INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

OPEN ACCESS

Received: 24.06.2021
Accepted: 17.10.2021
Published: 27.11.2021

Citation: Gardie B, Asemie S, Azezew K (2021) Anyuak Language Named Entity Recognition Using Deep Learning Approach. Indian Journal of Science and Technology 14(39): 2998-3006. https://doi.org/10.17485/IJST/v14i39.1163

* **Corresponding author**.

birie16@gmail.com
simegnewasemie@gmail.com
azeze2912@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Anyuak Language Named Entity Recognition Using Deep Learning Approach

Birhanu Gardie[1]*, **Smegnew Asemie**[1], **Kassahun Azezew**[1]

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

Abstract

Objectives: This study aims about the development of Anyuak language named entity recognition of its first kind. NER is a fundamental sub task in natural language processing and the high accuracy competence in NER system marks the effectiveness of the downstream tasks. Anyuak language named entity recognition concern is addressed by using a long short-term memory model to categorize tokens into predefined classes. **Methods:** A long short-term memory is used to model the NER for Anyuak language to detect and classify words into five predefined classes: Person, Time, Organization, Location, and Others (non-named entity words). Because of feature selection plays a vital role in long short-term memory framework, the experiment in this work were conducted to discover most suitable features for Anyuak NER tagging task. **Findings:** When we evaluated the experiment in cross-validation, we achieved a promising result of precision, recall, and F1-measure values of 98%, 90, and 94% respectively. From the experimental result, it is possible to determine that tag context, word features, part of speech tags, suffixes and prefixes are significant features in named entity recognition and classification for Anyuak language. **Novelty:** Finally we have contributed a new architecture for Anyuak NER which uses automatically features for Anyuak named entity recognition which are not dependent on other NLP tasks. We proved that deep learning models can be extended, trained and can work for Anuak languages.

Keywords: Named entity recognition in Anyuak; Recurrent neural network; long shortterm memory; Natural language processing; and deep learning

1 Introduction

Anyuak commonly spelled as "Anywa" is a language that has its place in the Western Nilotic division language of the Nilotic language family. In Ethiopia "Nilotic" denotes the Nilo-Saharan languages and their populations. Primarily, it is spoken in the South West of Ethiopia specifically in Gambella regional state and adjacent border areas of South East Sudan by the Anyuak community[1,2]. The Anyuak community constitutes 27% of the total population living in Gambella regional state, Ethiopia, and scattered along the river banks of Baro and its tributary the Jajjabe[3,4]. According to the Ethiopian 2007 National Census, the Anyuak community was numbering more than 89,000, they are agrarians, fishers, and hunters based in the fertile area of Gambella, Ethiopia[5].

Named entity recognition is a sub-task of natural language processing in identifying and classifying named entities in a text document, which is a key component in NLP systems, especially in information retrieval, machine translation, automatic document summarization, and question-answering[6]. These named entities are predefined and denote words or word phrases such as organization names, location names, time names, and person names[7,8]. For instance, the sentence **Kwäärö mar Juuc Atut Gambëëla Omød Ojulu Aöö ya adï̈c abäba** ("The president of Gambella Omod will come to Addis Ababa") holds the semantic relationships **Kwäärö** ("PRESIDENT") between named entities **Omød Ojulu** (Omod Ojulu) (PERSON) and **Gambëëla** (Gambella) (LOCATION). Extractions of such named entities (NE) are fundamental for ontology development and information extraction. In named entity recognition the first step is identifying proper nouns from a text corpus and the next is categorizing the identified noun into one of the classes defined such as time, location, organization, and person[6,9]. Properly detecting named entities in a given text corpus is a significant step towards knowing and processing a document especially in the domain of formal documents and news reports[10].

Accurate recognition and classification of named entities are very essential and create challenges to natural language processing researchers. The ambiguity level in named entity recognition system makes it tough problem to attain human performance. This involved issue of accurate recognition of named entities specifically addressed by information extraction developers. Additionally, the task for named entity recognition is vary from language to language. For instance, named entities in English language starts with block letters as a result it makes simple to identify names written in the language. But no such kind of rule coming to Anyuak language, as a result, NER in Anyuak language is becoming tougher and challenging as compared to English language. For instance, in this sentence: **omään Ker aa Ithoopia.** This statement is equivalent to the English version: omään left from Ethiopia. Here**,** **omään** is name of a person but the first letter is not block and **Ker** is a verb in the sentence but the first letter written is in block letter as the language have its own structure. Moreover, most of the existing works in NER are conducted for well-formed text corpus like news articles and these systems are stated to work poorly on informal text types such as tweets[11]. NER system for the Anyuak language has not been conducted so far. Thus the development of such kind of system for Anyuak language is crucial in easing the development of natural language applications in Anyuak language and that is what this work is intended for.

In our study, we focused on four named entities from the data we collected and if an entity is out of the four NE, it is classified in the ”others” class. We also try to present the uniqueness of Anyuak texts and analysis of errors in the experimental evaluation results. In this study we used to apply recurrent neural network technique for the development of the named entity recognizer system for the Anyuak language due it can directly learn from raw data directly, and in turn do not need a domain expert to input features manually[12]. In named entity recognition tasks which uses sequential data labelling, long short term memory (LSTM) model is used that can remove the problem of limited context that uses in any feed-forward model[13]. An LSTM is a recurrent neural network that has its cell blocks which have different components such as input, forget and output gate.

2 Related works

This Anyuak language NER is the first work of its kind so far, a language with many speakers in Gambella, Ethiopia. Several efforts in named entity recognition research work applied machine learning approaches. A lot of research works has been conducted in named entity recognition but it has been focused on Western and Asian country languages, while African language has been given little concentration; however Amharic language NER has been conducted by Moges[7] and Dagmawi[14] using conditional random field and neural word embedding as a feature respectively. Moges mined random sentences that constitute at a minimum of one entity from the text document and annotated manually the NEs into four tags which are person, organization, location, and others(non-named entities; almost 90% of his dataset). He realized recall, precision, and F1-measure values of 75.0%, 74.2%, and 74.6% respectively, for this named entity recognition work. In his work, regard to features, authors used POS tags, suffixes and prefixes which are not sufficient enough for informal texts such as tweets and sentiments.

In[14]potential word feature information is denoted as word vectors by applying a neural network from an unstructured Amharic text and these created features are used as features for Amharic NER system. From his experiment result, he achieved a 95.5% F-score using a support vector machine classifier. In this study, word features are automatically learned for Amharic named entity recognition which can substitute manually designed features. That can give better recognition performance while minimizing manual feature design efforts. Word vectors can capture syntactic and contextual information relations for Amharic words but authors used small text corpus for word vector creation to build the model.

Mandefro[15] a NER for Afan Oromo language by using a hybrid approach that contains rule-based and machine learning techniques. He obtained from his experimental result 77.41% recall, 75.8% precision, and F1-measure of 76.6% in two scenarios. Authors present a rule generation technique so that the NER model can learn the structure of Afan Oromo language named entities. And the NER system authors described are dependent on the nature of features than on maximizing the training data size but features are not automatically learned that leads in efficient accuracy performance. A neural network forward feeds model is conducted to detect and classify named entities using a limited amount of contextual words by Collobert[16]. Author's also present feature engineered and knowledge based named entity recognizer system that would constitute in domain knowledge, orthographic and gazetteers using supervised and unsupervised techniques. This is dependent on word embedding and part of speech tagging's. However, Nichols and Chiu[17], in advance improved the performance of named entity classification using character and word embedding features using long short-term memory that can eliminate the need for feature engineering. Authors present a technique of encoding of partial lexicon mappings in neural networks but it should be improved using more flexible lexicon applications to improve the performance with large amount of data to learn complex semantics from large amount of dataset.

Huang[18]has presented an extra multifaceted technique that is built upon bi-directional and long short-term memory models. This model uses the past and future input features efficiently. They also present a sentence level tag information to extract features from contextual relationship using conditional random field.

But in this work, identification of NEs makes use of various features such word features, contextual tags, tokens of speech tags, prefixes and suffixes. The experiment in this study is to specify and discover most important word features for Anyuak NER tagging task. We have developed a new architecture for Anyuak named entity recognition which uses automatically generated features for named entity identification which are not dependent on other NLP tasks. Deep learning models can be extended, trained and can work for Anuak languages too.

3 Annotation: Named entity tagging

We have annotated the dataset with the present four tags; named entity tags of person, location, time, organization, and other classified categories. It was carried out by a person having linguistic knowledge and proficiency. The corpus is taken from Walta Information Center, Dictionary and grammar of Anywa, the pedagogical grammar of Anywa which is delivered by the region of Gambella, and from grade 6 Anywa student textbook having 1353 tokens but all these tokens are not annotated with NEs. Among this number of words, 655 tokens are not annotated to named entities; 698 tokens are annotated with NEs. From this amount of data we have splitted by 80/20% for training and testing purpose respectively.

3.1 POS tagging

The part of speech tagging receives input as an arrangement of tokens of a sentence, allots POS tags to the words of that sentence. It specifies the grammatical structure of a token based on the token itself and the surrounding semantics. Such grammatical structures result from the POS tagger are verb adjective, noun preposition, etc. In general, the normal order of tokens during sentence structure in the main clause follows the SOV format: Subject Object Verb. However, the resulting sentence structure may vary as shown in the following cases

1. Subject + Verb
   Abïï atal. (The cloth dried).
2. Subject + complement (object or adverbial) + verb
   Obäng öö Kiwane. (Obang Will come soon).
   Gadët Ker aa øt-jwøk. (Gadet left for church)
3. Complement + Verb
   Tiete ki cengnge alwøe. (He washed his foots and hands, too).

The order of various complements does not follow a special order. It deems that they are ordered according to their concept reasons.

Cengnge cabuun lwøge. /Lwøk Cengnge ki cabuun**.** Hand soap with wash (Wash your hand with soap!)

3.2 Noun phrase word order

The noun always comes before the qualifiers. Qualifiers are ordered in the sequence: noun-adjective- demonstrative pronoun. The subject marker is added to the noun itself and the qualifiers as the base construct are repeated in all qualifiers of the noun.

Ngate ni bäär icë ajote? (Did you see that tall man?)
Nyane ni kwaar icë mïërö döc. (That red girl is very beautiful).

3.3 Noun Category

Nouns are words that are used to name or identify any of the categories of things, people, places, or ideas or a particular of one of these entities. Based on its contextual position, nouns are commonly found at the beginning of a sentence in Anyuak. For instance, words that are italic in the following sentences are nouns.

Diel lum anyame**.** (The sheep grazes grass).

Mana kwaar be mana jiera**.** (Red is my preference).

A noun can further be classified as a common or proper noun. A proper noun is a noun that will name a specific, usually a one-of-a-kind item. In Anyuak, it begins with a capital letter no matter where it occurs in a sentence. Example:

Obäng a dipööy**.** (Obang is a teacher).

Opënnö ena Gambëlla. (Baro is found in Gambella region).

4 Anyuak named entity recognition in RNN

RNN is a neural network approach designed for analyzing data streams through hidden units. In some applications such as text processing, speech recognition, the output depends on the previous operations. We have collected the dataset from student text books of Anyuak language and from various people who can speak it reasonably well. We have prepared the training text documents in the first stage as we observe in Figure 1, we conducted text tokenization that would serve as reference to the original data. Tokenization is the task of chopping it up into pieces, perhaps at the same time removing certain characters such as punctuation. For instance we have following sentence: "**Mana kwaar be mana jiera",** can be tokenized into "Mana", "be", "Kwaar", "jiera". After performing annotation we conduct named entity annotation with their proper names based on their class category, and the annotation of the functional elements of the speech such as adjectives, nouns, adverbs, verbs etc. are performed, additionally text external source features are considered in the annotation. Then training of the annotated text dataset to the model of the Named entity recognition is conducted, finally classification of entities according to their proper class category have performed.

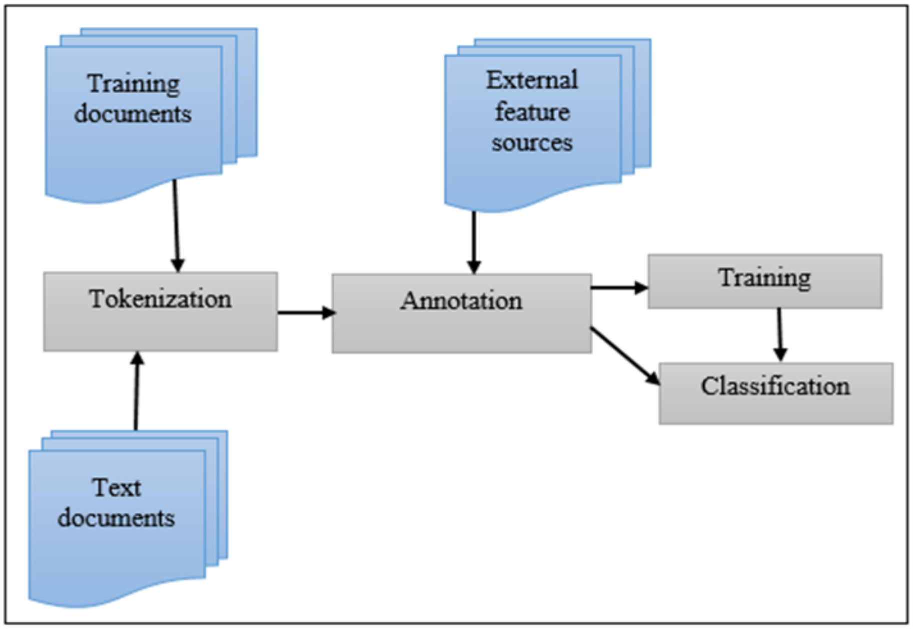
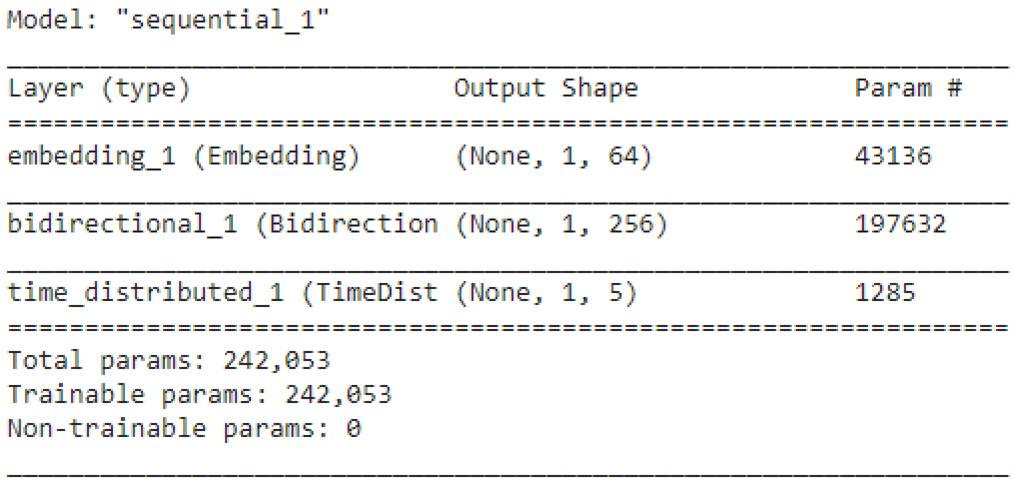


Fig 1. The Anyuak NER tagger pipeline

In our experiment for the named entity recognition of Anyuak language the model is summarized as in the following plot (Figure 2). In our NE recognition architecture we have worked on three layers (embedding, bidirectional LSTM, time distributed). Embedding layer is used to put the maximum padded sequence then transfer the tokens into a vector of n dimensions. Bidirectional LSTM layer takes results from the embedding layer. It synchronizes the results by forward and backward before passing to the following layer by summarizing or taking the average. Time distributed layer takes the output dimension from the previous layer then outputs the maximum tags and sequence length.



```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
embedding_1 (Embedding)      (None, 1, 64)             43136     
_________________________________________________________________
bidirectional_1 (Bidirection (None, 1, 256)            197632    
_________________________________________________________________
time_distributed_1 (TimeDist (None, 1, 5)              1285      
=================================================================
Total params: 242,053
Trainable params: 242,053
Non-trainable params: 0
_________________________________________________________________
```

Fig 2. Model summary of the experiment

5 Experimental results

For evaluation, the manually annotated document has to be created. For our relation extraction purpose, we prepared the annotated document with target relations. The general principle that we followed during evaluating the performance of the system is first we prepared manually annotated test data with target relation types. Then the test data is given to the proposed system and the system would predict the relations found between named entities. Then we manually check the output generated by the system against the corresponding manual tags and compute the performance. We prepared test data in a similar manner to the training data

5.1 Training phase

Once the text is parsed and everything is successful, it will be passed to the training phase. The main activity carried out in the training phase is BIO-Encoding which generates a word/tag from the parsed text. NE chunking creates a chunk from the order of words with a similar NE type. Model building estimates the model parameter values and generates a trained model. The non-named entity distribution is depicted as in the following figure. As we have observed from the following graph (Figure 3), it is the distribution of the named entities in the dataset used in our experiment. “PER” class have much entities relative to others where as “LOC”, and “ORG” classes have less entities in the dataset.

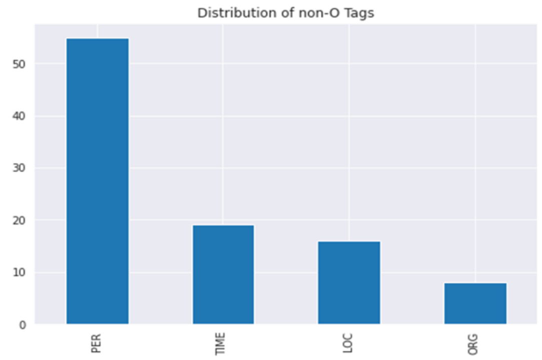


Fig 3. Non- O named entity-tag distribution

5.2 Results

We use Google Colaboratory to train our model which have a free Jupyter Notebook development environment with several pre-installed libraries such as TensorFlow, Keras. We used Keras with TensorFlow backend which is a free source CNN library written in python programming language. So as we trained the model with TensorFlow and Keras fit function and generated graphs of training and validation loss for each epoch value. We trained our work in different epoch parameters, the following figure result shows the training and validation accuracy done using 50 epoch size in the first scenario. The X-axis represents the number of epoch and the Y axis represents the accuracy of the algorithm in the given value.

To the best of our knowledge, we tried to develop the Anyuak language NER system which is the first work in the language using long short term memory techniques which can automatically extract features from words and sentences which is efficient in sequential labelling tasks in named entity recognition tasks. In(7),(14) works achieved good NEs identification using CRF machine learning techniques, in this approach feature engineering is challenging due it is not automatically extracted and it not sufficient in informal NE types. It needs extra knowledge to improve the performance accuracy. Work in(19) presents a Hindi named entity recognition using deep learning technique using pre- trained word embedding's to represent texts in a corpus and NER tags of the texts are determined by the applied annotated corpora to analyze word embedding's in the performance. In(20,21) introduced a comprehensive deep learning approach in NER system and claimed that it is far better NEs identifier than supervised and unsupervised techniques. But In our work, identification of NEs makes use of various features such word features, contextual tags, tokens of speech tags, prefixes and suffixes. Because of feature selection plays a vital role in long short-term memory framework, the experiment in this work were to determine and discover most suitable features for Anyuak NER tagging task.

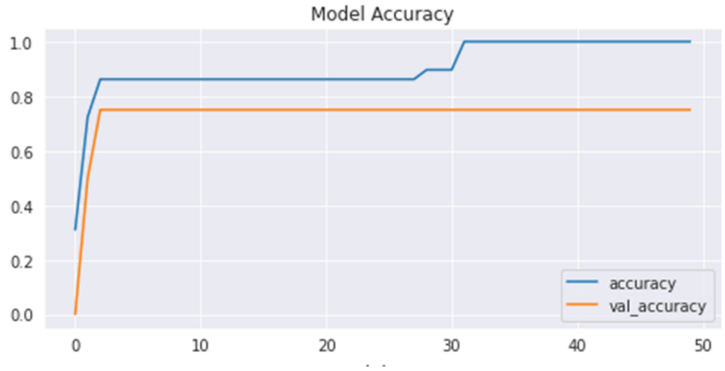


Fig 4. Training and validation accuracy

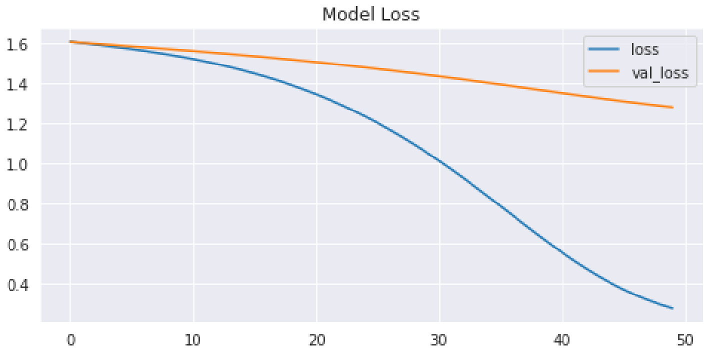


Fig 5. Training and validation loss

As depicted in the above two graphs (Figures 4 and 5), we can see the training and validation accuracy as it increases the validation accuracy were increased from 0.13-1, after this value the validation accuracy is almost by constant but have little changes. The model training accuracy increased from 0.3-0.86 constantly, there is accuracy improvement from 0.86 – 0.89. As shown in Figure 4 the training and validation accuracy have gaps and this comes from overfitting happened. In Figure 5, loss is improved from 0.30088 to 0.28043 and validation loss starts from 1.6 and improved to 1.13 in the specified epoch size. Here the validation loss and loss value have much higher gap and this is due to overfitting occurred which means the model has face difficult to generalize well. When we used the epoch parameters the training accuracy and validation accuracy value changes and it looks the following figure. Most of the time the training accuracy is higher than the validation accuracy.

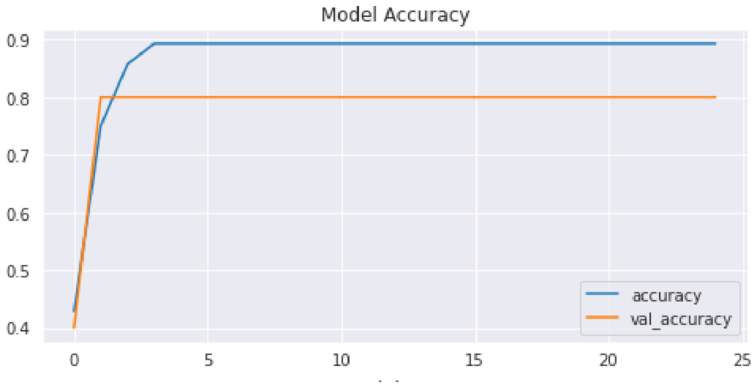


Fig 6. Training and validation accuracy using 25 Epochs

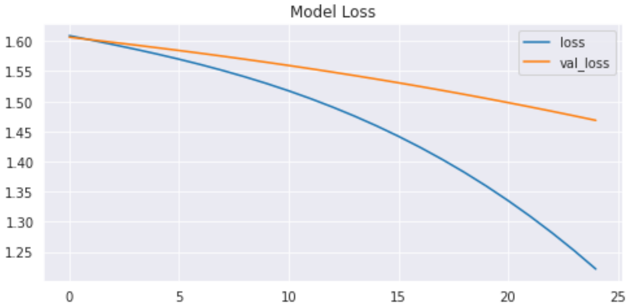


Fig 7. Training and validation loss using 25 epochs

In the above two Figures 6 and 7 which is tested in 25 epoch sizes, and the model accuracy started at 0.4 and reached at 0.89 whereas validation loss in Figure 7, started from 1.6 and reached at 1.4 and loss improved from loss improved from 1.24 to 1.21 in the specified epoch size. The higher gap in the first experiment with 25 epoch size is now improved which is addressed by applying dropouts. The following table presents the results of the named entity-tag of Anyuak which is of the first kind in the language. We have evaluated our result in the following performance evaluation metrics.

Table 1. Results from RNN and CRF classifier (scale 100%)

Classifier	Class	Precision	Recall	F1-measure
RNN	Location	0.00	0.00	0.00
	Person	1.00	0.80	0.89
	Others	0.97	1.00	0.98
	Accuracy			0.97
	Weighted Averages	0.97	0.97	0.97
CRF	Location	0.00	0.00	0.00
	Person	1.00	0.12	0.22
	Others	0.91	1.00	0.95
	Accuracy			0.91
	Weighted Averages	0.90	0.91	0.88

From the above Table 1, as we conduct the experiment in two different approaches, recurrent neural network and conditional random field. The results obtained from the experiments conducted shows the lower F-measures could be due to three possible reasons, the first one is the train test split used in these experiments discards 20% of training data for test. This affects the accuracy of the model. The other reason is that, deep neural networks need very large amount of data to give better

performance and the data set we used is not large enough.in fact the dataset we have used in the experiment is small data this is due to resource scarcity in the language. Finally the network parameters used in our experiments might not be in their optimized value. The accuracy obtained in the experiment is 97% in RNN and 91% in CRF. As we have observed in the table, precision, recall and f1-measure for location class is Zero due to the lack of named entities in the class. There is a lack of resources in the class, we have collected these data from student books in the language, and peoples who teaches the language in Gambella regional state, Ethiopia. Hence, there scarcity of the large amount of dataset resources for such kind of research investigation.

6 Conclusion

In this study, we tried to develop named entity recognition for the Anyuak language, which is the first kind of work in the language. It is an under-resourced language likewise other African languages. Several works in NER are conducted for well-formed text documents like news articles and these systems are stated to perform unwell on informal text types such as sentiments and tweets. The development of such kind of system for Anyuak language is crucial in easing the development of natural language applications in Anyuak language and that is what this work is intended for. In this work, we developed a part of speech tagging which is the first task in the named entity recognition work for the Anuak. Language autonomous feature is developed to extract Anyuak language named entities using long short term memory deep learning neural network model. We achieved a promising performance result after doing POS tagging and nurturing for training and NER to recurrent neural network identification and classification model. It would be well reasonable in the future to conduct research works in the language using a large corpus to develop language-specific features to enhance the performance. Additionally, it might be possible to develop and utilize word embedding's for the language.

References

1) Anuak language - Wikipedia. . Available from: https://en.wikipedia.org/wiki/Anuak_language.
2) Maraisa CH, da Silva Jose AG, Luiz CG, Emilio GA, de Oliveira Antonio C, de Carvalho Fernando IF. Correlations between chemistry components of caryopsis in oat genotypes cultivated in different environments. *African Journal of Agricultural Research*. 2015;10:4295–4305. Available from: https://dx.doi.org/10.5897/ajar2015.10079.
3) What is the meaning of Anuak, the name Anuak means, Anuak stands for. . Available from: https://thenamesdictionary.com/name-meanings/anuak/name-meaning-of-anuak.
4) Breaking the Cycle of Conflict in Gambella Region. 2003.
5) Anuak | Minority Rights Group. . Available from: https://minorityrights.org/minorities/anuak/.
6) Dou Y. Automatically Extracting Relations between Clinical Finding and Treatment from Clinical Texts. 2019.
7) Named Entity Recognition (NER) with keras and tensorflow | by Nasir Safdari | Towards Data Science. . Available from: https://towardsdatascience.com/named-entity-recognition-ner-meeting-industrys-requirement-by-applying-state-of-the-art-deep-698d2b3b4ede.
8) Named Entity Recognition (NER) with keras and tensorflow | by Nasir Safdari | Towards Data Science. . Available from: https://towardsdatascience.com/named-entity-recognition-ner-meeting-industrys-requirement-by-applying-state-of-the-art-deep-698d2b3b4ede.
9) Gamback B, Sikdar UK. Named entity recognition for Amharic using deep learning. *2017 IST-Africa Week Conference (IST-Africa)*. 2017. doi:10.23919/ISTAFRICA.2017.8102402.
10) Peng L, Gao D, Bai Y. A Study on Standardization of Security Evaluation Information for Chemical Processes Based on Deep Learning. *Processes*. 2021;9(5):832–832. Available from: https://dx.doi.org/10.3390/pr9050832.
11) Küçük D, Jacquet G, Steinberger R. Named entity recognition on Turkish tweets. *Proc 9th Int Conf Lang Resour Eval Lr*. 2014;p. 450–454.
12) 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.
13) Han X, Zhou F, Hao Z, Liu Q, Li Y, Qin Q. MAF-CNER:A Chinese Named Entity Recognition Model Based on Multifeature Adaptive Fusion. *Complexity*. 2021;2021(2):1–9. Available from: https://dx.doi.org/10.1155/2021/6696064.
14) Demissie D, Submitted T, Fulfillment P. Addis Ababa Institute of Technology School of Electrical and Computer Engineering Amharic Named Entity Recognition Using Neural Word Embedding as a Feature Amharic Named Entity Recognition Using Neural. . Available from: http://etd.aau.edu.et/handle/123456789/12386.
15) F. O. F. Computer and M. Sciences, "SCHOOL OF GRADUATE STUDIES FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES Named Entity Recognition for Afan Oromo Named Entity Recognition for Afan Oromo. 2010. Available from: http://etd.aau.edu.et/bitstream/handle/123456789/2440/Mandefro%20Legesse.pdf?sequence=1&isAllowed=y.
16) Collobert R, Weston J, Com J, Karlen M, Kavukcuoglu K, Kuksa P. Natural Language Processing (Almost) from Scratch. 2011. doi:10.5555/1953048.2078186.
17) Chiu JPC, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*. 2016;4:357–370. Available from: https://dx.doi.org/10.1162/tacl_a_00104.
18) Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging. *arxiv*. 2015. Available from: https://arxiv.org/abs/1508.01991.
19) Shah B, Kopparapu SK. A Deep Learning approach for Hindi Named Entity Recognition. *arxiv*. 2019. Available from: https://arxiv.org/abs/1911.01421.

20) Yadav V, Bethard S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. *arxiv*. 2019. Available from: https://arxiv.org/abs/1910.11470.

21) Benikova D, Yimam SM, Santhanam P, Biemann C. GermaNER : Free Open German Named Entity Recognition Tool. 2010;1:31–38. Available from: https://www.semanticscholar.org/paper/GermaNER%3A-Free-Open-German-Named-Entity-Recognition-Benikova-Yimam/f6f9cfbff9b77e3b43bc3c025859c3870d32fb61.