Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT

Mustafa Jarrar, Mohammed Khalilia, Sana Ghanem
Birzeit University
Palestine

Contributions

❖ Arabic NER Corpus (Wojood)

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

❖ BERT Model

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

Used AraBBERT-V2 pre-trained model (Antoun et al., 2020)

Existing Arabic NER Corpora

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**

Corpus Annotation

Annotation Process
- 2 experts + 12 annotators (intensive training)
- Done using Google Sheets
- Over 8 months

Annotation Phases:
Phase 1: each annotator was given ~46K tokens
Phase 2: experts reviewed all annotations
Phase 3: used a trained model to predict tags then reviewed differences (two iterations)

Example

Bank of Cairo Manager awarded one million pound to the Employees Union at Cairo University to support the 2021 budget

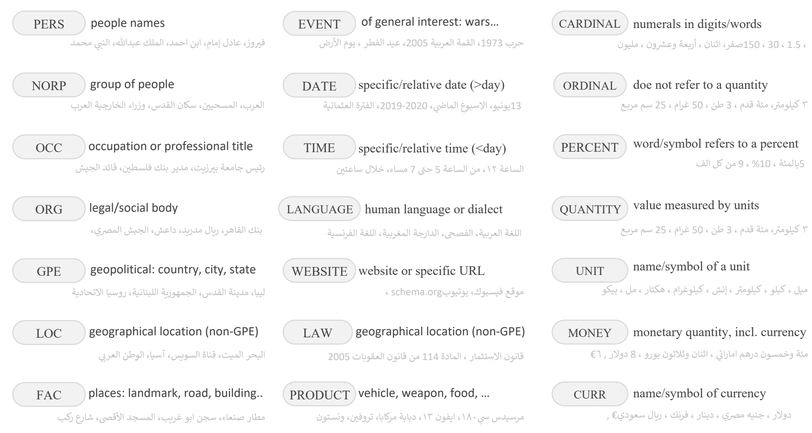
منح مدير بنك القاهرة مبلغ مليون جنية لائتلاف العاملين بجامعة القاهرة لدعم ميزانية ٢٠٢٢

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		

Raw Corpus

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

Annotation Guidelines



Named Entity Recognition

Dataset

- 22.5% are nested within other entity mentions
- 576 of the nested entities are of the same type

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**

NER Results

- Used AraBERT2 pre-trained Model
- Multiple classifiers, one for each entity type (21 classification layers)

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**

Downloads and Demo

https://ontology.birzeit.edu/wojood

Public (data, code, demo)

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

BIRZEIT UNIVERSITY