Arabic Named Entity Recognition using a Semi-Weakly Supervised Learning Approach

Ghassan Dib, Mohsen Shamas
May 2020

1 Abstract

Named Entity Recognition is a core field of Natural Language Processing. It is a subtask of information extraction and classification of named entities in text into pre-defined categories. We tackle the problem of Arabic NER by employing the architecture of Facebook's semi-supervised training framework.
Our approach is an improvement to the Arabic named entity recognition via deep co-learning which uses LSTM model. The model we used is the Bidirectional Encoder Representations from Transformers (BERT) model which we trained and tested on the same data sets.

2 Introduction

Arabic is one of the most spoken languages in the world with around 420 million native speakers. However, it is also a challenging language when it comes to some natural language processing (NLP) tasks such as named entity recognition (NER).NER is the task of extracting,locating and classifying named entities in a given piece of text.
Our main concern with NER is classifying proper nouns presented in Arabic text into three categories, namely: person, location, and organization.
Our main task is, given an Arabic text, to extract named entities, and to classify them into the three categories.
In this project, we do so by using the semi-supervised training framework. The architecture is defined by first training a teacher model on a small labeled data set. Then we use this model on a huge semi-labeled data set to predict the labels of its unlabeled data. The output is then fed to a student model to learn from it. Finally the student model is tested on 3 different data sets and scored relatively high.

In our approach, we will be using deep learning techniques to tackle the Arabic NER Problem. The model we will use to train our data is the BERT model introduced by Google in 2018 (mainly the AraBERT model), which is a transformer model. Unlike the famous LSTM model and other recurrent neural networks, Bert is attention-based model, which only relies on the attention of each word in the given text, and outputs probabilities of dependencies among the words, which will be used to predict the tag of the named entity in the text. The method we will be using to approach our model consists of three crucial steps:

- We will train a teacher model on a small amount of fully labeled data.
- We will use this model on a semi labeled data to predict the label of the unlabeled data.
- We will use the output to train a student model, and then train the student model on all the data including the fully labeled data to get a general-purpose Arabic NER model that can predict the labels of named entity given any text.

Related Work:

In general, we can talk about many different approaches that were performed regarding NER. The three main approaches are rule-based, machine learning based, and hybrid approaches. However, what we care about is the

machine learning based approaches. These approaches have been proposed for many different language datasets including English, Dutch, German, Spanish and Russian NER datasets. Those approaches were mostly done relying on support vector machines, meta-classifiers, and many other features. However, the most popular ones are studied using deep learning, especially Neural network models, mainly based on LSTM, and they obtained state-of-the-art performance on the above mentioned language datasets.
On the other hand, very few approaches were done for the Arabic NER, mainly one approach. This approach relied on "character-level neural networks and conditional random fields, in a fully-supervised fashion". However, the problem here was that this model was trained and tested using only one dataset, so it does not reflect any generalization.
Mr. Chadi Helwe does another main approach, where the model was based on "Unsupervised Arabic word embedding" using deep learning. This approach showed significant results.
The most important resource that we are going to rely on in our project is the "AraBERT: Transformer-based Model for Arabic Language Understanding". The model aims at pre-training BERT specifically for the Arabic language in the pursuit of achieving the same success that BERT did for the English language.

3 Data set

The dataset is divided into four main sets (about 12240 Arabic articles from Wikipedia). First, there is the training dataset (ANERCorp dataset), this dataset consists of 114926 labeled data (about 10880 articles), and it was used to initially train the classifiers in a completely supervised manner. Then there is the validating data, the NewsFANE Gold corpus validating data, this data consists of 71067 labeled data (about 1360 articles), and we used it for validation. Our approach was then tested on three different Arabic NER benchmarks.The first dataset we evaluated our model on is the AQMAR dataset, which is an annotated corpus for the task of ArabicNER. AQMAR consists of 2456 sentences from 28 articles from Arabic Wikipedia. The articles belong to four domains, namely history, science, sports, and technology. The second dataset is the NEWS dataset, which is also an annotated corpus for the task of Arabic NER constructed by Darwish. The NEWS dataset consists of 292 sentences that were retrieved from the RSS feed of the Arabic (Egypt) version of news.google.com from October 6, 2012. The corpus contains news from different sources and covers international and local news related to politics, finance, health, sports, entertainment, and technology. The third and final dataset we used for evaluation is the TWEETS dataset, also constructed by Darwish. The TWEETS dataset consists of 982 tweets that were randomly selected from tweets posted between November 23, 2011 and November 27, 2011. The tweets were retrieved from Twitter API using the query "lang:ar" (language=Arabic). Finally, we

used our trained model on a semi-labeled dataset in order to predict their labels. The semi-labeled data consists of 1617184 labeled and unlabeled data, each line contains the data, its label, and three probabilities that are used to choose the label according to the highest probability.

4 Model

To build a robust Arabic NER system, we propose to use deep learning based on weakly supervised Arabic word embeddings.

FACEBOOK COMPUTER VISION.

Our project was influenced by a research done by Facebook called "Billion-scale semi-supervised learning for state-of-the-art image and video classification". This approach, which is called semi-weak supervision, is a new way to combine the merits of two different training methods: semi-supervised learning and semi-weakly supervised learning. It opens the door to creating more accurate, efficient production classification models by using a teacher-student model training paradigm and billion-scale semi-weakly supervised data sets.
The human labeling of training data required for fully supervised approach cannot scale to all the possible visual concepts in the world. Labeling thousands of species of plants and animals, for example, is resource intensive and requires extensive domain expertise. Semi-supervised learning offers a different approach to decreasing AI systems' dependence on labeled data sets. The method trains a target model using large amounts of unlabeled data in combination with a small set of labeled examples.

TRANSFORMERS MODEL.

The model is called a Transformer and it makes use of several methods and mechanisms. Like LSTM, Transformer is an architecture for transforming one sequence into another one with the help of two parts (Encoder and Decoder); however, unlike LSTMs, the transformer model takes the tokens in batches, in our case full sentence, in one shot and predicts it which makes it much faster than the sequencial input of LSTMs.
The Transformer consists of two main components: a set of encoders chained together and a set of decoders chained together. The function of each encoder is to process its input vectors to generate what are known as encoding, which contain information about the parts of the inputs which are relevant to each other.

However, it differs from the previously described/existing sequence-to-sequence models because it does not imply any Recurrent Networks (GRU, LSTM, etc.).

BERT MODEL.

Unlike recent language representation models, Bidirectional Encoder Representations from Transformers (BERT) is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

To implement our architecture, we made use of the implementation of BERT For Token Classification by huggingface.co, and we used Arabert, a Pre-trained BERT model for the Arabic Language trained by the AUBMind group, for the task of Arabic NER.

Semi-weakly Supervised BERT

Our approach consists of four steps, based on the Billion-scale semi-supervised learning for state-of-the-art image and video classification by Facebook AI.
The framework consists of Four main steps. First we we trained a teacher BERT Model for token classification on the labeled dataset, namely 12240 Arabic articles retrieved from Wikipedia using the Python Package Wikipedia by Dr. Elbassuouni and Mr. Helwe. Then we use the teacher model to predict the labels of unlabeled data from the Partially Annotated Arabic NER Dataset. After that we trained a student model of the same architecture on the data of the predicted output, and finally, fine tune the student model by the labeled dataset. To get the best performance of the student of the model, we did hyper parameter tuning to find the best batch size and learning rate.

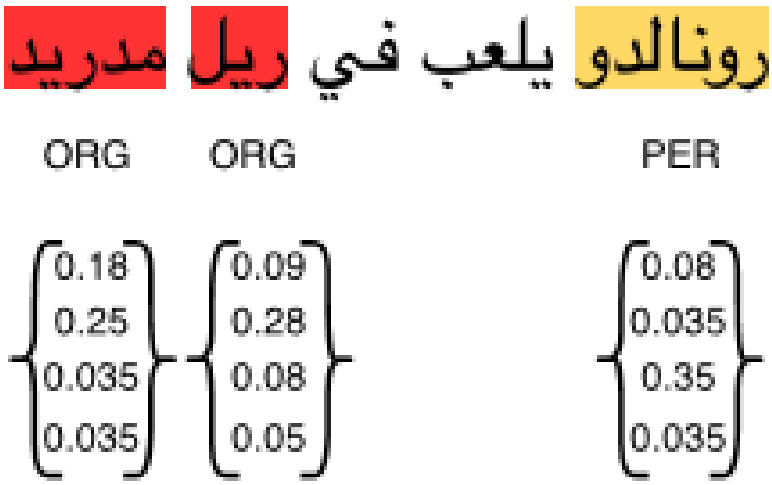


Fig. 1: Example of how the model works.

We can see in figure 1 an example of how the BERT model works for Arabic NER. For example, the word رونالدو scores a probability of 0.08 for the LOC, 0.035 for the ORG, and 0.35 for the PER, and a total Average of 0.035. Therefore, the word is labeled as PER because person feature got the highest score.

5 Results

In this section we will evaluate our semi-weakly supervised approach for the task of Arabic NER. We tested our approach, as mentioned above, on three different datasets and compared the results with the Arabic NER via Deep Co-learning approach, which was done using LSTM model, and our approach showed better results on both the AQMAR and NEWS datasets.

AQMAR Dataset

The first dataset we evaluated our model on is the AQMAR dataset, which is an annotated corpus for the task of ArabicNER. Table 1 shows the results of the BERT model and the LSTM model.
As can be seen from Table 1, the BERT weakly supervised approach using semi-labeled data outperforms the LSTM approach in terms of the average F-measure overall the classes with an average of 64.1 (the last column in Table1)

Model	LOC	ORG	PER	Avg
Arabert Semi-weakly Supervised Learning	67.5	32.3	73.5	64.1
LSTM Deep Co-learning: Semi-labeled Data	67.0	38.2	65.1	61.8
BERT Teacher model: labeled data	64.8	31.3	69.7	61.5

Table 1: F-measure after testing on AQMAR

NEWS Dataset

The second dataset is the NEWS dataset, which is also an annotated corpus for the task of Arabic NER.
Table 2 shows the results of BERT weakly supervised model and the LSTM deep Co-learning model.
As can be seen from Table 2, the BERT weakly supervised approach using semi-labeled data outperforms the LSTM approach in terms of the average F-measure overall the classes with an average of 77.0 (the last column in Table2)

Model	LOC	ORG	PER	Avg
Arabert Semi-weakly Supervised Learning	78.5	59.4	88.0	77.0
LSTM Deep Co-learning: Semi-labeled Data	81.6	52.7	82.4	74.1
BERT Teacher model: labeled data	76.4	54.3	84.2	73.2

Table 2: F-measure after testing on NEWS

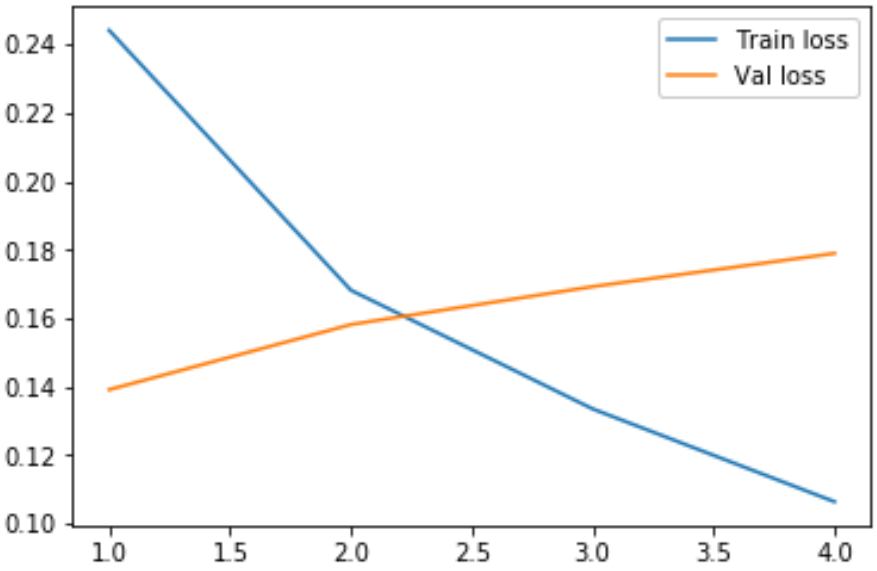
TWEETS Dataset

The third and final dataset we used for evaluation is the TWEETS dataset. Table 3 shows the results of the BERT and LSTM models on the TWEETS dataset.
As can be seen from Table 3, the LSTM approach using semi-labeled data outperforms the BERT weakly supervised approach in terms of the average F-measure overall the classes with an average of 59.2 (the last column in Table3)

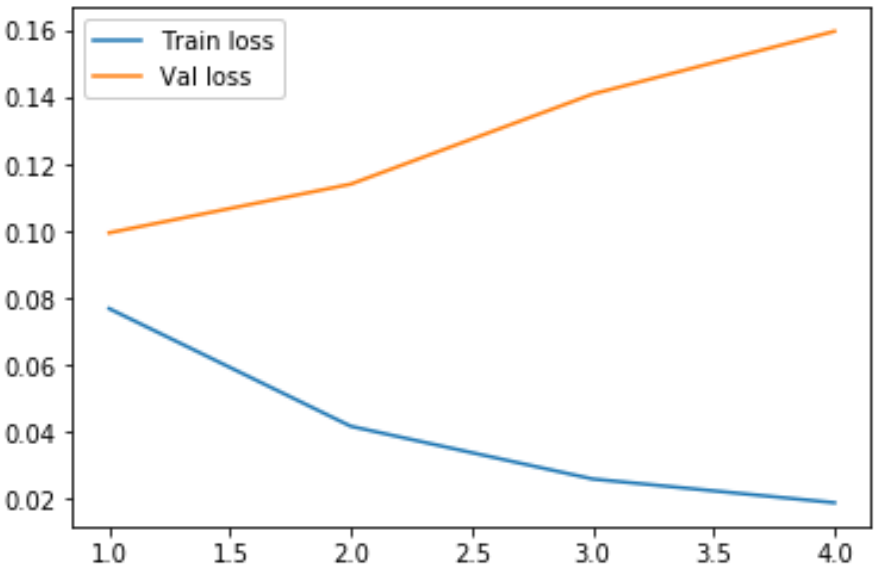
Model	LOC	ORG	PER	Avg
Arabert Semi-weakly Supervised Learning	62.7	40.0	63.6	58.5
LSTM Deep Co-learning: Semi-labeled Data	65.3	39.7	61.3	59.2
BERT Teacher model: labeled data	59.0	31.7	68.3	57.8

Table 3: F-measure after testing on TWEETS

In the three different datasets, we can notice the great improvement that was done from the teacher model to the AraBERT semi-labeled supervised model.



Graph1: The learning curve of the Teacher model.



Graph2: The learning curve of the Student model.

As we can see from the above graphs, that plots the predicted values of our two models, usually fine-tuning a bert model needs only a 1 or a couple iterations to have good results. However, we can notice that the graph suffer from over-fitting, this could be explained by the large number of iteration and training examples that we did.

6 Conclusion

In this paper, we presented a new approach to detect and classify named entities in any Arabic text. To target our weakly-supervised NER approach, we first used a fully labeled data to train our teacher model, predicted output for semilabeled data, and use it to train a student model. We then evaluated our approach using three different Arabic NER datasets and compared it to the deep-Co-learnig approach that is a state-of-the-art and baseline Arabic NER approach. Our semi-semi-weakly supervised approach significantly and consistently outperformed the other compared to approach on the AQMAR and NEWS datasets.

References

1. Arabic Named Entity Recognition via Deep Co-learning.
2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
3. Facebook: Billion-scale semi-supervised learning for image classification.

Big thanks for Mr. Chadi Helwe for his cooperation in making this work successful.