PerNest Tagging Guidelines

PerNest is the first Persian nested NER corpus. It includes entity types of persons, locations, and organizations. To represent the entities we used JoinedTL encoding, our new proposed CoNLL compatible labelling scheme which has the capability of tagging inner named entities. In JoinedTL, only the first token of each entity is tagged by its type and span, and the remaining tokens are left untagged (are tagged with a dash symbol (-)). In the case that two or more entities start from the same token, we join the tags by a vertical line symbol (|). The details of JoinedTL are presented in “Joined Type Length Encoding for Nested Named Entity Recognition”.

Applied guidelines to make PerNest:

1. Only proper nouns of persons, locations, and organizations are considered.
    a. Examples of proper nouns:
    مسعود الماسی، کوه دماوند، شیراز، مسجد گوهرشاد، سازمان بهداشت جهانی
    b. Examples of non-proper nouns which cannot be considered as a named entity:
    آقای رئیس جمهور، این دانشگاه، آن سازمان

2. Ministries, universities, governments, empires, races and ethnicities are considered as organizations. Examples:
    وزارت نفت، دانشگاه امیرکبیر، جمهوری اسلامی ایران، ساسانیان، ایرانیان، اعراب

3. Streets, squares, seas, oceans, mountains, jungles, parks, rivers, mosques, tombs, museums, cinemas, cities, towns, villages, countries, provinces and such other things are considered as locations.

4. Common nouns that determine a group of entities are **not** considered as named entities. So, we don’t tag these examples:
    بیمارستانهای آموزشی، شهرهای شمالی، مدارس غیرانتفاعی

5. We don’t tag entities by their meaning in the context, but tag them based on their common meaning. So, Iran is always taken as a location even in this sentence: “Iran and China signed an economic agreement.”

6. Honorifics that usually come before a person's name are considered as part of that entity. So, these examples are considered as entity of person type as a whole:

Example 1		Example 2		Example 3		Example 4		Example 5	
آقای	Per-2	دکتر	Per-2	رضا	Per-2	حجت	Per-3	احمد	Per-3
محمودی	-	شریعتی	-	خان	-	الاسلام	-	شاه	-
						کاشانی	-	قاجار	-

7. The words before or after entities that depict their type are considered as part of the entity. For example:

Example 1		Example 2		Example 3		Example 4		Example 5	
شهر تهران	Loc-2 -	استان تهران	Loc-2 -	رود نیل	Loc-2 -	رشته کوه البرز	Loc-3 - -	سلسله هخامنشیان	Org-2 -

8. Inner entities are also tagged in this corpus. Examples:

Example 1		Example 2		Example 3		Example 4		Example 5	
مسجد جامع اصفهان	Loc-3 - Loc-1	حرم حضرت عبدالعظیم	Loc-3 Per-2 -	جمهوری اسلامی ایران	Org-3 - Loc-1	اداره مخابرات استان تهران	Org-4 - Loc-2 -	موزه لوور پاریس	Loc-3 - Loc-1

9. If two or more entities of the same type come with a plural determiner, the whole phrase starting from the determiner to the last entity is considered as an outer entity, and all of the inner entity types that the determiner refers to is taken as inner entities. Example:

شهرهای	Loc-6
تهران	Loc-1
،	-
اصفهان	Loc-1
و	-
شیراز	Loc-1