Dịch máy

Lê Thanh Hương
Bộ môn Hệ thống Thông tin
Viện CNTT &TT – Trường ĐHBKHN
Email: huonglt-fit@mail.hut.edu.vn

Ví dụ

- Au sortir de la saison 97/98 et surtout au debut de cette saison 98/99…
- With leaving season 97/98 and especially at the beginning of this season 98/99…

Các vấn đề

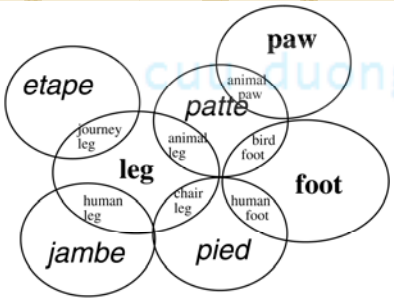
1. Xử lý sự giống và khác nhau giữa các ngôn ngữ
   - Hình vị: # số âm tiết/từ:
     - *Ngôn ngữ đơn âm tiết (tiếng Việt, Trung Quốc) – 1 tiếng/từ*
     - *Ngôn ngữ đa âm tiết* (Siberian Yupik), 1 từ = cả 1 câu
   - Mức độ phân chia âm tiết

Các vấn đề

2. Cú pháp: trật tự từ trong câu
   - *To Yukio; Yukio ne*
   - Tiếng Anh – tiếng Việt:
     - *The* (affix1) *red* (affix2) *flag* (head)
     - *Lá cờ* (head) *đỏ* (affix2) *ấy* (affix1)
3. Các nét riêng biệt

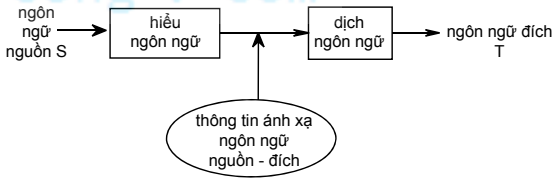
English	brother	Vietnamese	anh em
English	wall	German	wand (inside) mauer(outside)
German	berg	English	hill mountain

Không gian khái niệm



Khoảng trống từ vựng: tiếng Nhật không có từ nào nghĩa *privacy;* tiêgns Anh không có từ ứng với *yakoko* (lòng hiếu thảo)

Ba khối chính trong dịch máy



Hiểu ngôn ngữ

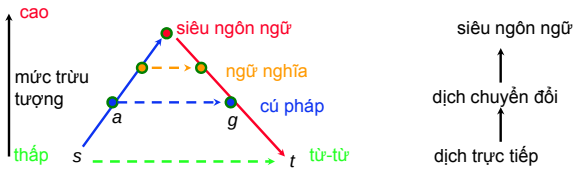
1. Nhập nhằng từ vựng:
   English: *book* - Spanish *libro, reservar*
⇒Sử dụng thông tin cú pháp

2. Nhập nhằng cú pháp:
   I saw the guy on the hill with the telescope

3. Nhập nhằng ngữ nghĩa:
- *E: While driving, John swerved & hit a tree*
  John's car
- *S: Minetras que John estaba manejando, se desvio y golpeop con un arbo*

7

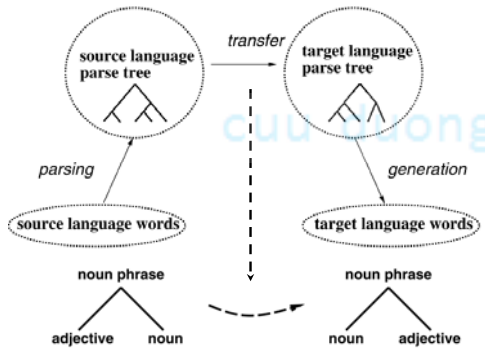
Các phương pháp dịch máy



a = *a(s)*
g = *f(a(s)); f* – hàm chuyển đổi
t=g(f(a(s)))

8

Sơ đồ chuyển đổi



9

Luật chuyển đổi

English to French:

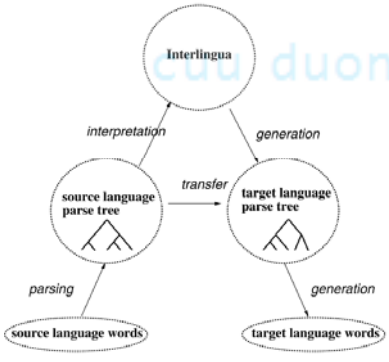
1. NP → Adjective$_1$ Noun$_2$
   ⇒
   NP → Noun$_2$ Adjective$_1$

Japanese to English:

2. Existential-There-Sentence → There$_1$ Verb$_2$ NP$_3$ Postnominal$_4$
   ⇒
   Sentence → (NP → NP$_3$ Relative-Clause$_4$) Verb$_2$
3. NP → NP$_1$ Relative Clause$_2$
   ⇒
   NP → Relative-Clause$_2$ NP$_1$

10

Sơ đồ chuyển đổi



11

Cách tiếp cận siêu ngôn ngữ: sử dụng nghĩa

- Chuyển đổi: các luật chuyển đổi từ ngôn ngữ này sang ngôn ngữ khác
- Đối tượng/sự kiện (ontology)

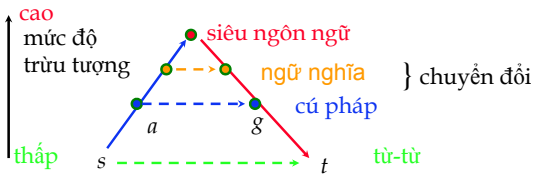
$$\begin{bmatrix} \text{event} & \text{gardening} \\ \text{agent} & \begin{bmatrix} \text{man} & \\ \text{number} & \text{sg} \\ \text{definiteness} & \text{indef} \end{bmatrix} \\ \text{aspect} & \text{progressive} \\ \text{tense} & \text{past} \end{bmatrix}$$

12

Dịch máy thống kê

13

Các kiểu dịch máy



14

ý tưởng

- Coi việc dịch như bài toán kênh có nhiễu

Input (Nguồn) → The channel (adds "noise") → "Noisy" Output (đích)

E: English words... → F: Les mots Anglais...

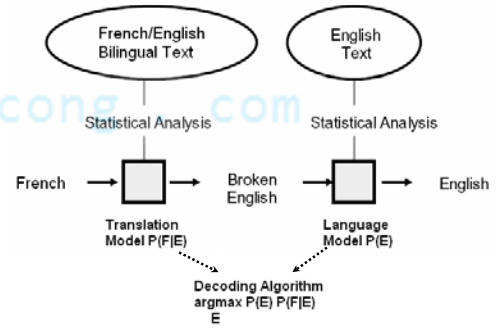
- Mô hình dịch: P(E|F) = P(F|E) P(E) / P(F)
- Khôi phục lại E khi biết F:

Sau khi đơn giản hóa (P(F) không đổi):

$$\text{argmax}_E P(E|F) = \text{argmax}_E P(F|E) P(E)$$

15

Dịch máy thống kê



16

Các yếu tố

- Mô hình ngôn ngữ - Language Model (LM): xác suất thấy 1 câu tiếng Anh (E) (xác suất tiền nghiệm):
  P(E)
- Mô hình dịch - Translation Model (TM): câu đích trong tiếng Pháp (F) khi có câu tiếng Anh:
  P(F|E)
- Thủ tục tìm kiếm:
  - Cho F, tìm E tốt nhất sử dụng mô hình ngôn ngữ LM và mô hình dịch TM.
- Vấn đề: thiếu dữ liệu!
  - Ta không thể tạo từ điển câu E ↔ F
  - Thậm thí bình thường ta không thấy 1 câu lặp lại 2 lần

17

Ý tưởng gióng hàng

- Mô hình dịch TM không quan tâm đến chuỗi đúng các từ tiếng Anh
- Sử dụng cách tiếp cận gán nhãn:
  - 1 từ tiếng Anh("tag") ~ 1 từ tiếng Pháp ("word")

→ không thực tế: thậm chí số từ trong 2 câu không bằng nhau

→ sử dụng "gióng hàng".

- Gióng hàng câu: tìm các nhóm câu trong 1 ngôn ngữ tương ứng với các nhóm câu khác trong ngôn ngữ khác

18

Giống hàng câu

The old man is happy. He has fished many times. His wife talks to him. The fish are jumping. The sharks await.

El viejo está feliz porque ha pescado muchos veces. Su mujer habla con él. Los tiburones esperan.

19

Giống hàng câu

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

20

Giống hàng câu

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

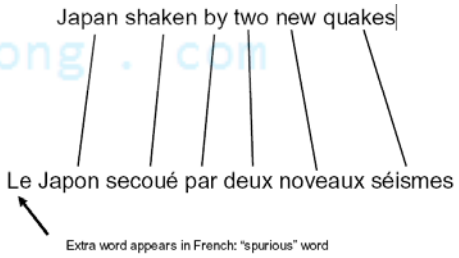
1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Khó khăn:

- Sự liên quan chéo: trật tự câu thay đổi khi dịch

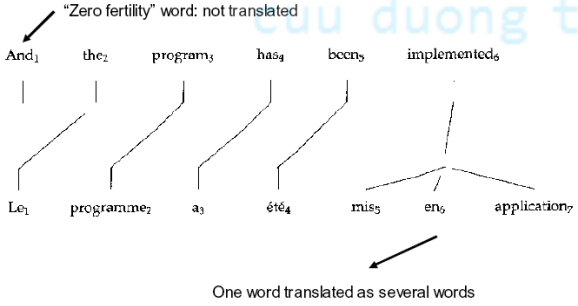
21

Giống hàng từ - Mức dễ



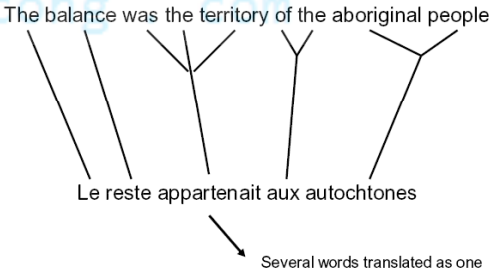
22

Giống hàng từ - Khó hơn



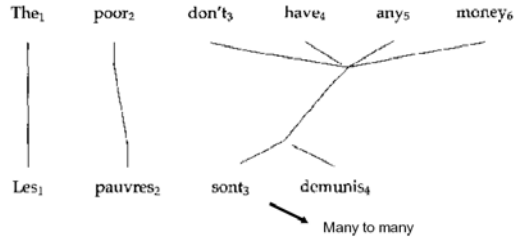
23

Giống hàng từ - Khó hơn



24

Gióng hàng từ - Khó



- A line group linking a minimal subset of words is called a 'cept' in the IBM work

25

Gióng hàng từ - Mã hóa

0 1 2 3 4 5 6

- e_0 And the program has been implemented
- f_0 Le programme a été mis en application

0 1 2 3 4 5 6 7

- Gán thông tin tuyến tính:
  - f_0(1) Le(2) programme(3) a(4) été(5) mis(6) en(6) application(6)
  - e_0 And(0) the(1) program(2) has(3) been(4) implemented(5,6,7)

26

Học việc gióng hàng từ sử dụng EM

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

27

Học việc gióng hàng từ sử dụng EM

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

28

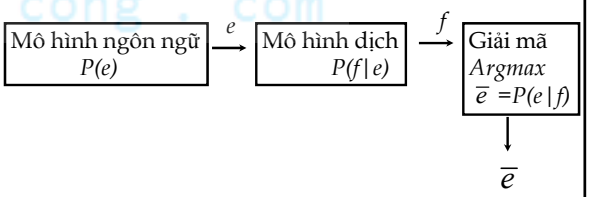
Học việc gióng hàng từ sử dụng EM

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

29

Kênh nhiễu



30

Các thành phần của mô hình dịch

- Giả thiết
  - Individual translations are independance
  - 1 từ tiếng Anh – n từ tiếng Pháp
  - 1 từ tiếng Pháp - (0-1) từ tiếng Anh

$$P(f \mid e) = \frac{1}{Z} \sum_{a_1}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} P(f_j \mid e_{a_j})$$

- f_j - *từ j* trong *f;*
- a_j - vị trí trong e được gióng hàng với f_j
- e_{aj} - từ trong *e* được gióng hàng với f_j
- Z là hằng số chuẩn hóa
- $a_j = 0$: từ *j* trong câu tiếng Pháp được gióng hàng với một từ rỗng (không dịch sang)
- m – độ dài của f

Ví dụ

- *P(Jean aime Marie| John loves Mary)*
- Gióng hàng*(Jean, John), (aime, loves), (Marie, Mary), ta có 3 xác suất*
  P(Jean|John) x *P(aime|loves)* x*P(Marie|Mary)*

Giải mã

$$\begin{aligned} \bar{e} &= \arg\max_e P(e \mid f) \\ &= \arg\max_e \frac{P(e)P(f \mid e)}{P(f)} \\ &= \arg\max_e P(e)P(f \mid e) \end{aligned}$$

Vấn đề: không gian tìm kiếm vô hạn

Mẹo:

- tìm kiếm dùng ngăn xếp: xây dựng dần, lưu trong stack các phần đã dịch
- sử dụng một số độ đo về độ phù hợp, vd., *chamber/house,* (nhưng có thể đi sai đường nếu 1 từ thường xuất hiện với từ khác, như *commune/house,* vì có *Chambre de Communes* (hạ nghị viện)

Thuật toán EM

E-step

- Khởi tạo giá trị $P(w_f/w_e)$ ngẫu nhiên
- Tính số lần tìm thấy w_f trong tiếng Pháp khi có w_e trong tiếng Anh

$$z_{w_f, w_e} = \sum_{(e,f) s.t. w_e = e, w_f = f} P(w_f \mid w_e)$$

M-step

- Đánh giá lại xác suất dịch prs từ giá trị z trên:

$$P(w_f \mid w_e) = \frac{z_{w_f, w_e}}{\sum_v z_{v, w_e}}$$

tổng được tính trên tất cả các từ tiếng Pháp v

Đánh giá

Đánh giá dựa trên tập ngữ liệu Hansard:

- 48% câu tiếng Pháp được dịch đúng
- 2 loại lỗi:
  - Dịch sai nghĩa:
    - Permettez que je donne un example à chambre
    - Let me give an example in the House (incorrect decoding)
    - (Let me give the House an example)
  - Dịch sai ngữ pháp:
    - Vous avez besoin de toute l'aide disponsible
    - You need all of the benefits available (ungrammatical decoding)
    - (You need all the help you can get)

Lý do

- Hiện tượng méo: từ tiếng Anh ở đầu câu được gióng hàng với từ tiếng Pháp ở cuối câu – hiện tượng này giảm xác suất gióng hàng
- Hiện tượng sinh (fertility): sự tương ứng giữa từ tiếng Anh và tiếng Pháp (1-to-1, 1-to-2, 1-to-0, …),
  - Vd, fertility(farmers) trong tập ngữ liệu = 2, vì từ này khi dịch sang tiếng Anh thường gồm 2 từ : les argiculteurs
  - To go → aller

Lý do

- Các giả thiết độc lập: các câu ngắn được ưu tiên hơn vì có ít xác suất hơn (khi nhân)
  ⇒ nhân kết quả với 1 hằng số tỉ lệ thuận với độ dài câu
- Phụ thuộc dữ liệu luyện: 1 thay đổi nhỏ trong dữ liệu luyện gây ra thay đổi lớn trong các giá trị ước lượng tham số
  Vd, *P(le|the)* thay đổi từ 0.610 xuống 0.497
- TÍnh hiệu quả. Bỏ các câu > 30 từ, vì làm không gian tìm kiếm tăng theo cấp số mũ
- Thiếu tri thức ngôn ngữ

37

Thiếu tri thức ngôn ngữ

- Không lưu thông tin về các ngữ: ví dụ không gióng hàng được "to go" và "aller"
- Không có ràng buộc cục bộ:
  Eg, is she a mathematician
- Âm vị. Các từ tạo bởi các âm vị khác nhau được coi là các ký hiệu riêng biệt
- Dữ liệu thưa. Các đánh giá cho các từ ít gặp không chính xác

38

Các hệ thống gióng hàng khác

- Các tập ngữ liệu sử dụng giả thiết:
  - Dữ liệu song song (dịch E ↔ F)
- Gióng hàng câu
  - Phát hiện câu
  - Gióng hàng câu
- Gióng hàng từ
  - Tách từ
  - Gióng hàng từ (với 1 số ràng buộc)

39

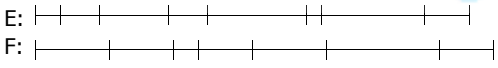
Phát hiện biên của câu

- Sử dụng luật, danh sách liệt kê:
  - Dấu kết thúc câu:
    - Dấu ngắt đoạn (nếu được đánh dấu)
    - 1 số ký tự: ?, !, ;
    - Vấn đề: dấu chấm '.'
      - Kết thúc câu (... left yesterday. He was heading to...)
      - Dấu chấm thập phân : 3.6 (three-point-six)
      - Dấu chấm hàng nghìn: 3.200
      - Viết tắt: cf., e.g., Calif., Mt., Mr.
      - Vân vân: ...
      - 1 số ngôn ngữ: 2nd ~ 2.
      - Ký hiệu đầu: A. B. Smith
- Phương pháp thống kê: vd Maximum Entropy

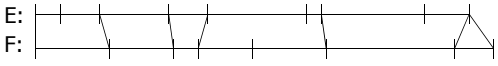
40

Gióng hàng câu

- Vấn đề với phát hiện biên của câu:



- **Đầu ra mong đợi**: Các phân mảnh với cùng số lượng mảnh liên tiếp nhau.
- Gióng hàng:



- Kết quả: 2-1, 1-1, 1-1, 2-2, 2-1, 0-1

41

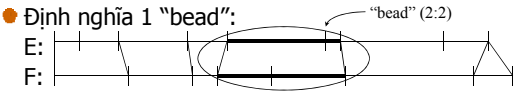
Các phương pháp gióng hàng

- Nhiều phương pháp (xác suất hoặc không)
  - Dựa trên độ dài ký tự
  - Dựa trên độ dài từ
  - "cùng gốc" (sử dụng nghĩa từ)
    - Sử dụng từ điển (F: prendre ~ E: make, take)
    - Sử dụng khoảng cách từ (độ tương tự): tên, số, từ vay mượn, từ gốc Latin
- Kết quả tốt nhất:
  - Thống kê, dựa trên từ hoặc dựa trên ký tự

42

Gióng hàng dựa trên độ dài

- Định nghĩa bài toán như việc tính xác suất:
  $\text{argmax}_A P(A|E,F) = \text{argmax}_A P(A,E,F)$ (E,F cố định)
- Định nghĩa 1 "bead":



- Lấy xấp xỉ:
  $P(A,E,F) \cong \Pi_{i=1..n} P(B_i)$,
  Trong đó B_i là 1 bead; $P(B_i)$ không phụ thuộc vào phần còn lại của E,F.

43

Nhiệm vụ gióng hàng

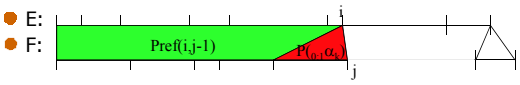
Định nghĩa:

- Cho $P(A,E,F) \cong \Pi_{i=1..n} P(B_i)$,
  tìm cách chia (E,F) thành n bead $B_{i=1..n}$, sao cho tối đa xác suất P(A,E,F) trên tập luyện.
- $B_i = {}_{p:q}\alpha_i$, với $p:q \in \{0:1,1:0,1:1,1:2,2:1,2:2\}$
  mô tả phép gióng hàng
- Pref(i,j) – xác suất của cách gióng hàng tốt nhất từ điểm đầu cho đến (i,j)

44

Định nghĩa đệ qui

- Khởi tạo: Pref(0,0) = 0.
- Pref(i,j) = max (
  Pref(i,j-1) $P({}_{0:1}\alpha_k)$, Pref(i-1,j) $P({}_{1:0}\alpha_k)$, Pref(i-1,j-1) $P({}_{1:1}\alpha_k)$,
  Pref(i-1,j-2) $P({}_{1:2}\alpha_k)$, Pref(i-2,j-1) $P({}_{2:1}\alpha_k)$, Pref(i-2,j-2) $P({}_{2:2}\alpha_k)$)



45

Xác suất của 1 Bead

- Định nghĩa $P({}_{p:q}\alpha_k)$:
  - k đề cập đến "bead" kế tiếp, với các đoạn của câu p và q, độ dài $l_{k,e}$ và $l_{k,f}$.
- Sử dụng phân bố chuẩn cho các độ dài khác nhau:
  $P({}_{p:q}\alpha_k) = P(\delta(l_{k,e},l_{k,f},\mu,\sigma^2),p:q) \cong P(\delta(l_{k,e},l_{k,f},\mu,\sigma^2))P(p:q)$
  $\delta(l_{k,e},l_{k,f},\mu,\sigma^2) = (l_{k,f} - \mu l_{k,e})/\sqrt{l_{k,e}\sigma^2}$
- Đánh giá P(p:q) từ tập dữ liệu nhỏ, hoặc đoán và đánh gía lại sau khi gióng hàng
- Từ có thể được dùng như dấu hiệu tốt hơn để định nghĩa $P({}_{p:q}a_k)$.

46

Gióng hàng từ

- Nếu chỉ dựa trên độ dài, không thực hiện được:
  - từ có thể bị đảo trật tự, các phép dịch thường có độ dài khác nhau
- Ý tưởng:
  - Đưa ra vài mô hình dịch đơn giản.
  - Tìm các tham số bằng cách xét tất cả các cách gióng hàng.
  - Sau khi có tham số, tìm cách gióng hàng tốt nhất khi có các tham số này.

47

Thuật toán gióng hàng từ

Khởi tạo với tập ngữ liệu gióng hàng câu.
Cho (E,F) là 1 cặp câu (là 1 bead).

1. Khởi tạo ngẫu nhiên p(f|e), f∈F, e∈E.
2. Đếm trên tập ngữ liệu:
   $c(f,e) = \Sigma_{(E,F);e\in E,f\in F} p(f|e)$
   với ∀ cặp gióng hàng (E,F), kiểm tra xem e có trong E và f có trong F không. Nếu đúng, bổ sung p(f|e).
3. Đánh giá lại:
   $p(f|e) = c(f,e) / c(e)$ $[c(e) = \Sigma_f c(f,e)]$
4. Lặp đến khi p(f|e) thay đổi ít.

48

Cách gióng hàng tốt nhất

Với mỗi cặp (E,F), tìm

A = argmax_A P(A|F,E) = argmax_A P(F,A|E)/P(F) = argmax_A P(F,A|E) = argmax_A (ε / (l+1)^m $\Pi_{j=1..m}$ p($f_j|e_{a_j}$)) = argmax_A $\Pi_{j=1..m}$p($f_j|e_{a_j}$)

- Sử sụng thuật toán lập trình động theo kiểuViterbi.
- Tính lại p(f|e)

49

Dịch máy sử dụng cú pháp

50

Tại sao dùng cú pháp

- Cần thông tin ngữ pháp
- Cần các ràng buộc khi sắp lại câu
- Khi chèn các từ chức năng vào câu, cần đặt ở vị trí chính xác
- Khi dịch từ cần sử dụng từ có cùng từ loại với nó

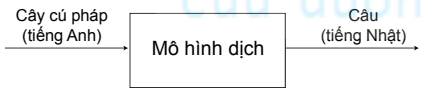
51

Yamada and Knight (2001): Lý do cần cú pháp

- He adores listening to music.

52

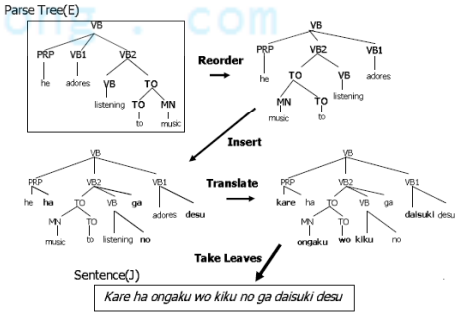
Mô hình dựa trên cú pháp



- Tiền xử lý câu tiếng Anh bằng bộ PTCP
- Thực hiện các phép tính xác suất trên cây cú pháp
  - Sắp lại trật tự các nút
  - Chèn nút mới vào
  - Dịch các từ ở lá

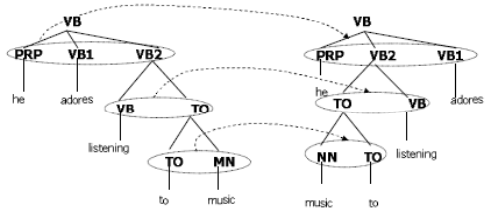
53

Cây cú pháp (Anh) → câu (Nhật)



54

1. Sắp lại trật tự



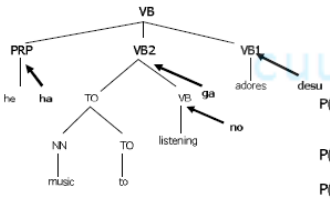
P(PRP VB1 VB2— PRP VB2 VB1) = 0.723
P(VB TO — TO VB) = 0.749
P(TO NN — NN TO) = 0.893

Đặc trưng điều kiện = dãy các nhãn con

Bảng tham số: sắp lại

Trật tự gốc	Sắp lại	P(Sắp lại| Trật tự gốc)
PRP VB1 VB2	PRP VB1 VB2	0.074
	PRP VB2 VB1	**0.723**
	VB1 PRP VB2	0.061
	VB1 VB2 PRP	0.037
	VB2 PRP VB1	0.083
	VB2 VB1 PRP	0.021
VB TO	VB TO	0.107
	TO VB	**0.893**
TO NN	TO NN	0.251
	NN TO	**0.749**

2. Chèn



P(none|TOP-VB) = 0.735

P(right|VB-PRP)* P(ha) = 0.652 * 0.219

P(right|VB-VB) * P (ga) = 0.252 * 0.062

P(none|TO-TO) = 0.900

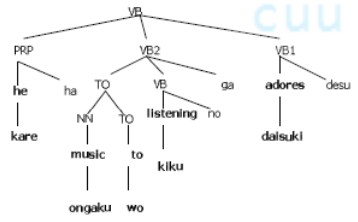
Đặc trưng điều kiện = nhãn cha & nhãn nút (vị trí) & none (là từ)

Bảng tham số: chèn

Parent label node level	TOP VB	VB VB	VB TO	TO TO	TO NN	TO NN
P (none)	0.735	0.687	0.344	0.700	0.900	0.800
P (left)	0.004	0.061	0.004	0.030	0.003	0.096
P (right)	0.260	0.252	0.652	0.261	0.097	0.104

W	P (insert-w)
ha	**0.219**
ta	0.131
wo	0.099
no	**0.094**
ni	0.080
te	0.079
ga	**0.062**
desu	**0.0007**

3. Dịch



P (he—— kare) = 0.952
P (music ——ongaku) =0.900
P (to— wo) = 0.038
P (listening —kiku) = 0.333
P (adore —daisuki) = 1.000

Đặc trưng điều kiện = từ (tiếng Anh)

Bảng tham số: Dịch

E	adores	he	listening	music	to
J	**daisuki 1.000**	**kare 0.952** NULL 0.016 nani 0.005 da 0.003 shi 0.003	**kiku 0.333** kii 0.333 mi 0.333	**ongaku 0.900** naru 0.100	**ni 0.216** NULL 0.204 to 0.133 no 0.046 wo 0.038

Ghi chú: Dịch thành NULL → xóa

Thử nghiệm

- Dữ liệu luyện: 2000 cặp câu J-E
- J: tách từ sử dụng Chasen
- E: PTCP sử dụng bộ PTCP Collins
  - Luyện trên 40000 câu từ Treebank, độ cx ~90%
- E: từ cây cú pháp, xác định trật tự từ và chuyển đổi (SVO <-> SOV)
- Luyện sử dụng EM: 20 vòng lặp

61

Kết quả

	Điểm trung bình	#câu
Y/K model	0.582	10
IBM model 5	0.431	0

- Điểm trung bình được tính trên 3 người với 50 câu
- ok(1.0), không chắc (0.5), sai (0.0)
- chỉ tính độ chính xác

62

Kết quả: giống hàng 1

Syntax-based Model

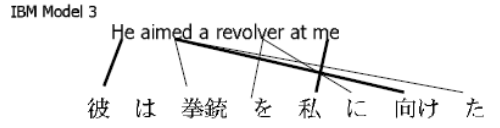
IBM Model 3

63

Kết quả: giống hàng 2

Syntax-based model

IBM Model 3



64

Một số hệ thống dịch máy trên Internet

- http://www.google.com/language_tools?hl=en
- http://www.systransoft.com/index.html
- http://babelfish.altavista.digital.com/

65