The 2022 International Conference on Language Resources and Evaluation (LREC 2022)
Marseille, France
20-25/6/2022

Wojood
Nested Arabic Named Entity Corpus and Recognition using BERT

Mustafa Jarrar **Mohammed Khalilia** **Sana Ghanem**

Birzeit University
Palestine

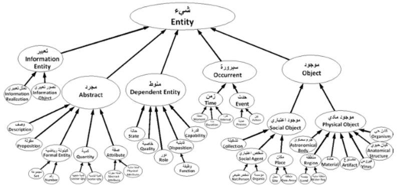
Lexical Resources at Birzeit University

Lexicographic Database

150 lexicons

Largest Arabic lexicographic database

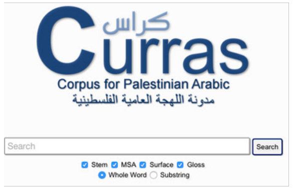
Arabic Ontology/Wordnet



Formal Arabic Wordnet

with ontologically clean content

Dialect Corpora



كراس

Curras

Corpus for Palestinian Arabic

مدونة اللهجة العامية الفلسطينية

Search | Search

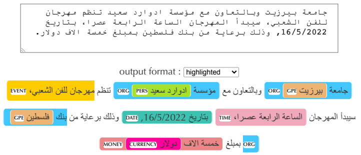
☑ Stem ☑ MSA ☑ Surface ☑ Gloss

● Whole Word ○ Substring

Annotated corpora

each word is annotated with many morph features

NLP library



جامعة بيرزيت وبالتعاون مع مؤسسة ادوارد سعيد تنظم مهرجان
للفن الشعبي، سيبدأ المهرجان الساعة الرابعة عصرا، بتاريخ
16/5/2022, وذلك برعاية من بنك فلسطين بمبلغ خمسة الاف دولار.

output format : highlighted

جامعة بيرزيت GPE ORG وبالتعاون مع مؤسسة ادوارد سعيد PERS ORG تنظم مهرجان للفن الشعبي، EVENT

سيبدأ المهرجان الساعة الرابعة عصرا، TIME بتاريخ 16/5/2022, DATE وذلك برعاية من بنك فلسطين GPE

ORG بمبلغ خمسة الاف دولار CURRENCY MONEY

APIs

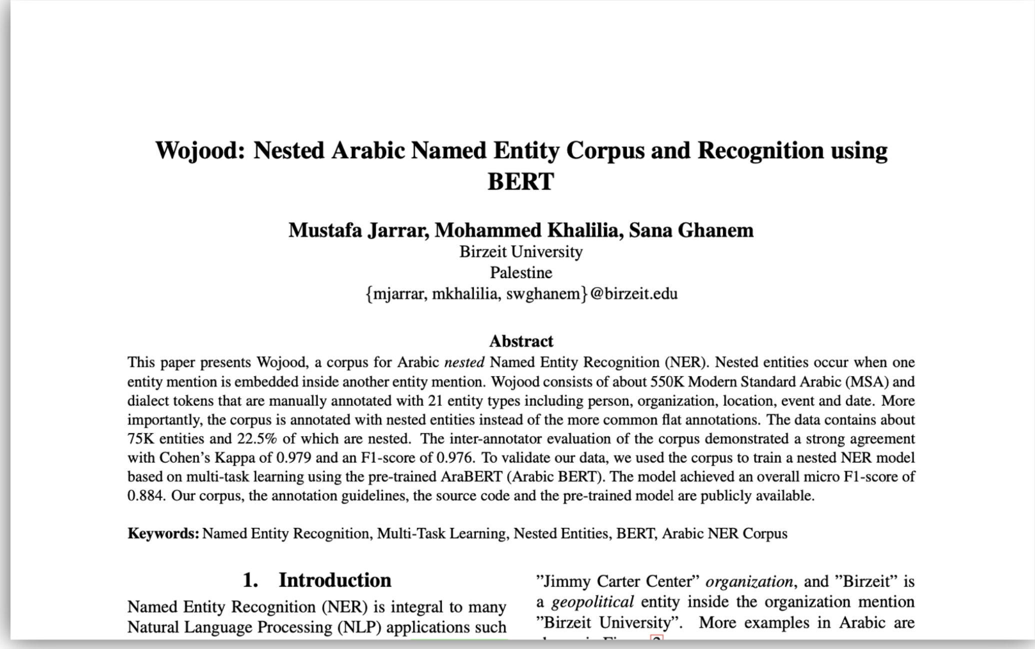
Linguistic Data, synonyms, tools, Nested named-entities, intents, …

WSD 84%

NER 88.4%

Big Linguistic Data Graph

https://ontology.birzeit.edu



Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT

Mustafa Jarrar, Mohammed Khalilia, Sana Ghanem
Birzeit University
Palestine
{mjarrar, mkhalilia, swghanem}@birzeit.edu

Abstract

This paper presents Wojood, a corpus for Arabic *nested* Named Entity Recognition (NER). Nested entities occur when one entity mention is embedded inside another entity mention. Wojood consists of about 550K Modern Standard Arabic (MSA) and dialect tokens that are manually annotated with 21 entity types including person, organization, location, event and date. More importantly, the corpus is annotated with nested entities instead of the more common flat annotations. The data contains about 75K entities and 22.5% of which are nested. The inter-annotator evaluation of the corpus demonstrated a strong agreement with Cohen's Kappa of 0.979 and an F1-score of 0.976. To validate our data, we used the corpus to train a nested NER model based on multi-task learning using the pre-trained AraBERT (Arabic BERT). The model achieved an overall micro F1-score of 0.884. Our corpus, the annotation guidelines, the source code and the pre-trained model are publicly available.

Keywords: Named Entity Recognition, Multi-Task Learning, Nested Entities, BERT, Arabic NER Corpus

1. Introduction

Named Entity Recognition (NER) is integral to many Natural Language Processing (NLP) applications such

”Jimmy Carter Center” *organization*, and ”Birzeit” is a *geopolitical* entity inside the organization mention ”Birzeit University”. More examples in Arabic are

Jarrar, M., Khalilia, M., Ghanem, S. (2022). Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France.

The problem

Flat NER

Sami	works	at	Jimmy	Carter	Center
B-PERS	O	O	B-ORG	I-ORG	I-ORG

✓mature

Nested NER

Sami	works	at	Jimmy	Carter	Center
B-PERS	O	O	B-PERS	I-PERS	
			B-ORG	I-ORG	I-ORG

!

Challenging: to build nested NER corpus and to train tags in BERT

Existing Arabic NER corpora are:

- Flat
- Small in size
- Limited in the number of NER classes
- Mostly limited to MSA

Our Contributions

❖ Wojood NER Corpus

Corpus	Nested?	size (tokens)	No. of entities	Entity classes	Which Arabic	Domain
Ontonotes5	No	300k	28k	18	MSA	News
ANERCorp	No	150k	11k	4	MSA	News
Canercorpus	No	258k	72k	14	Classic	Religion
AQMAR	No	74k	-	open	MSA	4 domains
Wojood Corpus	YES	550K	75K	21	MSA & Dialect	**Multi domains** Media, History, Culture, Health, Finance, ICT, Law, Elections, Politics, Migration, Terrorism

❖ Wojood NER BERT

- Multi-task learning (nested entities)
- 88.4% F1-score

Corpus Collection

Source - Topics	Sentences	Tokens
Web Articles[1] (MSA) Health, Finance, ICT, Law, Elections, Politics, Migration and Terrorism	9,053	258,102
Archive[2] (MSA) History and Culture	12,271	227,020
Social Media[3] (Dialect) General topics	5,653	65,342
Total	26,977	550,464

[1] un.org, hrw.org, msf.org, who.org, mipa.institute, elections.ps, sa.usembassy.gov, diplo- matie.ma, quora.com ….

[2] Awraq, Birzeit University Digital Palestinian Archive

[3] Palestinian and Lebanese dialect corpora

Annotation Guidelines

21 entity classes

PERS	People names	فيروز، عادل إمام، ابن احمد، الملك عبدالله، النبي محمد
NORP	Group of people	العرب، المسحيين، سكان القدس، وزراء الخارجية العرب
OCC	Occupation or professional title	رئيس جامعة بيرزيت، مدير بنك فلسطين، قائد الجيش
ORG	Legal/social body	بنك القاهر، ريال مدريد، داعش، الجيش المصري،
GPE	Geopolitical: country, city, state	ليبا، مدينة القدس، الجمهورية اللبنانية، روسيا الاتحادية
LOC	Geographical location (non-GPE)	البحر الميت، قناة السويس، آسيا، الوطن العربي
FAC	Places: landmark, road, building..	مطار صنعاء، سجن ابو غريب، المسجد الأقصى
EVENT	Events of general interest	حرب 1973، القمة العربية 2005، عيد الفطر ، يوم الأرض
DATE	Specific/relative date (>day)	13يونيو، 2019-2020، الفترة العثمانية
TIME	Specific/relative time (<day)	الساعة ١٢، من الساعة 5 حتى 7 مساء، خلال ساعتين
LANGUAGE	Human language or dialect	اللغة العربية، الفصحى، الدارجة المغربية، اللغة الفرنسية
WEBSITE	Website or specific URL	موقع فيسبوك، يوتيوبschema.org ،
LAW	Geographical location (non-GPE)	قانون الاستثمار ، المادة 114 من قانون العقوبات 2005
PRODUCT	Vehicle, weapon, food, …	مرسيدس سي١٨٠، ايفون ١٣، دبابة مركابا، تروفين
CARDINAL	Numerals in digits/words	، 1.5 ، 30 ، 150صفر، اثنان ، أربعة وعشرون ، مليون
ORDINAL	does not refer to a quantity	٣ كيلومتر، مئة قدم ، 3 طن ، 50 غرام ، 25 سم مربع
PERCENT	Word/symbol refers to a percent	5بالمئة ، 10% ، 9 من كل الف
QUANTITY	Value measured by units	٣ كيلومتر، مئة قدم ، 3 طن ، 50 غرام ، 25 سم مربع
UNIT	Name/symbol of a unit	ميل ، كيلو ، كيلومتر ، إنش ، كيلوغرام ، هكتار ، مل
MONEY	Monetary quantity, incl. currency	مئة وخمسون درهم اماراتي ، اثنان وثلاثون يورو ، 8 دولار
CURR	Name/symbol of currency	دولار ، جنيه مصري ، دينار ، فرنك ، ريال سعودي€ ,

Example

منح مدير بنك القاهرة مبلغ مليون جنية لائتلاف
العاملين بجامعة القاهرة لدعم ميزانية ٢٠٢٢

The manager of the Cairo Bank awarded one million pound to the Employees Union at Cairo University to support the 2022 budget

Token	Level 1	Level 2	Level 3
منح	O		
مدير	B-OCC		
بنك	I-OCC	B-ORG	
القاهرة	I-OCC	I-ORG	B-GPE
مبلغ	O		
مليون	B-MONEY		
جنية	I-MONEY	B-CURR	
لائتلاف	B-ORG		
العاملين	I-ORG		
بجامعة	I-ORG	B-ORG	
القاهرة	I-ORG	I-ORG	B-GPE
لدعم	O		
ميزانية	O		
٢٠٢٢	B-DATE		

Annotation Process

- 2 experts + 12 annotators (intensive training).
- Annotations were performed using Google Sheets.
- Took over 8 months to complete.

- Phases:

  Phase 1: each annotator was given ~46K tokens

  Phase 2: experts reviewed all annotations manually

  Phase 3: used a trained model to predict tags then reviewed differences (two iterations)

Wojood NER Corpus

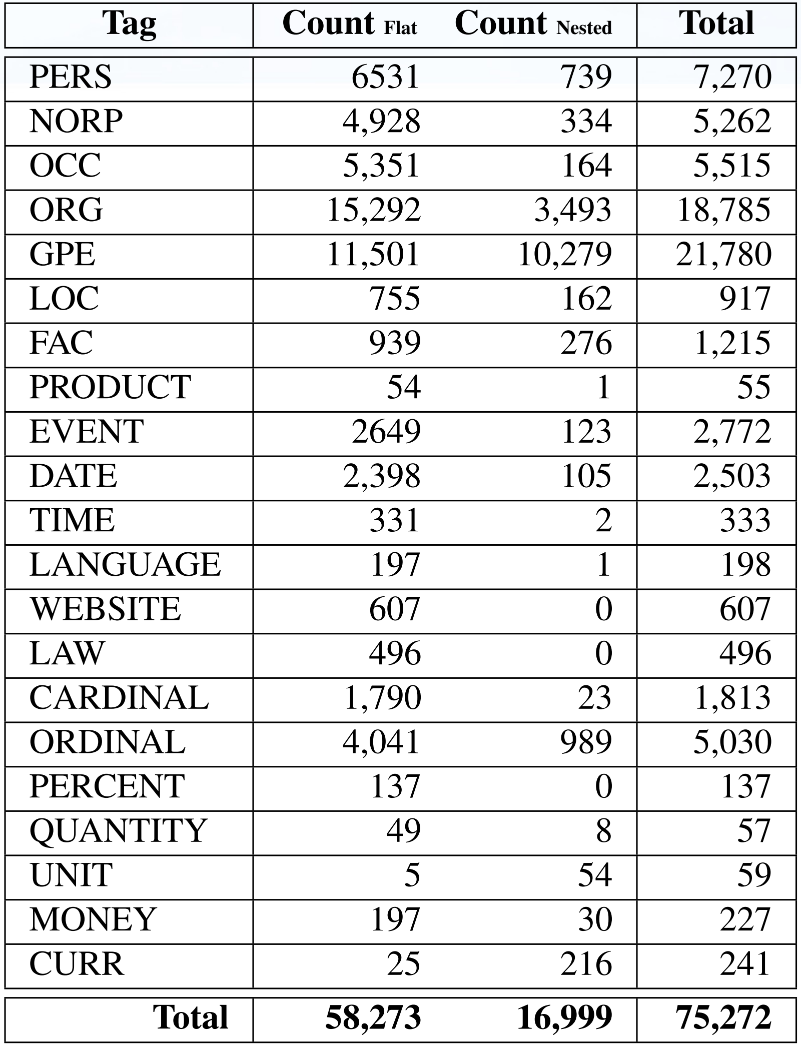
Statistics

Annotated Corpus

Counts of the flat, nested, and total of each entity type

22.5% are nested within other entity mentions

576 of the nested entities are of the same type (training challenge!)



Tag	Count Flat	Count Nested	Total
PERS	6531	739	7,270
NORP	4,928	334	5,262
OCC	5,351	164	5,515
ORG	15,292	3,493	18,785
GPE	11,501	10,279	21,780
LOC	755	162	917
FAC	939	276	1,215
PRODUCT	54	1	55
EVENT	2649	123	2,772
DATE	2,398	105	2,503
TIME	331	2	333
LANGUAGE	197	1	198
WEBSITE	607	0	607
LAW	496	0	496
CARDINAL	1,790	23	1,813
ORDINAL	4,041	989	5,030
PERCENT	137	0	137
QUANTITY	49	8	57
UNIT	5	54	59
MONEY	197	30	227
CURR	25	216	241
Total	**58,273**	**16,999**	**75,272**

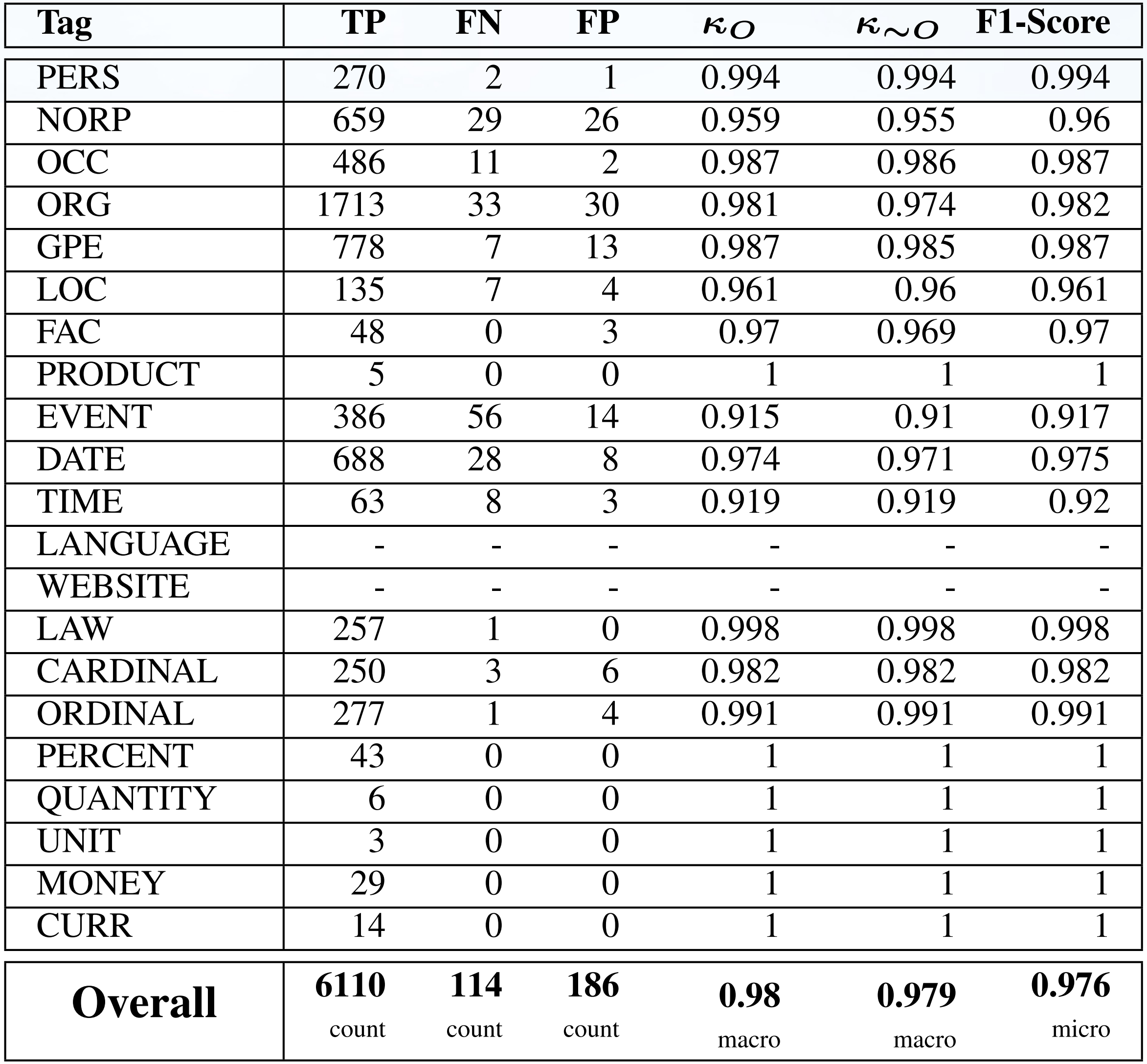
Inter-annotator agreement

IAA Evaluation

24K tokes (4.3% corpus) selected randomly

2k to each annotator A_1–A_2, A_2–A_3,…, A_{12}–A_1

See the annotation challenges in the article



Tag	TP	FN	FP	κ_O	$\kappa_{\sim O}$	F1-Score
PERS	270	2	1	0.994	0.994	0.994
NORP	659	29	26	0.959	0.955	0.96
OCC	486	11	2	0.987	0.986	0.987
ORG	1713	33	30	0.981	0.974	0.982
GPE	778	7	13	0.987	0.985	0.987
LOC	135	7	4	0.961	0.96	0.961
FAC	48	0	3	0.97	0.969	0.97
PRODUCT	5	0	0	1	1	1
EVENT	386	56	14	0.915	0.91	0.917
DATE	688	28	8	0.974	0.971	0.975
TIME	63	8	3	0.919	0.919	0.92
LANGUAGE	-	-	-	-	-	-
WEBSITE	-	-	-	-	-	-
LAW	257	1	0	0.998	0.998	0.998
CARDINAL	250	3	6	0.982	0.982	0.982
ORDINAL	277	1	4	0.991	0.991	0.991
PERCENT	43	0	0	1	1	1
QUANTITY	6	0	0	1	1	1
UNIT	3	0	0	1	1	1
MONEY	29	0	0	1	1	1
CURR	14	0	0	1	1	1
Overall	**6110** count	**114** count	**186** count	**0.98** macro	**0.979** macro	**0.976** micro

Named Entity Recognition (NER)

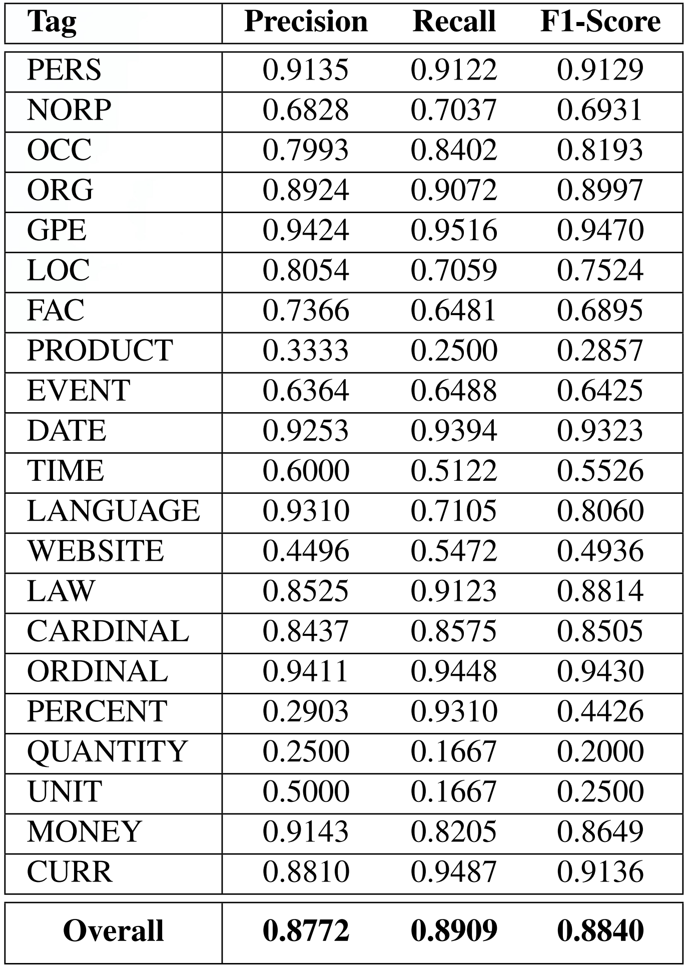
Multi-Task Learning for Nested NER

- Used AraBBERT-V2 pre-trained model (Antoun et al., 2020)
- The model consists of the sequence encoder and multiple classifiers, one for each entity type (21 classification layers)

Model Training

- Training (385K tokens, 70%), Validation (55K tokens, 10%) and Test (110K tokens, 20%).
- Learning Rate $\eta = 1e^{-3}$
- Batch size of 32, maximum of 20 epochs
- Converged around epoch nine

Nested NER Results

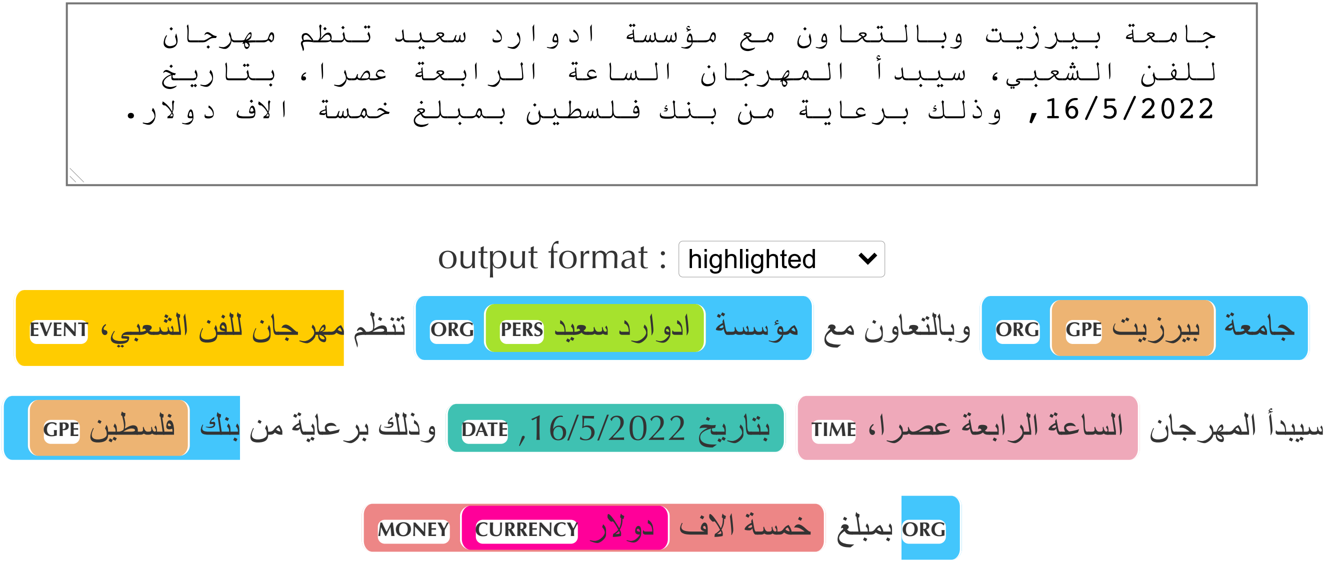


Tag	Precision	Recall	F1-Score
PERS	0.9135	0.9122	0.9129
NORP	0.6828	0.7037	0.6931
OCC	0.7993	0.8402	0.8193
ORG	0.8924	0.9072	0.8997
GPE	0.9424	0.9516	0.9470
LOC	0.8054	0.7059	0.7524
FAC	0.7366	0.6481	0.6895
PRODUCT	0.3333	0.2500	0.2857
EVENT	0.6364	0.6488	0.6425
DATE	0.9253	0.9394	0.9323
TIME	0.6000	0.5122	0.5526
LANGUAGE	0.9310	0.7105	0.8060
WEBSITE	0.4496	0.5472	0.4936
LAW	0.8525	0.9123	0.8814
CARDINAL	0.8437	0.8575	0.8505
ORDINAL	0.9411	0.9448	0.9430
PERCENT	0.2903	0.9310	0.4426
QUANTITY	0.2500	0.1667	0.2000
UNIT	0.5000	0.1667	0.2500
MONEY	0.9143	0.8205	0.8649
CURR	0.8810	0.9487	0.9136
Overall	**0.8772**	**0.8909**	**0.8840**

Entities of the same time are not supported yet

Downloads and Demo

https://ontology.birzeit.edu/wojood



جامعة بيرزيت وبالتعاون مع مؤسسة ادوارد سعيد تنظم مهرجان
للفن الشعبي، سيبدأ المهرجان الساعة الرابعة عصرا، بتاريخ
16/5/2022, وذلك برعاية من بنك فلسطين بمبلغ خمسة الاف دولار.

output format : highlighted

جامعة بيرزيت GPE ORG وبالتعاون مع مؤسسة ادوارد سعيد PERS ORG تنظم مهرجان للفن الشعبي، EVENT

سيبدأ المهرجان الساعة الرابعة عصرا، TIME بتاريخ 16/5/2022, DATE وذلك برعاية من بنك فلسطين GPE

ORG بمبلغ خمسة الاف دولار CURRENCY MONEY

Output formats: JSON IOB2, JSON entities, XML, highlighted

Summary

❖ Wojood NER Corpus

- Nested named entities
- 550K tokens (large)
- 75K named entities in the corpus
- 21 classes of entities
- MSA & dialect
- Multi-domain
- IAA: 97.9 Kappa, 97.6 F1

❖ Wojood NER BERT

- Multi-task learning (nested entities)
- 88.4% F1-score

Public (data, code, demo)

https://ontology.birzeit.edu/wojood

Thank You

References