**Wojood - Nested Arabic Named Entity Corpus**

Copyright © 2024 Birzeit University

𝐎𝐧𝐥𝐢𝐧𝐞: <https://sina.birzeit.edu/wojood/>

𝐃𝐨𝐰𝐧𝐥𝐨𝐚𝐝, 𝐆𝐮𝐢𝐝𝐞𝐥𝐢𝐧𝐞𝐬, 𝐚𝐧𝐝 𝐒𝐭𝐚𝐭𝐢𝐬𝐭𝐢𝐜𝐬: <https://sina.birzeit.edu/wojood/>

**Contact:** SinaLab for Computational Linguistics and Artificial Intelligence

Prof. Mustafa Jarrar ([sina@birzeit.edu](mailto:sina@birzeit.edu))

𝐕𝐞𝐫𝐬𝐢𝐨𝐧

▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬

Version 1.1 (Last modification: 20-7-2023)

𝐋𝐢𝐜𝐞𝐧𝐬𝐞:

▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬

See the attached (license.pdf) file

𝐃𝐚𝐭𝐚 𝐝𝐞𝐬𝐜𝐫𝐢𝐩𝐭𝐢𝐨𝐧:

▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬▬

There are two versions of the Wojood dataset (Nested and Flat), and each version is available in CoNLL format in TXT and CSV files. The data is available in three splits {train, val, and test}, 70%, %10, 20% respectively. See the article [[1](https://www.jarrar.info/publications/JKG22.pdf)] for the splitting details and statistics.

**1. Wojood Nested**

Wojood Nested means that the NER tags can overlap each other. For example, “Birzeit University” is tagged as ORG and “Birzeit” is tagged as GPE. We used 21 entity types (tagset). See the article [[1](https://www.jarrar.info/publications/JKG22.pdf)] for the tagset and the annotation guidelines.

**2. Wojood Flat**

Wojood flat means that the NER tags are not nested. It is automatically derived from the Wojood Nested. We used to topmost tag. For example, in the nested files, “Birzeit University” is tagged as ORG and “Birzeit” is tagged as GPE; however, it is only tagged as ORG in the flat version because it is topmost (i.e., longer span),

**CoNLL format:** means that each token and its tag(s) are stored in a line, and sentences are separated by an empty line.

**CSV format:** Each txt file consists of the following columns:

▸global\_row\_id: a unique id for each token in the whole corpus.

▸global\_sentence\_id: a unique id for each sentence in the whole corpus.

▸Sub\_corpus: the sub domain/corpus. There are 11 sub corpora/domains in Wojood: (1) Terrorism, (2) Foreign Policy (3) Security, (4) Law and Elections, (5) Migration, (6) Health Policy, (7) PCMA - Finance and Banking, (8) QUORA - ICT, (9) Awraq - History and Culture, (10) Curras - Palestinian Dialect, and (11) Lebanese - Lebanese dialect.

▸subcorpus\_sentence\_id: a unique id for each sentence in the within the sub\_corpus.

▸token\_position: the position of the token in within the sentence.

▸token: the token itself

▸tags: the tag(s). Nested tags are separated by space, and they are ordered as (longest tag first)

𝐂𝐢𝐭𝐚𝐭𝐢𝐨𝐧:

▬▬▬▬▬▬▬▬▬▬▬▬

Wojood was developed based on the following articles, which should be acknowledged and cited properly each time Wojood is used:

[1] Mustafa Jarrar, Mohammed Khalilia, Sana Ghanem: [Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT](https://www.jarrar.info/publications/JKG22.pdf). In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France. 2022