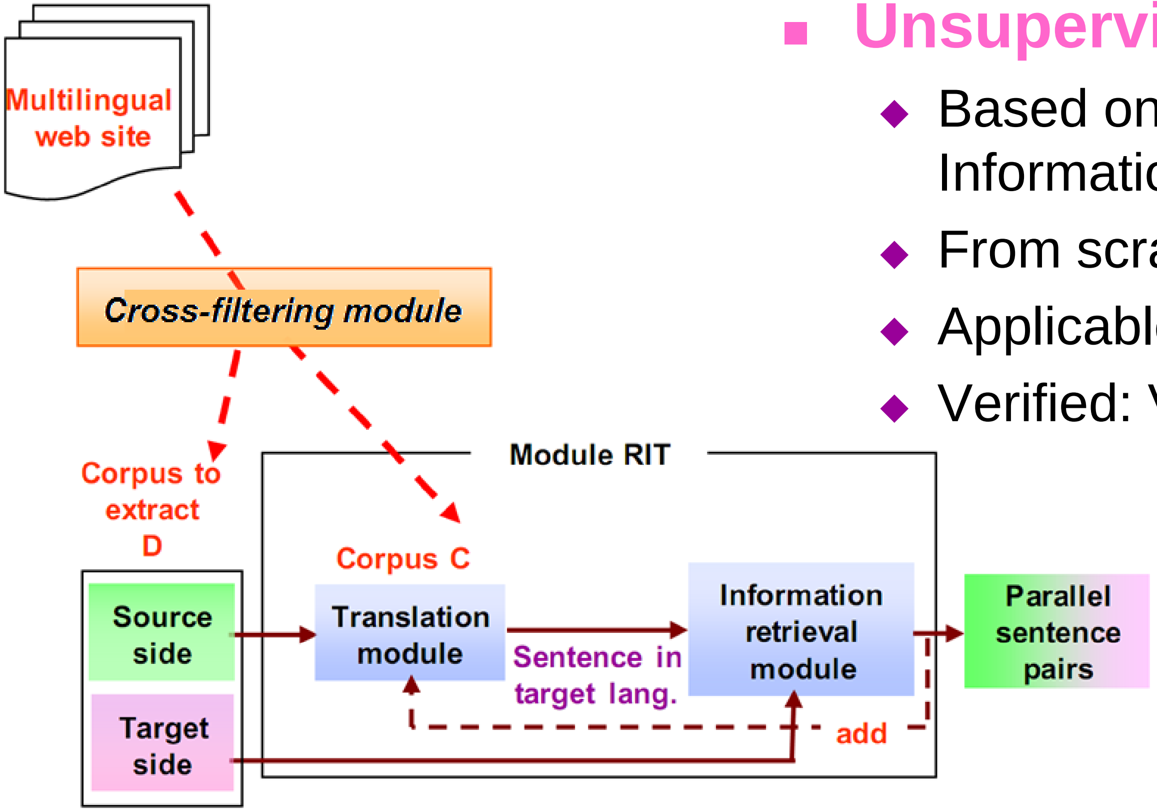
[5] PhD thesis of DO Thi-Ngoc-Diep. *Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée*, 2011, Université de Grenoble, LIG - MICA

Thank for your attention !

Our recent researches

Extracting parallel corpus

- **Fast collection method**
  - Lexical information, heuristic rules
  - Dictionary
  - Depend on language pairs



- **Unsupervised method**
  - Based on Cross Language Information Retrieval
  - From scratch
  - Applicable to any language pairs
  - Verified: VN-FR, VN-EN