



# NGỮ NGHĨA HỌC TÍNH TOÁN

## CHƯƠNG 4 – CẤU TRÚC BIỂU DIỄN DIỄN NGÔN

NGUYỄN TRỌNG CHÍNH



# TRÌNH BÀY

- 1. DIỄN NGÔN**
- 2. LÝ THUYẾT BIỂU DIỄN DIỄN NGÔN**
3. TÍNH TOÁN DRS VĂN BẢN
4. TÍNH TOÁN NGỮ NGHĨA VĂN BẢN



# 1. DIỄN NGÔN

---

## 1.1 DIỄN NGÔN LÀ GÌ?

# 1.1 DIỄN NGÔN LÀ GÌ?

Theo tác giả Đỗ Hữu Châu (2009):

Phát ngôn là câu được dùng ở những ngữ cảnh khác nhau.

Ví dụ: câu “23 giờ rồi!”





# 1.1 DIỄN NGÔN LÀ GÌ?

Theo tác giả Đỗ Hữu Châu (2009):

Diễn ngôn (discourse) là một quá trình sản sinh và liên kết các phát ngôn thành một chỉnh thể. Nó cũng là tên gọi của cái sản phẩm ngôn từ do quá trình đó tạo nên.

- Diễn ngôn còn được gọi là ngôn bản.
- Văn bản là một dạng diễn ngôn.



# 1.1 DIỄN NGÔN LÀ GÌ?

Ví dụ: xét hai đoạn văn bản sau:

*Thông thường chị em có thói quen sấy tóc ở nhiệt độ cao để tóc nhanh khô. Tuy nhiên các chuyên gia tóc cảnh báo việc làm này khiến tóc bị khô, giòn và dễ gãy, rụng.*

Và

*Tuy nhiên các chuyên gia tóc cảnh báo việc làm này khiến tóc bị khô, giòn và dễ gãy, rụng. Thông thường chị em có thói quen sấy tóc ở nhiệt độ cao để tóc nhanh khô.*



# 1.1 DIỄN NGÔN LÀ GÌ?

## Nhận xét:

- Một văn bản được trình bày theo một cấu trúc xác định nhằm biểu diễn ý của người viết.
- Ý nghĩa của một văn bản không đơn giản là sự cộng gộp ý nghĩa của từng câu.

⇒ Phân tích diễn ngôn nhằm mục đích xác định rõ mối liên kết của các câu trong văn bản từ đó xác định ý nghĩa của toàn văn bản.



# 1.1 DIỄN NGÔN LÀ GÌ?

Vấn đề của phân tích diễn ngôn là xác định ý nghĩa của một đoạn văn bản dựa trên ý nghĩa của mỗi câu trong văn bản.

Ví dụ: xác định ý nghĩa của các đoạn văn bản sau:

1. Hoa là học sinh. Hoa có một cây bút. Hoàng là em của Hoa
2. Hoa là học sinh. Cô ấy có một cây bút. Hoàng là em của cô ấy
3. Một người là học sinh. Cô ấy có một cây bút. Hoàng là em của cô ấy





# 1.1 DIỄN NGÔN LÀ GÌ?

1. Hoa là học sinh. Hoa có một cây bút. Hoàng là em của Hoa

Ta có:

- $||\text{Hoa là học sinh}|| = \text{học\_sinh}(\text{hoa})$
- $||\text{Hoa có một cây bút}|| = \exists X (\text{cây\_bút}(X) \wedge \text{có}(\text{hoa}, X))$
- $||\text{Hoàng là em của Hoa}|| = \text{em}(\text{hoàng}, \text{hoa})$

Ngữ nghĩa của đoạn văn bản theo logic vị từ bậc một là:

$$\text{học\_sinh}(\text{hoa}) \wedge \exists X (\text{cây\_bút}(X) \wedge \text{có}(\text{hoa}, X)) \wedge \text{em}(\text{hoàng}, \text{hoa})$$





# 1.1 DIỄN NGÔN LÀ GÌ?

2. Hoa là học sinh. **Cô ấy** có một cây bút. Hoàng là em của **cô ấy**

Cụm từ “**Cô ấy**” thay thế cho “**Hoa**”. Ta có thể thay thế như sau:

Hoa là học sinh. **Hoa** có một cây bút. Hoàng là em của **Hoa**

Ngữ nghĩa của đoạn văn bản theo logic vị từ bậc một:

$\text{học\_sinh}(\text{hoa}) \wedge \exists X (\text{cây\_bút}(X) \wedge \text{có}(\text{hoa}, X)) \wedge \text{em}(\text{hoàng}, \text{hoa})$





# 1.1 DIỄN NGÔN LÀ GÌ?

3. Một người là học sinh. **Cô ấy** có một cây bút. Hoàng là em của **cô ấy**

Cụm từ “**Cô ấy**” thay thế cho “**Một người**”. Nếu ta thực hiện việc thay thế như đoạn văn bản 2):

**Một người** là học sinh. **Một người** có một cây bút. Hoàng là em của **một người**

Ngữ nghĩa của đoạn văn bản theo logic vị từ bậc một:

$$\exists X (\text{người}(X) \wedge \text{học\_sinh}(X)) \wedge \exists Y \exists Z (\text{người}(Y) \wedge \text{cây\_bút}(Z) \wedge \text{có}(Y, Z)) \wedge \exists T (\text{người}(T) \wedge \text{em}(\text{hoàng}, T))$$





# 1. DIỄN NGÔN

---

## 1.2 BIỂU DIỄN DIỄN NGÔN



## 1.2 BIỂU DIỄN DIỄN NGÔN

Các lý thuyết:

- Lý thuyết cấu trúc tu từ (Rhetorical Structure Theory - RST)
- Lý thuyết biểu diễn diễn ngôn (Discourse Representation Theory - DRT)
- Lý thuyết biểu diễn phân đoạn diễn ngôn (Segmented Discourse Representation Theory – SDRT).



## 1.2 BIỂU DIỄN DIỄN NGÔN

Lý thuyết cấu trúc tu từ (Mann và Thompson, 1988):

- Một văn bản được cấu tạo từ nhiều đơn vị tu từ (Rhetoric Unit – RU)
- Các đơn vị tu từ liên kết với nhau theo những quan hệ tu từ đã được định nghĩa.

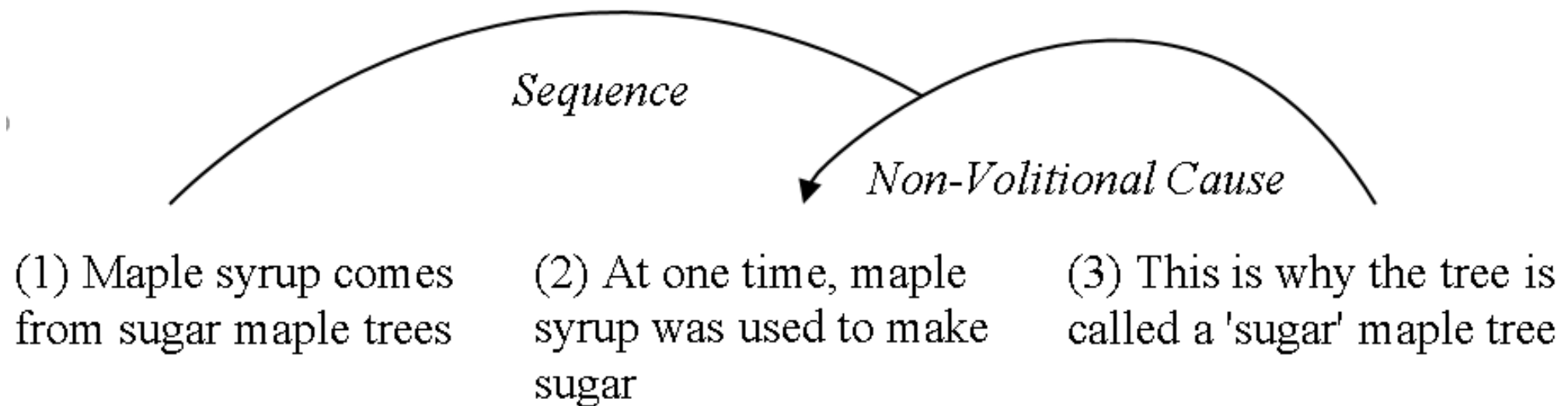
Đơn vị tu từ = đơn vị diễn ngôn (discourse unit)

Quan hệ tu từ = quan hệ diễn ngôn (discourse relation)

## 1.2 BIỂU DIỄN DIỄN NGÔN

Ví dụ: phân tích diễn ngôn theo RST cho đoạn văn bản sau.

*Maple syrup comes from sugar maple trees. At one time, maple syrup was used to make sugar. This is why the tree is called a 'sugar' maple tree. (Delmonte, 2008)*





## 1.2 BIỂU DIỄN DIỄN NGÔN

Lý thuyết biểu diễn diễn ngôn (Kamp, 1981):

- Sử dụng logic vị từ bậc một để biểu diễn ngữ nghĩa.
- Diễn ngôn của văn bản được tính toán từ ngữ nghĩa của từng câu.
- Nghĩa của một câu, được biểu diễn bằng một cấu trúc cơ bản, gọi là Discourse Representation Structure (DRS), có hai phần.

Sở chỉ 1, sở chỉ 2, ...
Mệnh đề 1 Mệnh đề 2 ...





## 1.2 BIỂU DIỄN DIỄN NGÔN

Lý thuyết biểu diễn diễn ngôn (Kamp, 1981):

- Quá trình tính toán diễn ngôn của văn bản sẽ giải quyết bài toán xác định đồng sở chỉ (Coreference Resolution).
- Phù hợp để tính toán ngữ nghĩa cho các đoạn văn miêu tả.

*Lý thuyết biểu diễn diễn ngôn là cơ sở để phân tích diễn ngôn trong môn học CS229 – Ngữ nghĩa học tính toán.*



## 1.2 BIỂU DIỄN DIỄN NGÔN

Segmented Discourse Representation Theory (Asher and Lascarides, 2003) :

- Mỗi mệnh đề, là một đơn vị diễn ngôn cơ bản, được biểu diễn bằng cấu trúc DRS theo lý thuyết biểu diễn diễn ngôn.
- Các câu liên kết với nhau theo các quan hệ tu từ theo lý thuyết cấu trúc tu từ.



# 1. DIỄN NGÔN

---

## 1.3 XÁC ĐỊNH ĐỒNG SỞ CHỈ



## 1.3 XÁC ĐỊNH ĐỒNG SỞ CHỈ

**Xác định đồng sở chỉ** (coreference resolution, hoặc pronoun resolution): là bài toán xác định từ ngữ được thay thế bởi các từ ngữ có chức năng như đại từ trong văn bản.

- Từ ngữ được thay thế được gọi là tiền ngữ (antecedent).
- Từ ngữ có chức năng như đại từ được gọi là hồi chỉ (anaphora).

Ví dụ: Lan có **một con mèo**. Lan rất thích **nó**.

Tiền ngữ: một con mèo. Hồi chỉ: nó.



## 1.3 XÁC ĐỊNH ĐỒNG SỞ CHỈ

Độ phức tạp của bài toán xác định đồng sở chỉ tùy thuộc vào văn cảnh.

Ví dụ:

- Lan có một con mèo. Lan rất thích nó.
- Anh Hải gặp một anh thanh niên. Anh ấy bỏ đi rất nhanh.



## 1.3 XÁC ĐỊNH ĐỒNG SỞ CHỈ

Trường hợp đoạn văn gồm hai câu:

Lan có một con mèo. Lan rất thích nó.

- Các tiền ngữ: **Lan** (chỉ người), **một con mèo** (chỉ vật).
- Các hồi chỉ: **nó**.

Đại từ nó có thể chỉ người và vật nhưng có tính trang trọng thấp nên ưu tiên để chỉ vật hơn chỉ người.

Xác định đồng sở chỉ, được cặp tiền ngữ “**một con mèo**” và hồi chỉ “**nó**”.



## 1.3 XÁC ĐỊNH ĐỒNG SỞ CHỈ

Trường hợp đoạn văn gồm hai câu:

Anh Hải gặp một anh thanh niên. Anh ấy bỏ đi rất nhanh.

- Các tiền ngữ: **anh Hải** (chỉ người), **một anh thanh niên** (chỉ người).
- Các hồi chỉ: **anh ấy** (chỉ người).

Trường hợp này có tính nhập nhằng với hai cặp:

- Tiền ngữ “**anh Hải**” và hồi chỉ “**anh ấy**”.
- Tiền ngữ “**một anh thanh niên**” và hồi chỉ “**anh ấy**”.





## 1.3 XÁC ĐỊNH ĐỒNG SỞ CHỈ

Xác định đồng sở chỉ là một bài toán quan trọng:

- Tăng hiệu quả dịch máy (machine translation). Ví dụ “nó” nên dịch là “he”, “she” hay “it”?
- Tăng hiệu quả trích chọn thông tin (information extraction). Ví dụ: “**Sản phẩm A đang được ưa chuộng. Nó có giá X.**”
- Tăng hiệu quả tóm tắt văn bản (text summarisation)
- Tăng hiệu quả hỏi-đáp tự động (question answering).

*Trong môn học này, Bài toán trên sẽ được giải quyết theo DRT.*





## 2. LÝ THUYẾT BIỂU DIỄN DIỄN NGÔN

### 2.1 CẤU TRÚC BIỂU DIỄN DIỄN NGÔN

## 2.1 CẤU TRÚC BIỂU DIỄN DIỄN NGÔN

**Cấu trúc biểu diễn diễn ngôn (DRS) gồm 2 phần:**

- Phần sở chỉ: chứa các hằng, các biến chỉ các thực thể.
- Phần điều kiện: mỗi dòng là một mệnh đề logic vị từ bậc một. Các mệnh đề trên từng dòng liên kết với nhau bằng phép hội.

Ví dụ: câu “*Bình gói một món quà*” được biểu diễn theo DRS:

bình, T
quà(T) gói(bình, T)

Hoặc theo dạng văn bản:

$\langle \{\text{bình}, T\}, \{\text{quà}(T), \text{gói}(\text{bình}, T)\} \rangle$



## 2.1 CẤU TRÚC BIỂU DIỄN DIỄN NGÔN

**Các quy tắc xác định cấu trúc biểu diễn diễn ngôn:**

- 1) Quy tắc 1: Nếu  $x_1, \dots, x_n$  là những hằng hoặc đại từ hồi chỉ ( $n \geq 0$ ) và  $\theta_1, \dots, \theta_m$  là những điều kiện ( $m \geq 0$ ) thì cấu trúc sau là một DRS:

$x_1, \dots, x_n$
$\theta_1$
.
.
$\theta_m$



## 2.1 CẤU TRÚC BIỂU DIỄN DIỄN NGÔN

**Các quy tắc xác định cấu trúc biểu diễn diễn ngôn:**

2) Nếu  $\tau_1$  và  $\tau_2$  là các biến hoặc hằng (term) thì  $\tau_1 = \tau_2$  là một điều kiện thuộc loại cơ bản.

Ví dụ: “Hòa có một con mèo mà anh ta quên nó” có ngữ nghĩa:

$$\exists T. \text{mèo}(a) \wedge \text{có}(\text{hòa}, T) \wedge \text{quên}(X, Y)$$

hòa, T, X, Y

mèo(a)

có(hòa, a)

thích(X, Y)

**X = hòa**

**Y = T**



## 2.1 CẤU TRÚC BIỂU DIỄN DIỄN NGÔN

**Các quy tắc xác định cấu trúc biểu diễn diễn ngôn:**

3) Nếu  $R$  là một quan hệ bậc  $n$  trong ngôn ngữ bậc một và  $\tau_1, \dots, \tau_n$  là những hằng hoặc biến thì  $R(\tau_1, \dots, \tau_n)$  là một điều kiện thuộc loại cơ bản.

Ví dụ: “**một con mèo**” có nghĩa:

$\exists T. \text{mèo}(T)$

T
mèo(T)



## 2.1 CẤU TRÚC BIỂU DIỄN DIỄN NGÔN

**Các quy tắc xác định cấu trúc biểu diễn diễn ngôn:**

- 4) Nếu  $\Delta_1$  và  $\Delta_2$  là những DRS thì  $\neg\Delta_1$ ,  $\Delta_1 \vee \Delta_2$  và  $\Delta_1 \rightarrow \Delta_2$  là những điều kiện thuộc loại phức.
- 5) Các kết quả tạo DRS khác với các quy tắc trên đều không hợp lệ.



## 2.1 CẤU TRÚC BIỂU DIỄN DIỄN NGÔN

Ví dụ: giả sử tồn tại một Ngôn ngữ bậc một phù hợp, hãy xác định cấu trúc DRS cho đoạn văn bản:

“Nam có một cây bút. Anh ấy cần một quyển vở”



## 2.1 CẤU TRÚC BIỂU DIỄN DIỄN NGÔN

Ngữ nghĩa của câu “Nam có một cây bút”:

$$\exists T. \text{bút}(T) \wedge \text{có}(\text{nam}, T)$$

- Các sở chỉ: **nam**, **T**.
- Các điều kiện: **bút(T)**, **có(nam, T)**.

Ta có DRS tương ứng:

<b>nam</b> , <b>a</b>
<b>bút(a)</b> <b>có(nam, a)</b>



## 2.1 CẤU TRÚC BIỂU DIỄN DIỄN NGÔN


Ngữ nghĩa của câu “Anh ấy cần một quyển vở”:

$$\exists b. \text{vở}(b) \wedge \text{cần}(X, b)$$

- Các sở chỉ:  $X, b$ .
- Có cặp hỏi chỉ:  $X$ , tiền ngữ: Nam
- Các điều kiện:  $\text{vở}(b)$ ,  $\text{cần}(X, b)$ ,  $X=\text{nam}$ .

Thêm các sở chỉ và điều kiện vào DRS trước:

nam, a
bút(a) có(nam, a)

$X$   
 $b$ 

 $\text{vở}(b)$   
 $\text{cần}(X, b)$   
 $X=\text{nam}$

nam, a, X, b
bút(a) có(nam, a) Vở(b) cần(X, b) X=nam



## 2. LÝ THUYẾT BIỂU DIỄN DIỄN NGÔN

---

### 2.2 TÍNH TOÁN DRS

## 2.2 TÍNH TOÁN DRS

### Phép toán kết hợp DRS:

Cho  $A = \langle D_A, C_A \rangle$  và  $B = \langle D_B, C_B \rangle$  là hai DRS. Phép toán kết hợp DRS, là phép toán hai ngôi, ký hiệu  $\otimes$ , xác định trên hai DRS A và B cho kết quả là một DRS  $T = \langle D_T, C_T \rangle$  sao cho:

$$T = A \otimes B \Leftrightarrow \begin{cases} D_T = D_A \cup D_B \\ C_T = C_A \cup C_B \end{cases}$$

## 2.2 TÍNH TOÁN DRS

Ví dụ: thực hiện phép toán kết hợp của hai DRS A và B bên dưới.

A =

nam, T
bút(T) có(nam, T)

B =

toàn, Q
vở(Q) có(toàn, Q)

Ta có:

nam, T	$\otimes$	toàn, Q	=	nam, a, toàn, b
bút(T) có(nam, T)		vở(Q) có(toàn, Q)		bút(a) có(nam, a) vở(b) có(toàn, b)

## 2.2 TÍNH TOÁN DRS

**Biểu thức kết hợp DRS:** là biểu thức có toán tử và toán hạng, lần lượt là phép toán kết hợp và các DRS, và cặp dấu “(” và “)” thể hiện cấp độ ưu tiên. Biểu thức kết hợp được tính toán theo thứ tự từ trái sang phải.

Ví dụ: Tính toán biểu thức kết hợp sau:

$$A = \langle \{X\}, \{\} \rangle \otimes \langle \{\}, \{\text{bút}(X)\} \rangle \otimes \langle \{\}, \{\text{mua}(V, X)\} \rangle$$

- $B = \langle \{X\}, \{\} \rangle \otimes \langle \{\}, \{\text{bút}(X)\} \rangle = \langle \{X\}, \{\text{bút}(X)\} \rangle$
- $A = B \otimes \langle \{\}, \{\text{mua}(V, X)\} \rangle = \langle \{X\}, \{\text{bút}(X), \text{mua}(V, X)\} \rangle$



## 2.2 TÍNH TOÁN DRS

**Xác định sở chỉ (reference resolution)** theo nguyên tắc: DRS  $\Delta_1$  chứa sở chỉ có thể xác định từ DRS  $\Delta_2$  khi và chỉ khi  $\Delta_1$  và  $\Delta_2$  trùng nhau hoặc  $\Delta_1$  chứa  $\Delta_2$ .

$\Delta_1$  chứa  $\Delta_2$  khi và chỉ khi:

- $\Delta_1$  trực tiếp chứa  $\Delta_2$ ; hoặc,
- Tồn tại DRS  $\Delta$  sao cho  $\Delta_1$  trực tiếp chứa  $\Delta$  và  $\Delta$  chứa  $\Delta_2$ .

Mỗi học sinh phải cố gắng học tập. Em ấy là tấm gương tốt cho các bạn





## 2.2 TÍNH TOÁN DRS

**Xác định sở chỉ (reference resolution): (tt)**

Cho  $\Delta_3$  là một DRS. DRS  $\Delta_1$  trực tiếp chứa DRS  $\Delta_2$  nếu và chỉ nếu một trong các điều kiện sau được thỏa:

- 1)  $\Delta_1$  chứa một điều kiện có dạng là  $\Delta_2$ .
- 2)  $\Delta_1$  chứa một điều kiện có dạng  $\Delta_2 \vee \Delta_3$  hoặc  $\Delta_3 \vee \Delta_2$ .
- 3)  $\Delta_1$  chứa một điều kiện có dạng  $\Delta_2 \Rightarrow \Delta_3$ .
- 4)  $\Delta_1 \Rightarrow \Delta_2$  là một điều kiện của  $\Delta_3$ .

## 2.2 TÍNH TOÁN DRS

### Xác định đồng sở chỉ:

Cho  $Y$  là hồi chỉ, DRS  $\Delta_2 = \langle \{Y, \dots\}, \{\dots\} \rangle$ ,  $a$  là một sở chỉ xuất hiện trong tất cả DRS có thể xác định từ  $\Delta_2$ . Khi đó, có thể thêm điều kiện  $Y=a$  vào tập điều kiện của  $\Delta_2$ .

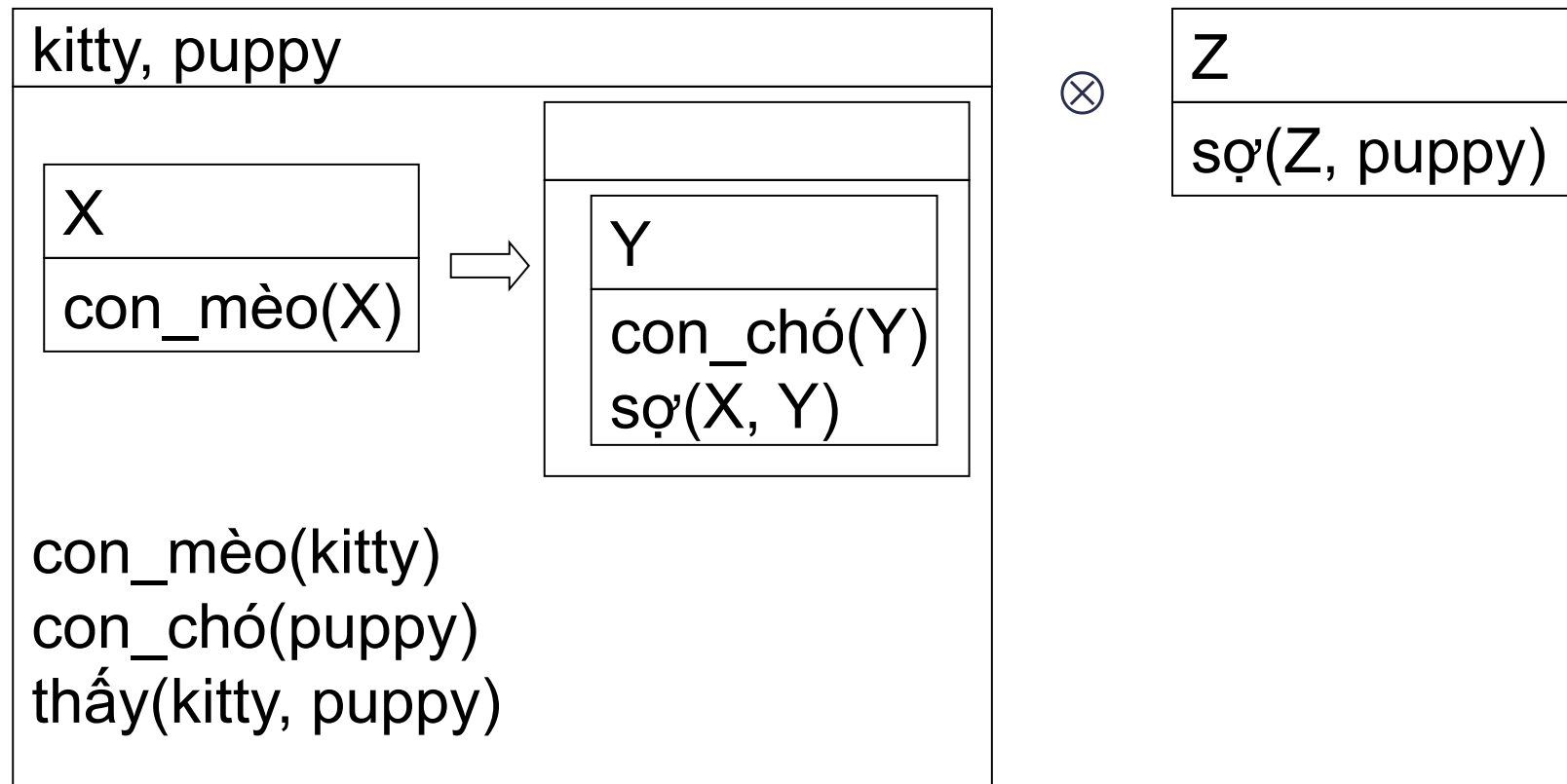
Trường hợp có nhiều sở chỉ, cần xét một số khía cạnh:

- Chỉ người hay chỉ vật.
- Giới tính nam (đực) hay nữ (cái).
- Số nhiều hay số ít.
- Vai trò chủ từ hay tân từ.



## 2.2 TÍNH TOÁN DRS

Ví dụ: tính biểu thức kết hợp sau.

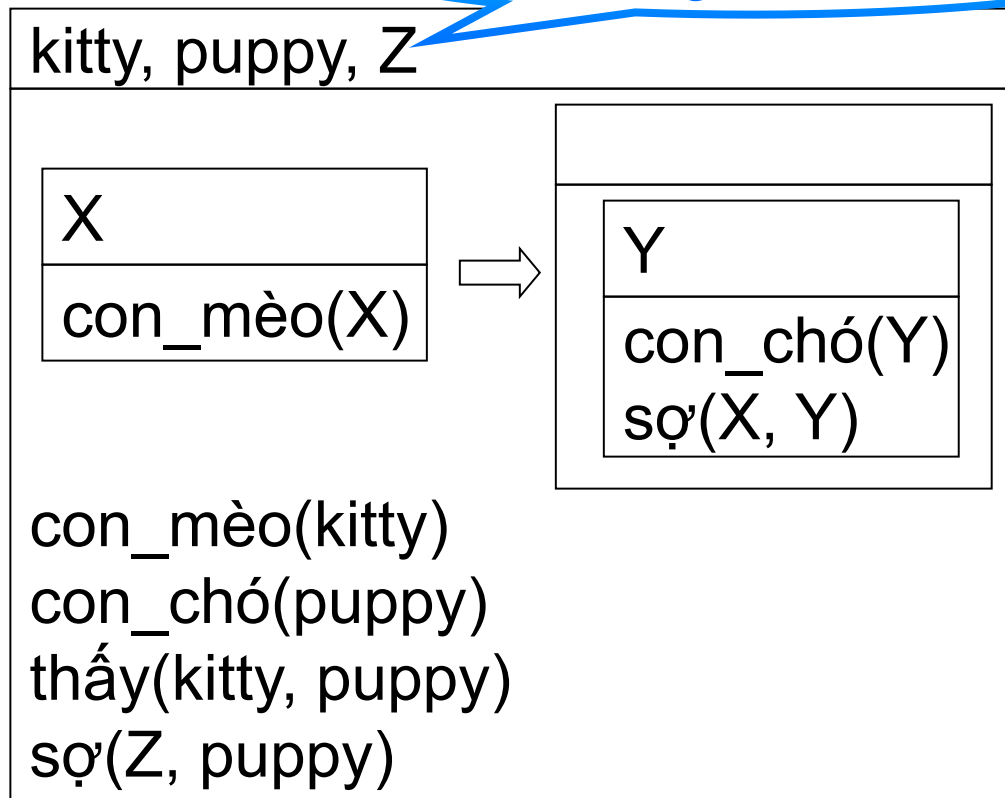




## 2.2 TÍNH TOÁN DRS

Kết quả

Đồng sở chỉ?



Z = kitty?

Z = puppy?

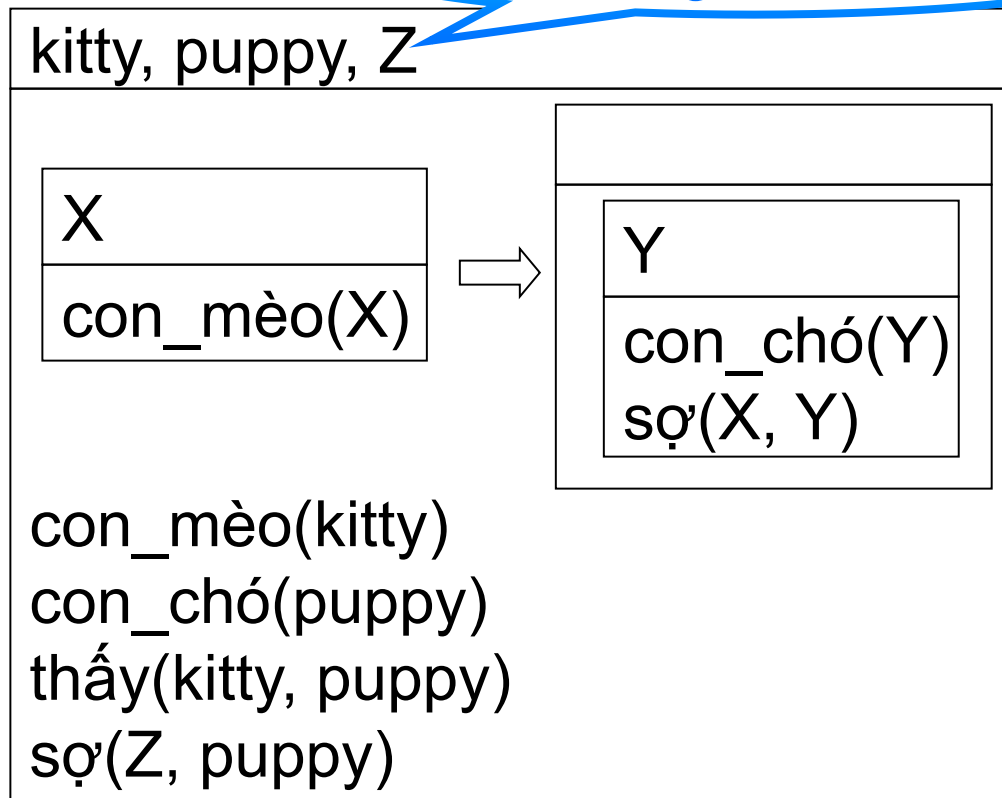
Z = X?

Z = Y?

## 2.2 TÍNH TOÁN DRS

Kết quả

Đồng sở chỉ?



Z = kitty ✓

Z = puppy ✗

Z = X ✗

Z = Y ✗

## 2.2 TÍNH TOÁN DRS

Trường hợp cần biểu diễn thêm một số khía cạnh của hằng và biến, chúng ta có thể sử dụng một cấu trúc.

Ví dụ: biểu diễn các hằng **kitty** (con mèo), **hoa** (một bạn nữ).

Quy ước (*có thể đặt tùy ý*):

- Thành phần thứ nhất biểu diễn tên,
- Thành phần thứ hai chỉ người/vật,
- Thành phần thứ ba biểu diễn giới tính.

Ta có: **kitty**: (**kitty**, vật, \_); **hoa**: (**hoa**, người, nữ).