English—Vietnamese Named Entity Guidelines

Version 1.0 – English

January 24, 2012

Created by Quoc Hung Ngo (hungnq@uit.edu.vn)

(Based largely on the MUC-7 NE Guidelines and Thai NE Guidelines)

http://code.google.com/p/evbcorpus/

TABLE OF CONTENTS

1 Introduction

An entity is some object in the world — for instance, a place, a person, or an organization. A named entity is a phrase that uniquely refers to that object by its proper name, acronym, nickname or abbreviation. Some examples of named entities follow:

Mt. Fuji	Núi Fuji
the Kremlin	Điện Kremlin

2 Entity Types

We will identify six types of named entities:

PERSON (PER): Person entities are limited to humans identified by name, nickname or alias.

ORGANIZATION (ORG): Organization entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure.

LOCATION (LOC): Location entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments.

TIME (TIM): Date/Time entities are limited to humans identified by name, nickname or alias.

PERCENTAGE (PCT): Percentage entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure.

MONEY (MON): Money entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments.

2.1 Person Names

People may be specified by name, nickname or alias. Family names should also be tagged as PERSON. Names of deceased people, as well as fictional

human characters appearing in movies, television, books and so on, should be tagged as PERSON entities. Religious deities should also be tagged as persons.

Titles, roles and honorifics such as "Mr." and "President" are ignored from the individual's name. For instance, in the following sentences, there are two separate entities marked:

- − Phó tổng thống [Cheney]$_{PER}$ ghé thăm nơi này.
- + Vice President [Cheney]$_{PER}$ visited this place.
- − Chủ tịch [Microsoft]$_{ORG}$ [Bill Gates]$_{PER}$ phát biểu rằng …
- + [Microsoft]$_{ORG}$ Chairman [Bill Gates]$_{PER}$ stated that …
- − Ông [Nguyễn Thiện Nhân]$_{PER}$ thăm các trường trước năm học mới .
- + Mr. [Nguyen Thien Nhan]$_{PER}$ visits schools before new academic-year.

2.2 Organization Names

Tag all proper name mentions of groups with a defined organizational structure. These include:

Businesses

- − [Công ty thể thao Bridgestone]$_{ORG}$ thu lợi nhận từ các sản phẩm xa xỉ.
- + [Bridgestone Sports Co.]$_{ORG}$ profits from luxury products.

Stock exchanges

- − Cổ phiếu [NASDAQ]$_{ORG}$
- + [NASDAQ]$_{ORG}$ shares

Multinational organizations

- − [Liên minh Châu Âu] $_{ORG}$ thể hiện …
- + [European Union] $_{ORG}$ represents …

Political parties

- − [Đảng Dân chủ GDP] $_{ORG}$
- + [GOP]$_{ORG}$ hopeful

Non-generic government entities

- − [Bộ Ngoại giao] $_{ORG}$
- + [the State Department] $_{ORG}$

Sports teams

- − [Câu lạc bộ MU] $_{ORG}$
- + [the Phillies] $_{ORG}$

Military groups

- Nhóm [Những con hổ Tamil]$_{ORG}$
+ [the Tamil Tigers]$_{ORG}$

Many other kinds of entities refer to facilities or buildings that are primarily defined by their established organizational structure, and can do things like issue statements, make decisions, hire people, raise money and so on. A mention of such an entity should be tagged as an ORGANIZATION when it functions like an ORG in the document. These include things like:

Churches and other religious institutions
- [Nhà thờ Đức Bà]$_{ORG}$
+ [Trinity Lutheran Church]$_{ORG}$

Hospitals
- [Bệnh viện Nhi Trung Ương]$_{ORG}$
+ The [National Hospital of Pediatrics]$_{ORG}$

Hotels
- [Khách sạn Bốn mùa]$_{ORG}$
+ [Four Seasons Hotel]$_{ORG}$

Museums
- [Bảo tàng Chứng tích Chiến tranh]$_{ORG}$
+ The [War Remnants Museum]$_{ORG}$

Universities
- [Đại học Chicago]$_{ORG}$
+ the [University of Chicago]$_{ORG}$

Government offices
- [Nhà Trắng]$_{ORG}$
+ the [White House]$_{ORG}$

Note that definite and indefinite determiners ‘the’ and ‘a’ are included in the annotation, except for cases when they quantify something other than the tagged entity, as in the following examples. This rule does not apply in the Vietnamese language.
+ A [Gulshan Hotel]$_{ORG}$ spokesman
+ the [U.S.] $_{ORG}$ Vice President

As in the above examples, this exception is particularly common when the tagged name is used in the pre-modifier (adjective) position.

General ORGANIZATION-like non-entities

General entity mentions such as "the university", "the national", "đại học", "chính phủ" should not be tagged, since these are not unique proper name references to specific entities.

2.3 Location Names

Examples of place-related strings that are tagged as LOCATION include named heavenly bodies, continents, countries, provinces, counties, cities, regions, districts, towns, villages, neighborhoods, airports, highways, street names, factories, manufacturing plants, street addresses, oceans, seas, straits, bays, channels, sounds, rivers, islands, lakes, national parks, mountains, fictional or mythical locations, and monumental structures, such as the Eiffel Tower, Washington Monument, Bến Nhà Rồng, Chùa Thiên Mụ. For instance:

- Người dân thường tụ tập ở [Nhà thờ Đức Bà]$_{LOC}$ vào mỗi dịp lễ lớn.
+ [The Walt Whitman Bridge]$_{LOC}$ remained closed repairs began on a 10-mile stretch of [the Alaskan Pipeline]$_{LOC}$
+ [The Garden State]$_{LOC}$ is known for its tomatoes.

These nicknames and aliases should be marked as any other country name. They should be tagged according to their meaning and treated variably as ORG or as LOC.

2.3.1 Compound expressions

There are several issues surrounding the expression of location names and which parts of a string to tag.

Compound expressions in which place names are separated by a comma in English should be tagged as separate instances of LOCATION.

+ [Kaohsiung] $_{LOC}$, [Taiwan] $_{LOC}$ [Washington] $_{LOC}$, [D.C.] $_{LOC}$

In Vietnamese, compound expressions in which place names are also separated by a comma should be tagged as separate instances of LOCATION.

- [Hà Nội]$_{LOC}$, [Việt Nam] $_{LOC}$

2.3.2 Designators

When a "designator" is customarily used as a regular part of a place name, that word should also be included in the extent of the LOCATION entity.

For example, include in the tagged string the word "River" in the name of a river, "Mountain" in the name of a mountain, "City" in the name of a city, etc., if such words are contained in the string.

- [Thành phố Hồ Chí Minh]$_{LOC}$

+ [HoChiMinh City]$_{LOC}$
− [Sông Sài Gòn]$_{LOC}$
+ [Saigon River]$_{LOC}$

2.3.3 Location modifiers and "semi-official" place names

Often times place names are modified by words like "Southern", "Lower", "West", "the former" and so on.

When these modifiers are part of a location's official name they should be tagged as part of the LOCATION name. For instance:

− [Thượng nguồn Mê-kong]$_{LOC}$ [Bắc Dakota]LOC
+ [Upper Volta]$_{LOC}$, [North Dakota]$_{LOC}$

Even if the place name does not have "official" status but has an agreed-upon definition and is in very frequent use, the string should be tagged as a LOCATION, as in:

− [Trung Đông]$_{LOC}$ [Ngân hàng Phương Tây]$_{LOC}$ [Đông Âu]$_{LOC}$
+ [the Middle East] $_{LOC}$ [the West Bank] $_{LOC}$ [Eastern Europe] $_{LOC}$

When these modifiers are not the official name of a place, or when the definition of the place might vary from person to person, do not tag the modifier as part of the LOCATION entity name.

+ [Mississippi River] $_{LOC}$ west bank former [Soviet Union] $_{LOC}$
+ [Gaul] $_{LOC}$, in present-day [France] $_{LOC}$ lower [Manhattan] $_{LOC}$
+ [Northern California] $_{LOC}$

These place names can sometimes be tricky. If you are not sure whether a modifier is part of an official name, you should include the modifier as part of the place name.

2.4 Time and Date

Both "absolute" time expressions and certain "relative" time expressions, as specified below (B.1.2), are to be tagged in MUC-7. Note that the tag itself does not differentiate between "absolute" and "relative" types, i.e., all time expressions are labeled with the same type of tag. The salient features of the time expressions that are marked is that whether absolute or relative, they can be anchored on a timeline; unanchored durations, for example, are not marked.

The TIME sub-type is defined as a temporal unit shorter than a full day, such as second, minute, or hour. The DATE sub-type is a temporal unit of a full day or longer. Both DATE and TIME expressions may be either absolute or

relative. Both absolute and relative times are tagged as TIME and absolute and relative dates are tagged as DATE.

2.4.1 Absolute Temporal Expressions

To be considered an absolute time expression, the expression must indicate a specific segment of time, as follows:

TIME-tagged expressions

- An expression of minutes must indicate a particular minute and hour, such as "20 minutes after 10", or "10 giờ 20".
- An expression of hours must indicate a particular hour, such as "midnight," "twelve o'clock noon," "noon" (not "mid-day," "morning").

DATE-tagged expressions

- An expression of days must indicate a particular day, such as "Monday," "10th of October" (not "first day of the month").
- An expression of seasons must indicate a particular season, such as "autumn" (not "next season").
- An expression of financial quarters or halves of the year must indicate which quarter or half, such as "fourth quarter," "first half." Note that there are no proper names, per se, representing these time periods. Nonetheless, these types of time expressions are important in the business domain and are therefore to be tagged.
- An expression of years must indicate a particular year, such as "1995" (not "the current year").
- An expression of decades must indicate a particular decade, such as "1980s" (not "the last 10 years").
- An expression of centuries must indicate a particular century, such as "the 20th century" (not "this century").

Temporal expressions are to be tagged as a single item. Contiguous subparts (month/day/year) are not to be separately tagged unless they are taggable expressions of two distinct TIMEX sub-types (date followed by time or time followed by date).

+ [twelve o'clock noon]$_{TIM}$
+ [5 p.m. EST]$_{TIM}$
+ [January 1990]$_{TIM}$

+ [fiscal 1989]$_{TIM}$
+ the [autumn]$_{TIM}$ report
+ [third quarter of 1991]$_{TIM}$
+ [the fourth quarter ended Sept. 30]$_{TIM}$
+ [the three months ended Sept. 30]$_{TIM}$
+ [the first half of fiscal 1990]$_{TIM}$
+ [first-half]$_{TIM}$ profit
+ [fiscal 1989's fourth quarter]$_{TIM}$
+ [4th period]$_{TIM}$
+ [1975]$_{TIM}$ World Series
+ [February 12]$_{TIM}$, [8 A.M.]$_{TIM}$
+ by [9 o'clock]$_{TIM}$ [Monday]$_{TIM}$

2.4.2 Relative Temporal-Expressions

A relative temporal expression (RTE) indicates a date relative to the date of the document ("yesterday", "today", etc.), or a portion of a temporal unit relative to the given temporal unit ("morning" as the initial part of a specified day). Taggable RTE's include compound temporal expressions containing a deictic marker followed by a time unit, such as "last month" or "next year". If a numeral is included in RTE's of this type, it falls within the scope of the taggable temporal expression ("last two months"). Note that sometimes the deictic marker is postposed, as in "10 years ago" and "four months later".

Note also that some RTE's lexicalize deictic markers and time units into a single word, such as "yesterday", which by itself constitutes a taggable expression, and that some RTE's can contain more than one deictic marker, such as "early this year" and "earlier this month." In addition, note that some of the expressions specifically defined as not being absolute temporal expressions are considered markable as relative temporal expressions.

Compound ("marker-plus-unit") temporal expressions, and their lexicalized equivalents, should be tagged as single items. However, if a lexicalized "marker-plus-unit" modifies a contiguous time unit of a different sub-type, they should be tagged as two items. Contrast the following two example markups:

− [tối hôm qua]$_{TIM}$
− [sang ngày mai]$_{TIM}$
+ [last night]$_{TIM}$
+ [yesterday evening]$_{TIM}$

2.4.3 Miscellaneous Temporal Non-Entities

Indefinite or vague date expressions with non-specific starting or stopping dates will not be tagged. Non-taggable expressions include:

- Vague Time Adverbials: eg. "now", "recently", ("bây giờ", "hiện nay", "gần đây"), etc.
- Indefinite Duration-of-Time Phrases: eg. "for the past few years" ("những năm qua", "những năm sắp tới")
- Time-Relative-to-Event Phrases: eg. "since the beginning of arms control negotiations"

2.5 Percentages

The entire string expressing the monetary or percentage value is to be tagged.

Both "absolute" time expressions and certain "relative" time expressions, as specified below (B.1.2), are to be tagged in MUC-7. Note that the tag itself does not differentiate between "absolute" and "relative" types, i.e., all time expressions are labeled with the same type of tag. The salient features of the time expressions that are marked is that whether absolute or relative, they can be anchored on a timeline; unanchored durations, for example, are not marked.

− khoảng [5%]$_{PCT}$
− khoảng [5 phần trăm]$_{PCT}$
− hơn [55%]$_{PCT}$
+ about [5%]$_{PCT}$
+ more than [55%]$_{PCT}$

2.6 Money

Modifiers that indicate the multiplied value of a number unit should be included in the tagged string, if the modifier is a substitute for a specific digit (or the indefinite article or other quantitative determiner) within the monetary expression.

− xấp xỉ [20 triệu New Pesos]$_{MON}$
− [5 triệu đồng]$_{MON}$
− [5,000,000 VNĐ]$_{MON}$
− hơn [90,000 đô-la]$_{MON}$
+ approximately [20 million New Pesos]$_{MON}$
+ over [$90,000]$_{MON}$

In this case, "several" is a substitute for some specific digit such as "3", or "4." Note that the expression remains grammatical if such a digit is substituted for the word "several", but that the expression "about 10 million New Pesos" does NOT remain grammatical if "about" is replaced by a digit. The indefinite article also can be substituted for "several," but not for "about," in the same examples.

- − [vài triệu New Pesos]$_{MON}$
- − [vài triệu đô-la]$_{MON}$
- + [several million New Pesos]$_{MON}$
- + [several million dollars]$_{MON}$

Miscellaneous Numeric Non-Entities

Numeric expressions that do not use currency terms to indicate money values and that do not use percentage terms to indicate percentages are not to be tagged.

+	12 points	[no markup]
+	unchanged at 95.05	[no markup]
+	1.5 times	[no markup]
+	about one-third of	[no markup]

2.7 Deciding among entity types

There are some situations where deciding what entity type to assign can be somewhat tricky.

ORG referring to LOC, LOC referring to ORG

Many organizations have not only an organizational structure, but a physical location. For instance, museums are primarily organizations but are also housed in a specific building or facility. So while we normally tag museums as ORG entities, there are cases when a particular example might function more like a LOCATION. In cases like this, annotators should tag the named entity based on the way it functions in the sentence.

For instance:

- − [Bảo tàng Guggenheim]$_{ORG}$ thông bảo về một cổ vật mới.
- − [Bảo tàng Guggenheim]$_{LOC}$ là điểm đến của nhiều du khách .
- − 30 người bị thương trong vụ đánh bom trước [Khách sạn Gulshan]$_{LOC}$ của thành phố.
- − Người phát ngôn [Khách sạn Gulshan]$_{ORG}$ gọi vụ tai nạn là một thảm kịch.
- + [The Guggenheim Museum]$_{ORG}$ announced a new acquisition.

+ [The Guggenheim Museum]$_{LOC}$ was designed by [Wright]$_{PER}$.
+ Thirty people were wounded in the bomb blast in front of the city's [Gulshan Hotel]$_{LOC}$.
+ A [Gulshan Hotel]$_{ORG}$ spokesman called the incident a tragedy.

Similarly, city, country, and other place names are frequently used to refer to organizations located in those places rather than the geographical places themselves. For instance:

− Hôm nay [Washington]$_{ORG}$ thông báo chính sách thuế mới.
+ [Washington] $_{ORG}$ announced a new tax policy today.

In this case, the name "Washington" and "ไซล" is used to refer to the US Government, located in Washington and the South Korean Government, located in Seoul, respectively. Because "Washington" and "ไซล" are referring to an organization entity in this example, it should be tagged as ORG.

Also, when the name of a unique structure or building (normally a location) is used to refer to the government or other organization housed in that facility, the name should be tagged as an ORG:

− [Pentagon]$_{ORG}$ phát đi thông điệp về vụ việc xảy ra.
+ [The Pentagon]$_{ORG}$ issued a statement about the incident.

The same logic applies to place names referring to sports teams. This rule does not apply to the Vietnamese language.

− [Boston]$_{ORG}$ thắng [New York]$_{ORG}$ tối qua trong hiệp đấu phụ.
+ [Boston]$_{ORG}$ beat [New York]$_{ORG}$ last night in extra innings.

In summary, for any cases where a place name is used to refer to an organization, you should tag the name based on function in the sentence:

ORG: used when the example primarily refers to the organizational structure, and is acting like an agent (issuing a statement, making a decision, hiring people, raising money, etc.)

LOC: used when the example primarily refers to the physical structure, rather than the people/groups who run it.

3 Difficult Cases

3.1 Expressions that refer to multiple entities

When a phrase refers to multiple named entities, mark each entity separately. For instance, this sentence contains two entities:

- [Trung Quốc]$_{LOC}$ và [Hàn Quốc]$_{LOC}$ ký thỏa thuận thương mại.
+ [China]$_{LOC}$ and [South Korea]$_{LOC}$ signed the trade agreement.

Similarly,

- [Jimmy]$_{PER}$ và [Rosalyn Carter]$_{PER}$
+ [Jimmy]$_{PER}$ and [Rosalyn Carter]$_{PER}$
- [North]$_{LOC}$ và [South America]$_{LOC}$
+ [North]$_{LOC}$ and [South America]$_{LOC}$

But be careful not to split apart proper names that contain a conjunction. For instance,

- the [Bộ Nông nghiệp và Phát triển Nông thôn]$_{ORG}$
+ [Ministry of Agriculture and Rural Development]$_{ORG}$

is the name of one organization and should be tagged as a single named entity (it's not “Ministry of Agriculture” and “Rural Development” and it’s not “Bộ Nông nghiệp” and “Phát triển Nông thôn” as separate names).

3.2 Nested Expressions

Recall that no nested expressions will be marked. When the name of one entity contains within it another entity name, do not pull out the name of the other entity and mark it separately. Only tag the larger entity. For instance

- [30 đô-la Mỹ] no markup for Mỹ alone
+ [Arthur Anderson Consulting]$_{ORG}$ no markup for Arthur Anderson alone
+ the [U.S. Customs Service]$_{ORG}$ no markup for U.S. alone
+ [U.S. $10 million]$_{ORG}$ no markup for U.S. or $10 alone

3.3 Entities as modifiers

If the entity name occurs in the form of an adjective you should also tag it. This rule does not apply to the Vietnamese language because Vietnamese does not have an adjectival form. However, its translation form is presented as a noun phrase.

- Các công ty [Mỹ]$_{LOC}$, công dân [Cuba]$_{LOC}$, và thức ăn [Trung Quốc]$_{LOC}$
+ the [American]$_{LOC}$ companies, [Cuban]$_{LOC}$ citizens, và [Chinese]$_{LOC}$ food

3.4 Possessives

When you encounter a possessive construction, tag the two parts individually as two separate names. For instance:

[Temple University's] [Graduate School of Business] [Canada's] [Parliament]

Keep in mind that annotation requires you to select whole words to tag as names, so you have to include the "'s" even though it's not part of the name.

3.5 Other types of names

Aliases, acronyms, nicknames and abbreviations for proper names should be tagged as a name:

− ACB [acronyms for Asia Commercial Bank or Ngân hàng Á Châu]
+ IBM [abbreviation for International Business Machines]
+ Big Blue [alias for International Business Machines]

4 What NOT to tag

4.1 Events

Do not tag event names, even if they refer to events that occur on a regular basis and are associated with institutional structures. However, the institutional structures themselves — steering committees, etc. — should be tagged.

− [Ủy ban Olympic]$_{ORG}$ [Organization]
− Thế vận hội Mùa đông [no markup]
+ [the Olympic Committee] $_{ORG}$ [Organization]
+ the Pan-American Games [no markup]

4.2 Artifacts and products

Miscellaneous types of proper names that are not to be tagged as named entities include artifacts, other products, and plural names that do not identify a single, unique entity. For instance,

− Taurus là mẫu xe mới nhất [no markup]
+ the Taurus is the latest car model [no markup]

5 Annotation Uncertainty

In some cases, you may encounter examples that you don't know how to handle. If so, you should proceed as follows:

- If it's an example not covered in the guidelines, note it in your copy of the guidelines and let your supervisor know about it.
- If it's an example where your language differs from the rules as written for English, note it in your copy of the guidelines and let your supervisor know about it.
- If it's a case where there's something wrong with the file you're working on, stop working on that file and let your supervisor know.

Your supervisor might not be available at the moment you have a question or issue, so it's important for you to write down the problem so it can be resolved later.

So that you can keep working even after you have a question about a particular example, we've created one more tag in the annotation tool: "No_Annotation". When you encounter a problem or have a question about a particular word or phrase and you can't get an immediate answer, label the item "No_Annotation". That will let us easily find it later when we try to resolve the problem.

You should also use the "No_Annotation" label in cases where there's some kind of problem with a single word or handful of words in the file – e.g., they're badly translated or the font isn't displaying properly.

If the whole file is problematic (i.e., poor translation, corrupted, font problems), stop working on it and let your supervisor know. If you are using

AWS to receive file assignments, you may simply mark the file as “Broken” in the file-assignment interface. If you use this approach, please be sure to include a descriptive comment when AWS asks what is wrong with the file.