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

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

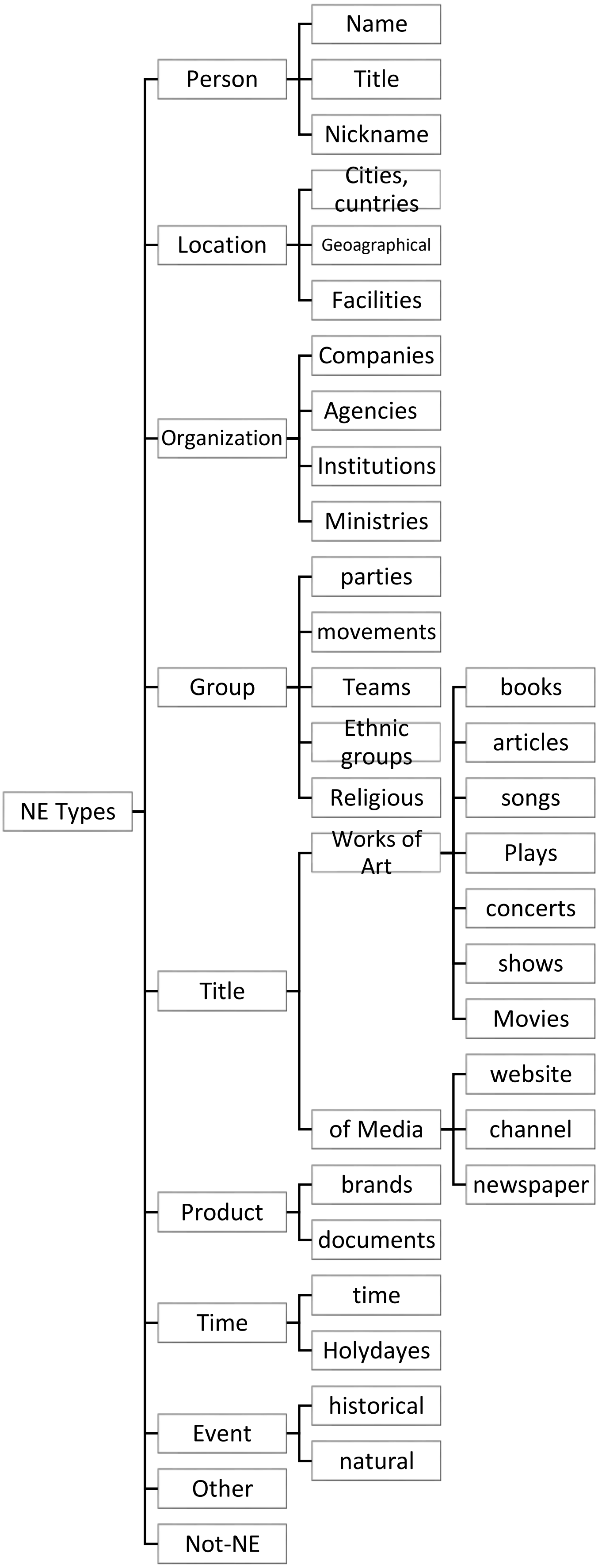
The NE has some different linguistic features, including lexical, semantic, syntactic and orthographic features. The following table summarizes the main linguistic features that could be used to identify the concept of NE.

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

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

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

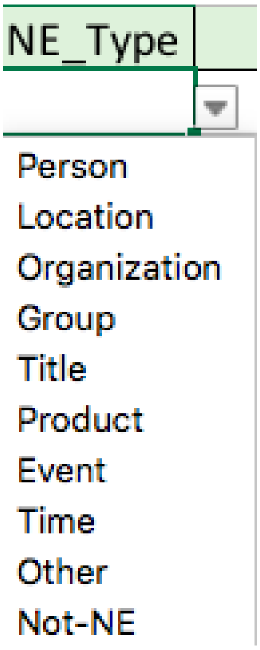
Word	***Annotation***	***Typo?***
الآن	Not-NE	Typo
أحل	Not-NE	Correct
ضيفا	Not-NE	Correct
على	Not-NE	Correct
قناة	B-TITLE	Correct
النيل	I-TITLE	Correct
للأخبار	I-TITLE	Correct

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

#	**NE Type**	**Beginning**	**Inside**	**Outside**
1	Person	B-PER	I-PER	
2	Location	B-LOC	I-LOC	
3	Organization	B-ORG	I-ORG	
4	Production	B-PROD	I-PROD	
5	Title	B-TITLE	I-TITLE	
6	Event	B-EVENT	I-EVENT	
7	Group	B-GROUP	I-GROUP	
8	Time	B-TIME	I-TIME	
9	Other	B-OTHER	I-OTHER	
10	Not_NE			O

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????)

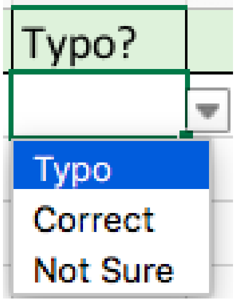


Figure ???

- 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 شوارزنيجر، السيدة عائشة ، المنفلوطي
- Nicknames (امير الاحزان، وردة في البستان، المصرية المحجبة، مصري قرفان)
    - اسماء الشهرة refer to concrete individuals (العندليب الاسمر، خادم الحرمين 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 (الله ، الرحيم، الغفور، الرحمن)
    - References to God (الرب، ربنا، ربكم، اللهم ، الهي)

<u>Notes:</u>

- الهانم، 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 حمزة...
- Pronouns are not NE.

2.2 Location (-LOC)

- Geographical or physical locations like:
    - Cities, provinces, states countries, address, facilities, etc. ايطاليا، نيويورك، الصعيد، ولاية اريزونا ، الخرطوم، المنطقة الشرقية ، المحافظة الغربية
- Mountain ranges, bodies of water, deserts (نياجرا فالز، نهر النيل
    - قناة السويس، هضبة نجد، صحراء سيناء ، البحر المتوسط)
    - نهر النيل ، شلالات نياجرا
- Facilities
    - Buildings: names of schools, airports, masjids, churches, etc. (مبنى الحزب الوطني, المتحف المصري، مطار القاهرة الدولي ، مدرسة التوفيقية الخيرية، كنيسة النبي العظيم يوحنا)

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

<u>Notes:</u>

- الاستاد الميدان ،, 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 الانسان ، 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 إبريل , ثورة ٢٥ يناير
- Sports Teams (منتخب اسبانيا، الاهلي ، الزمالك، المنتخب الالماني)
- Ethnic group (قبيلة العبابدة,
- Religious groups (الإخوان المسلمين، الشيعة....)
- 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 ، اولاد حارتنا)
- Title of songs (اهواك سيرة الحب، الاطلال،)
- Plays and concerts (مهرجان الموسيقي العربية ، العيال كبرت)
- TV shows and movies برنامج القاهرة اليوم ، برنامج ودن واحدة
- Media (websites, channels, newspapers, etc.) (موقع تويتر، جريدة الشروق المصرية)، جريدة الشرق الاوسط

2.6 Product (PROD)

"Product" includes the following subtasks

- Brands
    - Vehicles (bmwبيجو، ميتسوبيشي ، هونداي، لكزس،)
    - Weapons (اف ١٦)
    - Foods brands (ليز، امريكانا)

- Document:
    - LAW & Documents:
    - Named documents made into laws (التعديلات الدستورية، قانون الطوارئ، الدستور، الدستور قانون التظاهر، المبادئ الفوق دستورية

2.7 Time (-TIME)

“Time” includes the following subtypes:

- Months, days of the week, hours, etc.
    - Months (يوليو، رمضان ، شعبان، ايلول)
    - Weekdays (السبت، فرايداي)
- Holidays that happen periodically:
    - Religious holidays (عيد الميلاد المجيد ، ليلة القدر، ليلة المنتصف من شهر شعبان، الايام البيض ، عيد الأضحى، المولد النبوي، ايام عشر ذو الحجة)
    - National events (فورث اوف جولاي ، شم النسيم
    - Social Holydays (عيد الام، عيد الحب، كذبة ابريل ،)
    - (Periodically) sport events (الدوري المصري)

2.8 Event (-EVENT)

A well-known (one time)event that happened or will happened.

“Event” includes the following subtypes:

- Battles and wars (الحرب العالمية الثانية ، غزوة بدر، حرب اكتوبر، فتح مكة)
- Revolutions موقعة الجمل ، ثورة ٢٥ يناير : الثورات
- Hurricanes (اعصار ساندي، اعصار كاترينا)

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 محمور الخطيب as opposed of محمودالخطيب.
- **Splits**: Words that are split into several consecutive rows (i.e., the word has extra spaces) such as الحز ب 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 ة and Ha ه, or phonological variants which are used with synonymous MSA cognates. For example: مدرسه, انباء
- 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. غدا تنشر مقالتي : “بأي ذنب سجنت؟” تعليقا 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” البرنامج. 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. بالتفصيل في خلال الايام القادمة سيتم الكشف عن مصير البرنامج
    - However, the following unit clearly refers to the well-known TV show “The Program” البرنامج. According to that, برنامج البرنامج must be labeled NE. كل سنة ومصر طيبة، انتظرونا في حلقات الموسم الجديد من برنامج البرنامج في سبتمبر ان شاء الله
- Words must be annotated based on the meanings that they refer to in the context. Some examples are:
    - الامن الوطني .VS الامن الوطني

      الامن 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:

      هي منع العنف وحماية كل الدم المصري مسئولية الامن (NE)

      وهي مخالفة للدستور والقانون. المحاكمات العسكرية فشلت في تحقيق الامن (Not NE)
    - الكنيسة المصرية/الكنيسة .VS الكنيسة

      الكنيسة/الكنيسة المصرية considered NE only if it refers to the association that spreads Christian teachings. Examples: دستور يسحب استقلال القضاء، ويؤمم المحكمة الدستورية. يتعالى علي الأزهر ولا يستمع إلى الكنيسة (NE)

      مع عائلته كل يوم احد يحرص على زيارة الكنيسة (Not NE)
- Nationalities are NOT named entities.

  الراجل مصري - الوزير السوري are NOT NE

  الشعب المصري ، المصريون are NOT NE
- NE is a label for a unique entity, so words such as حفيدو ، الرجل أبو 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 إبريل , 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 <u>NOT</u> named entities.
- NE is a label for a unique entity.

Example

Word	***NE-Type***	***Typo***
يقوم	Not-Ne	Correct
الحزب	B-ORG	Correct
الوطني	I-ORG	Correct
باستئجار	Not-Ne	Correct
معلقين	Not-Ne	Correct
بمرتب	Not-Ne	Correct
1500	Not-Ne	Correct
جنيه	Not-Ne	Correct
شهريا	Not-Ne	Correct
لتحسين	Not-Ne	Correct
صورته	Not-Ne	Correct
منهم	Not-Ne	Correct
صاحب	Not-Ne	Correct
التعليق	Not-Ne	Correct
رقم	Not-Ne	Correct
25	Not-Ne	Correct