The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Boosting Arabic Named Entity Recognition Transliteration with Deep Learning

Manar Alkhatib,[1] Khaled Shaalan[2]
The British University in Dubai[1, 2]
Manar.Alkhatib@buid.ac.ae[1], Khaled.Shaalan@buid.ac.ae[2]

Abstract

The task of transliteration of named entities from one language into another is complicated and considered as one of the challenging tasks in machine translation (MT). To build a well performed transliteration system, we apply well-established techniques based on Hybrid Deep Learning. The system based on convolutional neural network (CNN) followed by Bi-LSTM and CRF. The proposed hybrid mechanism is examined on ANERCorp and Kalimat corpus. The results show that the neural machine translation approach can be employed to build efficient machine transliteration systems achieving state-ofthe-art results for Arabic – English language.

1. Introduction

Transliteration of Named Entities (NEs) is the process of phonetic translation for names from source language into a target language (Grundkiewicz 2018). It is an important part for many natural language processing tasks, and machine translation in particular.

Arabic is a language of rich vocabulary and complex morphology and syntax. The features and characteristics of the Arabic language pose many challenges particularly in NER. There has been a growing interest to tackle these challenges, and encourage the development of a productive and robust Arabic Named Entity Recognition system (Shaalan 2014).

The use of capital letters in the English language as indicators for start and end of NEs in a text, make it an easy task. Hence the capital letters in Arabic language not exist, the achievement of high performance in NER is complex and difficult (Benajiba and Rosso 2007; Benajiba, Diab, and Rosso 2008; Shaalan 2014).

Transliteration for NEs need different translation techniques than the other words of a text, so the post-editing step is more expensive when we face an error. Transliteration can handle the translation of words that are out of vocabulary in machine translation system (Kundu, Paul, and Pal 2018).

In this work, we explore the Neural Machine Translation (NMT) approach based on three stages. Firstly, we apply several preprocessing steps to clean the dataset (Tokenization, Stop Words Removal, Morphological Segmentation, and POS Tagging). Secondly, apply multi- feature extraction and selection. In the final stage, we apply the Named entity classifier algorithm to classify the data.

2. Related Work for Arabic Named Entity Recognition

Many related works have been conducted to recognize Arabic name entities. Generally, ANER systems were developed based on any one of the following approaches: machine learning based, deep learning based, and hybrid based. In this section, we present the most relevant surveys and comments on some of such works on ANER.

2.1 Machine Learning Based ANER

Machine-learning based approaches learn NE tagging decisions from annotated texts. The most common machine learning methods are Supervised, Semi-Supervised, and Unsupervised. These methods solve the problem of NER in classification task and require a huge annotated corpus. Some of the supervised techniques applied for NER are Support Vector Machine (Al- Ahmari and Al-Johar 2016), Conditional Random Fields, and Hidden Markov Model (Mouhcine, Mustapha, and Zouhir 2018), Genetic Algorithm, and Naive Bayesian Classifier (Alsayadi and ElKorany 2016), and Decision trees (Hamadou, Piton, and Fehri 2010). The problem with this machine learning approach is the ―Lack of Accuracy since a single classifier cannot produce sufficient results for training and testing of ANER.

2.2 Hybrid ANER

Hybrid approaches are the combination of the rule-based and machine-learning based approaches. The process of this hybrid approach flows from the rule-based system to the machine-learning approach or from machine learning to rule based approach. In the following literature, we have utilized challenges of a hybrid approach for NER in the Arabic language (Hamadou, Piton, and Fehri 2010; Boujelben, Jamoussi, and Ben Hamadou 2014). Mainly, a machine learning approach with a rule-based approach is used to improve the system performance. Genetic algorithm (ML algorithm) is used to extract and generate the most significant and exciting rules. It is suitable for searching problems, and it suffers from the memory resources and high computation that are required if the size of the individuals of the problem solution is increased. Hybrid (rule-based and machine-learning based approach) approaches do not always support for the large corpus. We require rules to solve this issue. All the approaches mentioned previously, have been reviewed in (Althobaiti, Kruschwitz, and Poesio 2015).

2.3 Deep Learning Based ANER

Deep Learning (DL) based ANER has been emerged currently. In (Ali, Tan, and Hussain 2018) combined Bidirectional Long Short-Term memory (Bi-LSTM) and CRF. Combination ANER approach on basis of deep learning must require huge volume of corpus. Both LSTM and CRF increase computational overhead for ANER system. In (Gridach and Haddad 2018), researchers have designed a novel architecture based on neural networks. It is a consolidation of bidirectional Gated Recurrent Unit (GRU) and Conditional Random Fields (CRFs). The minimal set of features were used in pre-trained character level and word level embedding. The most significant improvement in this work is the uses of minimal set of features for ANER. This work is failed to produce high recognition rate.

3. Proposed ANER Using CNN-Bi-LSTM -CRF

In this section, we firstly describe the problem statement. Then, we describe briefly of our proposed ANER transliteration system. Figure.1. indicates the proposed system architecture.

Figure 1: The main architecture of our NER neural network machine transliteration.

3.1 System Overview

Our proposed ANER translation comprised of five steps: Text- Pre-processing, Multi-Feature Extraction, Feature Selection, Entity Recognition, and translation. All CNN, Bi-LSTM and CRF are well defined classifiers. By combining them, we achieve an excellent performance regarding F-measure, precision, and recall.

3.2 Text Pre-processing

A predictable pre-processing phase consists of the following steps: tokenization, stop word removal, morphological analysis and POS tagging.

To do pre-processing of named entity, we used MADAMIRA, Morphological Analyzer and POS tagging (Pasha et al., 2014). In the following, we describe the pre-processing steps:

3.3 Multi-Feature Extraction

Arabic documents contain a large number of redundant and irrelevant words. Therefore, feature selection is highly required. We considered the set of features of Arabic language such as contextual features using -/+1 token window, lexical features using N-Grams n range from 1 to 3, and grammatical features (noun, verb, adjective). We used the TF-IDF term weighting scheme for feature selection to exploit the most important and discriminant features for Arabic Named Entity Recognition.

3.4 Classification

Existing deep learning approaches are not suitable for ANER. We propose a novel neural network architecture for sequence labeling. We use a combination of three algorithms, Bi-LSTM, CNN and CRF. These three techniques improve the classification process and easily handle the unknown words and ambiguity. We first use convolutional neural network (CNN) (LeCun et al., 1989), the multilayer supervised learning framework. Its purpose is to promote the training set. It provides sufficient training data; this feature enables the network to work efficiently regardless of different classes. Then we used bi-directional LSTM (BLSTM) to model context information of each word. Finally, the output vectors of BLSTM are an input to the CRF layer to decode the best label sequence.

4. Experimental Results and Analysis

In this section, experimental results are shown to verify the effectiveness of the proposed ANER extraction, and translation system. This section is subdivided into three sections (dataset description, evaluation metrics and comparative study). We compare the performance of our proposed in terms of Precision, Recall, and F-measure.

4.1 Dataset Description

For our approach presented, we train the model by developing our own NER corpus. The training corpus covered specific types of NEs that were not included previously. In particular, our NER corpus generate the identification of person name, location, date and time, and phone number.

The training dataset has been developed and tagged using our tag schema, in XML format. The total number of NEs/keywords included in our corpus is 55,760 keywords.

4.2 Testing Data Corpus

The ANERcorp corpus built by (Benajiba, 2007) used for evaluating many Arabic NER extraction systems. We compare the performance with three baseline systems the bi-direction LSTM, and BLSTM-CNNs, the combination of BLSTM with CNN to CRF. We used our own corpus as a training corpus, and KALIMAT Corpus for testing our model. KALIMAT corpus built by (El-Haj and Koulali, 2013), and consists of data collection articles fall into six categories: religion , sports culture, local-news, economy, and international-news.

4.3 Results on the Hybrid System

In this paper, a number of experiments conducted to examine the performance of the proposed system with the best previous works using F-measure.

According to the results shown in Table 1, adding the CRF layer for joint decoding of CNN-BLSTM models, had achieved significant improvements over BLSTM-CNN models for NER on all metrics.

We evaluated our transliteration method on the development dataset from the KALIMAT over the six categories as shown in Table 2. Results for sports and economy present the highest F score measures, while the international news had the lowest transliteration system. As per our knowledge, there is no available Arabic NER transliteration corpus to compare our model with it.

Model	Person	Location	Organization
CNN	85.4	87.3	85.3
CNN-BLSM	87.2	88.4	87.1
CNN-BLSTM-CRF	94.2	95.3	93.1

Table 1: Performance of our model together with three baseline systems

Category	F% Person	F%- Location	F% Organization
Culture	81.2	82.6	84.3
Economy	85.2	85.4	85.1
Religion	77.3	77.7	76.5
local –news	81.1	83.5	83.1
international news	79.1	79.7	79.1
Sports	86.4	85,8	86.2

Table 2: Performance of our transliteration model on the six categories.

Source	Technique	Person	Location	Organization
Benajiba & Rosso	ANERsys	52.2	86.7	46.5
Zaghouani	RENAR	55.3	85.2	47.0
Al-Ahmari	Rule Based Approach	63.4	89.1	56.5
Al Thobaiti	Maha Althobaiti	64.2	73.1	54.5
Our System	CNN-Bi-LSTM- CRF	93.7	95.2	95.3

Table 3: Comparative Results for F-Measure

4.4 Comparative Study

We evaluated the extraction system performance against other related works. Based on our search, the “ANERcorp” corpus created by (Benajiba and Rosso 2007) had been used for examining various systems. Therefore, we used it to evaluate our system’s performance with other systems results such as; “ANERsys 2.0” (Benajiba and Rosso 2007), “RENAR” (Zaghouani 2012), (Althobaiti et. al. 2015), and “Rule Based Approach”(Al-Ahmari and Al-Johar 2016) .

The dataset distribution is as follows: Person: 39%, Location: 30.4%, Organization: 20.6%, and Miscellaneous: 10%. In addition, we present the results for the proposed CNN-Bi-LSTM-CRF for better comparison.

The overall performance obtained for various categories such as a person, location, organization, and miscellaneous types in terms of F-measure are 93.7%,95.2%,95.3%, respectively as shown in table 3.

5 Conclusion

In this paper, a new approach is proposed to tackle the problem of ANER transliteration in an innovative way. The main difference between our approach and the previous ones is that our model uses a hybrid scheme that combines the three algorithms. The first algorithm is a CNN, which gives us the insight to move fully to neural network approaches, connected Bi-LSTM and the last one, is CRF to examine the performance of our model in comparison to other architectures, a well-known dataset is used, the ANERcorp dataset. The proposed model outperforms the current state-of-the-art models by considerable results.

Acknowledgment

This work is a part of a project undertaken at the British University in Dubai.

References

Al-Ahmari, S.S. and Al-Johar, B.A., 2016, July. Cross domains Arabic named entity recognition system. In First International Workshop on Pattern Recognition (Vol. 10011, p. 100111I). International Society for Optics and Photonics.

Ali, M.N., Tan, G. and Hussain, A., 2018. Bidirectional recurrent neural network approach for Arabic named entity recognition. Future Internet, 10(12), p.123.

Alsayadi, H.A. and ElKorany, A.M., 2016. Integrating semantic features for enhancing arabic named entity recognition. Int. J. Adv. Comput. Sci. Appl.(IJACSA), 7(3), p.2016.

Althobaiti, M., Kruschwitz, U. and Poesio, M., 2015. Combining minimally-supervised methods for arabic named entity recognition. Transactions of the Association for Computational Linguistics, 3, pp.243-255.

Benajiba, Y., Diab, M. and Rosso, P., 2008, October. Arabic named entity recognition using optimized feature sets. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (pp. 284-293).

Benajiba, Y. and Rosso, P., 2007, December. ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information. In IICAI (pp. 1814-1823).

Boujelben, I., Jamoussi, S. and Hamadou, A.B., 2014. A hybrid method for extracting relations between Arabic named entities. Journal of King Saud University-Computer and Information Sciences, 26(4), pp.425-440.

Darwish, K. and Magdy, W., 2014. Arabic information retrieval. Foundations and Trends® in Information Retrieval, 7(4), pp.239-342.

El-Haj, M. and Koulali, R., 2013. KALIMAT a multipurpose Arabic Corpus. In Second Workshop on Arabic Corpus Linguistics (WACL-2) (pp. 22-25).

Hamadou, A.B., Piton, O. and Fehri, H., 2010, May. Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform.

Grundkiewicz, R. and Heafield, K., 2018, July. Neural machine translation techniques for named entity transliteration. In Proceedings of the Seventh Named Entities Workshop (pp. 89-94).

Kundu, S., Paul, S. and Pal, S., 2018, July. A deep learning based approach to transliteration. In Proceedings of the seventh named entities workshop (pp. 79-83).

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. and Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), pp.541-551.

Mouhcine, R., Mustapha, A. and Zouhir, M., 2018. Recognition of cursive Arabic handwritten text using embedded training based on HMMs. Journal of Electrical Systems and Information Technology, 5(2), pp.245-251.

Pasha, A., Al-Badrashiny, M., Diab, M.T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O. and Roth, R., 2014, May. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In LREC (Vol. 14, pp. 1094-1101).

Shaalan, K., 2014. A survey of arabic named entity recognition and classification. Computational Linguistics, 40(2), pp.469-510.

Zaghouani, W., 2012. RENAR: A rule-based Arabic named entity recognition system. AC`M Transactions on Asian Language Information Processing (TALIP), 11(1), pp.1-13.