VietBioNER ANNOTATION GUIDELINES

INTRODUCTION

In this project, we focus on extracting information from documents related to tuberculosis in Vietnam. Specifically, we are interested in adverse drug reactions, multidrug resistant, diagnostic, therapeutic and preventive procedures in tuberculosis.

Henceforth, we denote that all expressions in a sentence that belong to the annotation category under discussion are enclosed in [square brackets]; expressions that could be considered as candidates for annotation, but should actually not be annotated are highlighted in red.

The task of an annotator is to strictly follow our detailed guidelines, described in the next section, to label the entities of interest within text. Moreover, the annotator should note some general rules as follows, during the annotating process.

- **Do not tag erroneous words.** Most texts have undergone optical character recognition (OCR) and thus contain a significant number of typos, e.g., "tìm **ơi** khuẩn lao âm tính", "Thời **sian** ho **truns** bình", "bệnh **phối**". For such erroneous words, please skip them.
- **Do not tag unclear cases.** If the annotator is not sure about a given mention, even after consulting some external sources, the corresponding mention should remain unlabelled.
- **Note sentences' boundary**. There are cases that a sentence is presented in different lines as shown in the below table. In such cases, we still tag "**lao phổi AFB(+) tái phát**" as a disease entity.

Line 1	Để đánh giá tình hình lao kháng thuốc Rifampicin trong nhóm **lao phổi**
Line 2	**AFB(+) tái phát** tại Cần Thơ là bao nhiêu

DETAILED GUIDELINES

In the scope of our project, we will annotate nine following categories of entities.

1. Organisation (ORG)

Names of specific organizations will be tagged in this category.

✓Include:

- Names of organizations (government or non-government), offices, unions
    - [Bộ Y tế]
    - [Tổ chức Y tế Thế giới]
    - [WHO]
    - [Các chi cục thuế, thú y, bảo vệ thực vật]

    - [Uỷ ban Nhân dân]
    - [Khoa Phổi Thận] [Bệnh viện Chợ Rẫy] (please note that there are two organisation entities in this case)
    - [khoa Giải phẫu bệnh] [Bệnh viện Nhân Dân Gia Định]
    - [UBND tỉnh Đồng Nai]- thông báo
    - [UBND thành phố] thông báo

- Tag all the common nouns (pre-words) with the proper nouns of the organization. Tag the whole phrase (including information about the specialty, characteristics, features, functionalities, proper name modifiers) of the organizations such as: đội tuyển, CLB, trường học, bệnh viện, công ty, xí nghiệp, siêu thị, etc.
    - [Bệnh viện Nhân Dân Gia Định]
    - [siêu thị Big C]
    - [Công ty TNHH Tiến Đạt]
    - [Đài Truyền hình Tp.Hồ Chí Minh]

- The common nouns indicating organizations of parties, government and their specialty will be included in one entity.
    - [Bộ Giáo dục và Đào tạo]
    - [Bệnh viện Lao và Bệnh phổi Cần Thơ]
    - [Trường Đại học Y Khoa Hà Nội]

- Foundations
    - [Quỹ quốc tế về bảo vệ thiên nhiên]
    - [Quỹ tiền tệ quốc tế] thông báo
    - [Quỹ Ford] trả lời phỏng vấn

✗Exclude: Plural form or unspecific entity:

- Các Chính phủ
- Các Nhà nước
- Các chi cục thuế
- một bệnh viện ở Hà Nội
- phòng nghiên cứu
- phòng thí nghiệm

2. Location (LOC)

Mentions of geographic locations, i.e., any identifiable point or area in the planet, ranging from continents, major bodies of water (e.g., oceans, rivers, lakes), named landforms, countries, states, cities, and towns, are marked up as geographic location entities. It should be noted that this type of mentions does not only include Vietnam geographic locations but also world-wide locations (outside of Vietnam).

✓Include: Instances of proper names and their abbreviations, except when used in the context as a political entity.

- ... thuộc [khu vực Đông Nam Á]
- ... còn có thể loại trừ bệnh này ở cả [châu Á] và [châu Phi]
- ... TẠI MỘT SỐ ĐỊA PHƯƠNG [PHÍA BẮC VIỆT NAM]
- [quận 3]
- [quận Tân Bình]
- [xã An Khê]
- [nước Mỹ]

✗ Exclude:

- Indefinite or definite quantifiers (hai, ba, bốn, các, những, một số, …) before the proper names will NOT be tagged.
    - các nước [Đông Nam Á]
    - các bang [miền Nam nước Mỹ]
    - một số nước [Bắc Âu]
    - Tại các tỉnh [Long An] , [Tiền Giang] , [An Giang]
    - Ba nước [Đông Dương] [Việt Nam], [Lào], [Campuchia]
    - Hai chợ [Tân Bình] và [Tân Định]
- If directional words are used not to mention a location or location direction, they will NOT be tagged:
    - gió mùa Đông Bắc
    - nhà xây hướng Tây Nam
- Unspecific entities:
    - thành phố này
    - cùng một quận
    - Vùng đất miền Trung
    - Khu vực miền Nam

✓ **Not tagging vs tagging**:

There are names which are always location (e.g river name, street name…), or organization (e.g Ministry name, party name…), but there are also names which can be tagged or depending on their contexts.

- Location name refers to its organizational body will be tagged
    - [tòa án tối cao] của [Nhật]
    - Chính sách mới của [Singapore]
- Location name which takes action, communication (announcement, agreement, approve, statement, denials, etc.) and decision-making, for example:
    - [Trung Quốc] tiết lộ rằng ...
    - [Bảo tàng Louvre] sẽ mở cửa lại chủ nhật này …
- Other context referring to organization bodies:
    - Quan hệ hợp tác giữa [Mỹ]] và [Việt Nam]
    - [Mỹ] bàn giao cho [Việt Nam] ...
- Location name follows organization name. There are two possible taggings:
    - name is part of the name: tag the whole as
        - [Bộ Y tế Nhật Bản]
        - [Đại học Quốc gia Tp.Hồ Chí Minh]
    - name is not part of the name: tag and separately:
        - [Đại học Harvard] của [Mỹ]

- [Đại học Stanford] ở [Mỹ]

3. DateTime (DTM)

✓Include: Time, a specific period of time (having specific beginning and ending)

- Seasons of the year
  - [mùa hè] [mùa hạ]
  - [mùa xuân]
  - [mùa đông]
  - [mùa thu]
  - Vụ [hè thu]
  - Vụ [đông xuân năm 2009-2010]

- Dates, months and years
  - [29 tháng bảy năm 2005]
  - [sáng 19 tháng 4]
  - [tháng 3]
  - [giữa tháng giêng 2004]
  -
  - [ngày 15/10]
  - [năm 2005]
  - Hà Nội [2005]
  - [thứ 2]
  - [Thứ Hai]
  - [thế kỷ 20]
  - [đầu thế kỷ XX]

- Special holidays or festivals in the year
  - [Ngày Quốc khánh]
  - [Ngày Quốc tế Lao động]
  - [Giỗ tổ Hùng Vương]
  - [Quốc tế Thiếu nhi]
  - [Tết Nguyên Đán]
  - [Tết dương lịch]
  - [Tết âm lịch]

- Specific time
  - [8 : 21 , (GMT+7) 15/11/2005]
  - [từ 1990-1992]
  - [Từ cuối 2002 đến nay]
  - [1995 - 2001]
  - [25 - 26/10]

- The name of the age:
  - [Thời Bảo Đại]
  - [Thời Phục Hưng]

• Combination of datetime:
    • [Ngày 6 và 7/11]
    • [Ngày 3 và 5]
    • [Ngày 3] và [ngày 5]

✗ Exclude:

- Quantifiers are NOT tagged in this case:
    - Các ngày [thứ bảy]
    - Các ngày [mùng một] hằng tháng
- Words used to introduce the datetime such as: khoảng, gần, ngay sau, ngay trước will not be tagged along with the datetime.
    - Khoảng [tháng bảy]
- Words describing the event, awards or programmes will be excluded
    - Mùa Oscar [2010]
    - Giải Nobel [2004]
- Words do not indicate specific time
    - Sáng sớm tinh sương
    - Tờ mờ sáng
    - Nhá nhem tối
    - Chập choạng tối
    - buổi sáng, buổi chiều, …
    - Đầu ngày, cuối ngày …
    - Từ trước đến nay
    - Sinh viên năm thứ 2

4. Symptom and Disease (SYM_DIS)

4.1. Disease

✓Include:

- All diseases, illness, inflammation, or disorder conditions related to human,
    - [bệnh lao] đã trở thành một vấn đề toàn cầu
    - bệnh nhân [lao phổi AFB(+) tái phát]
    - những trường hợp nghi mắc [lao đa kháng thuốc] hoặc [lao đồng nhiễm HIV]
    - ... có kèm [viêm gan siêu vi B]
    - ... đồng mắc [lao phổi] và [ung thư]
    - [bệnh mạn tính]: [loét dạ dày-tá tràng], [đái tháo đường], [suy thận mãn]
- Annotator can check the Vietnamese ICD10 for further reference about diseases: http://123.31.27.68/ICD/ICD10.htm

✗Exclude:

- Although they are rare in our document collection, in this work we do not tag diseases of plants or animals.

4.2. Symptom

✓Include: Altered physical appearance or behaviour as a probable result of injury and/or underlying pathological/disease process, and thus a sign of the disease process, rather than

disease or illness in itself. This category includes only symptoms that could be experienced, observed, and described by a patient directly.

- triệu chứng lâm sàng (hay gặp là [ho khạc đờm]SYM)
- [đau ngực]SYM, [khó thở]SYM
- Annotator can check the Vietnamese ICD10 for further reference about symptoms: http://123.31.27.68/ICD/ICD10.htm

✗Exclude: exclude diseases or illness as they have to be annotated as another category of entities

6. Diagnostic Procedure (DiaPro)

✓Include: A procedure method or technique used to determine the composition, quality, or concentration of a specimen, and which is carried out in a clinical laboratory [UMLS definition].

- ... chẩn đoán bằng [AFB (+)]
- ... [sinh thiết màng phổi bằng kim]

✗Exclude: exclude common nouns such as “phương pháp", "kỹ thuật"

- ... sử dụng kỹ thuật [GeneXpert]
- ... phương pháp [nhuộm lam tìm AFB]