NHẬN DẠNG THỰC THỂ CÓ TÊN

Với dữ liệu tiếng Việt

1. Giới thiệu

Nhận dạng thực thể có tên (*Named Entity Recognition* – NER) nhằm nhận biết các chuỗi từ trong văn bản là tên của một đối tượng nào đó, điển hình như tên người, tên tổ chức, tên địa danh, thời gian v.v. NER là nhiệm vụ đóng vai trò quan trọng trong các ứng dụng trích xuất thông tin, đã được quan tâm nghiên cứu trên thế giới từ đầu những năm 1990.

Từ năm 1995, hội thảo quốc tế chuyên đề Hiểu thông điệp (*Message Understanding Conference* - MUC) lần thứ 6 đã bắt đầu tổ chức đánh giá các hệ thống NER cho tiếng Anh. Tại hội thảo CoNLL năm 2002 và 2003, các hệ thống NER cho tiếng Hà Lan, Tây Ban Nha, Đức và Anh cũng được đánh giá. Trong các tác vụ đánh giá này, người ta xét 4 loại thực thể có tên: tên người, tên tổ chức, tên địa danh và các tên khác. Gần đây, vẫn tiếp tục có các cuộc thi về NER được tổ chức, ví dụ GermEval 2014 cho tiếng Đức.

Đối với tiếng Việt, đây là cuộc thi thứ hai nhằm đưa ra được một đánh giá khách quan về chất lượng các công cụ NER, khuyến khích phát triển các hệ thống trích rút thực thể có tên đạt độ chính xác cao. So với cuộc thi thứ nhất tại VLSP 2016, tập dữ liệu lần này đa dạng, phong phú hơn và được tập hợp theo một số lĩnh vực nhằm có các đánh giá chi tiết hơn về các hệ thống NER.

2. Mô tả nhiệm vụ

Phạm vi của cuộc thi đầu tiên này là đánh giá khả năng nhận dạng các thực thể có tên thuộc một trong ba loại: tên người, tên tổ chức và tên địa danh. Việc nhận dạng các loại thực thể có tên khác sẽ được đề cập đến trong các lần thi sau.

3. Dữ liệu

Dữ liệu là các bài báo, đăng trên các phương tiện truyền thông xã hội, không phải dữ liệu nhân tạo (do người làm dữ liệu sinh ra).

Trong đó, ba loại thực thể có tên được xác định tương thích với các loại thực thể mô tả trong CoNLL2003.

1. Tên địa lí (Địa danh - Location) bao gồm các thực thể có toạ độ địa lí nhất định, ghi lại được trên bản đồ:
   - Tên gọi các hành tinh: Mặt Trăng, Mặt Trời, Trái Đất…
   - Tên gọi các thực thể mang yếu tố địa lí tự nhiên và địa lí lịch sử (quốc gia, vùng lãnh thổ, châu lục), các vùng quần cư (làng, thị trấn, thành phố, tỉnh, giáo khu, giáo xứ), các điểm kinh tế (vùng nông nghiệp, khu công nghiệp)
   - Tên gọi các thực thể tự nhiên (đèo, núi, dãy núi, rừng, sông, suối, hồ, biển, vịnh, vũng, eo biển, đại dương, thung lũng, cao nguyên, đồng bằng, khu bảo tồn thiên nhiên, bãi biển, khu sinh thái, v.v.)
   - Tên gọi các thực thể là công trình xây dựng, công trình kiến trúc công cộng (cầu, đường, cảng, đập, lâu đài, tháp, quảng trường, bảo tàng, phòng trưng bày, hội trường, trường học, nhà trẻ, thư viện, bệnh viện, viện dưỡng lão, trung tâm y tế, nhà thờ, nhà xứ, tu viện, nhà ở, chung cư, kí túc xá, chợ, công viên, nhà hát, rạp chiếu phim, khu thể thao, bể bơi, trung tâm thanh

thiếu niên, khu cắm trại, doanh trại quân đội, nhà máy, sân bay, nhà ga, nhà kho, bãi đỗ xe, sân chơi, nghĩa trang, ...)
- Tên gọi địa điểm, địa chỉ thương mại (hiệu thuốc, quán rượu, nhà hàng, khách sạn, câu lạc bộ đêm, các địa điểm tổ chức âm nhạc, ...)
- Một số địa danh trừu tượng khác (Vườn Địa Đàng, Sông Ngân, Cầu Ô Thước…).

2. Tên tổ chức (Organization) bao gồm các loại tên sau:
   - Các cơ quan chính phủ (các bộ ngành, uỷ ban nhân dân, hội đồng nhân dân, toà án, cơ quan báo chí, hội nghề nghiệp, đoàn thể chính trị, phòng ban, ...)
   - Công ti (ngân hàng, thị trường chứng khoán, hãng phim, nhà sản xuất, hợp tác xã, phòng ban, ...)
   - Các thương hiệu
   - Các tổ chức chính trị (các đảng phái chính trị, các tổ chức khủng bố, ...)
   - Các ấn phẩm (các tạp chí, báo)
   - Các công ti âm nhạc (ban nhạc, dàn nhạc, đội hợp xướng ...)
   - Các tổ chức công cộng (trường học, tổ chức từ thiện)
   - Các tổ chức khác của con người (câu lạc bộ thể thao, các hiệp hội, nhà hát, công ti, tôn giáo, tổ chức thanh niên...)

3. Tên người (Person) bao hàm các loại tên sau:
   - Tên, tên đệm và họ của một người
   - Tên động vật và các nhân vật hư cấu
   - Các bí danh

Ví dụ về dữ liệu:

- Tên địa lí: Thành phố Hồ Chí Minh, Núi Bà Đen, Sông Bạch Đằng…
- Tên tổ chức: Công ty Formosa, Nhà máy thủy điện Hòa Bình…
- Tên người: tên riêng trong "ông Lân", "bà Hà"…

Một thực thể có thể chứa thực thể khác nhúng trong đó. Ví dụ "Uỷ ban nhân dân Thành phố Hà Nội" là tên tổ chức, trong đó có chứa tên địa danh "thành phố Hà Nội".

Dữ liệu huấn luyện gồm hai phần. Một phần gồm dữ liệu đã tách từ và có thể bổ sung thêm thông tin nhãn từ loại và nhãn phân cụm bằng các phần mềm có sẵn. Một phần là văn bản thô chỉ bổ sung thêm các thẻ đánh dấu các thực thể.

4. Định dạng của dữ liệu

4.1 Đối với dữ liệu huấn luyện đã tách từ và có thể bổ sung thông tin từ loại là nhãn phân cụm:

File dữ liệu huấn luyện chứa một văn bản đã tách từ và gán nhãn. Mỗi từ được đặt trên một dòng riêng biệt và mỗi câu được phân cách nhau bởi một dòng trống. Mỗi dòng bao gồm năm cột, các cột được cách nhau bởi một khoảng trắng:

1. Cột đầu tiên là một từ

2. Cột thứ hai là từ loại của từ
3. Cột thứ 3 là nhãn phân cụm cú pháp
4. Cột thứ 4 là nhãn thực thể
5. Cột thứ 5 là nhãn thực thể lồng

Nhãn thực thể được gán theo cấu trúc BIO như định dạng dữ liệu phân cụm CoNLL. Có 11 nhãn: B-PER và I-PER cho tên người, B-ORG và I-ORG cho tên tổ chức, B-LOC và I-LOC cho tên địa danh và O cho các phần tử khác.

Ví dụ đối với tiếng Việt:

Word	**POS**	**Phrase**	**Nhãn thực thể**	**Nhãn thực thể lồng**
Anh	N	B-NP	O	O
Thanh	NPP	B-NP	B-PER	O
là	V	B-VP	O	O
cán_bộ	N	B-NP	O	O
Uỷ_ban	N	B-NP	B-ORG	O
nhân_dân	N	I-NP	I-ORG	O
Thành_phố	N	I-NP	I-ORG	B-LOC
Hà_Nội	NPP	I-NP	I-ORG	I-LOC
.	CH	O	O	O

Trong đó, tập {N, NPP, V, E, .} là các nhãn từ loại, tập {B-NP, I-NP, B-VP, O} là các nhãn phân cụm cú pháp.

Lưu ý:

- Các thông tin về nhãn từ loại và nhãn phân cụm được xác định tự động bằng các phần mềm có sẵn nên có thể chứa lỗi.
- Đối với các nhãn thực thể, sử dụng hai loại nhãn chính là B-XXX và I-XXX. Trong đó, B-XXX dùng cho từ đầu tiên của thực thể XXX, I-XXX dùng cho các từ tiếp theo trong cụm thực thể XXX đó. Nhãn O dùng cho từ không thuộc bất cứ thực thể nào.

4.2. Đối với dữ liệu không gán nhãn từ

Ví dụ:

"Anh Thanh là cán bộ Uỷ ban nhân dân Thành phố Hà Nội."

Anh <ENAMEX TYPE="PERSON"> Thanh </ENAMEX> là cán bộ <ENAMEX TYPE="ORGANIZATION"> Uỷ ban nhân dân <ENAMEX TYPE="LOCATION"> Thành phố Hà Nội </ENAMEX> </ENAMEX> .

5. Đánh giá

Các hệ thống nhận dạng thực thể có tên sẽ được đánh giá thông qua độ đo F1 và độ chính xác (accuracy).

- ❖ Độ đo F:

$$F_1 = \frac{2 * P * R}{P + R}$$

Trong đó P là độ chính xác (precision), R là độ bao phủ (recall) được tính theo công thức sau.

$$P = \frac{NE-true}{NE-sys}$$

$$R = \frac{NE-true}{NE-ref}$$

Trong đó:

- ★ NE-ref: Là số thực thể trong dữ liệu gốc
- ★ NE-sys: Là số thực thể được đưa ra bởi hệ thống
- ★ NE-true: Là số thực thể được hệ thống gán nhãn đúng

- ❖ Accuracy

$$A = \frac{S[?][?][?]\,t[?][?][?]\,g[?][?]n\,nh[?][?]n\,[?][?][?][?]ng}{T[?][?][?]ng\,s[?][?][?]\,t[?][?][?]}$$

Accuracy chỉ áp dụng với dữ liệu đầu ra được xuất theo định dạng có gán nhãn từ vựng. Kết quả hệ thống sẽ được đánh giá trên cả 3 mức gán nhãn thực thể.

6. Tham khảo

http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html
http://www.clips.uantwerpen.be/conll2002/ner/
http://www.cnts.ua.ac.be/conll2003/ner/
http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt
https://sites.google.com/site/germeval2014ner/

http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book 7.html#HEADING18

Ghi chú thêm:

- Số lượng từ trong văn bản mong muốn: có gán nhãn từ loại: 250K từ
- Ko gán nhãn cho tên xuất hiện trong các cụm từ ko xác định một cá thể cụ thể, kiểu như trong "Máy tính Asus khá phổ biến".
  - http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book 7.html#HEADING18 A.1.7

HƯỚNG DẪN CỤ THỂ

(Vũ Xuân Lương – Vietlex)

Kí hiệu ▷: ngữ liệu trích nguyên văn.

Trong phần hướng dẫn này, để dễ hình dung, chúng tôi không trình bày cột nhãn từ loại và cột nhãn phân cụm cú pháp.

A. Tên người

○ Tên người (nhân danh) được xem là *tên riêng*. Viết hoa tên riêng là để chỉ ra rằng người đó chỉ có một mà thôi, không giống với người khác.

○ Thường đi trước tên người có các danh từ chung như *ông, bà, anh, chị, chú, bác, thằng, chủ tịch, giám đốc, trưởng phòng, v.v.* Các danh từ chung này được dùng để chỉ hoặc gọi người nào đó tuỳ theo mối quan hệ. Chẳng hạn, cùng một người tên *Thanh*, nhưng có người gọi *anh Thanh*, có người gọi *ông Thanh*, *giám đốc Thanh*, v.v. Chúng tôi cho rằng, các danh từ loại này không nằm trong cấu tạo tên người, vì chúng không có tính cố định.

1. Tên người là tên riêng chỉ từng cá nhân. Dạng đầy đủ, tên người gồm 3 thành phần: *họ* + *chữ đệm* + *tên*. Không phân biệt *họ, chữ đệm* và *tên* vì coi chúng đều được riêng hoá, vì thế viết hoa chữ cái đầu của các âm tiết.

Họ	**Chữ đệm**	**Tên**
Trần	Thánh	Tông
Cao	Bá	Quát
Hồ	Chí	Minh
Nguyễn	Thị Minh	Khai

▷ "thủ tướng Nguyễn Xuân Phúc"

a.

thủ tướng	O	O
Nguyễn	B-PER	O
Xuân	I-PER	O
Phúc	I-PER	O

b.

thủ tướng <PER>Nguyễn Xuân Phúc</PER>

2. Dạng rút gọn còn 2 thành phần: *họ* + *tên* hoặc *chữ đệm* + *tên*

Họ - Chữ đệm	**Tên**

Nguyễn	Trãi
Lê	Lợi
Thị	Nở
Chí	Phèo
Quang	Thọ

▹ "nhân vật Chí Phèo"

a.

nhân_vật	O	O
Chí	B-PER	O
Phèo	I-PER	O

b.

nhân vật <PER>Chí Phèo</PER>

3. Dạng rút còn 1 thành phần: *tên gọi*. Trường hợp có các ***danh từ chung là từ xưng hô*** đứng trước bộ phận ***tên*** thì các danh từ này *không được coi là thuộc tên người*. Trường hợp *danh từ chung chỉ chức vụ, công việc* được dùng để gọi thay cho tên người đảm nghiệm chức vụ, công việc đó trong một không gian cụ thể (bối cảnh của câu chuyện) thì *cũng được coi là tên người* (có thể viết hoa theo phong cách: ông Hàn, ông Lí…).

Từ xưng hô	**Tên**
anh	Thanh
ông	An
ông	Hàn
cụ	Bá
ông	Lí
bà	Tham
ông	Tuần

▹ "anh Thanh là cán bộ"

a.

anh	O	O
Thanh	B-PER	O
là	O	O
cán_bộ	O	O

b.

anh <PER>Thanh</PER> là cán bộ

▹ “chuyện đời cụ Bá”

a.

chuyện	O	O
đời	O	O
cụ	O	O
Bá	B-PER	O

b.

chuyện đời cụ <PER>Bá</PER>

▹ “chính phủ Obama”

a.

chính_phủ	O	O
Obama	B-PER	O

b.

chính phủ <PER>Obama</PER>

▹ “the Clinton government”

a.

the	O	O
Clinton	B-PER	O
government	O	O

b.

the <PER>Clinton</PER> government

4. Tên danh nhân, nhân vật lịch sử được cấu tạo bằng cách kết hợp giữa bộ phận là ***danh từ chung chỉ chức vụ, công việc hoặc từ tôn sùng*** - Y) với bộ phận là ***tên*** (CapWord - từ đỉnh**)** thì *được coi là tên người* (X). X = Y + CapWord.

Từ chức vụ, tôn sùng	**Tên**
Đề	Thám
Đội	Cấn
Nghị	Hách

Lí	Cường
Kí	Đạt
Tuần	Vi
Lang	Vinh
Trùm	Sủng
Ông	Gióng
Ông	Đùng
Bà	Đà
Bà	Trưng
Bà	Triệu
Thánh	Gióng
Đức Phật	Như Lai

▹ “con cháu Bà Trưng, Bà Triệu”

a.

con_cháu	O	O
Bà	B-PER	O
Trưng	I-PER	O
,	O	O
Bà	B-PER	O
Triệu	I-PER	O

b.

con cháu <PER>Bà Trưng</PER>, <PER>Bà Triệu</PER>

5. Tên hiệu, tên tự, bí danh, biệt danh cũng *được coi là tên người*.

▹ “Ức Trai là tên hiệu của Nguyễn Trãi”

a.

Ức	B-PER	O
Trai	I-PER	O
là	O	O
tên_hiệu	O	O
của	O	O
Nguyễn	B-PER	O
Trãi	I-PER	O

b.

<PER>Ức Trai</PER> là tên hiệu của <PER>Nguyễn Trãi</PER>

▷ “Hoàng Hoa Thám còn gọi là “Hùm thiêng Yên Thế”” [dấu ngoặc kép cũng được xét nằm trong thành phần tạo tên riêng]

a.

Hoàng	B-PER	O
Hoa	I-PER	O
Thám	I-PER	O
còn	O	O
gọi	O	O
là	O	O
“	B-PER	O
Hùm	I-PER	O
thiêng	I-PER	O
Yên_Thế	I-PER	B-LOC
”	I-PER	O

b.

<PER>Hoàng Hoa Thám</PER> còn gọi là “<PER>Hùm thiêng Yên Thế</PER>”

▷ “*trong vai **Bát A-ca Dận Tự** si tình*”

```
Bát     B-PER
A-ca    I-PER
Dật     I-PER
Tự      I-PER
```

[markup cả tên hiệu Bát A Ca]

B. Tên địa lí (Địa danh – Location)

1. Tên các quốc gia thuộc lĩnh vực Địa chính trị (Geo-Political), mang tính tri nhận phổ quát. Viết hoa tất cả chữ cái đầu của các yếu tố cấu tạo nên tên địa lí, trừ chữ “và” (and) có trong cấu trúc.

Nga, Anh, Pháp, Đức, Việt Nam, Thái Lan, Nhật Bản, Thuỵ Điển, Đan Mạch, Brazil, Canada, Chile, Colombia, Syria, Slovakia, Singapore, Bosnia and Herzegovina, v.v.

▷ “*nhà nước Việt Nam*”

```
nhà_nước    O
Việt_Nam    B-LOC
```

▷ "*nước Nga*"

```
nước        O
Nga         B-LOC
```

▷ "*công dân Bosnia và Herzegovina*" (Bosnia and Herzegovina)

```
công_dân    O
Bosnia      B-LOC
và          I-LOC
Herzegovina I-LOC
```

▷ "*the new England captain*"

```
the         O
new         O
England     B-LOC
captain     O
```

2. Tên địa phương được phân chia theo cấp khu vực địa lí của một nước (như *xã, phường, huyện, quận, hạt, tỉnh, thành phố, bang*...), cũng là các đơn vị địa danh hành chính của một nước, thuộc tri thức nền mang tính cộng đồng cao.

New York, Paris, Canbera, Jakacta, Bangkok, Hà Nội, Hải Phòng, Hà Nam, Hoà Bình, Thanh Hoá, Quảng Nam, Kon Tum, Đắc Lắc, Lâm Đồng, Đồng Nai, Tiền Giang, Cà Mau, Cầu Dền, Cầu Giấy, Ba Vì, Gia Lâm, Củ Chi, Nhà Bè, Kiến An, Hải Châu, An Khê, Bát Xát, Mèo Vạc, Tân Trào, v.v.

CHÚ Ý:

- Với các kiểu cấu tạo *Tỉnh Nam Định, Thành phố Nam Định, Thành phố Hà Nội, Thủ đô Hà Nội, Thành phố Hồ Chí Minh, Thành phố New York (New York City), v.v.* là bao hàm ý phân biệt vị thế, cấp độ của một địa danh. So sánh:

> **Tỉnh Nam Định**: tỉnh gồm Thành phố Nam Định và 9 huyện.
>
> **Thành phố Nam Định**: trung tâm của Tỉnh Nam Định, không bao gồm 9 huyện.

- Với các cấu tạo *Thành phố Hà Nội, Thủ đô Hà Nội, Hà Nội* thì cả 3 kiểu đều chỉ chung một thực thể, do vậy chúng đồng nhất với nhau về tri nhận (*Hà Nội* là cách gọi rút gọn của *Thành phố Hà Nội* hoặc *Thủ đô Hà Nội*). Trong khi, *Thành phố Hồ Chí Minh* và *Hồ Chí Minh* lại chỉ hai thực thể khác nhau: một chỉ *địa danh*, một chỉ *nhân danh*. Từ những lí do đó, chúng tôi cho rằng, yếu tố "thành phố, thủ đô, tỉnh, thị xã, thị trấn" là thành phần tham gia vào cấu tạo nên địa danh, và vì vậy coi cả khối *Thành phố Nam Định, Thành phố Hồ Chí Minh* là một đơn vị để phân biệt với *Tỉnh Nam Định* (bao hàm *Thành phố Nam Định*) và *Hồ Chí Minh* (chỉ người).

- Xử lí tương tự với các trường hợp: *Tỉnh Đồng Nai, Quận Cầu Giấy, Quận 3, Quận Hai Bà Trưng* (phân biệt với *Phố Hai Bà Trưng*), *Phường Minh Khai* (phân biệt với *Phố Minh Khai*), *Huyện Cầu Kè, Huyện Sông Cầu, Thị trấn Chợ Đồn, Thị xã Sông Công, v.v.*

> Tham khảo: **THÔNG TƯ Hướng dẫn thể thức và kỹ thuật trình bày văn bản hành chính** (Hà Nội, ngày 19 tháng 01 năm 2011)

"Trường hợp địa danh ghi trên văn bản của cơ quan thành phố thuộc tỉnh mà tên thành phố trùng với tên tỉnh thì ghi thêm hai chữ thành phố (TP.), ví dụ:

Văn bản của Ủy ban nhân dân thành phố Hà Tĩnh (tỉnh Hà Tĩnh) và của các phòng, ban thuộc thành phố: *TP. Hà Tĩnh,*

- Địa danh ghi trên văn bản của các cơ quan, tổ chức cấp huyện là tên của huyện, quận, thị xã, thành phố thuộc tỉnh, ví dụ:

Văn bản của Ủy ban nhân dân huyện Sóc Sơn (thành phố Hà Nội) và của các phòng, ban thuộc huyện: *Sóc Sơn,*

Văn bản của Ủy ban nhân dân quận Gò Vấp (thành phố Hồ Chí Minh), của các phòng, ban thuộc quận: *Gò Vấp,*

Văn bản của Ủy ban nhân dân thị xã Bà Rịa (tỉnh Bà Rịa-Vũng Tàu) và của các phòng, ban thuộc thị xã: *Bà Rịa,*

- Địa danh ghi trên văn bản của Hội đồng nhân dân, Ủy ban nhân dân và của các tổ chức cấp xã là tên của xã, phường, thị trấn đó, ví dụ:

Văn bản của Ủy ban nhân dân xã Kim Liên (huyện Nam Đàn, tỉnh Nghệ An): *Kim Liên,*

Văn bản của Ủy ban nhân dân phường Điện Biên Phủ (quận Ba Đình, TP. Hà Nội): *Phường Điện Biên Phủ,* "

▷ "*chủ tịch Thành phố Hà Nội*"

chủ_tịch	O
Thành_phố	B-LOC
Hà_Nội	I-LOC

▷ "*Thị xã Sông Công*"

Thị_xã	B-LOC
Sông_Công	I-LOC

[tổ hợp "Sông Công" đã trở thành một địa danh hành chính]

▷ "*phía tây Hà Nội*"

phía	O
tây	O
Hà_Nội	B-LOC

▷ "*in Panama City*"

in	O
Panama	B-LOC
City	I-LOC

▷ "*message from New York City*"

message	O	O
from	O	O
New	B-LOC	O
York	I-LOC	O

```
City            I-LOC   O
```

3. Tên các đơn vị dân cư dưới cấp xã, phường (X) được cấu tạo giữa một danh từ chung chỉ loại (Y: *thôn, xóm, bản, mường, khóm, đội, tổ, khối, ngõ, ngách, hẻm*) luôn luôn xuất hiện với một tên gọi chính để khu biệt (CapWord). Chúng tôi đề nghị xem X là tên công trình, khi X = Y + CapWord. Với tiếng nước ngoài thì thường X = CapWord + Y.

▹ "*thường trú: tổ 3, phường Tân Thịnh*"

```
Thường_trú  O
:           O
Tổ          B-LOC
3           I-LOC
,           O
phường      B-LOC
Tân_thịnh   I-LOC
```

[tổ: *tổ dân phố* nói tắt]

▹ "*ngụ Thôn 6, xã Ia Blang*"

```
ngụ         O
Thôn        B-LOC
6           I-LOC
,           O
xã          B-LOC
Ia_Blang    I-LOC
```

▹ "*đoạn qua khối Tân Sơn*"

```
đoạn        O
qua         O
khối        B-LOC
Tân_Sơn     I-LOC
```

4. Tên gọi chỉ thực thể địa lí tự nhiên (X) được cấu tạo giữa một danh từ chung chỉ loại của thực thể (Y: *núi, rừng, sông, suối, hồ, biển, vịnh, vũng, châu, đại dương, đại lục, đồng bằng, cao nguyên, thiên thể, v.v.*) luôn luôn xuất hiện với một tên gọi chính chỉ thực thể (CapWord) do nhu cầu phân biệt trong giao tiếp chi phối. Trong đó, CapWord thường là tên có lí do (theo một ý nghĩa, sự tích nào đó: Than Thở, Hoàn Kiếm, Kẻ Gỗ…; hoặc CapWord vốn là một tên gọi cụ thể khác).

Để đảm bảo tính khu biệt, tránh nhầm lẫn với các tên gọi có tính phổ biến khác, chúng tôi đề nghị xem X là tên địa lí, khi X = Y + CapWord. Với các trường hợp cấu tạo theo trật tự Hán-Việt thì coi cả khối là CapWord (Âu Châu, Á Châu, Thái Bình Dương, Ấn Độ Dương, Mộc Tinh, Thổ Tinh…). Với tiếng nước ngoài thì X = CapWord + Y.

Hồ Tây, Hồ Gươm, Hồ Hoàn Kiếm, Hồ Ba Bể, Hồ Than Thở, Hồ Kẻ Gỗ, Sông Hồng, Sông Cầu, Sông Thái Bình, Sông Amazon (Amazon River), Sông Mississippi (Mississippi River), Sông Thêm, Sherwood Forest (một khu rừng ở Anh)*, Đảo Cô Tô, Quần đảo Hoàng Sa, Vịnh Hạ Long, Vịnh Cam Ranh, Núi Nùng, Núi Đọ, Núi Ba Vì, Đồng bằng Sông Hồng, Cao nguyên Lâm Viên, Châu*

Âu, Châu Phi, Thái Bình Dương, Lục địa Á-Âu, Mặt Trời, Mặt Trăng, Trái Đất, Sao Hoả (Hoả Tinh), Sao Mộc (Mộc Tinh), Sao Thổ (Thổ Tinh), Sao Thiên Vương (Thiên Vương Tinh), v.v. [Trao đổi: có thể coi cả khối là CapWord cho nó dễ xử lí trong hệ thống???]

```
Sông            B-LOC
Công            I-LOC

Hồ              B-LOC
Gươm            I-LOC

Hồ              B-LOC
Hoàn_Kiếm       I-LOC

Quần_đảo        B-LOC
Hoàng_Sa        I-LOC

Thái_Bình_Dương B-LOC

Mộc_Tinh        B-LOC

Tiên_Vương _Tinh   B-LOC

Mississippi     B-LOC
River           I-LOC
```

- Trường hợp Châu Âu và Âu Châu hiện chưa xử lí thống nhất:

```
Châu            B-LOC
Âu              I-LOC

Âu_Châu         B-LOC
```

Theo chúng tôi, nên xử lí giống như Nam_Á, Đông_Á ở mục 4. Bởi vì "Âu, Á, Phi, Mỹ" không hoạt động độc lập, chúng luôn đi kèm với "Châu". Chúng chỉ xuất hiện không có "Châu" khi ở trong tổ hợp: *Hội nghị Á-Phi, nhạc Âu-Mỹ, Lục địa Á-Âu, ngoại ngữ Âu-Úc-Mỹ, v.v.,*

- Trường hợp dịch nghĩa như *Biển Đen, Biển Đỏ, Biển Chết* thì coi cả khối là CapWord.

```
Biển_Đen        B-LOC
Biển_Đỏ         B-LOC
Biển_Chết       B-LOC
```

5. Tên địa lí (X) được cấu tạo giữa một danh từ chỉ phương hướng (Y: *đông, tây, bắc, nam, phương*) với một tên gọi chính chỉ một vùng, một miền (CapWord: *Á, Phi, Mĩ, Dương, Bộ* (chỉ đất liền), *v.v.*). Với loại tên địa lí (X) được cấu tạo giữa hai danh từ chỉ phương hướng (Y: *đông nam, đông bắc, v.v.*) với một tên gọi chính chỉ một vùng, một miền (CapWord: *Á, Phi, Mĩ, Dương, v.v.*) thì có thể diễn ra 2 khả năng kết hợp: *1) Đông_Nam + Á* (khu vực phía Đông Nam

của Châu Á)*; 2) Đông + Nam_Á* (phía Đông của vùng Nám Á). Để tránh nhập nhằng về ý nghĩa, chúng tôi đề nghị xem cả khối là CapWord. Với tiếng nước ngoài thì X = CapWord + Y. Chẳng hạn: *Đông Á, Tây Á, Đông Âu, Tây Âu, Trung Mỹ, Trung Phi, Đông Dương, Đông Nam Á, Đông Bắc Á, Bắc Bán Cầu, Nam Bán Cầu, West Texas, Northern California, Northern Ireland, v.v.*

```
Nam_Bộ          B-LOC
Bắc_Bộ          B-LOC
Bắc_Bán_Cầu     B-LOC
Đông_Dương      B-LOC
Đông_Âu         B-LOC
Đông_Á          B-LOC
Nam_Á           B-LOC
Đông_Nam_Á      B-LOC
Đông_Nam_Bộ     B-LOC
Trung_Trung_Bộ  B-LOC

West        B-LOC
Texas       I-LOC

Northern    B-LOC
California  I-LOC

các         O
nước        O
Bắc         B-LOC
và          O
Nam_Mỹ      B-LOC
```

- Với trường hợp cấu tạo theo kiểu như *Miền Nam, Miền Bắc, Miền Trung, Miền Đông, Miền Tây, Miền Đông Nam Bộ, Miền Tây Nam Bộ, Xứ Lạng, v.v.* thì xử lí như sau:

```
Miền        B-LOC
Nam         I-LOC

Miền          B-LOC
Đông_Nam_Bộ   I-LOC

Miền          B-LOC
Tây_Nam_Bộ    I-LOC

Xứ          B-LOC
Lạng        I-LOC

Xứ          B-LOC
Thanh       I-LOC

Xứ          B-LOC
Hàn         I-LOC
```

```
Xứ          B-LOC
Wale        I-LOC

Xứ          B-LOC
Catalunya   I-LOC
```

- Không xác định kết hợp giữa hai hay ba từ chỉ hướng với nhau là LOCATION. Trừ trường hợp nói về vùng *Tây Bắc* và *Đông Bắc* là hai trong 3 tiểu vùng địa lí tự nhiên của Bắc Bộ Việt Nam (tiểu vùng còn lại là *Đồng bằng sông Hồng*) hoặc nói về vùng biên giới *Tây Nam* thì vẫn xác định là LOCATION.

```
phía    O
Bắc     O

phía    O
Nam     O

Đông    O
Đông    O
Bắc     O

Tây     O
Tây     O
Nam     O
```

6. Tên địa lí (X) được cấu tạo giữa một *danh từ chung* hoặc *danh từ chỉ hướng* (Y: *đông, tây, bắc, nam, phương, biển*) với một *danh từ chung* hoặc *danh từ chỉ hướng* (Z: *đông, tây, bắc, nam, phương*). Để đảm bảo tính khu biệt, chúng tôi đề nghị xem X là tên địa lí, khi X = Y + Z, và đồng nhất CapWord = X (coi cả khối là CapWord).

Phương Tây, Tây Phương, Đông Bắc, Tây Bắc, Đông Nam, Tây Nam, *Biển Đông, Biển Hồ, v.v.*

```
Tây_Phương      B-LOC
Phương_Tây      B-LOC
Trung_Bộ        B-LOC
Nam_Trung_Bộ    B-LOC
Biển_Đông       B-LOC
Biển_Hồ         B-LOC
```

7. Tên các công trình xây dựng, kiến trúc (X) được cấu tạo giữa một danh từ chung chỉ loại của công trình (Y: *cầu, đường, phố, đại lộ, cao tốc, chùa, tháp, sân, v.v.*) luôn luôn xuất hiện với một tên gọi chính chỉ công trình (CapWord). Trong đó, CapWord thường là tên có lí do (theo một ý nghĩa, sự tích nào đó). Để đảm bảo tính khu biệt, chúng tôi đề nghị xem X là tên công trình, khi X = Y + CapWord. Với tiếng nước ngoài thì thường X = CapWord + Y.

Phố Huế, Phố Cầu Gỗ, Đường Phạm Văn Đồng, Cầu Long Biên, Cầu Tràng Tiền, Cầu Bố, Cầu Si, Chùa Keo, Tháp Bút, Tháp Chàm, Tháp Eiffel (Eiffel Tower, Tour Eiffel), Đại lộ Thăng Long, Đường cao tốc Pháp Vân – Cầu Giẽ, Đường vành đai 3 Hà Nội, v.v.

▷ "*thăm Cầu Tràng Tiền*"

```
thăm          O       O
Cầu           B-LOC   O
Tràng_Tiền    I-LOC   O
```

▷ "*sống ở Phố Huế*"

```
sống          O       O
ở             O       O
Phố           B-LOC   O
Huế           I-LOC   O
```

▷ "*Đường Phạm Văn Đồng*"

```
Đường         B-LOC   O
Phạm_Văn_Đồng I-LOC   O
```

▷ "*trên Đường cao tốc Pháp Vân - Cầu Giẽ*" [dấu gạch ngang (-) được xem là thành phần cấu tạo tên địa danh]

```
trên          O       O
Đường_cao_tốc B-LOC   O
Pháp_Vân      I-LOC   O
-             I-LOC   O
Cầu_Giẽ       I-LOC   O
```

▷ "*tham quan Địa đạo Củ Chi*"

```
tham_quan     O       O
Địa_đạo       B-LOC   O
Củ_Chi        I-LOC   O
```

▷ "*tại Khu Đô thị Time City*"

```
tại           O       O
Khu           B-LOC   O
Đô thị        I-LOC   O
Time          I-LOC   O
City          I-LOC   O
```

▷ "*tại Time City*"

```
tại           O       O
Time          B-LOC   O
```

```
City            I-LOC   O
```

▷ "*For Wall Street…*"

```
For             O       O
Wall            B-LOC   O
Street          I-LOC   O
```

▷ "*Welcome to the Eiffel Tower*"

```
Welcome         O       O
to              O       O
the             O       O
Eiffel          B-LOC   O
Tower           I-LOC   O
```

▷ "*sân Thống Nhất*"

```
sân             B-LOC   O
Thống_Nhất      I-LOC   O
```

▷ "*sân Mỹ Đình*"

```
sân             B-LOC   O
Mỹ_Đình         I-LOC   O
```

▷ "*sân Bet365*"

```
sân             B-LOC   O
Bet365          I-LOC   O
```

▷ "*sân Wembley*"

```
sân             B-LOC   O
Wembley         I-LOC   O
```

▷ "*Wembley Stadium*"

```
Wembley         B-LOC   O
Stadium         I-LOC   O
```

- Tại sao lại có hiện tượng này???
  Từ điển Oxford:

```
Wembley Stadium
White Hart Lane
Old Trafford
```

- Cần thảo luận:

```
Đường          B-LOC   O
Phạm_Văn_Đồng I-LOC   O
```

Hay

```
Đường      B-LOC   O
Phạm       I-LOC   O
Văn        I-LOC   O
Đồng       I-LOC   O
```

- Chấp nhận:

```
Đường           B-LOC   O
Phạm_Văn_Đồng   I-LOC   O
```

8. Trường hợp tên địa chỉ số nhà cũng được xác định là LOCATION

▷ "*Số 14/27 Hoàng Dư Khương*"

```
Số                B-LOC
124/27            I-LOC
Hoàng_Dư_Khương I-LOC
```

▷ "*sống ở 83 Lý Nam Đế*"

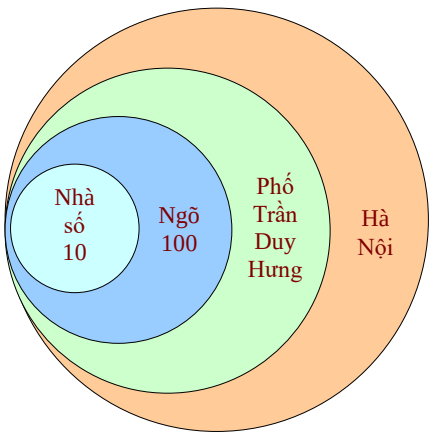
```
83          B-LOC
Lý_Nam_Đế   I-LOC
```

▷ "*Tôi đang ở Số nhà 83 Lý Nam Đế*"

```
Số          B-LOC
nhà         I-LOC
83          I-LOC
Lý_Nam_Đế   I-LOC
```

▷ "Nhà số 10, Ngõ 100, Phố Trần Duy Hưng, Thành phố Hà Nội."

```
Nhà             B-LOC
số              I-LOC
10              I-LOC
,               O
Ngõ             B-LOC
100             I-LOC
,               O
Phố             B-LOC
Trần_Duy_Hưng   I-LOC
,               O
Thành_phố       B-LOC
Hà_Nội          I-LOC
```



C. Tên tổ chức

1. X là tên tổ chức, khi X được tạo bởi một *danh từ chung* (Y) với một *danh từ chỉ tên gọi cụ thể* (CapWord). Danh từ chung có thể đứng trước hoặc đứng sau danh từ chỉ tên gọi cụ thể, và được coi như hai từ. X = Y + CapWord. Với các trường hợp cấu tạo theo trật tự Hán-Việt thì truyền thống coi cả khối là CapWord (*Phật giáo, Công giáo, Thiên Chúa giáo, Ấn Độ giáo,…*).

Đạo Phật, Đạo Thiên Chúa, Đạo Cao Đài, Đạo Kitô, Phật giáo, Công giáo, Thiên Chúa giáo, Ấn Độ giáo, v.v.

- Cần thảo luận đi đến thống nhất:

```
Đạo         B-ORG
Phật        I-ORG

Phật_giáo   B-ORG
```

Hay

```
Đạo         B-ORG
Phật        I-ORG

Phật        B-ORG
Giáo        I-ORG
```

Hay

```
Đạo_Phật    B-ORG
Phật_Giáo   B-ORG
```

Chấp nhận:

```
Đạo         B-ORG
Phật        I-ORG

Đạo         B-ORG
Thiên_Chúa  I-ORG

Đạo         B-ORG   O
Cao_Đài     I-ORG   LOC

Phật        B-ORG
Giáo        I-ORG

Thiên_Chúa  B-ORG
Giáo        I-ORG
```

2. **X** là tên tổ chức, khi **X** thoả mãn mô hình: X = A (bộ phận chỉ loại hình, cấp độ) + B (bộ phận chỉ nhiệm vụ, lĩnh vực đặc thù) + C (bộ phận khu biệt). Căn cứ vào các thành tố tạo nên tên tổ chức, viết hoa chữ cái đầu của mỗi thành tố. Ví dụ: *Trường Đại học Bách khoa Hà Nội, Trường Đại học Khoa học Tự nhiên Hà Nội, Nhà máy Thuốc lá Thăng Long, Công ti Xuất nhập khẩu Hải Hà, v.v.*

▷ "*Trường Đại học Bách khoa Hà Nội tuyển sinh*" [X = A + B + C]

```
Trường      B-ORG   O
Đại_học     I-ORG   O
Bách_khoa   I-ORG   O
Hà_Nội      I-ORG   B-LOC
tuyển_sinh  O       O
```

▷ "*Trường Đại học Hà Nội tuyển sinh*" [X = A + B + C]

```
Trường      B-ORG   O
Đại_học     I-ORG   O
Hà_Nội      I-ORG   B-LOC
tuyển_sinh  O       O
```

▷ "*Viện Đại học mở Hà Nội tuyển sinh*" [X = A + B + C]

```
Viện        B-ORG   O
Đại_học     I-ORG   O
Mở          I-ORG   O
Hà_Nội      I-ORG   B-LOC
tuyển_sinh  O       O
```

▷ "*Trường Đại học New York tại Brockport*" [X = A + B + C]

```
Trường      B-ORG   O
Đại_học     I-ORG   O
New         I-ORG   B-LOC
York        I-ORG   I-LOC
tại         O       O
Brockport   B-LOC   O
```

▷ "*Toà án nhân dân Tỉnh Bình Dương*" [X = A + B + C]

```
Toà_án      B-ORG   O
nhân_dân    I-ORG   O
Tỉnh        I-ORG   B-LOC
Bình_Dương  I-ORG   I-LOC
```

▷ "*Viện kiểm sát nhân dân Thành phố Đà Nẵng*" [X = A + B + C]

```
Viện                B-ORG  O
kiểm_sát_nhân_dân   I-ORG  O
Thành_phố           I-ORG  B-LOC
Đà_Nẵng             I-ORG  I-LOC
```

▷ "*Câu lạc bộ bóng đá Manchester United*" [X = A + B + C]

```
Câu_lạc_bộ    B-ORG   O
bóng_đá       I-ORG   O
Manchester    I-ORGB  LOC
United        I-ORG   O
```

▷ "*Đội bóng Manchester United*" [X = A + B + C]

```
Đội             B-ORG   O
bóng            I-ORG   O
Manchester      I-ORGB  LOC
United          I-ORG   O
```

▷ "*Xí nghiệp Thoát nước Số 3*" [X = A + B + C]

```
Xí_nghiệp   B-ORG
Thoát       I-ORG
nước        I-ORG
Số          I-ORG
3           I-ORG
```

▷ "*Hội Liên hiệp Phụ nữ Việt nam*" [X = A + B + C]

```
Hội          B-ORG   O
Liên_hiệp    I-ORG   O
Phụ_nữ       I-ORG   O
Việt_Nam     I-ORG   B-LOC
```

▷ "*Viện Kinh tế Kỹ thuật Thuốc lá Việt Nam*" [X = A + B + C]

```
Viện         B-ORG   O
Kinh_tế      I-ORG   O
Kỹ_thuật     I-ORG   O
Thuốc_lá     I-ORG   O
Việt_Nam     I-ORG   B-LOC
```

3. Trường hợp tên tổ chức không gây ra sự nhầm lẫn thì *xác định là ORG*.

a. Bộ là cơ quan trung ương của bộ máy nhà nước, lãnh đạo và quản lí một ngành công tác. Mỗi ngành công tác chỉ có một bộ duy nhất nên chỉ cần X = A (bộ phận chỉ cấp độ) + B (bộ phận chỉ nhiệm vụ đặc thù) là thoả mãn. Ví dụ: *Bộ Ngoại giao, Bộ Nội vụ, Bộ GDĐT, Bộ Y tế, v.v.*

▷ "*Bộ Ngoại giao ra thông báo*" [X = A + B]

```
Bộ           B-ORG
Ngoại_giao   I-ORG
ra           O
Thông_báo    O
```

b. Các trường hợp hiển nhiên là ORG do có sự khu biệt rõ ràng (có thể không cần xét đến mô hình).

▷ "*Viện kiểm sát nhân dân tối cao*" [X = A + B]

<ORG>Viện kiểm_sát nhân_dân tối_cao</ORG>

▷ "*Toà án nhân dân tối cao*" [X = B + C]

<ORG>Toà_án_nhân_dân tối_cao</ORG>

▷ "*Chủ tịch UBND TP HCM*" [X = B + C]
Chủ _tịch <ORG>UBND <LOC>TP HCM</LOC></ORG>

▷ "*Công an TP HCM báo cáo*" [X = B + C]
<ORG>Công_an <LOC>TP HCM</LOC></ORG> báo cáo

▷ "*Bộ Nông nghiệp và Phát triển nông thôn*" [X = A + B]
<ORG>Bộ Nông_nghiệp và Phát_triển Nông_thôn</ORG>

▷ "*Tập đoàn Hyundai của Hàn Quốc*" [X = A + B + C]
<ORG>Tập_đoàn Hyundai</ORG> của <LOC>Hàn_Quốc</LOC>

▷ "*Nhà máy Z129*"
<ORG>Nhà máy Z129</ORG>

4. Trường hợp tên tổ chức thuộc một tổ chức lớn hơn mà không gây ra sự nhầm lẫn thì *xác định là ORG*.

a. Tổng cục, cục là cơ quan quản lí một ngành chuyên môn thuộc quyền quản lí của bộ (tổng cục) hay tổng cục (cục).

- Nếu mỗi ngành công tác chuyên môn là tên duy nhất thuộc một bộ hay một tổng cục thì chỉ cần X = A (bộ phận chỉ cấp độ) + B (bộ phận chỉ nhiệm vụ đặc thù) là thoả mãn. Ví dụ: *Tổng cục Chính trị, Tổng cục Hậu cần, Tổng cục An ninh, Tổng cục Thống kê, Tổng cục Đường bộ, Tổng cục Thuế, Tổng cục Hải quan, Cục Điện Ảnh, Cục Quản lí Xuất nhập cảnh, Cục Quân y, v.v.*

▷ "*Tổng cục Chính trị*" [X = A + B]

```
Tổng_cục    B-ORG
Chính_trị   I-ORG
```

▷ "*Cục Quản lí và Khám chữa bệnh, Bộ Y tế*" [X = A + B]

```
Cục         B-ORG
Quản_lí     I-ORG
và          I-ORG
Khám        I-ORG
chữa        I-ORG
bệnh        I-ORG
,           O
Bộ          B-ORG   [X = A + B]
Y_tế        I-ORG
```

▷ "*Bộ Y tế Việt Nam*" [X = A + B + C]

```
Bộ          B-ORG   O
Y_tế        I-ORG   O
```

```
Việt_Nam    I-ORG    B-LOC
```

Chú ý: "Việt Nam" trong cụm "Bộ Y tế Việt Nam" là thành phần không bắt buộc, tuy nhiên khi cần phân biệt với nước khác thì vẫn được xác định là thành phần tham gia với "Bộ Y tế" để cấu tạo tên tổ chức. Khác với "Việt Nam" trong "Hội Liên hiệp Phụ nữ Việt Nam" và "Viện Kinh tế Kỹ thuật Thuốc lá Việt Nam", lại là thành phần cơ hữu của tên tổ chức mà không thể tách rời. Tương tự, "Hà Nội" và "Đà Nẵng" trong "Trường Đại học Bách khoa Hà Nội" và "Trường Đại học Bách khoa Đà Nẵng" cũng là thành phần không thể tách rời.

- Nếu mỗi ngành công tác chuyên môn không phải là tên duy nhất thuộc một bộ hay một tổng cục thì xử lí giống như mục "c". Ví dụ:

<ORG>Cục Công nghệ thông tin - <ORG>Bộ Tư pháp</ORG></ORG>
<ORG>Cục Công nghệ thông tin - <ORG>Bộ Giáo dục và Đào tạo</ORG></ORG>
<ORG>Cục Hợp tác quốc tế - <ORG>Bộ Giáo dục và Đào tạo</ORG></ORG>
<ORG>Cục Hợp tác quốc tế - <ORG>Bộ Văn hoá, Thể thao và Du lịch</ORG></ORG>
v.v.

b. Vụ là cơ quan chuyên môn trong một bộ, một tổng cục hoặc ngang bộ. Các bộ, tổng cục khác nhau có thể có tên một vụ chuyên môn giống nhau, vì vậy để đảm bảo tính khu biệt thì cần phải có có thành phần tên bộ hoặc tổng cục đứng sau, và phải gán nhãn lồng. Ví dụ:

<ORG>Vụ Kế hoạch Tài chính - <ORG>Bộ Y tế</ORG></ORG>
<ORG>Vụ Kế hoạch Tài chính - <ORG>Bộ Tư pháp</ORG></ORG>
<ORG>Vụ Kế hoạch Tài chính - <ORG>Bộ Y tế</ORG></ORG>
<ORG>Vụ Kế hoạch Tài chính - <ORG>Bộ Văn hoá, Thể thao và Du lịch</ORG></ORG>
<ORG>Vụ Kế hoạch Tài chính - <ORG>Tổng cục Thuỷ lợi</ORG></ORG>
<ORG>Vụ Kế hoạch Tài chính - <ORG>Tổng Cục Địa chất và Khoáng sản Việt Nam</ORG></ORG>

c. Trường hợp *khoa, phòng, ban, hội…* thuộc một tổ chức, khu vực thì chỉ gán nhãn ORG khi có đầy đủ cả tên của tổ chức, khu vực đó, và phải gán nhãn lồng. Nếu có thêm "dấu phẩy, dấu gạch ngang" xen vào để chỉ ranh giới của hai hay nhiều thành phần tạo nên tên tổ chức thì vẫn chấp nhận gán nhãn ORG. Ví dụ:

▷ "*Khoa Toán, Trường Đại học Tổng hợp Hà Nội*"

```
Khoa          B-ORG    O          O
Toán          I-ORG    O          O
,             I-ORG    O          O
Trường        I-ORG    B-ORG      O
Đại_học       I-ORG    I-ORG      O
Tổng_hợp      I-ORG    I-ORG      O
Hà_Nội        I-ORG    I-ORG      B-LOC
```

▷ "*Khoa Truyền thông và Văn hoá Đối ngoại, Học viện Ngoại giao*"

```
Khoa            B-ORG   O           O
Truyền_thông    I-ORG   O           O
và              I-ORG   O           O
Văn_hoá         I-ORG   O           O
Đối_ngoại       I-ORG   O           O
,               I-ORG   O           O
Học_viện        I-ORG   B-ORG       O
Ngoại_gia       I-ORG   I-ORG       O
```

d. Trường hợp có các từ "của, thuộc, tại" xen vào tổ hợp tên tổ chức thì tuỳ vào cấu tạo của từng tên tổ chức mà xử lí sao cho phù hợp. Ví dụ:

- "của, thuộc" không nằm trong cấu trúc tên tổ chức (thường dùng dấu phẩy hoặc dấu gạch ngang)

▷ "*Khoa Truyền thông và Văn hoá Đối ngoại của Học viện Ngoại giao*"

```
Khoa            B-ORG   O           O
Truyền_thông    I-ORG   O           O
và              I-ORG   O           O
Văn_hoá         I-ORG   O           O
Đối_ngoại       I-ORG   O           O
của             O       O           O
Học_viện        B-ORG   B-ORG       O
Ngoại_gia       I-ORG   I-ORG       O
```

▷ "*Khoa Truyền thông và Văn hoá Đối ngoại thuộc Học viện Ngoại giao*"

```
Khoa            B-ORG   O
Truyền_thông    I-ORG   O
và              I-ORG   O
Văn_hoá         I-ORG   O
Đối_ngoại       I-ORG   O
thuộc           O       O
Học_viện        B-ORG   O
Ngoại_gia       I-ORG   O
```

- "tại, in, of" nằm trong cấu trúc tên tổ chức

▷ "*Hội Sinh viên Việt Nam tại Đại học La Trobe*"

```
Hội             B-ORG   O           O
Sinh_viên       I-ORG   O           O
Việt_Nam        I-ORG   B-LOC       O
tại             I-ORG   O           O
Đại_học         I-ORG   B-ORG       O
```

```
La_Trobe    I-ORG   I-ORG    B-PER
```

▷ "*Hội người Việt Nam tại tỉnh Champasak*"

```
Hội         B-ORG   O
người       I-ORG   O
Việt_Nam    I-ORG   B-LOC
tại         I-ORG   O
tỉnh        I-ORG   B-LOC
Champasak   I-ORG   I-LOC
```

▷ "*Hội người Việt Nam tại Singapore* (Vietnammese Association in Singapore)"

```
Hội         B-ORG   O
người       I-ORG   O
Việt_Nam    I-ORG   B-LOC
tại         I-ORG   O
Singapore   I-ORG   B-LOC
```

▷ "(Vietnammese Association in Singapore)"

```
(           O       O
Vietnamese  B-ORG   O
Association I-ORG   O
in          I-ORG   O
Singapore   I-ORG   B-LOC
(           O       O
```

▷ "ngân hàng Bank of China"

```
ngân_hàng   O       O
Bank        B-ORG   O
of          I-ORG   O
China       I-ORG   B-LOC
```

e. Trường hợp *trường đại học* thuộc một một cấp đại học cao hơn thì gán nhãn ORG riêng cho từng cấp, không dùng nhãn lồng. Ví dụ:

▷ "*Trường Đại học Ngoại ngữ – Đại học Quốc gia Hà Nội*"

```
Trường      B-ORG   O
Đại_học     I-ORG   O
Ngoại_ngữ   I-ORG   O
-           O       O
Đại_học     B-ORG   O
Quốc_gia    I-ORG   O
Hà_Nội      I-ORG   B-LOC
```

▷ "*Đại học Khoa học xã hội và nhân văn, Đại học Quốc gia Hà Nội*"

```
Đại_học     B-ORG   O
```

```
Khoa_học     I-ORG   O
xã_hội       I-ORG   O
và           I-ORG   O
nhân_văn     I-ORG   O
,            O       O
Đại_học      B-ORG   O
Quốc_gia     I-ORG   O
Hà_Nội       I-ORG   B-LOC
```

f. Trường hợp viết không đầy đủ (viết rút gọn) tên tổ chức thì chấp nhận gán nhãn ORG cho phần rút gọn, nếu không có sự nhầm lẫn. Ví dụ:

▷ <ORG>Ban Chấp hành Trung ương</ORG>
[tên đầy đủ: *Ban Chấp hành Trung ương Đảng Cộng sản Việt Nam*]

▷ <ORG>Ban Bí thư</ORG>
[tên đầy đủ: *Ban Bí thư Ban Chấp hành Trung ương Đảng Cộng sản Việt Nam*]

▷ <ORG>Ủy ban Kiểm tra Trung ương</ORG>
[tên đầy đủ: *Uỷ ban Kiểm ta Trung ương Đảng Cộng sản Việt Nam*]

▷ <ORG>Ủy ban Tư pháp</ORG>
[tên đầy đủ: *Ủy ban Tư pháp Quốc hội*]

g. Trường hợp viết tên tổ chức có một bộ phân không chính xác thì không gán nhãn ORG cho phần không chính xác đó. Ví dụ:

▷ <ORG>Ủy ban Văn hóa giáo dục Thanh Thiếu niên & Nhi đồng</ORG> của Quốc hội
[tên chính xác: *Ủy ban Văn hóa giáo dục Thanh Thiếu niên và Nhi đồng Quốc hội*]

▷ <ORG>Ủy ban Tư pháp</ORG> của Quốc hội
[tên chính xác: *Ủy ban Tư pháp Quốc hội*]

h. Trường hợp chỉ 1 thực thể, nhưng do văn bản thể hiện không chuẩn mực nên dẫn đến xử lí khác nhau. Ví dụ:

- "<ORG>Trung tâm Chẩn đoán và Điều trị Ung bướu</ORG>, tại <ORG>Bệnh viện Quân y 175</ORG>" và
"<ORG>Trung tâm Ung bướu <ORG>Bệnh viện Quân y 175</ORG></ORG>"

- "<ORG>Tổng Liên đoàn Lao động <LOC>Việt Nam</LOC></ORG>" và

"<ORG>Tổng Liên đoàn Lao động (LĐLĐ) <LOC>Việt Nam</LOC></ORG>"

- "sinh viên tốt nghiệp <ORG>Đại học <PER>Harvard</PER></ORG>" và
"Nghiên cứu của <PER>Luke</PER> tại <ORG>Harvard</ORG> đưa ra một lời giải thích khác"
[Harvard là tên người, nhưng ở đây là chỉ tên Đại học Harvard]

- "<ORG>Phòng CSGT Đường bộ, Đường sắt <ORG>Công an <LOC>TP HCM</LOC></ORG></ORG>" và
"<ORG>Phòng CSGT Đường bộ - Đường sắt</ORG> (<ORG>PC67</ORG>) <ORG>Công an <LOC>TP HCM</LOC></ORG>" và
"Trưởng <ORG>PC67, <ORG>Công an <LOC>TP HCM</LOC></ORG></ORG>"

i. Không gán nhãn ORG cho các trường hợp sau:

- *Sở GDĐT, Sở Y tế, v.v.*: Sở là cơ quan quản lí một ngành chuyên môn của nhà nước ở cấp tỉnh và thành phố, nên tỉnh nào cũng có chung tên sở như vậy, do đó chúng trở thành danh từ chung. Chúng chỉ trở thành ORG khi có danh từ xác định đi sau (thường là địa danh).

- *Phòng GDĐT, Phòng LĐTBXH, Phòng khám Đa khoa, Phòng Kinh tế, Phòng Quản lý đô thị, Phòng Nông nghiệp và Phát triển nông thôn, Phòng Công Thương, v.v.*: Phòng là đơn vị hành chính, sự nghiệp của một quận, huyện, nên quận, huyện nào cũng có chung tên phòng như vậy. Chúng chỉ trở thành ORG khi có danh từ xác định đi sau (thường là địa danh).

- *Hội chữ thập đỏ* Xử lí như trường hợp *Sở, Phòng*.

▷ "*Sở Giáo dục Đào tạo*"

```
Sở          O
Giáo_dục    O
Đào_tạo     O
```

▷ "~~*báo cho*~~ *Sở LĐTB&XH* ~~*biết*~~"

```
Sở          O
LĐTB&XXH    O
```

▷ "*Phòng Y tế huyện*"

```
Phòng       O
Y_tế        O
huyện       O
```

▷ "*Uỷ ban Nhân dân xã*"

```
Uỷ_ban      O
Nhân_dân    O
xã          O
```

- *Hội đồng giáo dục, Ban đại diện cha mẹ học sinh, Ban đại diện phụ huynh, Ban phụ huynh, Ban giám khảo, Ban tổ chức, Ban Tổ chức Hội thảo…* là các tổ chức lâm thời, không có tính ổn định và xác định nên tạm thời không gán nhãn ORG.

j. Không gán nhãn ORG cho các trường hợp có mô hình X = A + C

▷ "~~*Các anh CN*~~ *Xí nghiệp 2* ~~*đang thi công công trình*~~" [X = A + C]

```
Xí_nghiệp   O
2           O
```

▷ "~~*anh Hải ở*~~ *Tổ 2* ~~*có con bị dị tật*~~" [X = A + C]

```
Tổ          O
2           O
```

[tổ: *tổ sản xuất* nói tắt]

k. Trường hợp viết tắt cả khối chữ mà không có sự phân tách thì không tách riêng ra các thành phần lồng nhau. Chẳng hạn:

"<ORG>MTTQVN</ORG>" (Mặt trận Tổ quốc Việt Nam)", thì không tách tiếp VN (Việt nam) ra khỏi khối MTTQVN.

D. Nhãn MISC (miscellaneous)

Nhãn MISC không được gán trong một cấu trúc nhãn lồng.

1. Nhãn MISC dùng để markup các trường hợp nhập nhằng giữa tên quốc gia (LOCITION) với các tên có nghĩa thuộc về quốc gia đó (trong tiếng Anh thì dựa vào hình thức biến hình của từ để xác định: danh từ → tính từ).

▷ "*tiếng Việt, người Việt, người Việt Nam, dân tộc Việt*"

```
tiếng       B-MISC
Việt        I-MISC

người     B-MISC
Việt        I-MISC

người     B-MISC
Việt_Nam  I-MISC

dân_tộc     B-MISC
Việt        I-MISC
```

▷ "*tiếng Đức, người Đức*"

```
tiếng       B-MISC
Đức         I-MISC

người    B-MISC
Đức         I-MISC
```

▷ "*tiếng Nga, người Nga, dân tộc Nga*"

```
tiếng       B-MISC
Nga         I-MISC

người    B-MISC
Nga         I-MISC

dân_tộc     B-MISC
Nga         I-MISC
```

2. Dùng để markup tên tác phẩm, tên sự kiện có chứa tên người, tên địa lí (nội dung thông báo không trực tiếp nói về tên người, tên địa lí).

```
sleep           O
over            O
a               O
film            O
than            O
over            O
“               O
Michael         B-MISC
Collins         I-MISC
”               O

in              O
“               O
Schindler       B-MISC
's              I-MISC
List            I-MISC
”               O

film            O
“               O
The             B-MISC
Crying          I-MISC
Game            I-MISC
”               O
```

trong phim "<MISC>Hoa Thiên Cốt</MISC>" (tên bộ phim, cũng là tên nhân vật)

phim truyền hình </MISC>Bao Thanh Thiên</MISC>

"<MISC>Sở Kiều Truyện</MISC>" (tên bộ phim nói về cuộc đời của nhân vật *Sở Kiều*)

"<MISC>Ngày mai Mai cưới</MISC>" (tên bài hát nói về nhân vật tên *Mai*)

"<MISC>Việt Nam ơi</MISC>" (tên bài hát ca ngợi đất nước Việt Nam)

<MISC>Hoa hậu Đại dương Việt Nam 2017</MISC>

<MISC>the Venice Film Festival</MISC>

3. Dùng để markup tên sản phẩm, ứng dụng có gắn tên thương hiệu

<MISC>Samsung Galaxy Note 8</MISC>

<MISC>Samsung Pay</MISC>

<MISC>Apple Music</MISC>

<MISC>Apple Store</MISC>

<MISC>Apple Pay</MISC>

<MISC>Vivo V7+</MISC>

<MISC>SoundMax AH-323</MISC>

<MISC>Boeing 737 Max</MISC>

<MISC>máy bay Boeing</MISC>

4. Tại sao lại có sự khác nhau ở ví dụ sau:

<MISC>Người Việt Nam</MISC> ưu tiên dùng hàng <LOC>Việt Nam</LOC>
<MISC>Người Việt</MISC> ưu tiên dùng <MISC>hàng Việt</MISC>

Với thao tác chèn yếu tố, ta sẽ thấy có sự khác biệt:

hàng của Việt Nam → chấp nhận → *Việt Nam* gán nhãn LOC
người của Việt Nam → không chấp nhận → *người Việt Nam* gán nhãn MISC
người của Việt → không chấp nhận → *người Việt* gán nhãn MISC
hàng của Việt → không chấp nhận → *hàng Việt* gán nhãn MISC

E. XỬ LÍ CÁC VẤN ĐỀ NHẬP NHẰNG

Quy tắc lựa chọn:

Quy tắc 1: Căn cứ vào cấu trúc bề mặt (hình thức cấu tạo của thực thể) và tri thức nền (background knowledge) để xác định nhãn thực thể.

Quy tắc 2: Dựa vào ngữ cảnh và tri thức nền để xác định nhãn thực thể.

2a) Nếu X là thực thể thuộc nhãn Y, nhưng phải suy ra từ cấu trúc bề sâu (nghĩa nội tại được suy ra từ kết hợp cú pháp) mới có thể hiểu được X quy chiếu về nhãn Z, thì vẫn *gán nhãn Y cho X*.

2b) **Nếu** X là thực thể thuộc nhãn Y, nhưng dựa vào ngữ cảnh và tri thức nền ai cũng hiểu X được quy chiếu về nhãn Z (hiện tượng đa nghĩa), thì *gán nhãn Z cho X*.

Quy tắc 3: Ưu tiên chọn nghĩa trội khi thấy việc xác định theo ngữ cảnh hoặc theo bộ nhãn kém thuyết phục. Để xử lí được trường hợp này cần phải có thêm các nhãn ngữ nghĩa khác. Trước mắt, đành thoả hiệp xử lí theo hướng có lợi nhất.

1. Nhập nhằng về cấu trúc

▷ "*các cửa khẩu Cầu Treo (Hà Tĩnh) và Ma Lù Thàng (Lai Châu)*"

```
các          O      O
cửa_khẩu     O      O
Cầu_Treo     B-LOC  O
(            O      O
Hà_Tĩnh      B-LOC  O
)            O      O
và           O      O
Ma_Lù_Thàng  B-LOC  O
(            O      O
Lai_Châu     B-LOC  O
)            O      O
```

[hiểu là *Cửa khẩu Cầu Treo* và *Cửa khẩu Ma Lù Thàng*, xử lí theo *Quy tắc 1*]

▷ "*học sinh Trường THPT Hà Nội - Amsterdam*"

```
học_sinh     O      O
Trường       B-ORG  O
THPT         I-ORG  O
Hà_Nội       I-ORG  B-LOC
-            I-ORG  O
Amsterdam    I-ORG  B-LOC
```

[dấu gạch ngang (-) được xác định là thành phần trong cấu trúc, xử lí theo *Quy tắc 1*]

▷ "*học tại Đại học Công nghệ Nanyang - Singapore*"

```
học          O      O
tại          O      O
Đại_học      B-ORG  O
Công_nghệ    I-ORG  O
Nanyang      I-ORG  B-LOC
-            O      O
Singapore    B-LOC  O
```

[hiểu là *Đại học Công nghệ Nanyang* ở Singapore, xử lí theo *Quy tắc 1*]

▷ "*đường sắt nối liền Bắc – Nam*"

```
đường_sắt    O      O
nối_liền     O      O
```

```
Bắc          B-LOC   O
-            O       O
Nam          B-LOC   O
```
[hiểu là đường sắt nối liền giữa *Miền Bắc* với *Miền Nam*, xử lí theo *Quy tắc 1*]

▷ "*tuyến Đường sắt Bắc-Nam*"

```
tuyến        O       O
Đường_sắt    B-LOC   O
Bắc          I-LOC   B-LOC
-            I-LOC   O
Nam          I-LOC   B-LOC
```
[*Đường sắt Bắc-Nam* là một công trình, xử lí theo *Quy tắc 1*]

▷ "*các nước Bắc và Nam Mĩ*"

```
các          O       O
nước         O       O
Bắc          B-LOC   O
và           O       O
Nam          B-LOC   O
Mĩ           I-LOC   O
```
[hiểu là các nước *Bắc Mĩ* và *Nam Mĩ*, xử lí theo *Quy tắc 1*]

▷ "*Phó trưởng khoa Nhi (BV Đa khoa Hà Đông) cho biết*"

```
Phó          O       O
trưởng       O       O
khoa_Nhi     O       O
(            O       O
BV           B-ORG   O
Đa_khoa      I-ORG   O
Hà_Đông      I-ORG   B-LOC
)            O       O
cho          O       O
biết         O       O
```
[*Phó trưởng khoa Nhi (BV Đa khoa Hà Đông)* - hiểu là: "Phó trưởng khoa" của "Khoa Nhi BV Đa khoa Hà Đông", ở đây bị chập cấu trúc (bị lược bỏ 1 chữ 'khoa') nên chọn cách xử lí theo *Quy tắc 1*]

▷ "*nói 5 thứ tiếng Đức, Pháp, Serbia, Anh và Slovenia*"

```
nói          O       O
5            O       O
thứ          O       O
tiếng        O       O
Đức          B-LOC   O
,            O       O
Pháp         B-LOC   O
```

```
,           O       O
Serbia      B-LOC   O
,           O       O
Anh         B-LOC   O
và          O       O
Slovenia    B-LOC   O
```

[hiểu là *tiếng Đức, tiếng Pháp, ...*, xử lí theo *Quy tắc 1*]

▷ "*UBND huyện An Dương và huyện Thủy Nguyên*"

```
UBND        B-ORG   O
huyện       I-ORG   B-LOC
An_Dương    I-ORG   I-LOC
và          O       O
huyện       B-LOC   O
Thuỷ_Nguyên I-LOC   O
```

[hiểu là *UBND huyện An Dương* và *UBND huyện Thủy Nguyên*, xử lí theo *Quy tắc 2a*]

▷ "*CLB Shanghai Shenhua*" [CLB Thượng Hải Trung Hoa]

```
CLB         B-ORG   O
Shanghai    I-ORG   B-LOC
Shenhua     I-ORG   B-LOC
```

[thiếu bộ phận quan trọng chỉ lĩnh vực đặc thù: *Shanghai Greenland Shenhua Football Club*, xử lí theo *Quy tắc 2b*]

▷ "*chủ tịch Shanghai Shenhua*"

```
chủ_tịch    O       O
Shanghai    B-ORG   B-LOC
Shenhua     I-ORG   B-LOC
```

[hiểu là chủ tịch *Shanghai Greenland Shenhua Football Club*, xử lí theo *Quy tắc 2b*]

▷ "*gia nhập Man United*"

```
gia_nhập    O       O
Man         B-ORG   B-LOC
United      I-ORG   O
```

[hiểu là *Manchester United Football Club*, xử lí theo *Quy tắc 2b*]

▷ "*U16 Việt nam*"

```
U16         B-ORG   O
Việt_Nam    I-ORG   B-LOC
```

[hiểu là *Đội tuyển Bóng đá U16 Việt Nam*, xử lí theo *Quy tắc 2b*]

▷ "*đội tuyển Việt nam*"

```
đội_tuyển   O       O
Việt_Nam    B-LOC   O
```

[thiếu bộ phận quan trọng chỉ lĩnh vực đặc thù: *bóng đá, bóng chuyền, bơi, điền kinh…*, xử lí theo *Quy tắc 1*]

2. Nhập nhằng về ngữ nghĩa

2.1. Nhập nhằng giữa tên người (PERSON) và tên đường phố (LOCATION)

▷ "*cố thủ tướng Phạm Văn Đồng*"

```
cố_thủ_tướng        O      O
Phạm                B-PER  O
Văn                 I-PER  O
Đồng                I-PER  O
```

▷ "*Đường Phạm Văn Đồng*"

```
Đường               B-LOC  O
Phạm_Văn_Đồng       I-LOC  O
```

▷ "*chủ tịch Hồ Chí Minh*"

```
chủ_tịch            O      O
Hồ                  B-PER  O
Chí                 I-PER  O
Minh                I-PER  O
```

▷ "*sống ở Thành phố Hồ Chí Minh*"

```
sống                O      O
ở                   O      O
Thành_phố           B-LOC  O
Hồ_Chí_Minh         I-LOC  O
```

▷ "*giao lộ Âu Cơ, Trường Chinh*"

```
giao_lộ          B-NP    O
Âu_Cơ            B-LOC   O
,                O       O
Trường_Chinh     I-LOC   O
```

2.2. Nhập nhằng giữa tên người (PERSON) và tên bệnh

▷ "*Bệnh Down*"

```
Bệnh        B-MISC
Down        I-MISC
```

▷ "*Bệnh Down*"

```
Bệnh        B-MISC
Alzheimer   I-MISC
```

▷ "*Hội chứng Stevens - Johnson*"

```
Hội_chứng   B-MISC
Stevens     I-MISC
-           I-MISC
Johnson     I-MISC
```

2.3. Nhập nhằng giữa PERSON, LOCATION với tên phương tiện giao thông

Do chưa gán nhãn cho các phương tiện, sản phẩm nên tạm thời gán nhãn MISC kết hợp nhãn LOC, PER:

▷ "*tàu Hoàng Đạt I*"

```
Tàu         B-MISC
Hoàng_Đạt   I-MISC
I           I-MISC
```

▷ "*tàu ngầm Hoàng Sa*"

```
Tàu_ngầm    B-MISC
Hoàng_Sa    I-MISC
```

▷ "*Chiến hạm Đinh Tiên Hoàng*"

```
Chiến_hạm       B-MISC
Đinh_Tiên_Hoàng I-MISC
```

▷ "*sửa chữa tàu Côn Sơn, Chí Linh*"

```
Sửa_chữa    O
tàu         B-MISC
Côn_Sơn     I-MISC
,           O
Chí_Linh    B-LOC
```

2.4. Nhập nhằng giữa ORGANIZATION và LOCATION

▷ "*Ông Toàn đến thăm Lâm trường Vĩnh Hảo*"

```
Ông          O       O
Toàn         B-PER   O
đến          O       O
thăm         O       O
Lâm_trường   B-LOC   O
Vĩnh_Hảo     I-LOC   O
```

▷ "*Ông Toàn là nhân viên của Lâm trường Vĩnh Hảo*"

```
Ông          O       O
Toàn         B-PER   O
```

```
là          O       O
nhân_viên   O       O
của         O       O
Lâm_trường  B-ORG   O
Vĩnh_Hảo    I-ORG   O
```

▷ "*Họ đứng trước cổng Nhà hát Thăng Long*"

```
Họ          O       O
đứng        O       O
trước       O       O
cổng        O       O
Nhà_hát     B-LOC   O
Thăng_Long  I-LOC   O
```

▷ "*Ca sĩ Tấn Minh làm giám đốc Nhà hát Thăng Long*"

```
Ca_sĩ       O       O
Tấn         B-PER   O
Minh        I-PER   O
làm         O       O
giám_đốc    O       O
Nhà_hát     B-ORG   O
Thăng_Long  I-ORG   O
```

▷ "~~*diễn ra*~~ *tại khách sạn Pierre & Vacances*"

```
tại         O       O
khách_sạn   B-LOC   O
Pierre      I-LOC   O
&           I-LOC   O
Vacances    I-LOC   O
```

▷ "*nhân viên khách sạn Pierre & Vacances*"

```
nhân_viên   O       O
khách_sạn   B-ORG   O
Pierre      I-ORG   O
&           I-ORG   O
Vacances    I-ORG   O
```

▷ "*NASDAQ*" [a stock exchange – chỉ một sàn giao dịch chứng khoán]

```
NASDAQ      B-ORG
```

▷ "*Hyundai of Korea, Inc.*" [Tập đoàn Hyundai Hàn Quốc]

```
Hyundai  B-ORG   O
of       I-ORG   O
Korea    I-ORG   O
,        I-ORG   O
Inc.     I-ORG   O
```

▷ "*Hyundai, Inc. of Korea*" [Tập đoàn Hyundai của Hàn Quốc]

```
Hyundai     B-ORG
,           I-ORG
Inc.        I-ORG
of          O
Korea       B-LOC
```

▷ "*University of California in Los Angeles*" [Đại học California ở Los Angeles]

```
University B-ORG
of         I-ORG
California I-ORG
in         O
Los        B-LOC
Angeles    B-LOC
```

▷ "*Ngân hàng Nhà nước ở Đà Nẵng*"

```
Ngân_hàng  B-ORG
Nhà_nước   I-ORG
ở          O
Đà_Nẵng    LOC
```

▷ "*tại Cảng Quốc tế Long An ~~... đã diễn ra lễ công bố~~*"

```
tại        O
Cảng       B-LOC
Quốc_tế    I-LOC
Long_An    I-LOC
```

▷ "*~~lễ ký kết hợp tác chiến lược giữa ...~~ De Heus và Cảng Quốc tế Long An*"

```
De         B-ORG
Heus       I-ORG
và         O
Cảng       B-ORG
Quốc_tế    I-ORG
Long_An    I-ORG   B-LOC
```

▷ "*~~Next week~~ Kansai Electric Power and Kansai International Airport ~~are likely to launch 10-year dollar deals~~*"

"Tuần tới, Kansai Electric Power và Sân bay Quốc tế Kansai có thể tung ra thương mại đồng đôla 10 năm"

```
Kansai          B-ORG   B-LOC
Electric        I-ORG   O
Power           I-ORG   O
and             O       O
Kansai          B-ORG   B-LOC
International   I-ORG   O
Airport         I-ORG   O
```

2.5. Nhập nhằng giữa ORGANIZATION và tên sản phẩm

Đối với các trường hợp lưỡng khả, vừa chỉ tên một *tờ báo*, vừa chỉ tên một *cơ quan báo chí* thì ưu tiên gán nhãn ORG, khi mà chưa có nhãn gán cho tên sản phẩm (tác phẩm).

▷ "*Theo Thời báo Kinh tế SG*"

```
Theo        O       O
Thời_báo    B-ORG   O
Kinh_tế     I-ORG   O
SG          I-ORG   LOC
```

[Markup cả khối "Thời báo Kinh tế SG" - tên tờ báo, cũng chỉ tên một cơ quan báo chí]

▷ "*Tờ báo Tuổi Trẻ đăng tin*"

```
Tờ_báo      O       O
Tuổi_Trẻ    B-ORG   O
đăng        O       O
tin         O       O
```

[Tuổi Trẻ - tên tờ báo, cũng chỉ tên một cơ quan báo chí, tên đầy đủ: Báo Tuổi Trẻ]

▷ "*Tờ Tuổi Trẻ đăng tin*"

```
Tờ          O       O
Tuổi_Trẻ    B-ORG   O
đăng        O       O
tin         O       O
```

▷ "*Báo Tuổi Trẻ đăng tin*"

```
Báo         B-ORG   O
Tuổi_Trẻ    I-ORG   O
đăng        O       O
tin         O       O
```

▷ "*phóng viên Báo Sài Gòn Giải Phóng*"

```
phóng_viên  O       O
Báo         B-ORG   O
Sài_Gòn     I-ORG   LOC
Giải_Phóng  I-ORG   O
```

▷ "*phóng viên báo Người Đưa Tin*"

```
phóng_viên      O       O
báo             B-ORG   O
Người Đưa Tin   I-ORG   O
```

[báo Người Đưa Tin - tên tờ báo, cũng chỉ tên một cơ quan báo chí, tên đầy đủ: Báo điện tử Người Đưa Tin]

▷ "*Theo VietNam Net*"

```
Theo          O
VietNam Net   B-ORG
```

[VietNam Net - tên tờ báo, cũng chỉ tên một cơ quan báo chí, tên đầy đủ: Báo điện tử VietNam Net]

▷ "*Theo vietnamnet.vn*"

```
Theo            O
vietnamnet.vn   O
```

[vietnamnet.vn - là một URL]

▷ "*Theo VnExpress*"

```
Theo       O
VnExpress  B-ORG
```

[VnExpress - tên tờ báo, cũng chỉ tên một cơ quan báo chí, tên đầy đủ: Báo điện tử Tin nhanh - VnExpress]

▷ "*Theo Báo Mới*"

```
Theo       O
Báo_Mới    O
```

[Báo Mới - là một website tổng hợp thông tin (phần mềm), không phải là cơ quan báo chí]

▷ "*Tạo tài khoản trên www.facebook.com*"

```
Tạo               O
tài_khoản         O
trên              O
www.facebook.com  O
```

[www.facebook.com là một URL]

▷ "*Tạo tài khoản để đăng nhập facebook*"

```
Tạo        O
tài_khoản  O
để         O
đăng_nhập  O
Facebook   O
```

[Facebook là một trang web (sản phẩm phần mềm)]

▷ "*Facebook là trang mạng xã hội*"

```
Facebook   O
là         O
một        O
trang_mạng O
xã_hội     O
```

[Facebook là một trang web (sản phẩm phần mềm)]

▷ "*Facebook bị kiện*"

```
Facebook   B-ORG
bị         O
kiện       O
```

[Facebook được dùng lâm thời là tên một công ti]

▷ "*trên Khoẻ 365*"

```
trên       O
Khoẻ 365   O
```

[Khoẻ 365 (Sức Khoẻ 365) là chuyên trang của Báo Người Đưa Tin]

▷ "*Ảnh: Wikkihow*"

```
Ảnh        O
Wikihow    O
```

[Wikihow là một trang web (sản phẩm phần mềm)]

▷ "*Ảnh: Warya Post*"

```
Ảnh        O
Warya Post O
```

[Warya Post là một trang web (sản phẩm phần mềm)]

▷ "*Tham gia diễn đàn NEU Confession*"

```
Tham_gia   O
diễn_đàn   O
NEU        B-ORG
confession O
```

[NEU là tên viết tắt của Trường Đại học Kinh tế Quốc dân]

▷ "*áo thun Adidas*"

```
áo_thun    O
Adidas     B-ORG
```

[áo thun thể thao của hãng Adidas]

▷ "*giày Nike*"

```
giày     O
Nike     B-ORG
```

[giày thể thao của hãng Nike]

▷ "*vợt Wilson*"

```
vợt      O
Wilson   B-ORG
```

[vợt tennis của hãng Wilson]

▷ "*điện thoại Nokia*"

```
điện_thoại O
Nokia      B-ORG
```

[điện thoại của hãng Nokia]

▷ "*đồng hồ Rolex*"

```
đồng_hồ    O
Rolex      B-ORG
```

[Rollex lúc đầu là tên chỉ một loại (type) đồng hồ của hãng Montres Rolex, nhưng hiện nay Rollex cũng là tên hãng]

2.6. Nhập nhằng giữa tên mã chứng khoán và tên viết tắt của ORGANIZATION

Tên mã chứng khoán của một công ti nếu trùng với tên viết tắt của công ti đó thì gán ngãn ORG.

▷ "*cổ phiếu ACB*"

```
Cổ_phiếu   O       O
ACB        B-ORG   O
```

▷ "*FLC ~~sau 2 phiên tăng nhẹ đã quay đầu giảm~~*"

```
FLC   B-ORG   O
```

2.7. Nhập nhằng giữa LOCATION, PERSON và ORGANIZATION trong thể thao

a) Nếu tên câu lạc bộ bóng đá viết rút gọn mà trùng với tên địa lí, tên người (như *Chelsea, Southampton, Barcelona, Bình Dương, Nguyễn Trãi,Phương Anh, v.v.*) thì được gán nhãn ORG (xử lí theo *Quy tắc 2b*).

▷ "*đương kim vô địch Chelsea*"

```
đương_kim   O
vô_địch     O
Chelsea     B-ORG
```

▷ "*MU ~~đều gặp những đối thủ nhẹ ký hơn là~~ Crystal Palace và Southampton*"

```
MU          B-ORG
…           O
Crystal     B-ORG
Palace      I-ORG
và          O
Southampton B-ORG
```

▷ "*Bình Dương thắng Hà Nội 2-0*"

```
Bình_Dương  B-ORG
thắng       O
Hà_Nội      B-ORG
2-0         O
```

▷ "*Nguyễn Trãi là đội bóng phong trào*"

```
Nguyễn_Trãi B-ORG
là          O
đội         O
bóng        O
phong_trào  O
```

Tham khảo TĐ Oxford:

Chelsea: **1** a district of London, England, on the north bank of the Thames west of Westminster. In the 19th and early 20th centuries many artists lived there and it had a reputation as an artistic area. ... **2** an English football club, based in Fulham, West London, which has been successful in Britain and Europe at various times since the 1950s.

Crystal Palace: **1** the Crystal Palace a very large building made of glass and metal, designed in nine days by Joseph Paxton (1801–65) for the Great Exhibition of 1851. … **2** a football club in south London.

Manchester City: a football team from Manchester, England. It was formed in 1894 and has had several wins in the FA Cup, as well as many other successes in Britain and Europe.

Manchester United: a football team from Manchester, England, with a ground at Old Trafford. It was formed in 1902 and has won the FA Cup more times than any other team, as well as having many other successes in Britain and Europe.

Southampton: a city on the south coast of England. It is one of Britain's most important ports. …

b) Trường hợp tên đội tuyển bóng đá quốc gia rút gọn trùng với tên nước (quốc gia) thì gán nhãn LOC (xử lí theo *Quy tắc 1*).

▷ "*tuyển VN ~~nằm chung bảng~~ với Myanmar, Campuchia, Đông Timor và Malaysia*"

```
tuyển       O
VN          B-LOC
…           O
với         O
Myanmar     B-LOC
,           O
Campuchia   B-LOC
,           O
Đông        B-LOC
Timor       I-LOC
và          O
Malaysia    B-LOC
```

▷ "*ĐTQG Việt Nam*"

```
ĐTQG        O
```

```
Việt Nam    B-LOC
```

2.8. Nhập nhằng giữa LOCATION và các tên khác

Nhập nhằng giữa *tên nước, thuộc về nước/quốc gia, chính thể quốc gia* (hình thức tổ chức của một nhà nước), *nhà cầm quyền* (những người nắm quyền quản lí nhà nước ở cấp trung ương nói chung).

1) Chỉ tên nước:

tổ chức tại Nga
căn cứ trên đất Mỹ

2) Thuộc về nước/quốc gia:

thời thuộc Pháp ((thuộc sự cai trị của nhà cầm quyền Pháp)
thực dân Pháp (thuộc tầng lớp bóc lột, thống trị ở nước Pháp)
chiến sĩ Nga
quân nhân Nga
quân Mỹ
lính Mỹ
pháo thủ Mỹ
thương binh Mỹ

3) Chỉ chính thể:

Cộng hoà Pháp
Cộng hoà Liên Bang Nga
Việt Nam Dân chủ Cộng hoà
Cộng hoà XHCN Việt Nam
Việt Nam Cộng hoà

4) Chỉ nhà cầm quyền:

Nga tiếp tục cảnh báo Mỹ về Triều Tiên
bị Nga xoá sổ
phía Mỹ không đồng ý
quan hệ Mỹ-Liên Xô

Khảo sát ví dụ:

"*Với lưới lửa phòng không dày đặc này, Israel sẽ không còn ngang nhiên tấn công Syria, điều này giúp Nga lấy lại được vị thế trước đồng minh của mình.*"

- Về nghĩa thì "Israel" chỉ chung chính phủ và quân đội chứ không phải chỉ đất nước Israel (lãnh thổ Israel: Location). Bởi vì Location thì không thể "ngang nhiên tấn công Syria" được. Trong khi "Syria" lại hoàn toàn được hiểu là lãnh thổ Syria (Location). Tương tự, "Nga" ở đây cũng được hiểu là chỉ toàn thể nhân dân Nga chứ không phải chỉ lãnh thổ Nga (Location).

- Như vậy, nếu sét "Israel" và "Nga" là ORG, còn "Syria" là LOC thì ta thấy nó vừa phi hình thức, vừa không đúng về mặt ngữ nghĩa của khái niệm ORGANIZATION. Để nhận ra được sự khác nhau giữa chúng đòi hỏi phải dựa vào nhiều loại nhãn ngữ nghĩa khác nữa. Trong khi chưa gán được nhãn ngữ nghĩa khác, tạm thời chấp nhận *địa danh hoá tất cả các trường hợp theo kiểu nêu trên*, nhất loạt gán nhãn LOCATION, trừ một số trường hợp sẽ gán nhãn MISC (kiểu như <MISC>Việt Nam Cộng hoà</MISC>).

F. CHÚ Ý CÁC TRƯỜNG HỢP KHÁC

1. Từ vay mượn (loan word) **và từ chuyển** (code-switching)**.**

▷ Từ vay mượn tiếng nước ngoài (không phải là PER, LOC, ORG) theo đường phiên âm, chuyển tự hoặc để nguyên dạng và được sử dụng như là từ của tiếng Việt (được từ điển thu thập) thì nhất loạt gán nhãn POS theo dạng Xb (X là từ loại nào đấy, b là viết tắt của từ "borrowed"). Từ nguyên dạng tiếng nước ngoài được đưa vào văn bản nhưng không được sử dụng phổ biến trong tiếng Việt (không/chưa được từ điển thu thập) thì nhất loạt gán nhãn POS là FW (foreign word).

múa	V	B-VP
sexy	Ab	B-AP
casino	Nb	B-NP
…	CH	O
sex	Nb	B-NP
truy_cập	V	B-VP
Internet	Nb	B-NP
xuất_trình	V	B-VP
passport	FW	B-NP
made	FW	O
in	FW	O
Vietname	NNP	B-NP

2. Tiếng Anh

▷ "Mr. Harry Schearer"
Mr. <PER>Harry Schearer</PER>

▷ "John Doe, Jr." [John Doe, Junior: John Doe Con]
<PER>John Doe, Jr.</PER>

▷ “giải thưởng Nôben” [the Nobel prize]
[không markup]

▷ “máy tính Macintosh” [Macintosh computers]
[không markup]

▷ “New Zealand”

New	NNP	B-NP	B-LOC
Zealand	NNP	I-NP	I-LOC

▷ “New York”

New	NNP	I-NP	B-LOC
York	NNP	I-NP	I-LOC

▷ “Wall Street”

Wall	NNP	I-NP	B-LOC
Street	NNP	I-NP	I-LOC

▷ “Ellis Park”

Ellis	NNP	I-NP	B-LOC
Park	NNP	I-NP	I-LOC

▷ “Grace Road”

Grace	NNP	B-NP	B-LOC
Road	NNP	I-NP	I-LOC

▷ “South Africa”

South	NNP	B-NP	B-LOC
Africa	NNP	I-NP	I-LOC

▷ “White House”

White	NNP	B-NP	B-LOC
House	NNP	I-NP	I-LOC

▷ “Chittagong Cement”

Chittagong	NNP	B-NP	B-ORG
Cement	NNP	I-NP	I-ORG

▷ “Minor Counties XI”

Minor	NNP	B-NP	B-ORG
Counties	NNPS	I-NP	I-ORG
XI	NNP	I-NP	I-ORG

▷ “Dynamo Moscow”

Dynamo	NNP	B-NP	B-ORG
Moscow	NNP	I-NP	I-ORG

▷ “Boston Res Sox”

Boston	NNP	B-NP	B-ORG
Red	NNP	I-NP	I-ORG
Sox	NNP	I-NP	I-ORG

▷ “Party of Democratic Action”

Party	NNP	B-NP	B-ORG
of	IN	I-PP	I-ORG
Democratic	NNP	I-NP	I-ORG
Action	NNP	I-NP	I-ORG

▷ “Test and County Cricket Board”

Test	NNP	B-NP	B-ORG
and	CC	I-NP	I-ORG
County	NNP	I-NP	I-ORG
Cricket	NNP	I-NP	I-ORG
Board	NNP	I-NP	I-ORG

▷ “Duke of Norfolk 's XI”

Duke	NNP	B-NP	B-ORG
of	IN	I-PP	I-ORG
Norfolk	NNP	I-NP	I-ORG
's	POS	I-NP	I-ORG
XI	NNP	I-NP	I-ORG

Japanese	JJ	I-NP	I-MISC
English	JJ	I-NP	I-MISC