Named Entity Annotation Guidelines for MSA

Abdelati Hawwari, Fahad AlGhamdi, Rehab Ibrahim, and Mona Diab
CARE4LANG lab
http://care4lang1.seas.gwu.edu/
George Washington University

Table of Contents
1 Introduction......................................................................................................................... 2
  1.1 What is a Named-Entity................................................................................................. 2
      1.1.1 NE features........................................................................................................... 2
      1.1.2 NE subtypes ...................................................................................................... 2
  1.2 Modern Standard Arabic (MSA) .................................................................................. 4
  1.3 What is your Task? ......................................................................................................... 5
2 NE Categories .................................................................................................................... 8
  2.1 Person.......................................................................................................................... 8
  2.2 Location ....................................................................................................................... 8
  2.3 Organization ................................................................................................................ 9
  2.4 Group .......................................................................................................................... 9
  2.5 Title (عنوان) ............................................................................................................. 9
  2.6 Product ....................................................................................................................... 9
  2.7 Time ............................................................................................................................ 10
  2.8 Event .......................................................................................................................... 10
  2.9 Other .......................................................................................................................... 10
  2.10 Not-NE ..................................................................................................................... 10
3 Typo .................................................................................................................................. 11
  3.1 Definition .................................................................................................................... 11
  3.2 Types of typo ............................................................................................................. 11
  3.3 How to Annotate a TYPO case? ................................................................................. 11
  3.4 Notes ........................................................................................................................ 11
4 General Notes.................................................................................................................. 12
1 Introduction

1.1 What is a Named-Entity

Named Entity (NE) is a word or multi word that represents names of a unique entity such as people’s names, countries and places, organizations, companies, websites, etc.

1.1.1 NE features

The NE has some different linguistic features, including lexical, semantic, syntactic and orthographic features. The following table summarizes the main linguistic features that could be used to identify the concept of NE.

<table>
<thead>
<tr>
<th>#</th>
<th>Feature Type</th>
<th>Description</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lexical features</td>
<td>NE is considered as an encyclopedic unit (not a lexical or a terminological unit)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Semantic features</td>
<td>A name represents one object.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Syntactic features</td>
<td>Serves as a syntactic unit, with a syntactic function, even it is made up of multi word unit</td>
<td>لإم تؤكد ( وكالة أنباء الشرق الأوسط) الخبر</td>
</tr>
<tr>
<td>4</td>
<td>Orthographic features</td>
<td>Starts with a capital letter (in some languages)</td>
<td>For Arabic, you can try to translate it into English, as a test,</td>
</tr>
</tbody>
</table>

*A summary of Linguistic features for the NE*

1.1.2 NE subtypes

The following table summarizes the subtypes of NE. (You will find a detailed information in section 2).
NE Types summary

Person
- Name
- Title
- Nickname
- Cities, countries
- Geographical
- Facilities
- Companies
- Agencies
- Institutions
- Ministries
- parties
- movements

Location

Organization

Group
- Teams
- Ethnic groups
- Religious
- Works of Art

Title

Product

Time

Event

Other

Not-NE
1.2 Modern Standard Arabic (MSA)

MSA is the formal language of Arabic typically used in formal settings such as newspapers, political speeches, religious sermons, etc.

- All the data is mainly in Modern Standard Arabic (MSA).
- Some words are transliterated into the Latin alphabet (for example, حلوه أوي could be written as 7elwa 2awy). These are considered Arabic words even though they have been transliterated. Thus, they must be annotated accordingly.
- Classical Arabic (i.e. Quranic verses, Hadith, poetry, proverbs) is all considered MSA for the purposes of this project.

Example

الآن أحل ضيفا على قناة النيل للأخبار

<table>
<thead>
<tr>
<th>Word</th>
<th>Annotation</th>
<th>Typo?</th>
</tr>
</thead>
<tbody>
<tr>
<td>الآن</td>
<td>Not-NE</td>
<td>Typo</td>
</tr>
<tr>
<td>أحل</td>
<td>Not-NE</td>
<td>Correct</td>
</tr>
<tr>
<td>ضيفا</td>
<td>Not-NE</td>
<td>Correct</td>
</tr>
<tr>
<td>على</td>
<td>Not-NE</td>
<td>Correct</td>
</tr>
<tr>
<td>قناة</td>
<td>B-TITLE</td>
<td>Correct</td>
</tr>
<tr>
<td>النيل</td>
<td>I-TITLE</td>
<td>Correct</td>
</tr>
<tr>
<td>للأخبار</td>
<td>I-TITLE</td>
<td>Correct</td>
</tr>
</tbody>
</table>
1.3 What is your Task?
Your task is to 1) identify the NE words in tweets and, 2) tag each NE into 10 labels based on the description given to each label. (See Table# below), and 3) identify the beginning, the inside and words that are not member inside the multi word NE (NE chunk)

1.3.1 Identifying the NE
- You will be given a sentence as a whole (text), the sentence ID and then each token (space separated unit) as a row entry with a unique ID.
- You will fill out two columns: Annotation and Correct/Typo
- In “Annotation” column, you will find 10 possible labels represent the NE Subtypes, which are Person, Location, Organization, Group, Title, Product, Event, Time, Other and Not-NE (See pages).

1.3.2 Multi-word NE Annotation: The IOB tagging format
The NE may comprise more than one word. In case of multi-word NE, you need to identify three things: 1) the beginning of the NE Chunk, words that are parts of /member in the NE Chunk, and words that is inside the NE Chunk but not parts of /member in the NE Chunk
The IOB tagging format (I is short for inside, O is short for outside, and is short for beginning)
The used tagging format (IOB) is as follows:
- I-tag indicates that the tag is inside the NE chunk.
- O-tag indicates that a token is not a part of the NE chunk.
- **B-tag** is used in the beginning of every Multi-word NE (NE chunk), to indicates the beginning of an NE chunk.

The following table presents the NE tags in IOB tag format:

<table>
<thead>
<tr>
<th>#</th>
<th>NE Type</th>
<th>Beginning</th>
<th>Inside</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Person</td>
<td>B-PER</td>
<td>I-PER</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Location</td>
<td>B-LOC</td>
<td>I-LOC</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Organization</td>
<td>B-ORG</td>
<td>I-ORG</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Production</td>
<td>B-PROD</td>
<td>I-PROD</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Title</td>
<td>B-TITLE</td>
<td>I-TITLE</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Event</td>
<td>B-EVENT</td>
<td>I-EVENT</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Group</td>
<td>B-GROUP</td>
<td>I-GROUP</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Time</td>
<td>B-TIME</td>
<td>I-TIME</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Other</td>
<td>B-OTHER</td>
<td>I-OTHER</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Not_NE</td>
<td></td>
<td></td>
<td>O</td>
</tr>
</tbody>
</table>

*A summary of NE in IOB tag format*

1) **Typo tagging**

In “Correct/Typo” column: the default tag is Correct. You will either change the cell it to Typo if you believe the word is a typo or you leave it Correct (See pages????)

- You will have to annotate all the words according to these 10 labels even the ones you marked as typos.
- Each row entry is a word in a tweet. Thus, to understand the context of each word, you have to read the whole sentence before you start to annotate the individual words.
DO NOT JUDGE THE WORDS IN ISOLATION. Each word should be judged based on the CONTEXT.
2 NE Categories

Each category of a Named Entity is further divided into subcategories as shown in Section 2-1, below.
You need to choose, as a second level of NEs, an NE subcategory for each NE and tag it with 10 possible labels represent the NE Subtypes, which are Person, Location, Organization, Group, Title, Product, Event, Time, Other and Not-NE (–PER)

2.1 Person (–PER)
- Proper nouns, nicknames
  - First, last and full names
    - النبي دانيال ، السيدة عائشة، محمود شكري، محمد عبدالسلام، ارنولد شوارننجر، السيدة عائشة، المنفلوطي
  - Nicknames
    - أمير الازهر، وردة في البستان، المصرية المحجبة، مصرى فرفان
    - اسماء الشهرة refer to concrete individuals
      - الشريفين ، نجمة الجماهير
  - Titles of (specific) persons
    - وزير الداخلية ، الرئيس ، رئيس الوزراء، مفتي الازهر
      - ابو أحمد، ابو خليفة ، ابو العراء
  - God names could be tagged – in this task - as “Person”
    - 99 names of God
    - References to God

Notes:
- الأستاذ ،الدكتور، المهندس الأستاذ ،الدكتور، and their abbreviations are NOT NE, if it appear alone without names. It could be parts of a multi-word NE, for example: الدكتور محمد البرادعي، المهندس ممدوح جمزة...
- Pronouns are not NE.

2.2 Location (–LOC)
- Geographical or physical locations like:
  - Cities, provinces, states countries, address, facilities, etc.
    - نيويورك، الصعيد، ولاية اريزونا ، الخريطوم، المنطقة الشرقية ، المحافظة الغربية
  - Mountain ranges, bodies of water, deserts
    - قناة السويس، هضبة نجد، صحراء سيناء ، البحر المتوسط
  - نهر النيل ، شلالات نيارجا
- Facilities
  - Buildings: names of schools, airports, masjids, churches, etc.

Notes:
- مبنى الحزب الوطني، المتحف المصري، مطار القاهرة الدولي ، مدرسة

Highways, bridges, tunnels and squares (شارع محمد علي، كوبري ٦ أكتوبر، ميدان التحرير - كوبري الجامعة)

Notes:
- The stadium, the field, for example, are considered NE if they refer to specific places that are known to all (الاستاد = استاد القاهرة، الميدان = ميدان التحرير)

2.3 Organization (−ORG)
Organization include any legal personality entity that has employees, including:
- Companies (مايكروسوفت، جوجل، فيس بوك)
- Agencies (البنك الدولي)
- Institutions (جامعة القاهرة، AUC)
- International Organizations (منظمة حقوق الإنسان، الأمم المتحدة)
- Ministries (وزارة التجارة، وزارة الخارجية)
- Others (القضاء المصري، البيت الأبيض)

2.4 Group (−GROUP)
A group of people that is not an organization which has a unique name.

“Title” includes the following subtypes:
- Political parties & movements (الحزب الوطني، جبهة الإنقاذ، حركة تمرد، اللجنة الثلاثية، حركة ٦ أكتوبر)
- Sports Teams (منتخب إسبانيا، الأهلي، الزمالك، المنتخب الإندونيسي)
- Ethnic group (قبيلة العبابدة)
- Religious groups (الإخوان المسلمين)
- board, council or senate (مجلس الشعب، المجلس الأعلى للقوات المسلحة

2.5 Title (عنوان) (−TITLE)
“Title” includes the following subtypes:
- Titles of books, articles (جدد حياتك، ذا سيكرت، أولاد حارتنا)
- Title of songs (اهواك سيرة الحرب، الاطالات،)
- Plays and concerts (مهرجان الموسيقي العربي، العيال كبرت)
- TV shows and movies (برنامج القاهرة اليوم، برنامج ودن واحدة)
- Media (websites, channels, newspapers, etc.) (موقع تويتر، جريدة الشرق، جريدة الشرق الأوسط)

2.6 Product (PROD)
“Product” includes the following subtasks
- Brands
  o Vehicles (bmw، لكزس، هونداي، ميتسوبيشي، بيجو)
  o Weapons (اف ١٥)
  o Foods brands (ليز، أمريكانا)
2.7 Time (–TIME)
“Time” includes the following subtypes:
- Months, days of the week, hours, etc.
  - Months (يوليو، رمضان، شعبان، أيول) (السنت، فرادي) 
  - Weekdays (السبت، فرايداي) 
- Holidays that happen periodically:
  - Religious holidays (عيد الميلاد المجيد، ليلة القدر، ليلة المنتصف من شهر (شعبان، الأيام البيض، عيد الأضحى، المولد النبوي، أيام عشر ذو الحجة) (فوريه أف جولي، شم النسيم) 
  - National events (عيد الأمل، عيد الحب، كذبة ابريل، (الدوري المصري) 
  - Social Holydays (عيد الأم، عيد الحب، كذبة ابريل، (Periodically) sport events

2.8 Event (–EVENT)
A well-known (one time) event that happened or will happen.
“Event” includes the following subtypes:
- Battles and wars (الحرب العالمية الثانية، غزوة بدر، حرب أكتوبر، فتح مكة) (موسومة الجمل، ثورة 25 يناير: الثورات) (اعصار ساندي، اعصار كاترينا) 
- Revolutions 
- Hurricanes 

2.9 Other (–OTHER)
NE that cannot be classified to any of the other NE subtypes labels, could be tagged as “Other”

Note:
Before tagging an NE with “Other” please make sure that the word you are tagging is NE, and no existing tag fits it.

2.10 Not-NE (O)
This tag is given to any word that is Not an NE (Not-NE).
3 Typo

3.1 Definition
Typo is a word that you recognize, but you believe the author misspelled.
As we are working on Twitter data, some NEs may be misspelled, or not written
in a proper manner, you need to tag the NE if it has a typo.

3.2 Types of typo
- **Misspelling**: Words that are misspelled by flipping letters, missing letters,
or/and adding extra letters such as محمود الخطيب as opposed of محمود الخطيب.
- **Splits**: Words that are split into several consecutive rows (i.e., the word
has extra spaces) such as الحزب الوطني > hgpbf hgmxkd
- **Merges**: Words that are combined where they should not be. No space
between the words so the row entry consists of multiple words such as حزب الحريةوالعدالة > حزب الحرية والعدالة
  - Letters flipped,
  - Missing letters
  - Additional letters

3.3 How to Annotate a TYPO case?
It is binary decision, if the word is recognizable by you in context but you believe
it has a typo then you should mark it as a typo. Otherwise leave it as correct,
which is the default setting for this criterion.

➢ For misspelling cases, you will annotate the word per the 10 labels
explained previously.
➢ For Splits cases, you will annotate both parts with the same label per the
10 labels explained previously.
➢ For merge cases, you will annotate the first word in the cell only.
   Exception. If the first word’s label is “others,” then you need to annotate
   the second word.

3.4 Notes
- The following cases are not considered typos:
  o Missing/Extra hamza, inconsistent (with respect to an MSA cognate)
    use of Ta-Marbuta ۰ and Ha ۷, or phonological variants which are used
    with synonymous MSA cognates. For example: مدرس، انترنت
  o Typo is not a label, so you still need to annotate the typo word, as belonging
to one of the possible 10 cases of annotations described above.
    o Do not leave the label column blank in typo cases.
- Digits should be labeled as NE if they are part of NE such as ثورة ٢٥ يناير
4 General Notes
- **DO NOT** tag NE that initiating with # or @ (words initiating with #)
- Multi-word NE overrides all its NE elements. Ex: (شارع (محمد محمود)) = Location. An NE can’t include more than one NE type.
- All words or phrases that belong to the NE categories must be considered NE no matter what language they are written in.
- The words of titles are all NE no matter how long the title is (articles, books, songs, etc.). For example, all words in red are NEs in the following sentence. غداً تنشر مقالتي: "بأي ذنب سجنت؟" تعليقاً على ظلم الفتيات
- Words are judged based on the sentence being annotated only. In the example below, one may label البرنامج NE only because he or she knows or assumes that it refers to the well-known TV show called “The Program”. In this case, the judgment is wrong since nothing in the unit (sentence) proves this assumption. That is, words in a unit must be judged based on the unit itself regardless of the background. According to that, البرنامج in this unit must be labeled as MSA based on the context.
- Words must be annotated based on the meanings that they refer to in the context. Some examples are:
  - الأمن الوطني VS. الأمن الوطني المصري considered NE if it refers to the agency that is responsible to insure security in the country.
  - However, it is NOT NE if it means security in general. Examples:
    - مسؤولية الأمن هي منع العنف وحماية كل الدم المصري (NE)
    - وهي مخالف للدستور والقانون. المجابهة العسكرية فشلت في تحقيق الأمن (Not NE)
  - الكنيسة المصرية/الكنيسة الكاثوليكية considered NE only if it refers to the association that spreads Christian teachings. Examples:
    - دستور يسحب استقلال القضاء، ويؤمّم المحكمة الدستورية. يتعالى على الأزهر ولا يستمع إلى الكنيسة (NE)
    - مع عائلته كل يوم احد يحرص على زيارة الكنيسة (Not NE)
- Nationalities are NOT named entities.
  - الوزير السوري - الراجل مصري are NOT NE
  - المصريون، الشعب المصري are NOT NE
- NE is a label for a unique entity, so words such as أبو الرجل، حفيدو are not NE as they are general words.
- Digits should be labeled as NE if they are part of NE such as ثورة ٢٥ يناير
- NE overrides in cases such as حركة ٦ أبريل , ثورة ٢٥ يناير. That is, the digits must be annotated as NEs even though “other” is what is normally used to annotate digits.
- Names written in English should be tagged NE such as: Google, Facebook.
- The words of titles are all NE no matter how long the title is (articles, books, songs, etc.).
- Words are judged based on the sentence being annotated only.
- Words must be annotated based on the meanings that they refer to in the context.
- Nationalities are NOT named entities.
- NE is a label for a unique entity.

**Example**

<table>
<thead>
<tr>
<th>Word</th>
<th>NE-Type</th>
<th>Typo</th>
</tr>
</thead>
<tbody>
<tr>
<td>يقوم</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>الحزب</td>
<td>B-ORG</td>
<td>Correct</td>
</tr>
<tr>
<td>الوطني</td>
<td>I-ORG</td>
<td>Correct</td>
</tr>
<tr>
<td>باستثمار</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>معلقين</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>بمرتب</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>١٥٠٠</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>جنيه</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>شهرها</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>لتحسين</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>صورته</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>منهم</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>صاحب</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>التعليق</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>رقم</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
<tr>
<td>رقم 25</td>
<td>Not-Ne</td>
<td>Correct</td>
</tr>
</tbody>
</table>