INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

# Afan-Oromo Named Entity Recognition Using Bidirectional RNN

**Birhanu Gardie**[1], **Zemedkun Solomon**[1]*

**1** School of Computing and informatics, Mizan Tepi University, Ethiopia

OPEN ACCESS

**Received:** 15-01-2022

**Accepted:** 11-03-2022

**Published:** 02.05.2022

**Citation:** Gardie B, Solomon Z (2022) Afan-Oromo Named Entity Recognition Using Bidirectional RNN. Indian Journal of Science and Technology 15(16): 736-741. https://doi.org/10.17485/IJST/v15i16.123

* **Corresponding author**.

zemeds@mtu.edu.et

**Funding:** Article processing fee is defrayed partially by Indian Society for Education and Environment

**Competing Interests:** None



Published By Indian Society for Education and Environment (iSee)

**ISSN**
Print: 0974-6846
Electronic: 0974-5645

## Abstract

**Objectives:** This work aims about the development of Afan-Oromo language named entity recognition which widely used in question answering, information extraction and information retrieval aimed at categorizing and predicting tokens of a given corpus into predefined named entity classes like organization, location person and others (non-named entity tags). **Methods:** In this work, a bidirectional long-short term memory technique is used to model the Afan-Oromo language NER system to recognize and classify words into their named entity classes. **Findings:** While we evaluated the experiment in cross-validation, we attained a result of precision, recall and f1-measure values 96.7%, 96.2% and 97.3% respectively. We have collected the data from Ethiopian broadcasting Corporation (Afan-Oromo program). Therefore, a newly annotated dataset having 12,479 instances is used for this study. **Novelty:** Finally we have contributed by boosting a NER system for Afan-Oromo language which is independent of other natural language processing tasks. We proved bidirectional long-short term memory approach can be extended, trained can work for Afan-Oromo language.

**Keywords:** Bidirectional long shortterm memory; Natural language processing; Softmax; recurrent neural network; Afan-Oromo named entity recognition

## 1 Introduction

Named entity recognition task is usually an essential step in natural language processing tasks[(1)]. It is used in several applications such as text summarization, question-answering, information extraction and machine translation. NER target is to detect mentions of rigid designators from text corpus belonging to semantic types such as location, person, organization etc.[(2–4)]. For instance in the statement **"Pireezidaant Obaamaa Pirezidaantii Ameerikaa osoo aangoorra jiranii Heerooshiimaa daawwatan isa jalqabaa ta'aniiru"** (President Obama became the first sitting American president to visit Hiroshima) holds the named entities of "**Obaamaa**" (person), "**Ameerikaa**" (location), and "**Heerooshiimaa**" (location) too. Additionally, named entity recognition is a basic process in constructing ontology or in building relationship graphs[(1)]. In recent years, Afan-Oromo NER (ANER) systems has become a challenging task and is receiving an increasing attention from recent researchers due to the limited availability of annotated datasets. Afan-Oromo (while translated it means the

Oromo language ) is one of the most languages which have a large number of speakers in Ethiopia [5]. It is a family of Cushitic language [6], written using the Latin alphabet which have around 50 million speakers in Ethiopia, Somalia, Egypt and Kenya [7]. Named entity recognition researches have been conducted for Amharic language and Afan-Oromo too using Hidden Markov Model, Conditional random field model in Ethiopia. Meanwhile, in the deep learning technique there is an approach that is proven to have the highest state-of-the-art performance in the case of named entity recognition, namely Bidirectional Long Short Term Memory (BLSTM). Bidirectional LSTM combines the previous context and the next context by processing data from two directions. Currently, recurrent neural network such as LSTM is taking a prominent place in named entity recognition because of its capability of constructing relationship in neighboring words.

Because of its significant role in several Natural Language Processing (NLP) tasks, named entity recognition system have been an active research issue in the past twenty years and gains researchers attention to the present day [8,9]. Several researchers have been performed research works on named entity recognition, unfortunately many of those research works are specific to high resourced Western and Asian country languages such as French, English, Hindi, Arabic, Chinese and other many European country languages. Despite its large number of speakers, only few research works has been done in Afan-Oromo language. A prototype of bidirectional RNN (B-RNN) is explicitly built on Long Short-Term Memory (LSTM).To understand the context better and resolve ambiguities in the text, bidirectional recurrent structures are utilized which learns the information from the previous and future time stamps. The two types of connections are one going forward in time and the other going backward in time [10]. These connections help in learning the previous and future representations respectively. The module in this bidirectional recurrent structure could be a recurrent neural network, long-short term memory. Bi-directional LSTM uses the information contained in whole sentences. We hypothesized that long sentences could contain information unrelated with the target entities, and hence, in domains with long sentences, such as the biomedical literature, the utilization of local information rather than whole sentences may help improve precision [11]. In this paper, based on the success of using machine learning architectures for NER task, for resource rich languages like English, we follow a simple yet effective approach of refining previously proven successful bidirectional long-short term memory models for Afan-Oromo language. The idea is to use sparse bidirectional long-short term memory architecture which allows to learn the model parameters in low-resource scenario [12]. The architecture geared towards low resource data has also the advantage that it allows not only using less resources in terms of computing time and power but also shows an improvement over the existing models for the Afan-Oromo NER task. Specifically, the main contribution of this paper is the use of two basic learning architectures in a hierarchical stack; the first model (BiLSTM) in the stack helps in estimating an initial NER output which is then fed to the second model in the stack to obtain an improved NER over the NER output given by the first model in the stack. Using this kind of hierarchical architecture, we show experimentally that there is an improvement in Afan-Oromo named entity recognition performance over the base bidirectional long-short term memory model by appending a small amount of network model parameters to the base BiLSTM model architecture. We believe that these kinds of modifications or integration of different network models help to improve Afan-Oromo NER performance especially in low resource conditions. Finally, we have contributed in adopting BiLSTM for AONER tasks and realizing the state-of-the-art results on the Afan-Oromo corpus without the need of feature engineering.

## 2 Related works

Most of the research works that have been presented in the past 20 years in named entity recognition systems cover both the supervised and unsupervised machine learning techniques with text feature engineering which are costly and time taking due to the manual rule designs. In the recent years, with the advent of deep learning approaches has contributed importantly to address this problem [13]. In recent years, deep learning models based on neural networks have made significant breakthroughs in performing various natural language processing tasks. Compared with traditional machine learning methods, the neural network model can automatically extract features and carry out end-to-end training; thus, the neural network model can achieve better results when performing NER. In [14] proposed a NER model based on a neural network. The model uses a CNN to extract features and fuse other linguistic features, such as part of speech tagging, even when only using word-level representation. While named entity recognition has rich in literature, it was not until 2016, several works for NER in Ethiopian languages saw prominence. With the advancement in the deep learning approach in proposed a method to categories Hindi language named entities in a given document corpus without language explicit rules. They used Bi-directional LSTM model to categorize words into its named entity class. But the major limitation in this work lies greedily tag decoding, which means that the input of the current stage needs the output of the previous stage. This technique may have a significant impact on the speed and parallelization. In [15] presented NER in Chinese clinical text using deep learning technique. In this work, authors combine knowledge-driven dictionary method with data driven deep learning technique for the Chinese clinical NER system using two various architectures to integrate feature vectors with character embedding's to conduct the task. In [16] develop a NER with context aware dictionary knowledge through combining the dictionary matching features with the hidden representation using

the LSTM-CRF technique. They used CoNLL-2003 dataset which is a widely used benchmark dataset for NER systems. Authors used an entity dictionary presented by which it is derived from the Wikipedia database which contains 297, 073, 139 named entities. In this work, we integrate various network models to improve Afan-Oromo named entity recognition performance particularly for the less resource rich language. We adopt BiLSTM for Afan-Oromo language NER tasks to confirm the state-of-the-art results in the language without the need of feature engineering.

**Table 1. Number of instances in each named entity class**

| Tag | Values |
|---|---|
| O | 10055 |
| I-ORG | 880 |
| I-LOC | 664 |
| B-ORG | 325 |
| I-PER | 255 |
| B-PER | 197 |
| B-LOC | 103 |

The dataset is collected from Ethiopian Broadcasting Corporation (EBC), Afan-Oromo program news articles annotated with the following possible named entity tags 'I-ORG', 'I-LOC', 'B-ORG', 'I-PER', 'B-PER', 'B-LOC', 'O'. Based on the classification schema B-Entity and I-Entity are used for the beginning and inside position of the named entity tag. Here 'B-Entity' is used whenever a new entity is beginning (regardless if it is longer than one word or not). 'I-Entity' is castoff for the words after 'B-Entity', when an entity spans over more than one word. The text is splitted word wise and every word is annotated with its POS and entity-tag. In total there are seven various token-level-tags, representing four classes.

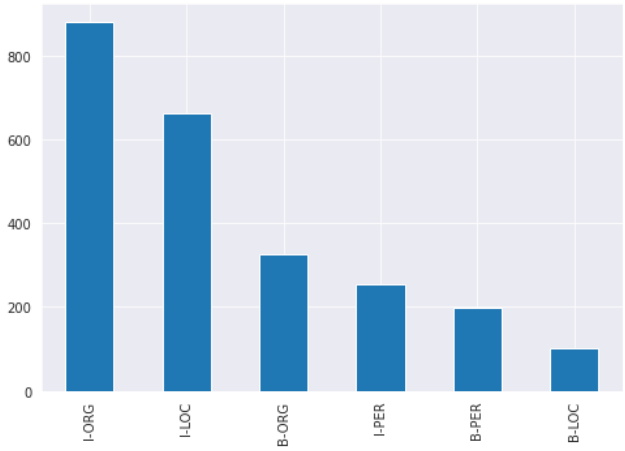


**Fig 1.** Non-O (named entity) tag distribution

For this experiment work, we have applied CoNLL"s BIO encoding structure for annotating the training and testing data. CoNLL has been [17] presented an allowed sequence of tags in which each named entity text should be tolerated by. The Parser will confirm and verify the data whether it is structurally annotated or not. For instance, in the following two tags generated for the following sentence "**Yunivarsitiin Adaamaa magaalaa Adaamaatti argama**." In English **("Adama University is found in Adama city")** the annotation structure based on the left part is legal while the annotation structure on the right part do not abide the rule of CoNLL 2002 encoding structure. Because, it has to be headed through B-LOC label or it has to be annotated as a B-LOC annotation tag.

Yunivarsitiin B-ORG Yunivarsitiin B-ORG
**Adaamaa I-ORG Adaamaa I-ORG**
Magaalaa O magaalaa O
**Adaamaatti B-LOC Adaamaatti I-LOC**
Argama O argama O
. O . O

Once the annotated training data is verified by the parser and everything is found to be correct, the data would be made ready for generating the tokens and the tags, token/tag sequence, which is done by the BIO encoder.

## 3 Bidirectional recurrent neural network

Bidirectional LSTM is a sequence processing model, which comprises two LSTMs: one taking the input in forward direction and the second in a backward direction [18]. BiLSTM can effectively maximize the amount of information available to the network, improving the context available to the algorithm [19] [20]. Learning character level information using bidirectional LSTM enables automatically extract the task specific information at character level without handcraft features like prefix and suffix of a token. Additionally, this has been found to be important to handle the vocabulary issues of NER Tasks. As depicted in Figure 2, a text corpus annotated with tags is used as an input for the new proposed AONER architecture. In our Afan-Oromo language NE recognition model we have conduct an experiment using three layers (embedding, bidirectional LSTM, time distributed with classifier). In embedding layer is used to put the maximum padded sequence then transfer words inputs into a vector of dimensions. Bidirectional LSTM layer takes results from the embedding layer. It coordinates the results through forward and backward before passing to the next layer by summarizing or taking the average. Time distributed layer earnings the output dimension from the previous layer then outputs the maximum tags and sequence length. Finally, softmax classifier recognize and classify the named entity tags into their corresponding class.

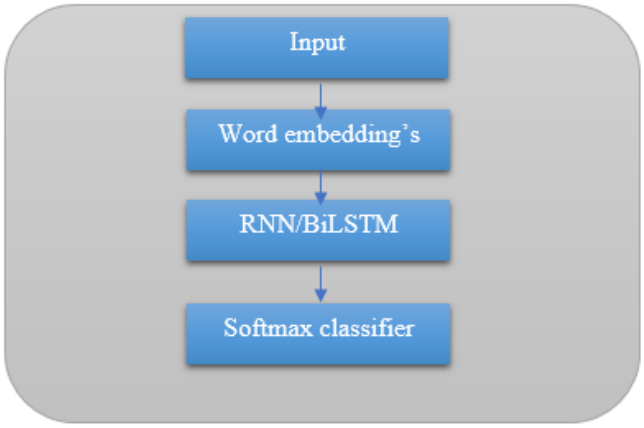


**Fig 2.** The proposed model

## 4 Experiment and discussion

In this work, for the experiment part we have used google Colaboratory [21] online platform to train our proposed model which have Jupyter Notebook to develop the NER system for Afan-Oromo language having pre-installed libraries like TensorFlow, and Keras. We have used Keras with TensorFlow backend which is an open source liberary written in python programming language. To evaluate the performance of the new AONER architecture and to show the impact of directionality and word embedding's, first we run our experiment using BiLSTM deep learning method, then with conditional random field (CRF). Notably, as we investigated in Table 2 below, the performance of BiLSTM approach is better realize than CRF approach. BiLSTM can handle the sequence labeling problem, effectively without requiring additional information.

In the following Figure 3, BiLSTM approach shows higher performance accuracy over CRF technique. As a part from the impact of word embedding's that is a powerful tool to learn the representation of tokens in the corpus and the capability to perform efficiently on various natural language processing tasks, word embedding improves the model and provide an accurate word representations. The embedding layer dimension is fixed to 64, the size of the hidden layer is aligned to 256. The synchronization of the forward and backward BiLSTM provided a dimension of 256 that is used as hidden activation function. Its result is feed into a softmax classifier output layer to generate probabilities for each four tags.

In this experiment, we have used two classifiers which shows a high performance accuracy with an average F1-measure results of 97%, 94% for BiLSTM and CRF respectively. BiLSTM classifier outperforms by 3% than random conditional field (CRF)

**Table 2. Results of BiLSTM and CRF Classifiers (scale 100%)**

| Model | Class | Precision | Recall | F1-Measure |
|---|---|---|---|---|
| BiLSTM | Organization | 0.96 | 0.95 | 0.94 |
| | Person | 0.98 | 0.97 | 0.98 |
| | Location | 0.94 | 0.95 | 0.98 |
| | Others | 0.99 | 0.98 | 1.00 |
| | Weighted averages | 0.967 | 0.962 | 0.973 |
| CRF | Organization | 0.91 | 0.93 | 0.95 |
| | Person | 0.96 | 0.96 | 0.95 |
| | Location | 0.91 | 0.97 | 0.96 |
| | Others | 0.94 | 0.92 | 0.91 |
| | Weighted averages | 0.932 | 0.942 | 0.946 |

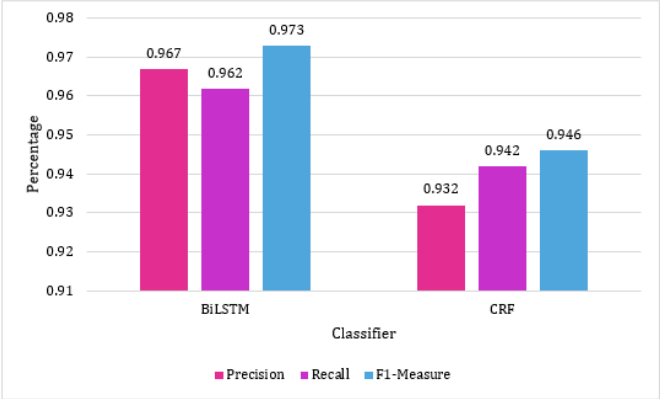


**Fig 3.** Classifier (BiLSTM and CRF) results

classifier. The lower in F1-measure might have three possible reasons. The first reason is due to the training and testing dataset is splitted used in this work discards 20% of the training data for testing the classifiers which highly impacts on the accuracy of the model. The other cause is that, deep learning techniques need large number of training data to provide a higher performance result, hence the dataset we used in this training experiment is not large enough. As a final reason the parameters used in the training network might not be in their improved value. Through changing the recurrent neuron cells of Bidirectional recurrent neural network, our model is trained within 100 iterations for the two techniques. For evaluation training and testing data is splitted into 80% of the text used for training and 20% text is used for testing. A total of 12,479 instances from organization, location, person and others is used as training data. Adam optimizer with categorical cross entropy objective function is used in this experiment. As a final point multilayer perceptron bidirectional neural network is developed to verify the regularity our experiment results. We used an input layer which have 120 neurons, a hidden layer enclosing four layers which have 120 neurons too with rectified linear unit activation and finally an output layer having 5 neurons with softmax classifier. To conclude our model computes the mean F1-measure for tokens in the corpus. The proposed model has enhancements which boost the identification efficiency and accuracy performance. To the best of our knowledge, in this study we contributed an Afan-Oromo NER system model using bidirectional long-short term memory by considering the NER problem in the language without using manual feature engineering and develop word embedding's which consists word and character that enables the model to make a good word representation in Afan-Oromo NER system. This work differs from other existing works for Afan-Oromo language is by the proposed model shits from the traditional machine learning techniques to deep neural network approach and bidirectional long-short term memory unit uses word and character embedding's an input.

## 5 Conclusion

A new Afan-Oromo NER architecture is boosted for identification and classification of named entities into their four predefined class including others (non-named entity class) which are person, organization, location and others. AONER is developed in

this study because of the language morphological ambiguity and writing style. Hence, natural language processing downstream task needs a large number of preprocessing and feature engineering steps. In this work we conduct an experiment using a bidirectional long-short term memory for AONER system. Without applying any manual feature engineering and other preprocessing steps, we try to address the issues of named entity recognition for the Afan-Oromo language corpus. We find that BiLSTM approach are very significant in identifying Afan-Oromo language named entities and can powerfully several other techniques which are based on manually engineered features and rule based systems. The integration of pre-trained text embedding's allows the system to gain considerable enhancements named entity recognition tasks and realize better results in F1-measures of 97.3% and 94.6% for bidirectional long-short term memory and conditional random field respectively. In general, BiLSTM outperforms, but there is still room to improve more with large Afan-Oromo language annotated corpus.

## References


1) Ngo QH, Kechadi T, Le-Khac NA. Domain Specific Entity Recognition With Semantic-Based Deep Learning Approach. *IEEE Access*. 2021;9:152892–152902. Available from: https://dx.doi.org/10.1109/access.2021.3128178.
2) Li J, Sun A, Han J, Li C. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*. 2022;34(1):50–70. Available from: https://dx.doi.org/10.1109/tkde.2020.2981314.
3) Le TA, Burtsev MS. A Deep Neural Network Model for the Task of Named Entity Recognition. *Social Psychology and Society*. 2019;9(1):8–13. doi:10.18178/ijmlc.2019.9.1.758.
4) Deng N, Fu H, Chen X. Named Entity Recognition of Traditional Chinese Medicine Patents Based on BiLSTM-CRF. *Wireless Communications and Mobile Computing*. 2021;2021(1):1–12. Available from: https://dx.doi.org/10.1155/2021/6696205.
5) Abafogi AA. Information Engineering and Electronic Business. *International Journal of Information Engineering and Electronic Business*. 2021;5:51–59. doi:10.5815/ijieeb.2021.05.05.
6) Alemayehu O, Fenet B. Review on gendered perspective of households participation in agricultural activities in Ethiopia. *Journal of Agricultural Extension and Rural Development*. 2019;11(1):1–10. Available from: https://dx.doi.org/10.5897/jaerd2018.0985.
7) Oromo Language (Afaan Oromoo) | Beekan Erena. 2021. Available from: https://scholar.harvard.edu/erena/oromo-language-afaan-oromoo.
8) Bazi IE, Laachfoubi N. Arabic named entity recognition using deep learning approach. *International Journal of Electrical and Computer Engineering (IJECE)*. 2019;9(3):2025–2025. Available from: https://dx.doi.org/10.11591/ijece.v9i3.pp2025-2032.
9) Gardie B, Asemie S, Azezew K. Anyuak Language Named Entity Recognition Using Deep Learning Approach. *Indian Journal of Science and Technology*. 2021;14(39):2998–3006. Available from: https://dx.doi.org/10.17485/ijst/v14i39.1163.
10) Huang W, Hu D, Deng Z, Nie J. EURASIP Journal on Image and Video Processing Named entity recognition for Chinese judgment documents based on BiLSTM and CRF. *Journal of Image Video Process*. 2020;2020:52–52. doi:10.1186/s13640-020-00539-x.
11) Wei H, Gao M, Zhou A, Chen F, Qu W, Wang C, et al. Named Entity Recognition From Biomedical Texts Using a Fusion Attention-Based BiLSTM-CRF. *IEEE Access*. 2019;7:73627–73636. Available from: https://dx.doi.org/10.1109/access.2019.2920734.
12) Zhang R, Zhao P, Guo W, Wang R, Lu W. Medical named entity recognition based on dilated convolutional neural network. *Cognitive Robotics*. 2022;2:13–20. Available from: https://dx.doi.org/10.1016/j.cogr.2021.11.002.
13) van Toledo C, van Dijk F, Spruit M. Dutch Named Entity Recognition and De-Identification Methods for the Human Resource Domain. *International Journal on Natural Language Computing*. 2020;9(6):23–34. Available from: https://dx.doi.org/10.5121/ijnlc.2020.9602.
14) Yadav V, Bethard S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. 2019. Available from: http://2016.bionlp-st.org/tasks/bb2.
15) Wang Q, Zhou Y, Ruan T, Gao D, Xia Y, He P. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. *Journal of Biomedical Informatics*. 2019;92:103133–103133. Available from: https://dx.doi.org/10.1016/j.jbi.2019.103133.
16) Wu C, Wu F, Qi T, Huang Y. Named Entity Recognition with Context-Aware Dictionary Knowledge. *Lecture Notes in Computer Science*. 2020;p. 129–143. doi:10.1007/978-3-030-63031-7_10.
17) Tjong EF, Sang K. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. 2002. Available from: https://aclanthology.org/W02-2024.
18) Cho M, Ha J, Park C, Park S. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. *Journal of Biomedical Informatics*. 2020;103:103381–103381. Available from: https://dx.doi.org/10.1016/j.jbi.2020.103381.
19) Complete Guide To Bidirectional LSTM (With Python Codes. 2022. Available from: https://analyticsindiamag.com/complete-guide-to-bidirectional-lstm-with-python-codes/.
20) Alhamid M. LSTM and Bidirectional LSTM for Regression. 2022. Available from: https://towardsdatascience.com/lstm-and-bidirectional-lstm-for-regression-4fddf910c655.
21) Welcome To Colaboratory - Colaboratory. 2022. Available from: https://colab.research.google.com/?utm_source=scs-index.