Học viện Công nghệ Bưu chính Viễn thông
Khoa Công nghệ thông tin 1

Xử lý ngôn ngữ tự nhiên
(Natural Language Processing-NLP)

Một số ứng dụng

Ngô Xuân Bách

Nội dung

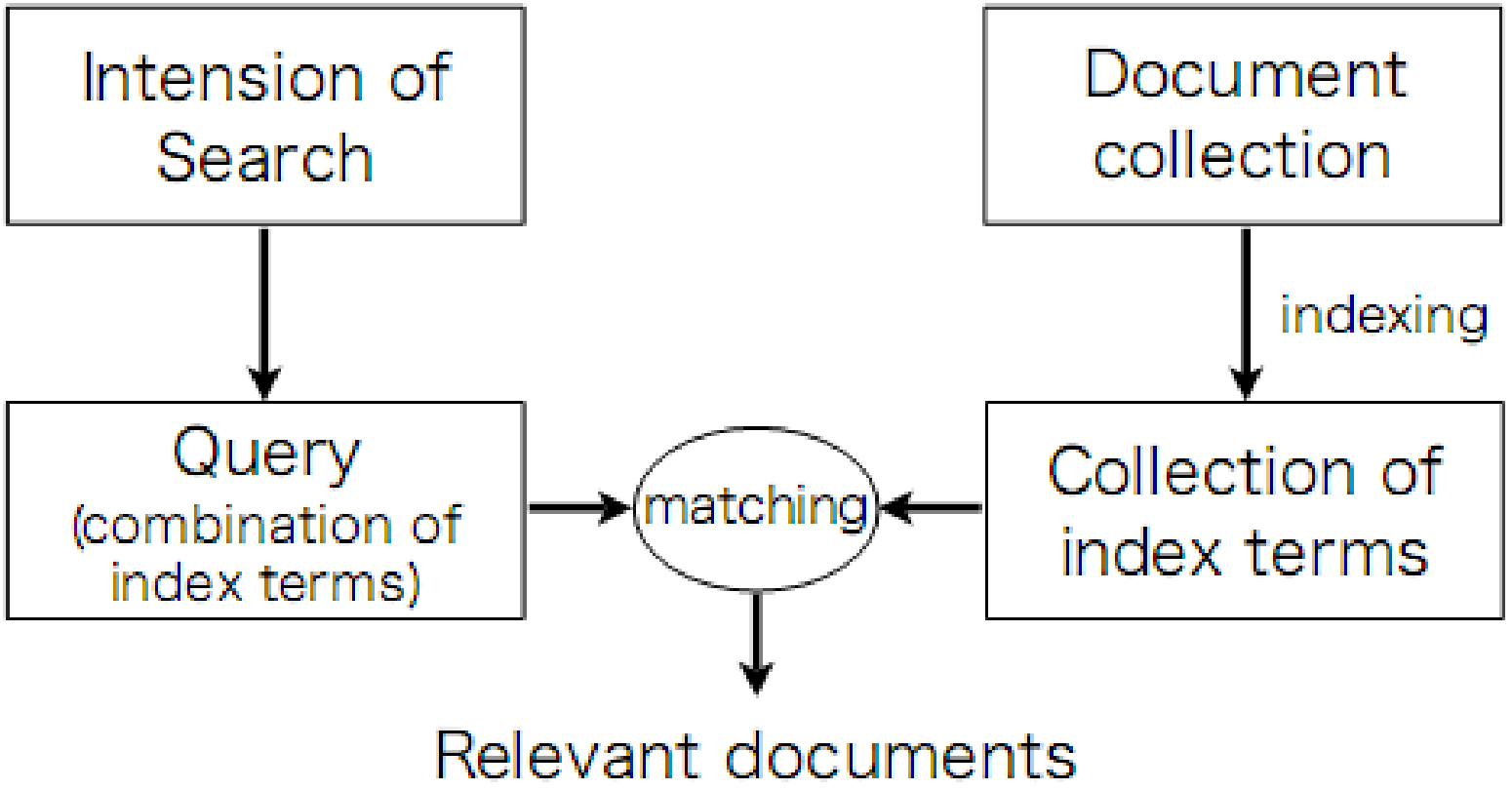
- Truy xuất thông tin (Information retrieval)
- Trích chọn thông tin (Information extraction)

Truy xuất thông tin

- Information Retrieval (IR)
- IR theo nghĩa rộng
  - Tìm kiếm thông tin từ các nguồn để giải quyết vấn đề
- IR theo nghĩa hẹp
  - Tìm kiếm các tài liệu (documents) phù hợp với một câu truy vấn (query) của người dùng trong một tập các tài liệu
    - Tài liệu hợp lệ; tài liệu chứa câu trả lời truy vấn của người dùng
  - Truy xuất văn bản

Truy xuất thông tin

- So khớp với từ chỉ mục (index term)



Câu truy vấn (Query)

- Một từ chỉ mục hoặc kết hợp các từ chỉ mục
- Cách tạo câu truy vấn
  - Sử dụng trực tiếp từ các từ chỉ mục
    - Thường theo dạng logic
    - V.d $(t_a$ and not $t_b)$ or t_c
  - Sử dụng ngôn ngữ tự nhiên
    - Yêu cầu chuyển đổi tự động sang từ chỉ mục
    - V.d. "I want to know how to make cake"
      → make and cake

Đánh chỉ mục (Indexing)

- Thủ tục trích ra các từ chỉ mục từ các tài liệu
  - Đánh chỉ mục nên được thực hiện tự động
    - Bởi vì số lượng các tài liệu là khá lớn
  - Cần phân tích hình vị
- Đơn vị của từ chỉ mục
  - Từ (cake, recipe, ingredient)
  - Cụm từ (recipe for cake, ingredient of case)
  - Khó xác định được đơn vị phù hợp
    - Nhìn chung, đơn vị từ hay được sử dụng nhất làm từ chỉ mục

Từ dừng (stop word)

- Từ dừng là gì?
  - Từ không phải là từ chỉ mục
- Cụ thể
  - Từ chức năng (functional words)
    - Tiếng Anh: deteminer, preposition, v.v
  - Không phải là từ mang nội dung (content word)
    - Là những từ mang ngữ nghĩa như danh từ, động từ
- Những từ xuất hiện trong nhiều văn bản, do đó nó không cung cấp ngữ cảnh hiệu quả cho IR

So khớp

- Chỉ mục ngược
- Mô hình không gian vector (Vector Space Model – VSM)

Chỉ mục ngược

- Xây dựng danh sách từ khóa chỉ mục cho mỗi tài liệu

	Novel	Story	Book review	Mystery
D_1	**1**	**0**	**0**	**0**
D_2	**1**	**0**	**1**	**1**
D_3	**1**	**1**	**0**	**0**
D_4	**0**	**0**	**1**	**0**

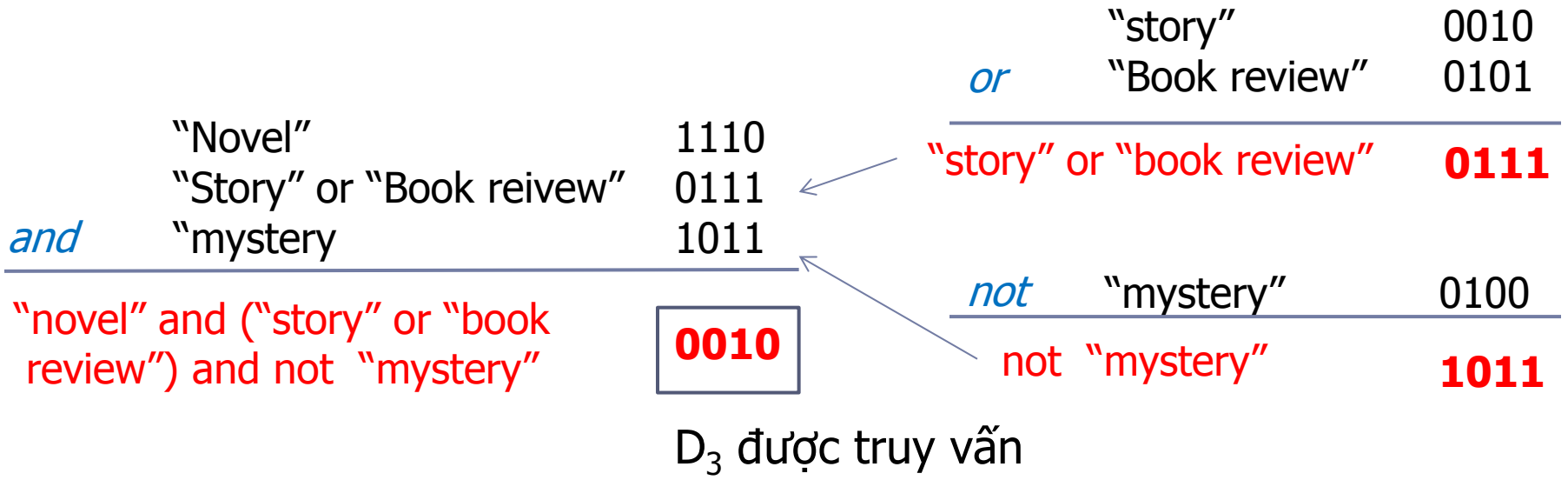
Chỉ mục ngược

- Đảo ngược ma trận
  - Dễ dàng biết được danh sách tài liệu chứa một từ chỉ mục

	D_1	D_2	D_3	D_4
Novel	**1**	**1**	**1**	**0**
Story	**0**	**0**	**1**	**0**
Book review	**0**	**1**	**0**	**1**
Mystery	**0**	**1**	**0**	**0**

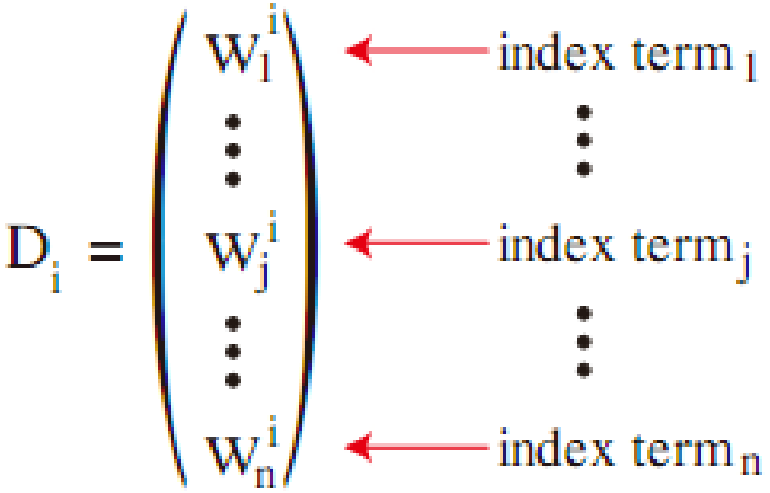
Chỉ mục ngược

- Khi cho một câu truy vấn ở dạng logic
  - Hàng trong ma trận chỉ mục ngược được coi như là một vector
  - Các tài liệu được truy vấn bằng thao tác ở mức bit trên vector
- "Novel" and ("Story" or "Book review") and not "mystery"



Mô hình không gian vector

- Cả tài liệu và câu truy vấn được biểu diễn dưới dạng vector
  - Vector tài liệu = D_i, vector truy vấn = Q
  - Tính độ tương tự giữa hai vector
    - Trả về tài liệu D_i có độ tương tự gần nhất với Q
- Vector
  - W^i_j là trọng số của từ chỉ mục



Tính trọng số cho từ chỉ mục

- Cách tính trọng số đơn giản
  - 1 nếu từ xuất hiện trong văn bản, ngược lại là 0
- TF.IDF
  - TF (term frequency – tần suất xuất hiện của từ)
    - tf_j^i: tần suất của từ chỉ mục j trong văn bản i
    - Trong cùng một văn bản, một từ xuất hiện càng nhiều thì từ đó càng quan trọng cho IR

Tính trọng số cho từ chỉ mục

- TF.IDF (tiếp)
  - IDF (inverse document frequency – tần suất văn bản nghịch đảo)
    - $idf = \log\frac{N}{df_i}$
    - df_i: tần suất xuất hiện của văn bản
      (số lượng văn bản chứa từ chỉ mục j)
  - Nếu một từ xuất hiện trong nhiều văn bản thì từ đó trở thành không quan trọng trong bài toán IR
- Trọng số của từ chỉ mục

$$w_j{}^i = tf_j{}^i . idf_j = tf_j{}^i . \log\frac{N}{df_i}$$

Độ tương tự giữa 2 vector

- Độ tương tự (similarity): sim(D_i, Q)
  - Truy vấn n văn bản có độ tương tự lớn nhất
- Ví dụ độ tương tự
  - Tích vector

$$Sim(D_i, Q) \stackrel{def}{=} D_i \cdot Q = \begin{pmatrix} w_1^i \\ \vdots \\ w_n^i \end{pmatrix} \cdot \begin{pmatrix} q_1 \\ \vdots \\ q_n \end{pmatrix} = \sum_j w_j^i q_j$$

  - Độ đo cosine

$$Sim(D_i, Q) \stackrel{def}{=} \cos\theta = \frac{D_i \cdot Q}{|D_i|\,|Q|}$$

Đánh giá truy xuất văn bản

- Hệ thống truy xuất văn bản điển hình
  - Đầu vào nhận một truy vấn Q
  - Truy xuất n văn bản phù hợp với Q
    - Chọn n văn bản theo thứ tự của $sim(Di, Q)$
  - Có thể thay đổi số lượng output (số văn bản được truy vấn) dễ dàng

Đánh giá truy xuất văn bản

- Tiêu chí đánh giá
  - Precision (P)

$$\frac{\#\ văn\ bản\ hợp\ lệ\ hệ\ thống\ trả\ về}{\#văn\ bản\ hệ\ thống\ trả\ về}$$

  - Recall (R)

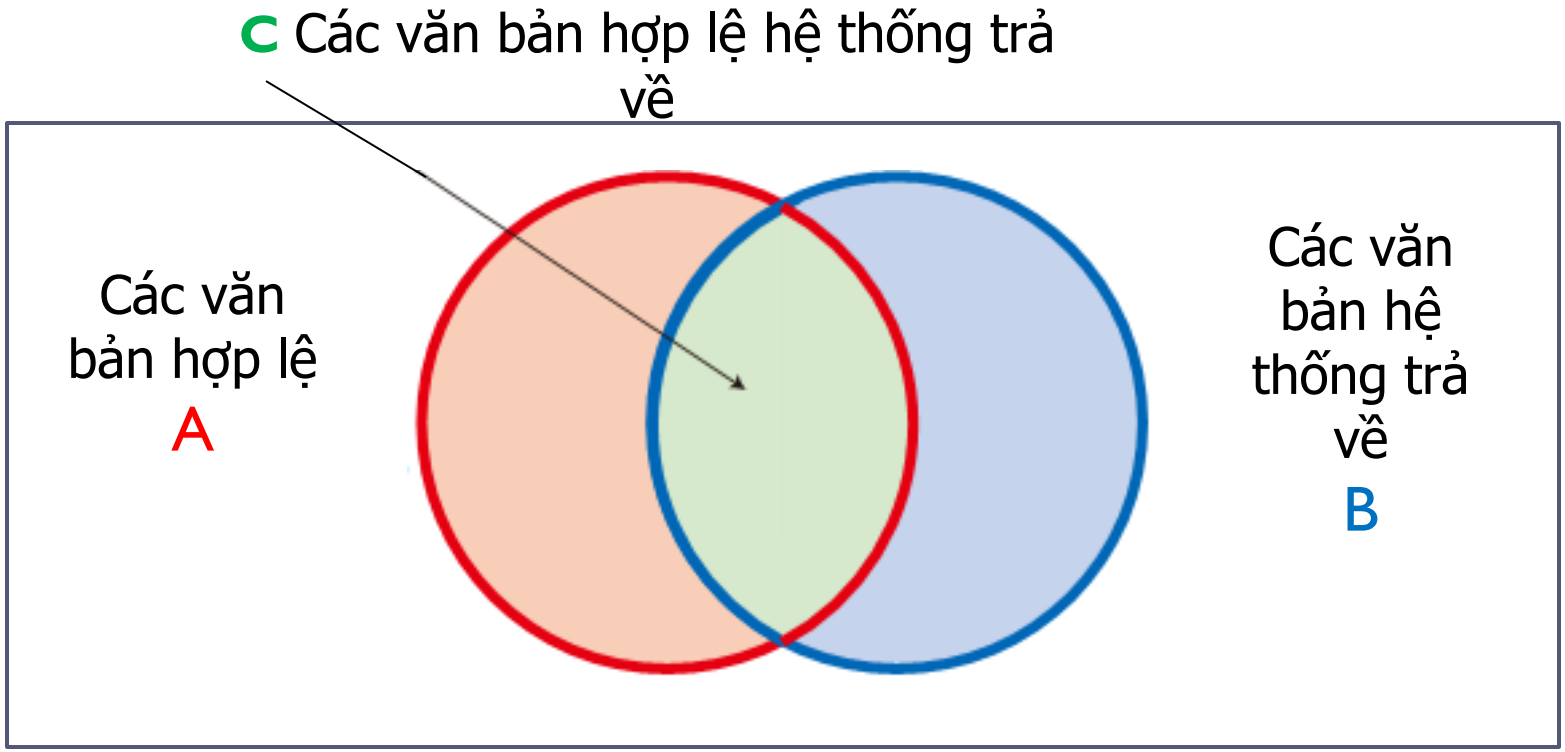
$$\frac{\#văn\ bản\ hợp\ lệ\ hệ\ thống\ trả\ về}{\#văn\ bản\ hợp\ lệ\ trong\ toàn\ bộ\ tập}$$

  - F-measure

$$F = \frac{2PR}{P + R}$$

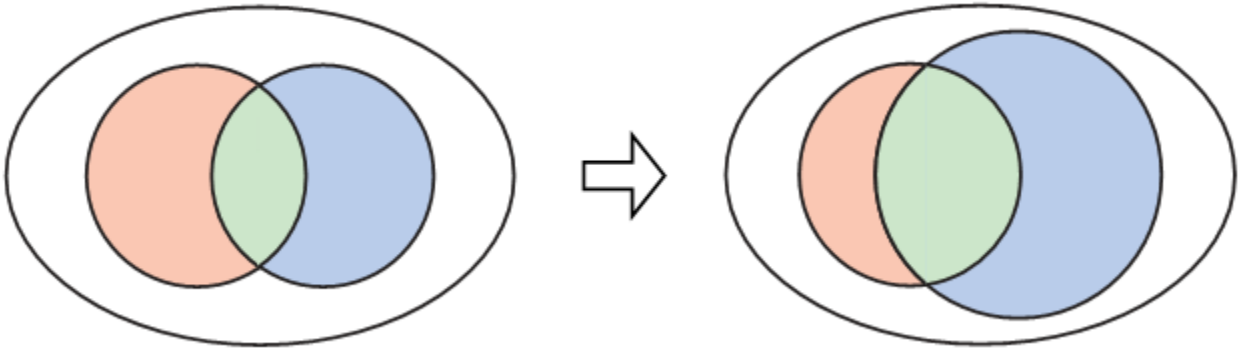
Precision và Recall

- Precision = C/B
- Recall = C/A



Precision và Recall

- Cần phải trade-off
- Nếu hệ thống trả về nhiều văn bản hơn
  - Precision sẽ thấp trong khi recall sẽ cao hơn



Precision và Recall

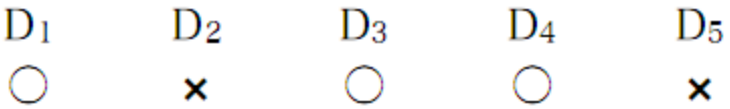
- Trường hợp cần quan tâm hơn tới Precision
  - Khi cần hệ thống chỉ hiển thị các văn bản hợp lệ cho người dùng
  - Ví dụ: máy tìm kiếm Web (Web search engine)
- Trường hợp cần quan tâm hơn tới Recall
  - Khi cần hệ thống truy xuất hầu hết các văn bản hợp lệ
  - Ví dụ: Patent text retrieval
- Các trường hợp cần xem xét cả hai
  - Đánh giá bằng độ đo F

Cải tiến truy xuất văn bản

- Hướng tới truy xuất văn bản chuẩn xác hơn
  - Phản hồi hợp lệ (relevance feedback)
  - Mở rộng truy vấn (query expansion)

Relevance Feedback

- Hiếm khi có kết quả tốt với một lần tìm kiếm
  → Tìm kiếm tương tác với người dùng
- Luồng (Flow)
  - Đầu tiên hệ thống truy xuất các văn bản
    - Hệ thống trả về n văn bản cho người dùng
  - Người dùng đánh giá xem các văn bản trả về có hợp lệ hay không



Relevance Feedback

- Flow (tiếp)
  - Vector truy vấn Q được chỉnh lại như sau

$$Q' = Q + \frac{1}{|R|}\sum_{D_i \in R} D_i - \frac{1}{|N|}\sum_{D_i \in N} D_i$$

    - R: tập các văn bản hợp lệ do người dùng đánh giá
    - N: tập các văn bản không hợp lệ do người dùng đánh giá
  - Tìm kiếm lại với vector truy vấn mới Q'
  - Lặp lại thủ tục trên

Relevance Feedback

- Tính hiệu quả của Relevance Feedback
  - Các văn bản tương tự với các văn bản hợp lệ sẽ được truy vấn mới
  - Các văn bản tương tự với các văn bản không hợp lệ sẽ không được truy vấn
  - Khả năng sẽ cải tiến precision/recall
- Giả relevance feedback
  - Người dùng không đánh giá các văn bản hợp lệ
  - Thay vào đó, coi n văn bản đầu tiên được trả về là các văn bản hợp lệ
  - Thủ tục được thực hiện hoàn toàn tự động

Mở rộng truy vấn

- Các cách biểu diễn khác nhau trong ngôn ngữ tự nhiên
  - Khi ta có một truy vấn chứa "car"
  - Ta không thể truy vấn các văn bản có từ "cars automobile automobiles auto", v.v...
- Mở rộng truy vấn là gì
  - Thủ tục tự động thêm các từ liên quan vào câu truy vấn
  - Q = (car)

    → Q = (car, cars, automobile, automobiles, auto)

  - Khả năng sẽ cải tiến được recall

Mở rộng truy vấn

- Nên thêm vào các loại từ nào?
  - Biến thể (Variant)
    - interest, interesting, interested
  - Đồng nghĩa (Synonym)
    - Human being, people
  - Hypernym
    - Beer → liquor
  - Hypornym
    - Liquor → beer, wine, whisky
- Thường sử dụng các loại từ điển

Tổng kết

- Các phương pháp truy xuất văn bản
  - Biểu diễn văn bản bằng từ chỉ mục
  - Chỉ mục ngược
  - Mô hình không gian vector
    - Tính trọng số cho từ bằng TF.IDF
- Tiêu chí đánh giá
  - Precision, Recall, F-measure
- Cải tiến truy xuất văn bản
  - Relevance feedback
  - Mở rộng truy vấn

Trích chọn thông tin

Trích chọn thông tin

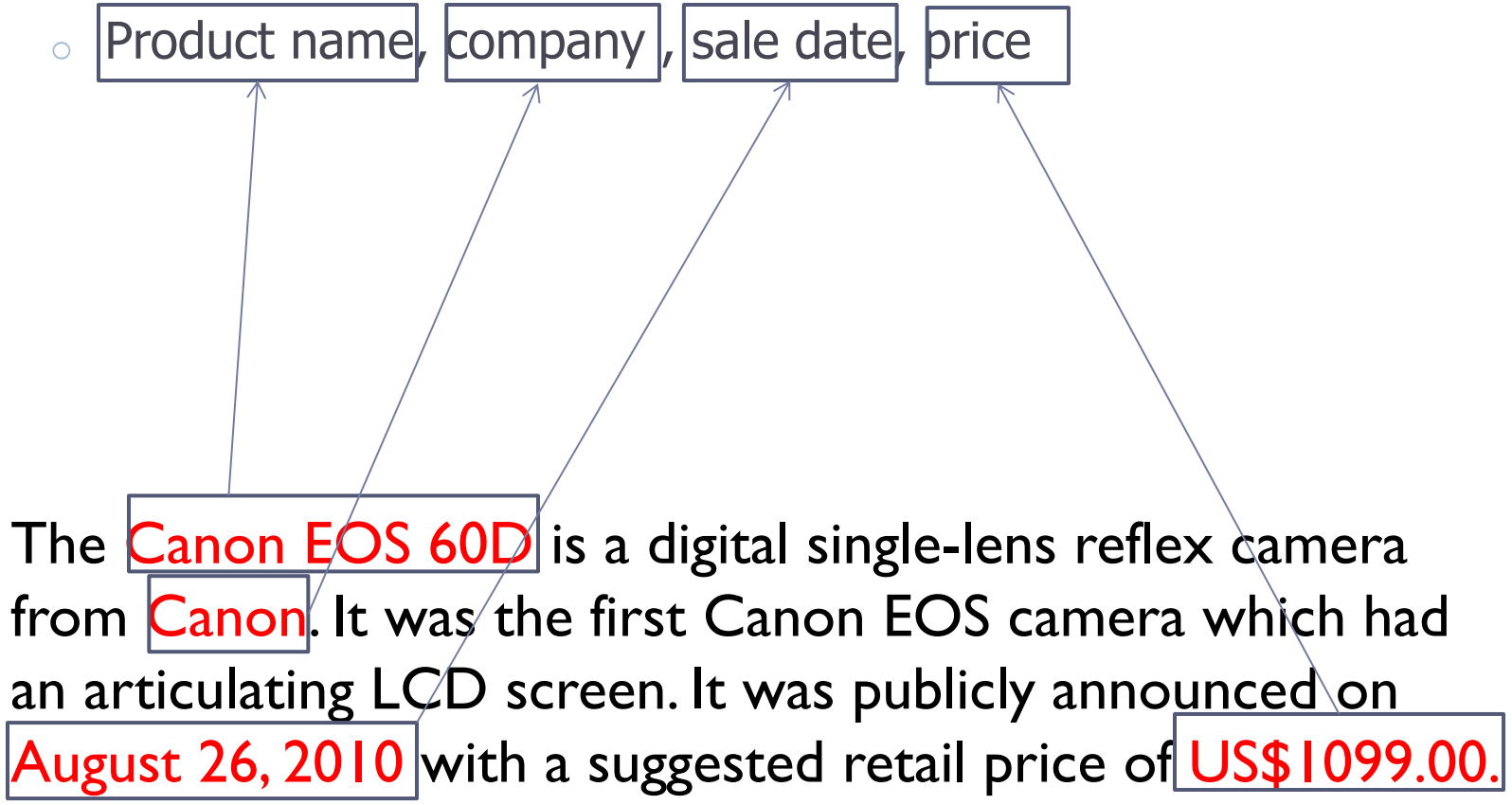
- Trích chọn thông tin (IE)
  - Là bài toán trích chọn trực tiếp thông tin mong muốn từ văn bản

(on the other hand)

- Truy xuất thông tin – Information Retrieval (truy vấn văn bản – Text Retrieval)
  - Là bài toán tìm ra các văn bản chứa thông tin mong muốn được mô tả trước

Trích chọn thông tin

- Ví dụ thông tin mong muốn
  - Product name, company, sale date, price



Trích chọn thông tin sử dụng Frame

- Thông tin được trích chọn được định nghĩa trước trong một khung (frame)
- Ví dụ frame

Slot	Slot value
Company	Canon
Sale date	August 26th, 2010
Product name	Canon EOS 60D
Price	US$1099.00

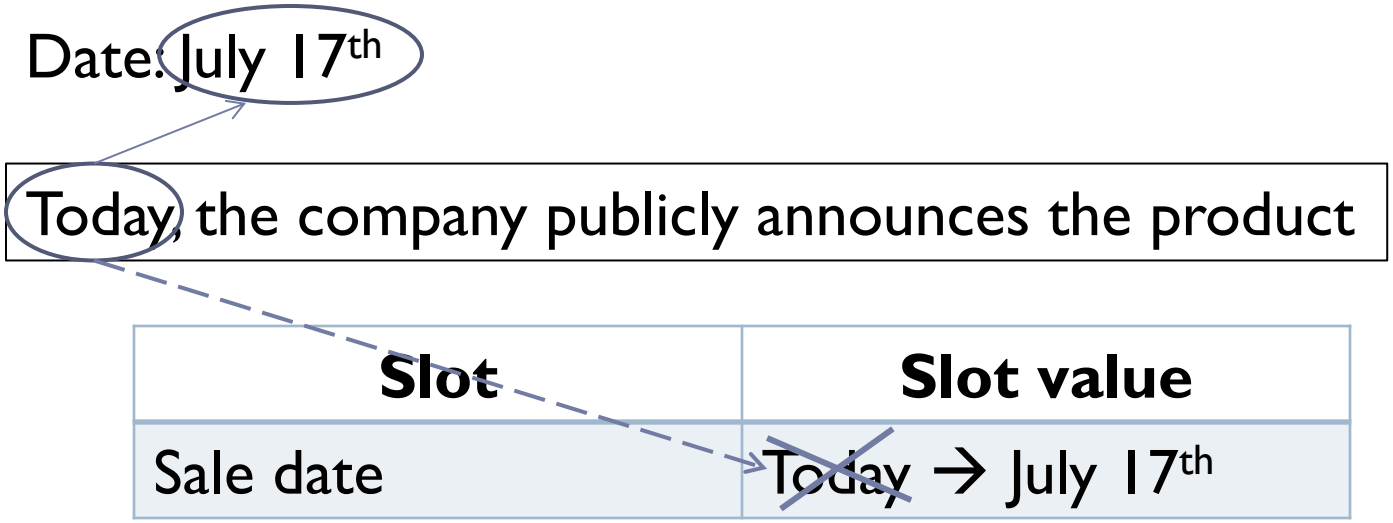
Các kỹ thuật IE

- IE bằng so khớp mẫu (pattern matching)
  - Ví dụ mẫu
    - <product_name> is a product of <company>
    - <product_name> is publicly announced on <sale_date>
    - Ta giả sử rằng thông tin cần trích chọn tuân theo một số kiểu nhất định nào đó
  - Tạo mẫu
    - Xây dựng bằng tay
    - Xây dựng tự động từ kho ngữ liệu

Các kỹ thuật IE

- Sử dụng thông tin chứ không phải văn bản (text)
  - Ngày bán

    → Xác định được ngày bán sản phẩm



Hệ trả lời câu hỏi

- Hệ trả lời câu hỏi là gì?
  - Hệ Question Answering (QA)
  - Hệ thống tìm ra câu trả lời cho mỗi câu hỏi của người dùng từ một tập các văn bản
  - Input: câu hỏi (một câu)
    - Where is the capital of East Timor?
  - Output: câu trả lời
    - Dili
  - Nguồn tri thức: tập các văn bản (document collection)
    - Newspaper
    - www

Hệ QA

- QA khác với truy vấn thông tin (IR)
  - IR: output là văn bản chứa câu trả lời
    - Người dùng tự tìm câu trả lời trong văn bản trả về
  - QA: output là câu trả lời

Hệ QA

- QA khác với trích chọn thông tin (IE)
  - IE: Chỉ thông tin giới hạn (limited information) được trích chọn
    - Sử dụng frames
  - QA: Không có giới hạn về thông tin được truy vấn
    - Người dùng có thể hỏi bất kỳ câu hỏi nào
    - Trên thực tế, câu hỏi về facts thường được quan tâm nhiều nhất (factoid questions)
  - IE: Văn bản chứa câu trả lời được xác định trước
  - QA: Văn bản chứa câu trả lời được truy vấn trước

Các bước trong hệ QA

- Phân tích câu hỏi – Question Analysis
  - Để hiểu được ý của người dùng
- Truy vấn văn bản – Text Retrieval
  - Truy vấn các văn bản chứa câu trả lời từ một tập văn bản
- Trích chọn câu trả lời – Answer Extraction
  - Trích ra câu trả lời từ các văn bản được truy vấn

Phân tích câu hỏi

- Input và Output
  - Input: câu hỏi
  - Output: loại câu truy vấn (query type), truy vấn (query)
- Thủ tục
  - Xác định loại câu truy vấn (ý của câu hỏi) – (hỏi cái gì)
    - Where is the capital of East Timor? → hỏi về location
    - When did World War II start? → hỏi về time
    - Which company developed Play Station? → hỏi về organization
  - Trích chọn từ khóa từ câu hỏi
    - Để tạo truy vấn cho phần truy vấn văn bản – text retrieval

Truy vấn văn bản

- Input và Output
  - Input: câu truy vấn
  - Output: văn bản chứa (được coi là chứa) câu trả lời
    - Nhìn chung có nhiều văn bản được truy vấn
- Thủ tục
  - Sử dụng các kỹ thuật cho truy vấn văn bản

Trích chọn câu trả lời

- Input và Output
  - Input: các văn bản, loại câu truy vấn
  - Output: câu trả lời
- Thủ tục
  - Áp dụng các kỹ thuật IE
    - Pattern matching
      - Khó chuẩn bị pattern trước
    - Tạo pattern từ câu hỏi

Question: Where is the capital of East Timor?

⇩

Pattern: The capital of East Timor is <answer>.

Trích chọn câu trả lời

- Thủ tục (tiếp)
  - Sử dụng trích chọn tên thực thể
    - Hầu hết câu trả lời là danh từ riêng
    - Dựa vào loại câu hỏi, ta có thể khoanh vùng được các loại danh từ riêng cho trích chọn câu trả lời
    - Ví dụ: Nếu người dùng hỏi về địa điểm thì chỉ các tên về địa điểm có khả năng là câu trả lời

Trích chọn tên thực thể

- NE hoặc NEE – Name Entity Extraction
- NEE là gì?
  - Trích ra các danh từ riêng trong văn bản
  - Xác định loại của danh từ riêng
    - Location, person, org, product name, v.v…
- Bằng cách nào?
  - Không sử dụng được từ điển danh từ riêng (không sẵn có)
  - Lấy ngữ cảnh từ các từ xung quanh
    - "Ông, bà, cô, dì, …" thường đứng trước tên người
    - "công ty, tổ chức …" thường đứng trước tên org
  - Có thể thu thập thông tin về từ ngữ cảnh này từ kho dữ liệu