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

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

Existing Arabic NER Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Nested?</th>
<th>Size (tokens)</th>
<th>No. of entities</th>
<th>Entity classes</th>
<th>Arabic</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArabicCore</td>
<td>No</td>
<td>350K</td>
<td>76</td>
<td>10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Wojood</td>
<td>Yes</td>
<td>550K</td>
<td>75K</td>
<td>21</td>
<td>N/A</td>
<td>Multi-domain</td>
</tr>
</tbody>
</table>

Corpus Annotation

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

Raw Corpus

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

Total                        26,977     550,464

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

Named Entity Recognition

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

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

Downloads and Demo
https://ontology.birzeit.edu/wojood
Public (data, code, demo)