ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Dịch máy

Viện CNTT &TT – Trường ĐHBKHN

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

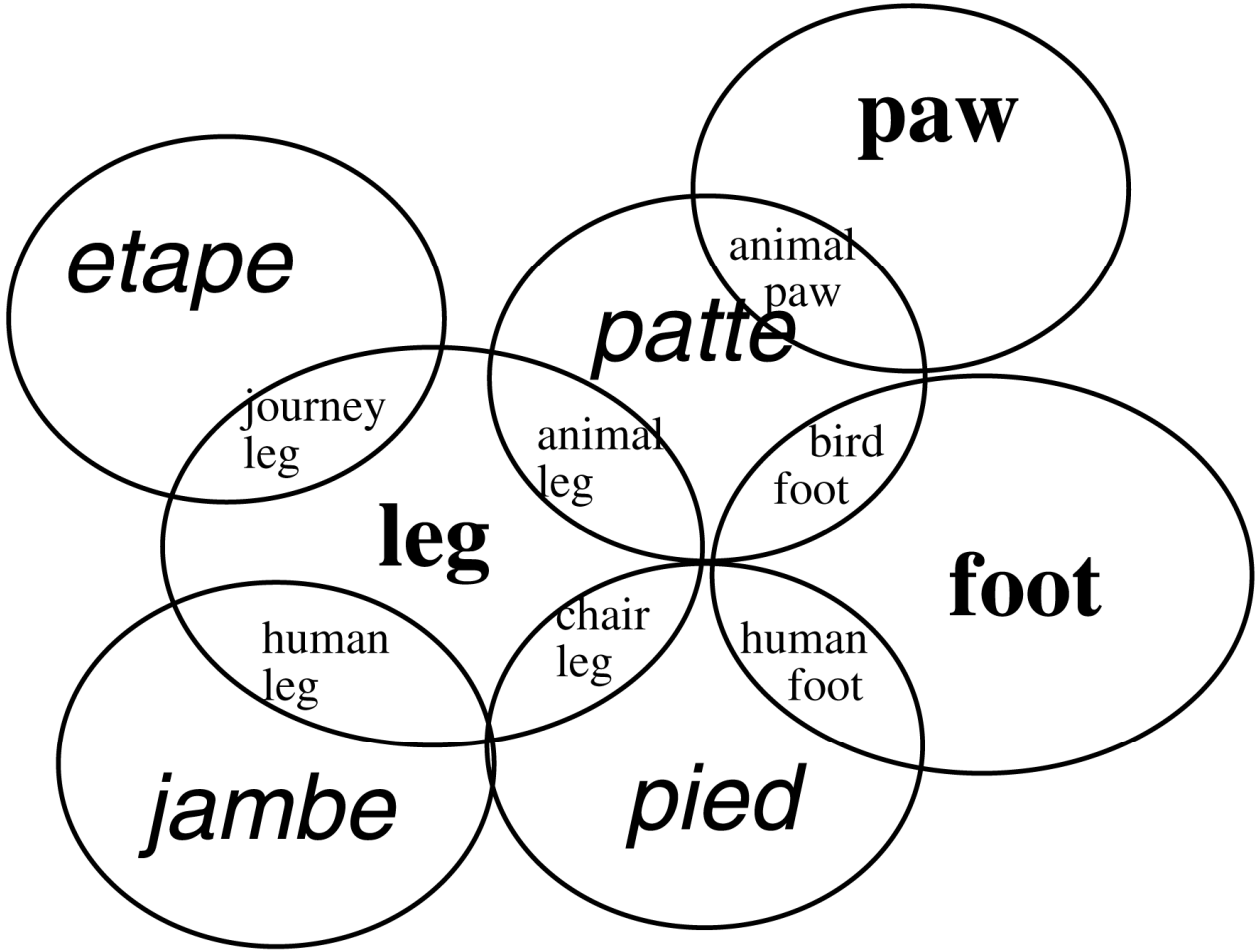
Các vấn đề

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

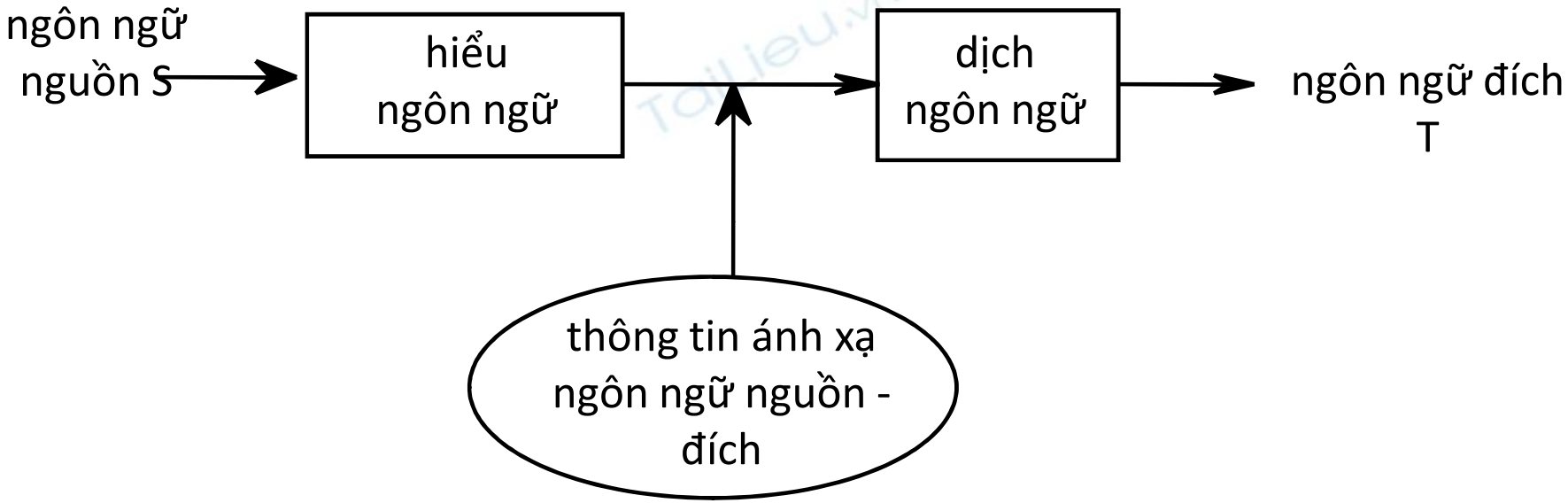
English	brother	Vietnamese	anh em
English	wall	German	wand (inside) mauer(outside)
German	berg	English	hill mountain

Không gian khái niệm



Khoảng trống từ vựng: tiếng Nhật không có từ nào nghĩa *privacy;* tiếng 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*

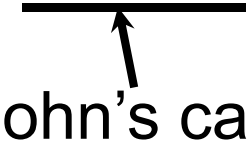
$\Rightarrow$ Sử dụng thông tin cú pháp

2. Nhập nhằng cú pháp:

I saw the guy on the hill with the telescope

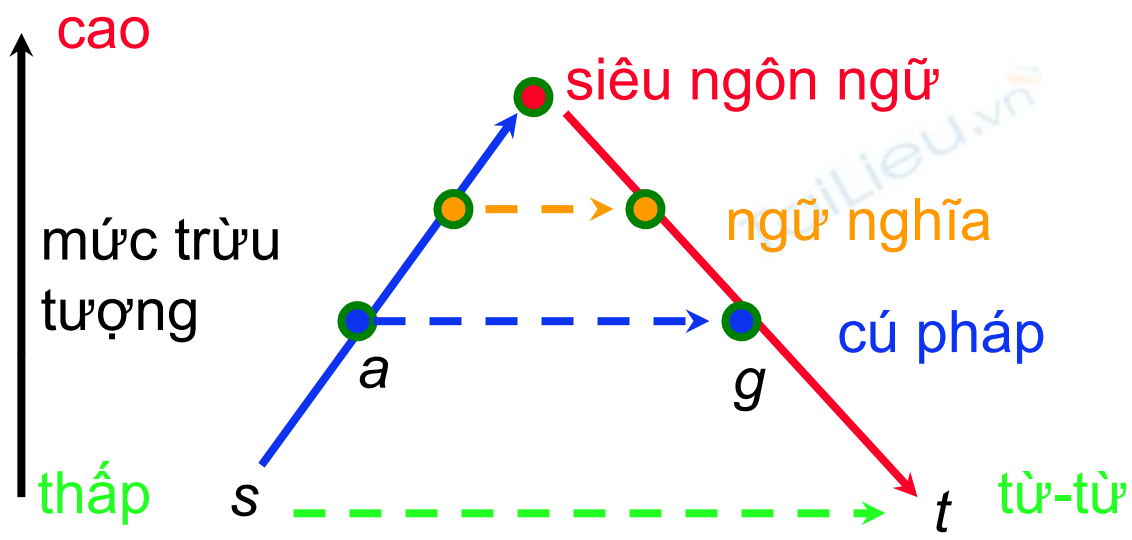
3. Nhập nhằng ngữ nghĩa:

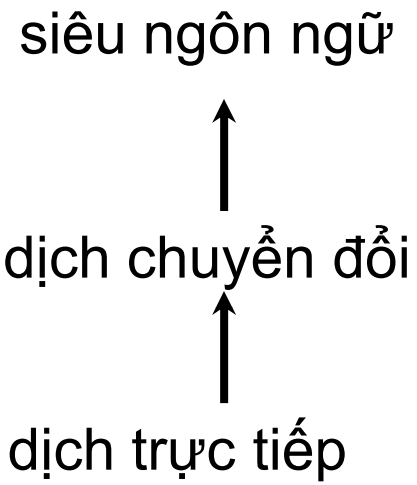
- *E: While driving, John swerved & hit a tree*



- *S: Minetras que John estaba manejando, se desvio y golpeop con un arbo*

Các phương pháp dịch máy



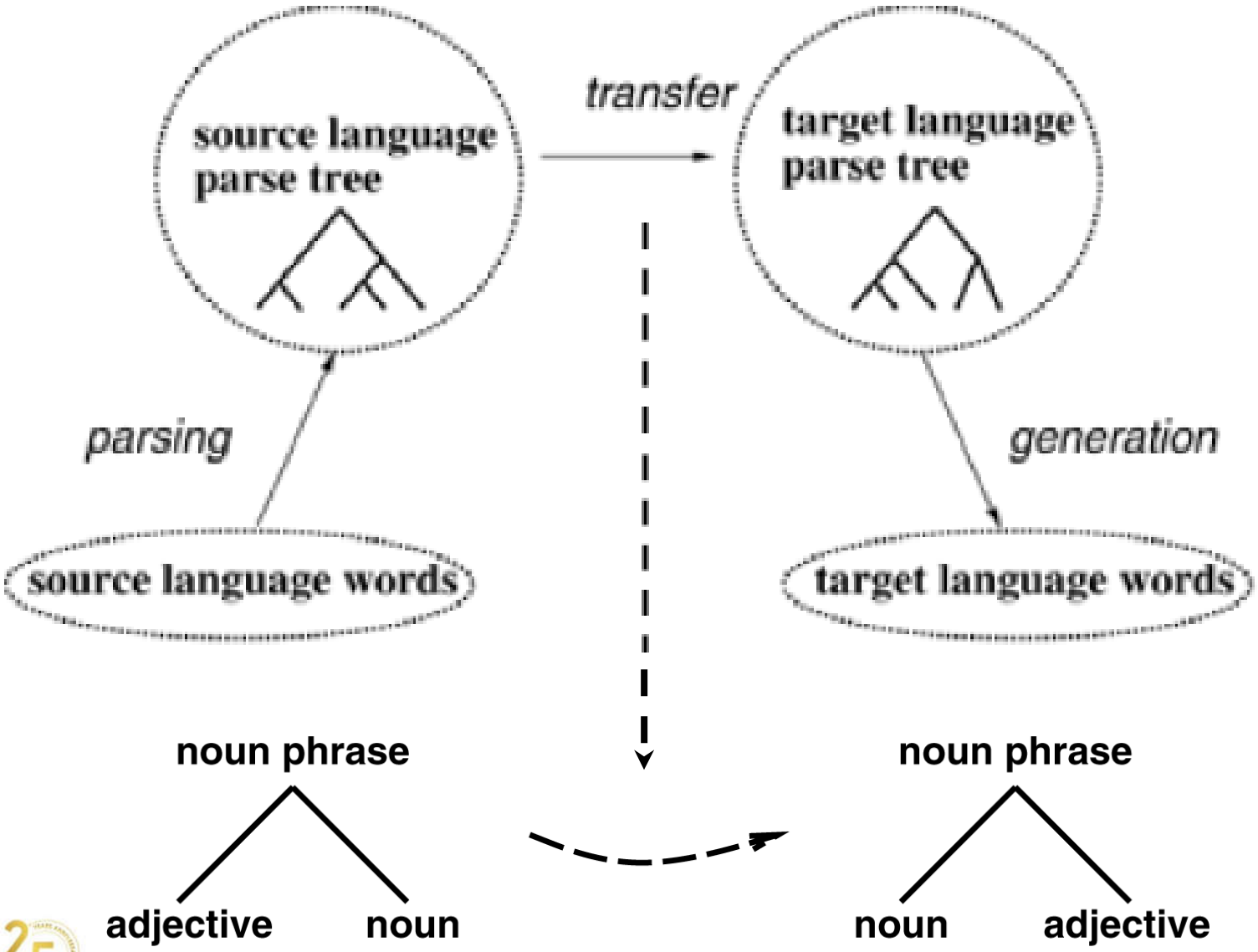


a = *a(s)*

g = *f(a(s)); f* – hàm chuyển đổi

t=g(f(a(s)))

Sơ đồ chuyển đổi



Luật chuyển đổi

English to French:

1. $NP \rightarrow Adjective_1\ Noun_2$

   $\Rightarrow$

   $NP \rightarrow Noun_2\ Adjective_1$

Japanese to English:

2. $Existential\text{-}There\text{-}Sentence \rightarrow There_1\ Verb_2\ NP_3\ Postnominal_4$

   $\Rightarrow$

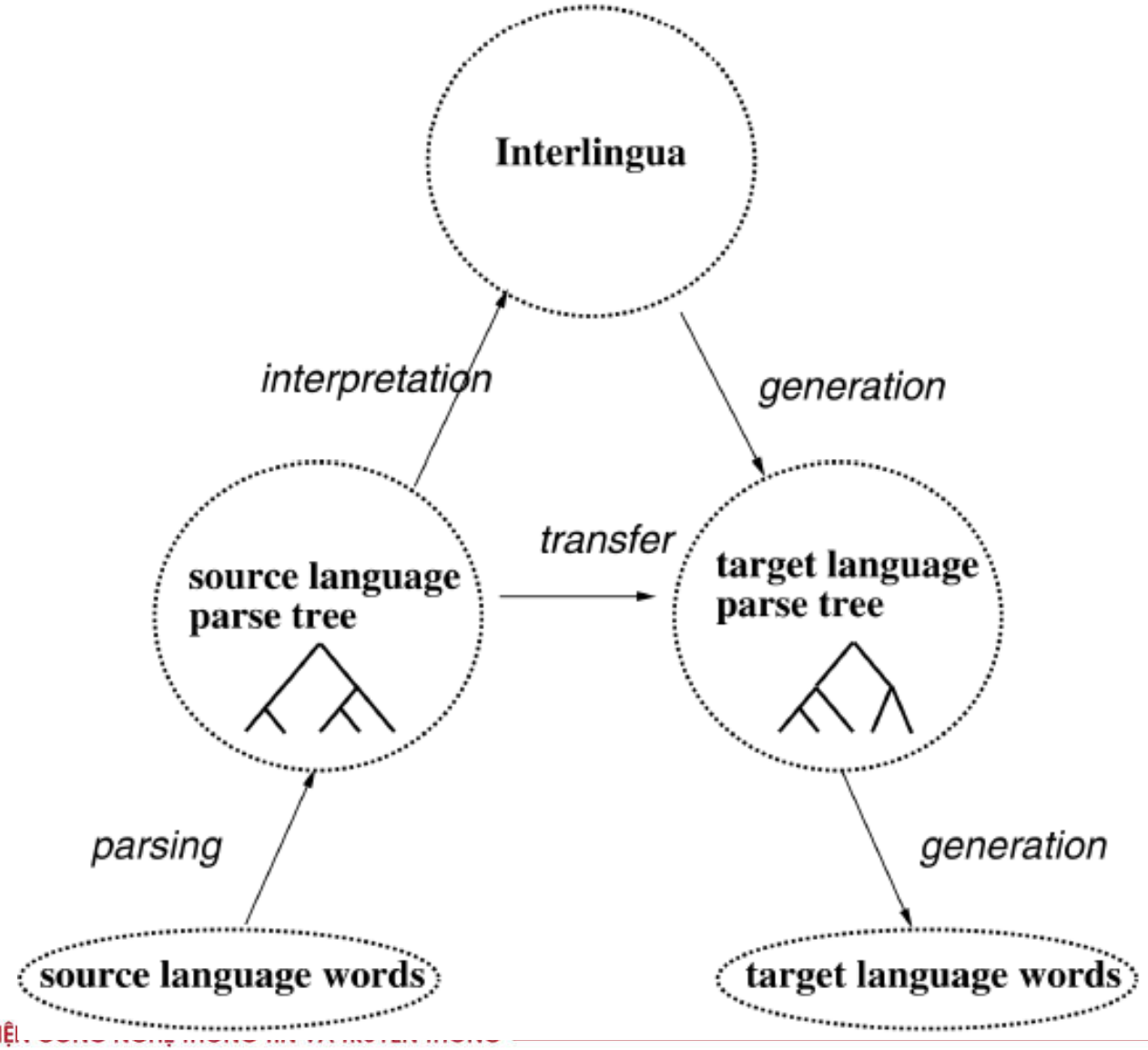
   $Sentence \rightarrow (NP \rightarrow NP_3\ Relative\text{-}Clause_4)\ Verb_2$

3. $NP \rightarrow NP_1\ Relative\ Clause_2$

   $\Rightarrow$

   $NP \rightarrow Relative\text{-}Clause_2\ NP_1$

Sơ đồ chuyển đổi



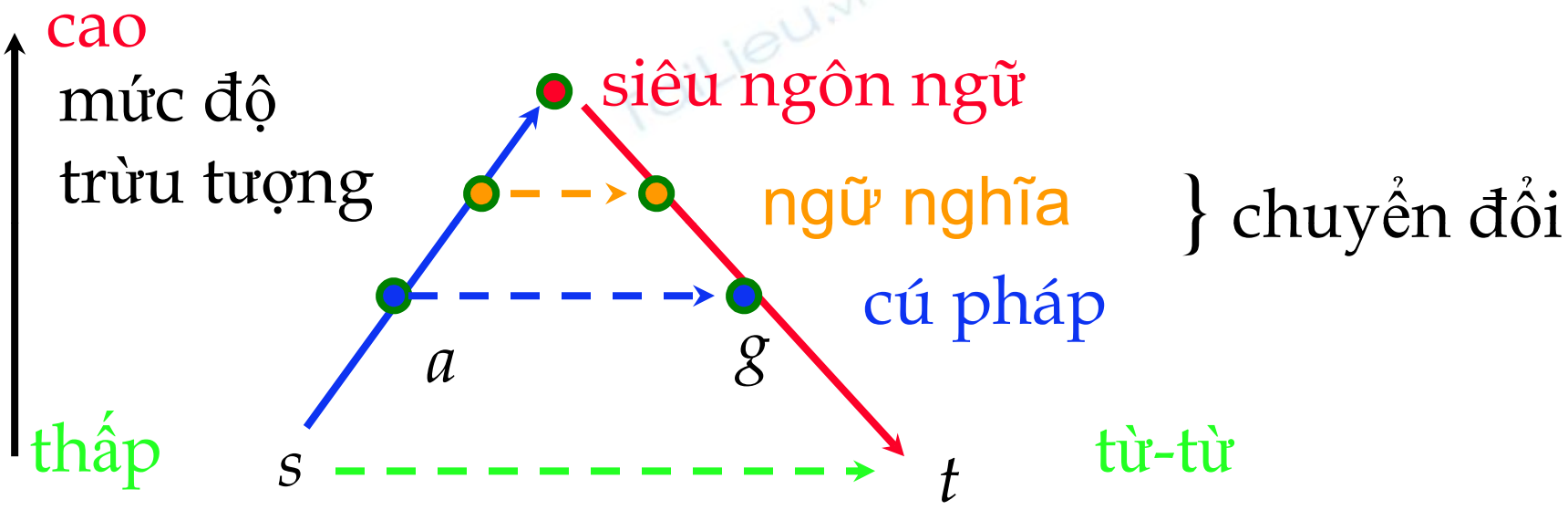
Cách tiếp cận siêu ngôn ngữ: sử dụng nghĩa

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

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

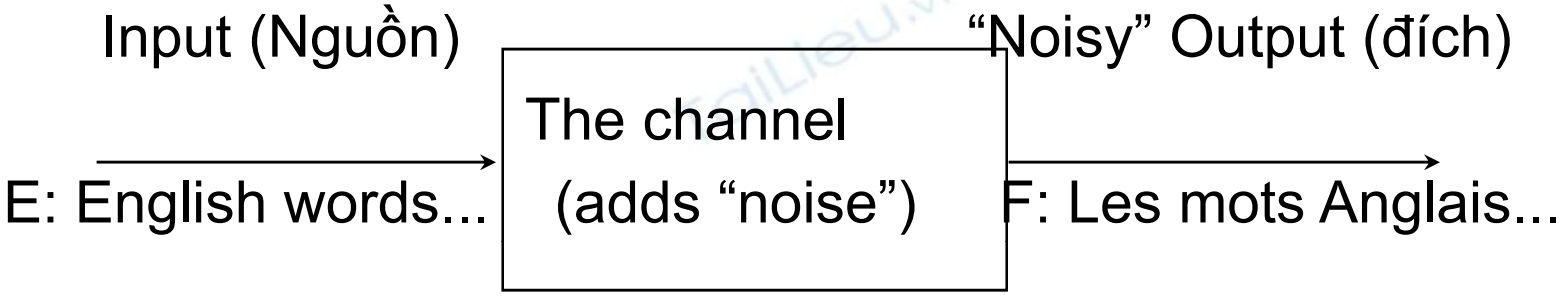
Dịch máy thống kê

Các kiểu dịch máy



ý tưởng

- Coi việc dịch như bài toán kênh có nhiễu

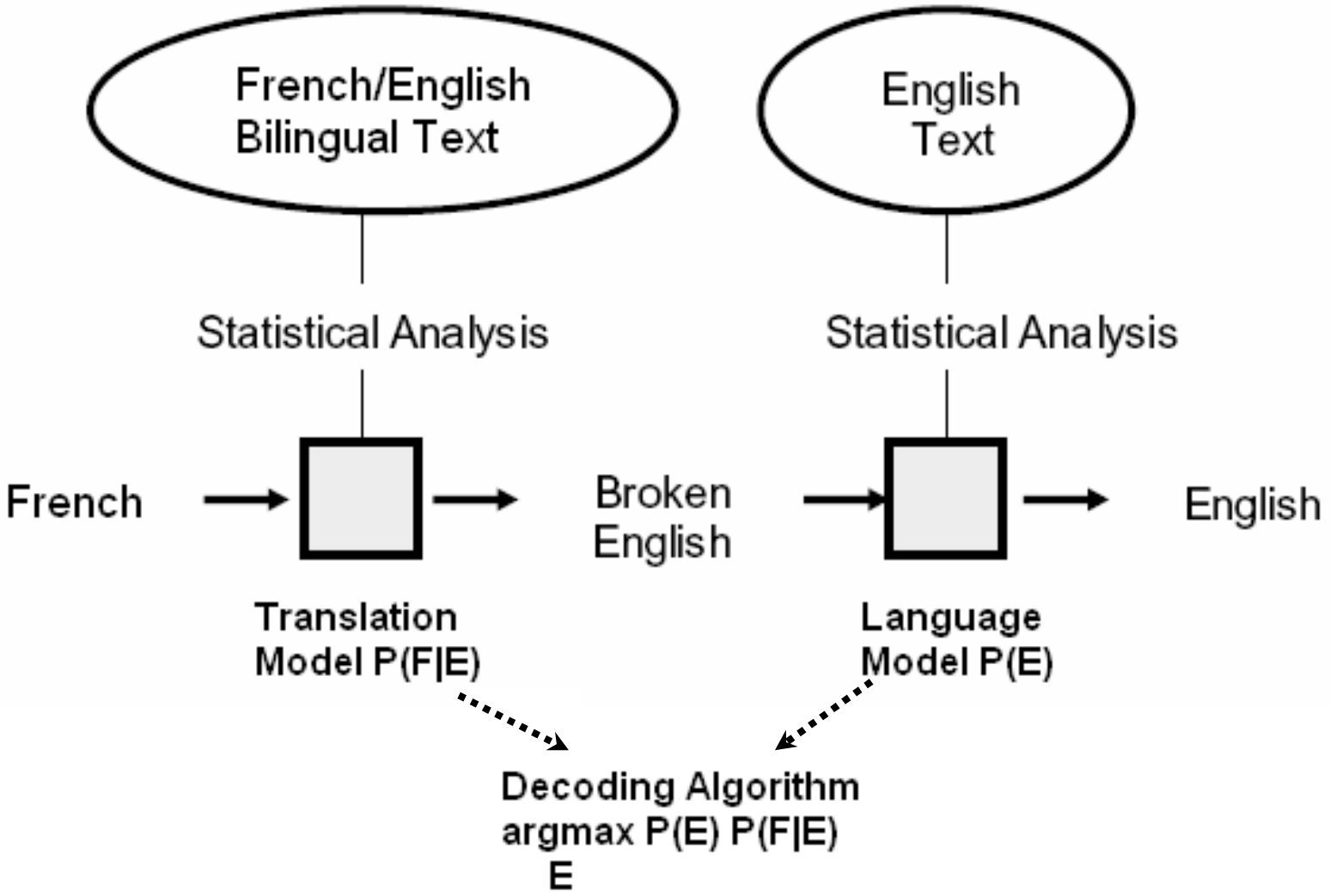


- Mô hình dịch: $P(E|F) = P(F|E)\,P(E) / P(F)$
- Khôi phục lại E khi biết F:

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

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

Dịch máy thống kê



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 $E \leftrightarrow F$
  - Thậm thí bình thường ta không thấy 1 câu lặp lại 2 lần

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”.

Ý tưởng gióng hàng

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

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.