AIM: To recognize the entity of Arabic words

Challenges in Arabic text entity recognition

- The Arabic language allows for multiple spellings of the same word, all of which refer to the same meaning. This creates an ambiguity where multiple spellings correspond to a single word.
- The use of capitalization is an important orthographic feature that can help NER systems identify types of entities, such as proper names and abbreviations, with greater accuracy.
- Arabic language has a complex morphology due to its highly agglutinative nature, where words are formed by combining prefixes, lemma, and suffixes in various ways, resulting in multiple combinations.
- Diacritics, also known as short vowels, play a crucial role in the pronunciation and disambiguation of Arabic words. However, Arabic texts are usually written without diacritics.
- The absence of linguistic tools: Building a solid NER system requires sufficient information resources, especially annotated corpora, and top-notch lexicons/gazetteers. However, because to the lack of annotations, there are fewer Arabic linguistic resources available for Arabic NER than there are for varieties like English and Chinese, which have extensive labelled corpora.

What is unique in this paper?

A lot of research has been done in this field but when we talked about accuracy, No one has 100%. This paper is also haven't 100 % accuracy, but it has much better performance as compared to other method.

This article introduces two methods that were devised for performing Arabic named entity recognition (ANER). The first method involves using a conventional machine learning technique that utilizes conditional random fields (CRF) trained on a predetermined set of syntactic and morphological features. In contrast, the second method employs a deep neural network known as a Bidirectional Long Short-Term Memory with CRF (BI-LSTM-CRF) model.

Application of Named Entity Recognition (NER)

Affective news analysis, paraphrase identification, machine translation, entity linking, and disambiguation are just a few of the many NLP and semantic computing applications that significantly rely on NER.

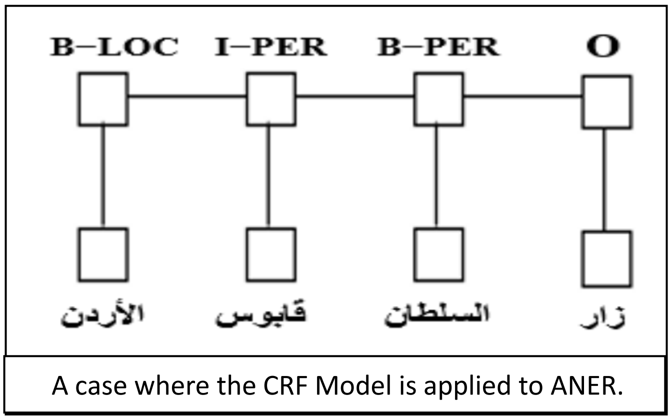
CRF with predefined features

The Conditional Random Fields (CRF) is a graphical model without a specific direction, where each word from the input sentence is represented with a corresponding tag in the output sentence. It predicts the entity of text by using neighbouring words which has predefine tag.

In this paper the CRF model is trained to predict three types of words entity/tag: Person (PER), Location (LOC), and organisation (ORG). Those words which aren't fall in above categories fall in other object(O).

IOB encoding is a way to indicate Named Entity Recognition (NER) tags in a sequence. In this encoding, B-tag denotes the start of an entity, I-tag indicates any intermediate entity, and O-tag represents any other non-entity in the sequence.

There was a total of 21 finely detailed entities utilized to depict the three primary categories of entities namely, PER, LOC, and ORG. These entities were employed in 36207 instances. See the below TABLE for the detailed representati on of entities.



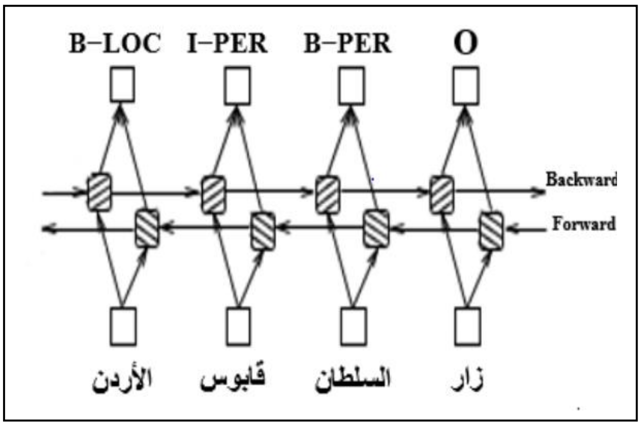
A case where the CRF Model is applied to ANER.

Main class	Sub-class
Person (PER)	Scientist (2178) Engineer (259) Religious-PER (2710) Businessman (864) Artist (2514) Police (1324) Group (3565) Athlete (1293) Politician (6453)
Organization (ORG)	Media (1194) Government (1785) Sports (1134) Medical-Science (185) Non-Governmental (2014) Commercial (2182) Entertainment (160) Educational (1250) Religious-ORG (194)
Location (LOC)	Waterbody (2928) Celestial (1067) Land-Region-Natural (954)
Total (3)	21 (36207)

Bi-LSTM-CRF Model

The Bi-LSTM model has the ability to capture information from both past and future input features within a specific time period, using a window approach. This is achieved by leveraging the forward pass state to learn from past input features and the backward pass state to learn from future input features.

To train the model, the back-propagation through time (BPTT) algorithm was employed. The process of tagging an Arabic sequence for ANER using a Bi-LSTM model is depicted in Figure. In this approach, every word within the sentence is assigned a label indicating one of the three entities mentioned previously, or alternatively labelled 'O' to signify non-entity recognition. This model uses a sentence-level tag to forecast potential tags with the aid of the CRF layer and inherits from the Bi-LSTM model the capacity to learn past and futuristic input attributes.



The pseudocode provided in Algorithm 1 served as the basis for training the Bi-LSTM-CRF model. The following are the four steps for Bi-LSTM-CRF:

Algorithm 1 The Bi-LSTM-CRF model training algorithm.

```
for i belong to epoch:
    for j belong batch:
        LSTM forward pass
        LSTM backward pass
        CRF layer forward and backward pass
        parameters update
    end batch for loop
end epoch for loop
```

TRAINING AND RESULTS

Models Parameter		
CRF Model	Optimizer C1 C2 Max iteration	IBFGs 0.1 0.1 25000
BI-LSTM-CRF Model	Optimizer Learning rate Hidden layer Epoch Batch size	RMSprop 0.1 300 10 100

The implemented models, CRF and Bi-LSTM-CRF, were trained on the same dataset with identical training and testing data sizes.

Precision is a metric that indicates the proportion of correctly extracted named entities identified by the system, relative to the total number of named entities extracted. This can be calculated using the following formula:

$$\boxed{\text{Precision} = \frac{\text{NumberOfCorrectlyExtractedEntities}}{\text{NumberOfN amedEntitiesExtracted}}}$$

On the other hand, recall is a metric that quantifies the proportion of named entities correctly identified by the system, relative to the total number of named entities present in the dataset. This can be calculated using the following formula:

$$\boxed{\text{Recall} = \frac{\text{NumberOfCorrectN amedEntities}}{\text{TotalNumberOfN amedEntities}}}$$

Finally, depending on the precision and recall values, the F1-measure is calculated as follows:

$$\boxed{\text{F1} - \text{Score} = 2\,\frac{(\text{P recision} * \text{Recall})}{(\text{P recision} + \text{Recall})}}$$

	Precision	Recall	F1	Support	Accuracy
PER	62.75%	52.50%	57.17%	4008	
ORG	70.34%	39.45%	51.11%	1927	
LOC	74.50%	45.25%	56.35%	956	
Overall	75.00%	48.05%	58.00%	6891	65.16%

Table 1: THE CRF MODEL'S PROPOSED PERFORMANCE RESULTS.

	Precision	Recall	F1	Support	Accuracy
PER	70.50%	75.05%	72.05%	4008	
ORG	69.34%	73.11%	73.11%	1927	
LOC	71.50%	74.35%	74.35%	956	
Overall	70.00%	76.05%	73.00%	6891	75.73%

Table 2: THE BI-LSTM-CRF MODEL'S PROPOSED PERFORMANCE RESULTS.

The result show that the Bi-LSTM-CRF model outperformed the CRF model in terms of accuracy for approach classification. The overall accuracy for Bi-LSTM-CRF was 75.73%, while it was only 65.16% for the CRF model.

Conclusion:

A classical machine learning model based on CRF that is trained using morphological and syntactic characteristics and a DNN model based on Bi-LSTM-CRF that is learned using word-level representations are both evaluated side by side in this article. The models were evaluated on the same reference dataset, which contains fine-grained named entities of Organization, Person and Location categories. The evaluation results indicate that the Bi-LSTM-CRF model give better result than the traditional CRF model, achieving a 15% improvement in performance.

Reference:

- https://ieeexplore.ieee.org/document/8554623/metrics#metrics
- https://arxiv.org/pdf/2302.03512.pdf