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.

**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 paper presents two approaches that were developed for Arabic named entity recognition (ANER). The first approach is based on a traditional machine learning method of using the conditional random fields (CRF) trained with predefined set of syntactic and morphological features. whereas, in the second model we applied a deep neural network by using a Bidirectional Long Short-Term Memory with CRF (BI-LSTM-CRF) model.

**Application of Named Entity Recognition (NER)**

Affective news analysis, semantic labelling, paraphrase identification etc.

**CRF with predefined features**

Conditional Random Fields (CRF) is an undirected graph model in which each word in the input sentence is represented using a corresponding tag in the output sentence. It predicts the entity of words 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.

A picture containing text, clock

Description automatically generated

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.

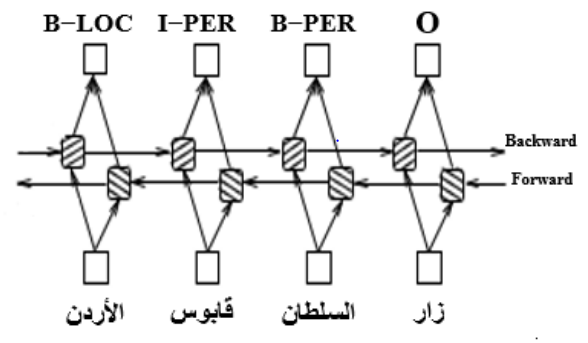
An example of using the CRF Model for ANER.

|  |  |
| --- | --- |
| Main class | Sub-class |
| Person (PER) | Athlete (1293)  Politician (6453)  Scientist (2178)  Businessman (864)  Artist (2514)  Religious-PER (2710)  Police (1324)  Group (3565)  Engineer (259) |
| Organization (ORG) | Government (1785)  Non-Governmental (2014)  Educational (1250)  Media (1194)  Commercial (2182)  Sports (1134)  Religious-ORG (194)  Entertainment (160)  Medical-Science (185) |
| Location (LOC) | Waterbody (2928)  Land-Region-Natural (954)  Celestial (1067) |
| 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 inherits the ability of learning past and futuristic input features from the Bi-LSTM model and uses a sentence-level tag to predict the possible tags with the help of the CRF layer.



The Bi-LSTM-CRF model was trained based on the pseudo code presented in Algorithm 1. The four steps for Bi-LSTM-CRF are as follows:

|  |
| --- |
| Algorithm 1 The Bi-LSTM-CRF model training algorithm. |
| 1: for each epoch do  2: for each batch do  3: LSTM model forward pass  4: LSTM model backward pass  5: CRF layer forward and backward pass  6: update parameters  7: end for  8: end for |

* The Bi-LSTM forward pass
* The Bi-LSTM backward pass
* The CRF forward and back-ward passes, then
* The parameters for the Bi-LSTM-CRF are updated.

**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 (i.e. CRF and Bi-LSTM-CRF) were trained using the same dataset, with the same size of training and testing data.

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:

Precision

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:

Recall =

Finally, the F1-measure is computed based on the precision and recall values as follows:

F1 – Score = 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 PERFORMANCE RESULTS OF THE PROPOSED CRF MODEL

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 PERFORMANCE RESULTS OF THE PROPOSED BI-LSTM-CRF MODEL

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:**

This paper presents a comparative evaluation of two models for ANER - a traditional machine learning model based on CRF, which is trained using morphological and syntactic features, and a DNN model based on Bi-LSTM-CRF, which is trained using word-level representations. The models were evaluated on the same reference dataset, which contains fine-grained named entities of Person, Location, and Organization categories. The evaluation results indicate that the Bi-LSTM-CRF model outperforms 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>