International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V70I1P101

Volume 70 Issue 1, 1-3, January 2022

Original Article

A Comparative Study on Various Deep Learning Techniques for Arabic NLP Syntactic Tasks

Shaima A Abushaala[1], Mohammed M Elsheh [2]

[1]*Misurata University*
[2]*Libya Academy-Misurata*

Received Date: 07 December 2021
Revised Date: 09 January 2022
Accepted Date: 21 January 2022

Abstract - *It is well known that there are three basic tasks in Natural language processing(NLP) (Tokenization, Part-Of-Speech tagging, Named Entity Recognition), which in turn can be divided into two levels, lexical and syntactic. The former level includes tokenization. The latter level includes part of speech (POS) and the named entity recognition (NER) tasks. Recently, deep learning has been shown to perform well in various natural language processing tasks such as POS, NER, sentiment analysis, language modelling, and other tasks. In addition, it performs well without the need for manually designed external resources or time-consuming feature engineering. In this study, the focus is on using Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BLSTM), Bidirectional Long Short-Term Memory with Conditional Random Field (BLSTM-CRF), and Long Short-Term Memory with Conditional Random Field (LSTM-CRF) deep learning techniques for tasks in Syntactic level and comparing their performance. The models are trained and tested by using the KALIMAT corpus. The obtained results show that a BLSTM-CRF model overcame the other models in the NER task. As for the POS task, the BLSTM-CRF model obtained the highest F1-score compared to the other models.*

Keywords — *Natural Language Processing, Deep learning, Part-of-Speech tagging, Named-Entity Recognition.*

I. INTRODUCTION

NLP is a branch of artificial intelligence that aims to make computers understand human languages and interact with them. Because of the importance of language in humans lives, the idea of giving computers the ability to process human languages has been around since the emergence of the idea of computers themselves [1]. In recent times, the applications related to the field of NLP have increased. It also involved other scientific branches such as linguistics was spoken speech processing, in addition to statistics and others. NLP has spread widely in recent times, including information extraction and analysis, translation, the answer to the question, and other applications.

However, there are three basic tasks in NLP (token, part of speech, name entity recognition) that can be divided into two lexical and grammatical levels. The former includes tokenization, and the latter includes a portion of Speech (POS) and Name Entity Recognition (NER) tasks

Deep learning is a new area of machine learning, uses 'deep' artificial neural networks, such as Recurrent Neural Networks (RNN), Deep Neural Networks (DNN), Convolution Neural Networks (CNN), and Gated Recurrent Unit (GRU). Recently, deep learning has gained significant importance in many applications and shows it is the ability to handle many complicated tasks. Therefore, it can be adopted as a basic approach to NLP. Many studies have been demonstrated superior deep learning on traditional methods of NLP.

In this paper, we aim to facilitate the selection of the most efficient model for POS and NER tasks by comparing the performance among various techniques based on deep learning techniques to determine the best technique for each task.

II. BACKGROUND KNOWLEDGE

A. Challenges In Arabic Natural Language Processing

The Arabic language is one of the most widely used languages on social media. It is spoken by about 300 million people[3]. Therefore, the necessity for developing NLP systems for the Arabic language becomes very crucial in order to communicate with those people by absorbing the Arabic language in applications that correspond to their nature and properties. However, there are three forms of the Arabic language.

- Classical Arabic (CA): It is used in ancient historical texts.
- Modern Standard Arabic (MSA): It is what we study and use in news, media and translation.
- Dialectal Arabic (DA): It is what we use in everyday speech among people.

Since it is one of the world's richest vocabulary languages, Arabic natural language processing faces several challenges. For example, one word may have many meanings, as well as the shape of the letters where different drawings depending on their location in the world. Also, words expressions where there are complex overlaps and details when expressing sentences. This makes the task of processing the Arabic language is difficult. Moreover, the shortage of scientific resources from labelled data and research dedicated to the Arabic NLP. The preceding reasons make parsing and building computer applications that can handle the Arabic language is a complicated task.

B. Tasks In Natural Language Processing

NLP is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for achieving human-like language processing for a range of tasks or applications. Language Levels explain what actually happens in the natural language processing system[4].

These levels can be divided into a lexical level (at the level of a single word), a syntax level (at the syntax level and grammatically syntactic), and a semantic level at the context and final meaning in the language[4]. Usually, we use the information we have obtained from a higher level of processing to help with a lower level of analysis. For example, the lexical level provides sufficient information and knowledge about each word to be used in the syntactic level that analyzes the words in the sentence.

The lexical level can be considered as the stage of pre-processing and preparing the natural language text for use at the syntactic level. We may need to divide the word into a set of small parts (Abstraction of the word precedents and suffixes), such as "مدنـ"، "ـها." "مدنها"

The syntactic level focuses on dividing the sentence and the identification of its templates. In addition, the fragmentation of words in sentences is considered to find the grammatical relationship between the words, and this requires both a grammar and a parser. At this level, there are two fundamental tasks, namely, Part-of-speech tagging and Named-entity-recognition.

C. Part-of-Speech Tagging

Part of speech tagging (POS) aims to extract sections of speech from the text, such as the act of the present, the action of a thing, the status of an effect .. etc. Each word in the text is mapped into the section it represents (providing information on each word and its neighbours). The classification of the speech parts is based on pre-established grammar.

There are several deep learning approaches for POS tagging. The Recurrent Neural Networks model outperformed in the POS tagging task from the study [2]. In this study, we are going to explore four different types of Recurrent Neural networks, including LSTM, BLSTM, LSTM-CRF, BLSTM-CRFas can be seen in Table 1.

Table 1. List of Model Names And Description In Pos Tagging and Ner Experiment

Model	Description
LSTM	Long Short-Term Memory
BLSTM	Bidirectional Long Short-Term Memory
LSTM-CRF	Long Short-Term Memory with ConditionalnRandom Field
BLSTM-CRF	Bidirectional Long Short-Term Memory with Conditional Random Field

D. Name-Entity Recognition

The task of named entity recognition (NER) is an important stage in NLP. This stage provides very useful information, which aims to extract information from speech such as names of people, institutions, places (countries and cities), products, dates, events etc. This can be a difficult task for the Arabic language because it is a language of rich form.

Long Short Term Memory (LSTM) is designed to overcome the long-term dependency problem. This is achieved by using the activation function of the memory cell to figure out what data is important and should be remembered and looped back into the network and what data can be forgotten. It is a virtual behaviour practically reminiscent of information for long periods of time. LSTM is characterized by the presence of four layers that interact in a very special way, rather than a single layer, as is the case in the pattern of recurrent neural networks[5].

Bidirectional Long Short-Term Memory(BLSTM) is a bidirectional structure. It is added to the LSTM model to increase the amount of entered information. The output layer can adopt information from both previous and future time-steps when two LSTM layers of opposite directions are connected together[5].

Conditional random fields (CRF) is one of the statistical modelling methods that are often applied in identifying patterns and implementing sequential dependencies in predictions, such as: named entity recognition, part of speech tagging, noise reduction, and gene prediction. Conditional random fields are integrated into the models LSTM and BLSTM as an output layer to include dependencies output labels into the models because a usual softmax output layer does not take dependencies across output labels into consideration[5].

III. EXPERIMENTS AND RESULTS

A. Dataset

The corpus used for this study is part of about 4,481,775 a word from the KALIMAT Corpus that was collected from the Omani Al-Watan newspaper. It is used to train and test the models. The version of the KALIMAT 1.0 corpus can be obtained from SOURCE FORGE[6]. The data set that was

adopted contains 105,107 words. This data was divided into a training and testing group with a ratio of 90% to 10%, respectively. The training set contains 94,597 words, and the testing set contains 10510 words. Then, the training set is divided later when writing the code into a training set and a verification set of 80% to 10%, respectively.

There are 30 part-of-speech classes and 6 named-entity classes in the dataset. Named-entity types in this dataset for Person Names, location Names, and organization Names For non-named-entities, the 'O' tag is used to represent them.

B. Preprocessing

The data that is used in this work is semi-processing which needs only transliteration, i.e. converting Arabic characters into Latin characters often referred to as "Arabizi".Because some programs do not support Arabization, therefore, it is necessary to use the transliteration method. This method is a Buckwalter transliteration system [7], used to write Arabic characters using Latin ASCII characters and vice versa.

This research concerns comparing LSTM, BLSTM, LSTM-CRF, BLSTM-CRF deep learning techniques on data for Arabic NLP Lexical and Syntactic Tasks. Four experiments were conducted the on KALIMAT 1.0 corpus for POS and NER tasks.

a) Part-Of-Speech Tagging

In this experiment, LSTM, LSTM-CRF, BLSTM, and BLSTM-CRF were compared by testing with the testing dataset. The evaluation model was focused on the F1-score to find out the best POS models. The results for 40 epochs are shown in Table 2.

Table 2. F1-Score of POS Models Tested By Test Datasets For 40 Epochs. The highest F-Score Is Bolded.

Models	Measurement
	F1
LSTM	80.2%
BLSTM	80.4%
LSTM-CRF	79.7%
BLSTM-CRF	**81.1%**

Through the results shown in Table, the two models, BLSTM and BLSTM-CRF, achieved the highest rates. But the BLSTM-CRF model outperformed the BLSTM model due to the fact that the CRF model is suitable for recognizing the sequence pattern. As a result, The models make use of the CRF property to achieve more accurate predictions.

b) Name-Entity Recognition

In this experiment, after training LSTM, LSTM-CRF, BLSTM, and BLSTM-CRF models by using the training dataset collected from the KALIMAT corpus. The models were evaluated on the test dataset by calculating the accuracy F1-score. The results for 40 epochs are shown in Table 3.

Table 3. F1-Score of NER Models Tested By Test Datasets For 40 Epochs. The highest F-Score Is Bolded.

Models	Measurement
	F1
LSTM	68.0%
BLSTM	67.5%
LSTM-CRF	68.1%
BLSTM-CRF	69.8%

Considering the F1-scors for NER that are shown in Table 3, and as in the previous task in POS, the BLSTM-CRF model achieved the best score.

According to F1-score results that achieved at POS and NER tasks, it is noted that the superiority of most of the models with CRF over models without CRF since the CRF model is suitable for recognizing the sequence pattern. As a result, the models use the CRF property to achieve more accurate predictions.

IV. CONCLUSION

In this study, we evaluated the performance of models LSTM, BLSTM, LSTM-CRF and BLSTM-CRF on the data of the Arabic language for the two tasks POS and NER to find out the best model for each task. Part of the KALIMAT corpus data was used to train the models. This is data converted from Arabic characters into Latin characters because some programs do not support Arabization. The method used for this processing is a Buckwalter transliteration system to write Arabic characters using Latin ASCII characters and vice versa.

After completing the training of the models in both tasks, experiments were conducted on the four models using the testing dataset. After testing and evaluating the models, the results for a POS task showed that the BLSTM-CRF model outperformed the other models, obtaining the highest F1-score as the F1-scored 81.1%. As for the results of the NER task, the BLSTM-CRF model obtained a higher F1-score, 69.8%.

REFERENCES

[1] Jurafsky, D. and H. James, Martin: Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics., (2008) Prentice-Hall, Englewood Cliffs.

[2] Huang, Z., W. Xu, and K. Yu, Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991, (2015).

[3] Cheriet, M. and M. Beldjehem, , Visual processing of Arabic handwriting: challenges and new directions, Summit on Arabic and Chinese handwriting (SACH'06), Washington-DC, USA, (2006) 129-136.

[4] Liddy, E.D., Natural language processing, (2001).

[5] Ma, X. and E. Hovy, End-to-end sequence labelling via bi-directional lstm-cnns-crf, arXiv preprint arXiv:1603.01354, (2016).

[6] Dr Mo El-Haj, R.K., KALIMAT a Multipurpose Arabic Corpus, 20132015-04-09Available from: https://sourceforge.net/projects/kalimat/files/?source=navbar.

[7] Habash, N., A. Soudi, and T. Buckwalter, On Arabic transliteration ,in Arabic computational morphology, (2007) 15-22.