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
- Annotated corpora
  - each word is annotated with many morph features

NLP library
- 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 as Jimmy Carter Center organization, and "Birzeit" is a geopolitical entity inside the organization mention "Birzeit University". More examples in Arabic are...
The problem

Flat NER

Existing Arabic NER corpora are:

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

Nested NER

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

**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** (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 |

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
### 21 entity classes

| Entity Class | Description |
|--------------|-------------|
| **PERS** | People names |
| **NORP** | Group of people |
| **OCC** | Occupation or professional title |
| **ORG** | Legal/social body |
| **GPE** | Geopolitical: country, city, state |
| **LOC** | Geographical location (non-GPE) |
| **FAC** | Places: landmark, road, building.. |
| **EVENT** | Events of general interest |
| **DATE** | Specific/relative date (>day) |
| **TIME** | Specific/relative time (<day) |
| **LANGUAGE** | Human language or dialect |
| **WEBSITE** | Website or specific URL |
| **CARDINAL** | Numerals in digits/words |
| **ORDINAL** | Does not refer to a quantity |
| **PERCENT** | Word/symbol refers to a percent |
| **QUANTITY** | Value measured by units |
| **UNIT** | Name/symbol of a unit |
| **MONEY** | Monetary quantity, incl. currency |

**Examples:**
- **PERS:** فوزوز، عادل إمام، ابن أحمد، الملك عبد الله، النبي محمد
- **NORP:** العرب، المسيحيين، سكان القدس، وزراء الخارجية العرب
- **OCC:** رئيس جامعة بيرزيت، مدير بنك فلسطين، قائد الجيش
- **ORG:** بنك القاهرة، ريال مدريد، داعش، الجيش المصري
- **GPE:** ليبيا، مدينة القدس، الجمهورية اللبنانية، روسيا الاتحادية
- **LOC:** البحر الميت، قناة السويس، آسيا، الوطن العربي
- **FAC:** مطار صنعاء، سجن أبو غريب، المسجد الأقصى
- **EVENT:** حرب 1973، القمة العربية 2005، عبد الفطر، يوم الأرض
- **DATE:** 13 يونيو، 2020-2019، الفترة العثمانية
- **TIME:** الساعة 12، من الساعة 5 حتى 7 مساء، خلال ساعات
- **LANGUAGE:** اللغة العربية، الفصيحي، الادارة الموازية، اللغة الفرنسية
- **WEBSITE:** موقع فيسبوك، جوجل، باحث
- **CARDINAL:** 1.5، 30، 150 صفر، اثنان، أربعة، وعشرون، مليون
- **ORDINAL:** 32 كيلومتر، منتهى 3، حفر، 50 غرام، 25 سم مربع
- **PERCENT:** 5 بالمائة، 10%، 9 من كل الف
- **QUANTITY:** 3 كيلومتر، منتهى 3، حفر، 50 غرام، 25 سم مربع
- **UNIT:** ميل، كيلو، كيلوغرام، إنش، كيلوغرام، هكتار، مل
- **MONEY:** مئة وخمسون درهم اماراتي، اثنان وثمانون يورو، 8 دولار
The manager of the Cairo Bank awarded one million pound to the Employees Union at Cairo University to support the 2022 budget.
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)
The NER experts, was conducted in three phases over eight months: Phase I, Phase II, and Phase III.

12 students who achieved above 90% in the quiz were selected to annotate, which we designed to be hard. The annotation process, managed by two NER experts, was conducted in three phases over eight months: Phase I, Phase II, and Phase III.

The experts reviewed all annotations in the corpus in Phase I. The experts then carefully re-read all annotations with the annotations suggested by the model and asked the annotators to verify the mistakes and provide feedback on the annotations. For tokens that are annotated with multiple tags, we annotate the corpus. We compared the manual annotations with the annotations suggested by the model and asked the annotators to verify the mistakes and provide feedback on the annotations.

The data was annotated with 21 entity types, see Table 3 for a short description and examples for each entity type.

Counts of the flat, nested, and total of each entity type are as follows:

- 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!).

The data was annotated with 21 entity types, see Table 3 for a short description and examples for each entity type.

Counts of the flat, nested, and total of each entity type are as follows:

- 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 |
## IAA Evaluation

24K tokens (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   | $\kappa_O$ | $\kappa_{\sim O}$ | 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 |
|----------|-------------|------------|-----------|------------|-------------|-------------|

The results in Table 6 illustrate a very strong agreement. We believe that the reason of why the overall inter-annotator agreement is so high is due to entity boundaries. Although the and DATE have the most disagreements. We found that the counts show that EVENT, NORP, TIME, LOC, FAC account the count (i.e., weight) of each tag. The overall F1-score of all annotations is the micro-averaged agreement on the "O" tag. Calculating the F1-score for a certain tag by:

$$F1 = \frac{2TP}{2TP + FP + FN}$$

where $TP$ is the number of tokens correctly annotated by both annotators, $FP$ the number of tokens annotated by the first annotator, but not by the second, and $FN$ the number of tokens annotated by the second, but not by the first. Note that $FP$ counts as false positives, while $FN$ counts as false negatives if the first annotator disagrees with the second, we count those as false positives. Otherwise, if the first annotator agrees with the second, we count them as true positives.

To calculate the F1-score for a specific tag such as PERS by both annotators. Otherwise, if the first annotator disagrees with the second, we count those as false negatives.

Calculating F1-score agreement on the "O" label. The chance (i.e., weight) defined as the agreement between annotators of all annotations. The average of all such, the second, however, we did not include the annotations on the "O" label. Agreements are counted as pair-wise matches at the token level; thus, if a token is annotated, e.g., as ORG by one annotator and as O by another, we counted only the tokens labeled with ORG as true positives.

The inter-annotator agreement on the "O" label, however, we did not include the annotations on the "O" label. Agreements are counted as pair-wise matches at the token level; thus, if a token is annotated, e.g., as ORG by one annotator and as O by another, we counted only the tokens labeled with ORG as true positives.

The model did not. Some events might be important to annotate in the USA, while others ignored the event completely. In addendum to the inter-annotator agreement, we developed a model for nested NER to further validated the data and showcase its accuracy.
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

James Birzeit University will cooperate with Edoardo Saeid to organize an exhibition of popular art on 5/16/2022, which will be sponsored by Bank Palestine for a sum of five thousand dollars.

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