INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Named Entity Recognition for Hadiyya Language using BiLSTM-CRF Model

Desalegn Ashebir[1], **Getahun Tadesse**[2]*

1 Department of Software Engineering, Mizan Tepi University, Ethiopia
2 Department of Computer Science, Mizan Tepi University, Ethiopia

OPEN ACCESS

Received: 19-05-2022

Accepted: 10-09-2022

Published: 21-12-2022

Citation: Ashebir D, Tadesse G (2022) Named Entity Recognition for Hadiyya Language using BiLSTM-CRF Model. Indian Journal of Science and Technology 15(47): 2612-2618. https://doi.org/10.17485/IJST/v15i47.1090

* **Corresponding author**.

getahun@mtu.edu.et

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: This work aims to the development of Hadiyya language named entity recognition which is widely used in text summarization, machine translation, and information retrieval to categorizing and predicting tokens of a given corpus into predefined named entity classes **Method :** In this paper, a method combining Bidirectional Long Short-Term Memory neural network with Conditional Random Field (BiLSTM-CRF) is proposed to automatically recognize entities of Hadiyya language (Location, time, person, geography and other non-name entity) from annotated Hadiyya language corpus, the experiment in this work was conducted to discover the most suitable features for Hadiyya NER system. We have collected the data from Department of Hadiya Language & Literature (DHLL) at Wachemo University, Ethiopia. Hadiyya TV, and Hadiyya Media Network (HMN) Therefore, a newly annotated dataset having 5,148 instances is used for this study. We have used 70 % for training and 30% for testing Hadiyya NER system. **Finding:** after training and validating BiLSTM-CRF model using the collected dataset we have obtained a result of precision, recall and f1-measure values of 95.49%, 94.93%, and 95.21% respectively. **Novelty**: Finally, we have contributed by hybrid NER system in Hadiyya language to obtain state-of-the-art result which is independent of other natural language processing tasks.

Keywords: Conditional Random Forest; Hadiyya Language; Long Short-Term Memory; Hadiyya Media Network

1 Introduction

Named Entity Recognition (NER) is frequently utilized in downstream applications of Natural Language Processing (NLP) and artificial intelligence in the 21st century such as text summarization, machine translation, and information retrieval[1]. With the development of AI technology, time and labor costs have significantly reduced the recognition of name entity recognition NER[2,3]. The research fields of named entities include journalism, biology, and medicine. Each domain has its characteristics. In the language area, the identification of named entities in language can be extracted from text corpus according to semantic types such as Person, Place, Time, Organization, politics, and Geography[1,3]. Named Entity Recognition (NER) is a complex sequence labeling task that requires a comprehension understanding of the orthographic and distributional representation of words. For example, in the Hadiyya language statement **“Doktor Aabiyyi Ahmed Bax quuxo daadeeshi na kaagain lasaan chonne Eertiira maroohane”** (doctor abiye Ahmed will visit Eertiira at the end of this month) holds the name entity of “**Aabiyyi**” (person), “**maroohane**” (Time), “**Eertiira**” (location,). “**Lasaan**” (Time). Various researches in Name entity recognition conducted on rich-resource languages like English, Afaan Oromo, Amharic, wolaita etc. however, Hadiyya languages NER research is almost an initial first kindly. In recent years, Hadiyya NER (HNER) systems has become a challenging task and is receiving an increasing attention for the development of the language from recent researchers due to the limited availability of annotated dataset[(4)]. Hadiyya language is one of the most languages which have a large number of speakers in Ethiopia[(4)].

Hadiyya is the language of the Hadiya people of Ethiopia. It is a Highland East Cushitic language of the Afro-Asiatic family. Most speakers live in the Hadiya Zone of the Southern Nations, Nationalities, and Peoples of Ethiopia Region[(4)], Hadiyya language is known by its speakers as Hadiyyisa or Hadiyyisa, It has 1.4 million total speakers and is a mother tongue to 1.25 million of them according to 2007 census of Ethiopia with 600,000 monolingual speakers[(4)]. It is one of the languages in Afro-Asiatic language family predominantly spoken in West Asia/Middle East, Horn of Africa, North Africa, and parts of the Sahel. Hadiyya Language written using the Latin alphabet[(4)].

Traditional machine learning requires the statistics, handcrafted feature extraction and analysis of readers to dig out features that impact the task. Neural network models have been employed in many natural language processing jobs in recent years as computer technology has advanced. Neural network models do not rely on feature engineering, reducing time and labor expenses. The expression of word vectors has provided named Entities with a significant development impetus. Neural network model that allows word vectors to collect more semantic information is regularly updated, and the representation of word vectors can express more semantic information than manually derived features[(5)].

The Bi-LSTM model and the CRF model are two popular NER models. The Bi-LSTM model eliminates the problem of long-term reliance, allows the model to learn more distant data, and allows it to collect contextual data. The CRF model not only uses internal information but also uses contextual information to mark a location. The main contribution of this paper is the combination of two models Bi-LSTM-CRF in Hadiyya language name entity recognition. the Bi-LSTM-CRF model combines the advantages of the two models(concatenating the output of forward and backward LSTM) and is also a mainstream model for NER[(5)]. The Bi-LSTM model is used to learn information from previous and future time stamps. The BiLSTM output is then fed into a linear chain CRF, which can generate predictions using this improved context to ensure that the final prediction result is valid, the Bi-LSTM layer can include some constraints. During training data, the CRF layer can learn these limitations automatically[(2)]. We believe that this combination or integration of these two models helps to improve the performance of Hadiyya language entity name recognition.

1.1 Related works

Most of the named entity recognition research presented in the last 20 years has included both supervised and unsupervised machine learning techniques with text handcrafted feature extraction, which is costly and time-consuming due to human rule designs. However currently within the development and high contribution of deep learning some entity name recognition problems has address more effective ways[(6–8)]. Deep learning models performed better by removing manual feature extraction and given state of the art result[(9)].

In[(2)] Proposed Afan-Oromo named entity recognition using Bidirectional LSTM to classify into 7 different classes, they used 12,479 instances to recognize and classify. After training and validating the model they achieved a result of precision, recall, and f1-measure values of 96.7%, 96.2%, and 97.3% respectively. In this study double Lstm cell is costly, takes longer to train, requires more memory to train, and also easier to overfitting and what the gap is It has control over deciding when to let the input enter the neuron. It has control over deciding when to remember what was computed in the previous time.

In[(3)] Proposed Afaan Oromo entity name recognition by hybrid machine learning and pattern matching approaches. They used a dataset of 44,120 words out of which around 7809 are NEs. To tian and evaluate the performance of the hybrid model they distributed datasets are into 90% for training and 10% for testing, finally they achieved Precision, Recall, and F1-measure that is 86.37, 85.66, and 86.01 respectively. In this paper they didn’t fairly distribute dataset this leads to dataset imbalanced in each entity class and finally extracted learnable features based on manual feature engineering this leads to fill local minima in testing of the model as we have seen the result.

In[(8)] Proposed a bidirectional LSTM technique for predicting positive or negative sentiment in Arabic text. The model’s performance was evaluated using six benchmark datasets and a total of 61582 datasets. Finally, they compared the model to CNN and LSTM techniques and discovered that the average sentiment prediction accuracy in Arabic text across the six datasets was 83. %.

In[10] proposed an RNN approach to identify name entity in Anyuak language to categorize into 4 entity and they achieved a result of precision, recall, and F1-measure values of 98%, 90, and 94% respectively, however they applied a single LSTM layer when they train the model, and this model only preserves information of the past because the only inputs it has seen are from the past not future so learnable feature may not extracted effectively in this case and the author didn't show how to solve fixed sequence to sequence case and if the input and output have the same size. But when we apply bidirectional LSTM the input run in two ways, one from past to future and one from future to past. That means every component of an input sequence has information from the past and present.

In[11] Proposed an approach for Hindi name entity recognition using a deep learning approach. They used a dataset available from (IITH, 2008) that are prepared during ICJNLP 2008 workshop in South and southeast Asian Languages. It consists of 19822 annotated sentences, 34193 unique tokens, 490368 total tokens and 12 categories of entities, and one negative entity class other. Finally, they achieved a recognition test accuracy of 73%

In[12] Proposed Deep learning approach to show NER in Chinese clinical literature. In this paper, the authors use two different architectures to integrate feature vectors with character embedding to conduct the task using a knowledge-driven dictionary method and a data-driven deep learning strategy for the Chinese clinical NER system.

2 Methodology

2.1 Name Entity Tagging

We have annotated the dataset into five tags: Organization, Person, Location, Time and other non-entity name tags, it has been done by linguist with extensive knowledge and experience. The Parts of Speech (POS) tagging method takes tokens from a phrase as input and assigns tags to the words in that sentence.

Bidirectional LSTM is a sequence processing model that consists of two LSTMs: one that takes the input ahead and the other that takes it backward LSTM is to save previously entered information or effectively maximize the amount of information available to the architecture[9,13]. BiLSTM is aware of the context information in the word sequence, and CRF aids in improving sentence labeling accuracy. The total recognition accuracy will be increased by combining the advantages of BiLSTM and CRF[14] in this paper we proposed a Bi-LSTM-CRF model for Hadiyya Language name entity recognition. Our proposed model can be divided into the BiLSTM layer and the CRF layer. The function of the Bi-LSTM layer is used to extract contextual information through the input words and word vectors determine the probability of a certain type of entity making the prediction. CRF layer is used to consider the correlation between tags. The process for recognizing Hadiyya language entities using the BiLSTM-CRF model structure is shown in Figure 2. The following are the primary steps:

2.2 Input layer

At the present time the input in most NER models are words, so, the effect of entity recognition depends heavily on the effect of word segmentation, instead of characters, which contain a lot of linguistic information. Studies have revealed that word-based NER generally performs better than character-based methods[15]. In this paper A sentence containing n-words is one-hot encoded as an n × v-dimension matrix, denoted as W = (w1, w2, $\cdots$, wn) where w i represents the vector of the Ith word of the sentence input into the input layers.

2.3 BiLSTM Layer

This layer is used to extract sentence features. As shown in Figure 2, every word vector Xi= (x1, x2, $\cdots$, xn) is taken as the input of the BiLSTM Layer in both the forward and backward direction

2.4 CRF Layer

This layer carries out sentence-level sequence labeling to ensure the generation of the globally optimal labeling sequence, the labels follow strict internal syntax, and this is extremely easy for the CRF to learn. For NER, there are several ways of encoding the output, but they typically encode at least: Beginning, inside, and Outside of an entity in this Hadiyya name entity recognition, and these can only be in a syntactically well-defined order. CRF will very quickly catch[1].

Generally, BiLSTM might be unsure if it should place on a position or one position later and end up outputting both of them because they are conditionally independent. CRF layer that knows that this is unlikely and enforces the internal logic of the tags and would output this highly improve the performance of Hadiyya language name entity recognition.

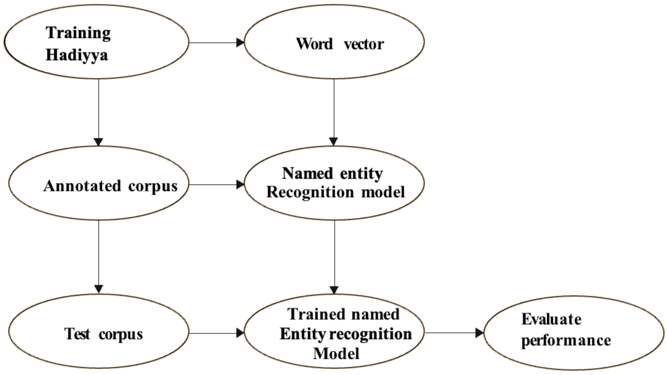

Fig 1. Flow chart of Hadiyya Language NER experiment

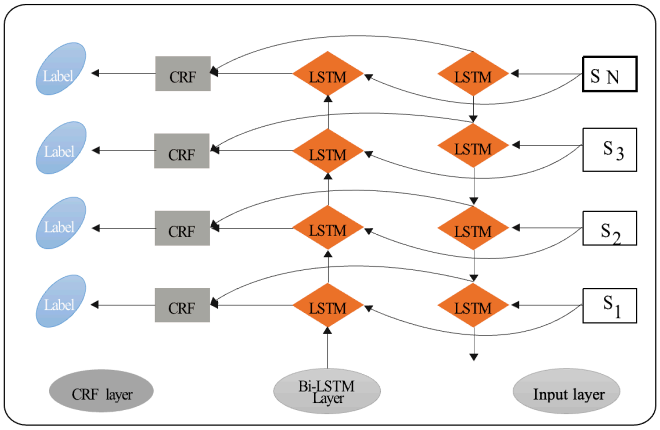

Fig 2. Bi-LSTM-CRF model structure in Hadiyya Name entity recognition

3 Result and Discussion

We have collected the data from Department of Hadiya Language & Literature (DHLL) at Wachemo University, Ethiopia. Hadiyya TV, and Hadiyya Media Network Therefore, a newly annotated dataset having 5,148 instances is used for this study as clearly show Table 1. We have used 70 % for training and 30% for testing Hadiyya NER system. The proposed model is implemented with Keras (Tensor Flow as a backend) using Python programming language. We have selected different hyper-parameter using empirical optimization, after train and validate the model empirically we have achieved state-of-the-art result on the following parameter: epochs at 20, and recurrent dropout of 0.5, initial learning rate 0.001, optimizer Adam, batch size 20, classifier SoftMax.

Table 1. Dataset distribution belong to 5 entity name class

Data [‘Tag’]	Value counts [‘no words’]
Other(O)	4504
Person (PER)	376
Time (TIME)	133
Location (LOC)	112
Organization (ORG)	59
Total word count	5148 from 232 sentences

As clearly show Figure 3, the validation loss and validation accuracy do not oscillate with the training loss and training accuracy at each epoch. However, it demonstrates that our model is not overfitting because it learns features well at training and is well generalized at testing: the validation loss is falling rather than increasing, and the difference between training accuracy,

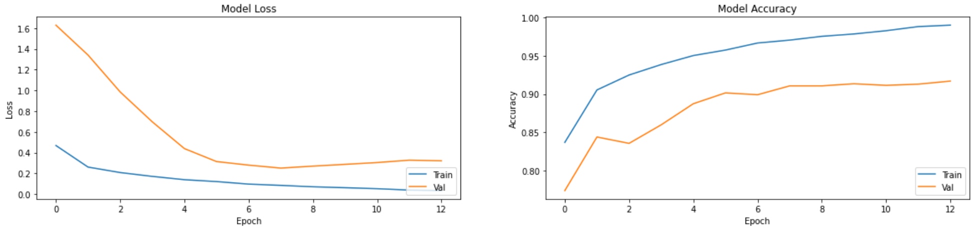

Fig 3. Hadiyya language entity name recognition BiLSTM-CRF model curve

validation accuracy, and validation loss is small. As a result, we can state that our model's generalization capability improved significantly, the loss on the validation loss was only marginally higher than the training loss.

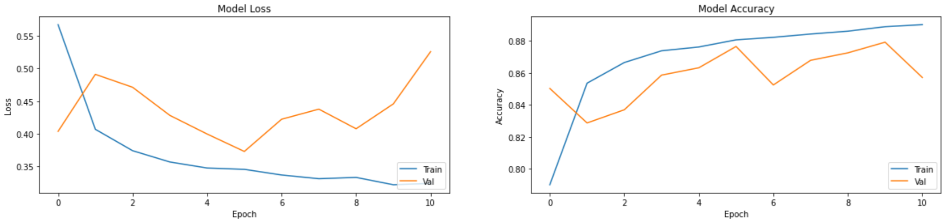

Fig 4. Hadiyya language entity name recognition CRF model curve

As clearly show Figure 4 the validation accuracy nearly stagnated after 1-3 epochs and only rarely rose at specific epochs. The validation accuracy increased linearly with loss at first, but it did not rise much after that. The validation loss indicates overfitting; it declined linearly, comparable to validation accuracy, but then began to climb after 4-5 epochs. This indicates that the model was not successful memorizing the data during training and not well generalize in testing.

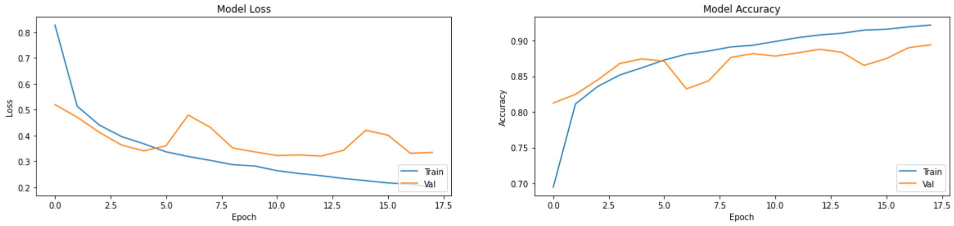

Fig 5. Hadiyya language entity name recognition LSTM model curve

As clearly shown Figure 5 training and validation accuracy increases while training and validation loss decreases nearly linearly up to epoch 5. Between epoch 14 and 15 training accuracies increase validation accuracies decrease too much and runs constantly while training and validation loss increases and run constantly. These two signs are typical indication of overfitting, this case happen due to small number of dataset in different class entity.

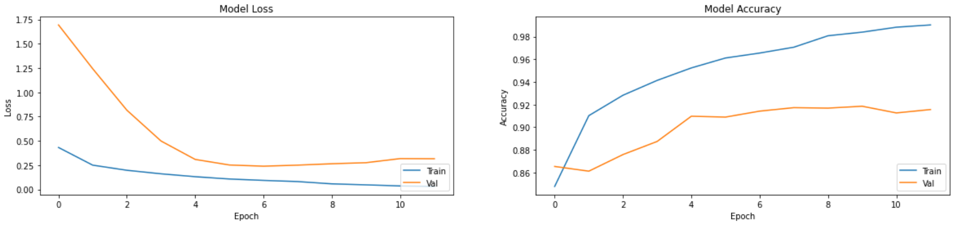

Fig 6. Hadiyya language entity name recognition BiLSTM model curve

As shown in the training loss and accuracy curve in Figure 6, training accuracy is greater than validation accuracy throughout the curve. However, the gap between the training accuracy and validation accuracy is lower as compared to Figures 5 and 6, In addition, the training loss is much smaller than validation loss throughout the curve. This show that overfitting decrease in some instants. That indicates that the model was successful in memorizing the data.

Table 2. Result of BiLSTM-CRF for each class entity Hadiyyaname entity recognition

Model	Class	Precision (%)	Recall (%)	F1 (%)
BiLSTM-CRF	O	99.38	98.56	99.56
	PER	97.56	96.61	97.08
	TIME	94.48	96.78	95.63
	LOC	94.28	92.86	93.57
	ORG	91.75	89.84	90.79

Table 3. Comparison of the selected model on Hadiyya nameentity recognition

Model	Precision (%)	Recall (%)	F1 (%)
CRF	87.25	85.76	86.49
LSTM	88.34	87.61	87.97
BiLSTM	90.45	89.72	90.08
BiLSTM-CRF	95.49	94.93	95.21

As clearly shown in Table 2, the proposed BiLSTM-CRF model classifier performed well in four entity class (location, time, person, other non-name entity class) but minimum precision, recall and F1 measure in organization, the reason are due to small number of tag words in this class but deep learning approach needs large amount of data to train and validate the model, even if the dataset little small BiLSTM takes into account the present input as well as (some) previous inputs, with each previous input being given this improve the performance of the model.

As clearly shown in Table 3, Bi-LSTM-CRF Although the SoftMax layer can extract the NER result from the context information generated by the BiLSTM layer, the result received directly through the BiLSTM layer only considers the context information. The dependencies between tags are not taken into account in the Bi-LSTM layer's output result this improve the performance of this model. To evaluate the dependency link between markers, the CRF model can learn some global based constraint information through corpus training. As a result, the Bi-LSTM-CRF model is used. The Bi-LSTM layer can be used to extract text context information in order to forecast the label, and the CRF layer can be used to add constraint rules to verify that the final recognition result is reasonable, by introducing relevant theories and experimental findings, The following are the benefits of using a legally NER approach based on a word-level neural network: (1) Compared to the traditional method, the deep learning method avoids artificial feature engineering design and solves the dimensional disaster problem caused by sparse data in the traditional method; (2) the model obtains contextual information using the Bi-LSTM-CRF model, which solves the long-distance dependence problem of ordinary models.

Table 4. Comparison of the proposed model to related works

Reference	**Model**	**Datasets**	**Test Accuracy (%)**
(2)	Bidirectional LSTM	They used a newly annotated dataset of 12,479 words used belongs to 7 class	a result of precision, recall, and f1-measure values 96.7%, 96.2% and 97.3% respectively
(3)	They used hybrid machine learning and pattern matching approaches	A dataset of size 44,120 words out of which around 7809 are NEs that were collected from 3 news websites.	A result of Precision, Recall, and F1-measure are 86.37, 85.66, and 86.01 respectively
(8)	Bidirectional LSTM	evaluated using six benchmark datasets and a total of 61582 datasets	83.15 test accuracy across six benchmark datasets
(10)	RNN approach	They didn't clearly show the number of datasets used in the experiment to classify into four class and other class	A result of precision, recall, and F1-measure values of 98%, 90, and 94% respectively
(11)	Deep learning approach	They used 490368 total tokens and 12 categories of entities and other one negative entity class	Finally, they achieved a recognition test accuracy of 73%

Continued on next page

Table 4 continued

Hadiyya Language ENR	BiLSTM-CRF	We have used 5148 total words belonging to five class	Finally, we achieved a result of precision, recall and f1-measure values of 95.49%, 94.93%, and 95.21% respectively

4 Conclusion

Named Entity Recognition is an important NLP task that has not been explored before for Hadiyya language its first kind. In this paper, we used hybrid BiLSTM-CRF architecture by reducing the size of the network to accommodate the lack of annotated data (low-resource). To achieve our aims, data had been collected from Department of Hadiya Language & Literature (DHLL) at Wachemo University, Ethiopia, Hadiyya media network (HMN), Hadiyya TV. Our corpus contains 5148 words belongs 5 name entities class. We have used a hybrid BiLSTM-CRF model. Finally, we achieved a result of precision, recall, and f1-measure values of 95.49%, 94.93%, and 95.21% respectively the model might be incorporated in advanced language-processing applications like Machine translation, information retrieval, and opinion mining for the language. Not just this, but the model also had great significance for mass media those using the language by extracting NE from unstructured text inputs. 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 Attention-based BiLSTM-CRF model.

References

1) Patil N, Patil A, Pawar BV. Named Entity Recognition using Conditional Random Fields. *Procedia Computer Science*. 2020;167:1181–1188. Available from: https://doi.org/10.1016/j.procs.2020.03.431.
2) Gardie B, Solomon Z. Afan-Oromo Named Entity Recognition Using Bidirectional RNN. *Indian Journal of Science and Technology*. 2022;15(16):736–741. Available from: https://doi.org/10.17485/IJST/v15i16.123.
3) Abafogi A. Boosting Afaan Oromo Named Entity Recognition with Multiple Methods. *Int J Inf Eng Electron Bus*. 2021;13(5):51–59. Available from: http://dx.doi.org/10.5815/ijieeb.2021.05.05.
4) Hadiyya. Hadiyya (Hadiyyisa) Language Orthography - Alphabet and Writing - Themes on the Hadiya People of Ethiopia. . Available from: https://hadiyajourney.com/state-of-hadiyya-hadiyyisa-language-of-ethiopia/.
5) Xu H, Hu B. Legal Text Recognition Using LSTM-CRF Deep Learning Model. *Computational Intelligence and Neuroscience*. 2022;2022:1–10. Available from: https://doi.org/10.1155/2022/9933929.
6) 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://doi.org/10.1155/2021/6696205.
7) 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://doi:10.1109/ACCESS.2019.2920734.
8) Elfaik H, Nfaoui EH. Deep Bidirectional LSTM Network Learning-Based Sentiment Analysis for Arabic Text. *Journal of Intelligent Systems*. 2020;30(1):395–412. Available from: https://doi.org/10.1515/jisys-2020-0021.
9) Huang W, Hu D, Deng Z, Nie J. Named entity recognition for Chinese judgment documents based on BiLSTM and CRF. *EURASIP Journal on Image and Video Processing*. 2020;2020. Available from: https://doi.org/10.1186/s13640-020-00539-x.
10) 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://doi.org/10.17485/IJST/v14i39.1163.
11) Shah B, Kopparapu SK. A Deep Learning approach for Hindi Named Entity Recognition. 2019. Available from: http://arxiv.org/abs/1911.01421.
12) Wu G, Tang G, Wang Z, Zhang Z, Wang Z. An Attention-Based BiLSTM-CRF Model for Chinese Clinic Named Entity Recognition. *IEEE Access*. 2019;7:113942–113949. Available from: https://doi.org/10.1109/ACCESS.2019.2935223.
13) 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. Available from: https://doi.org/10.1016/j.jbi.2020.103381.
14) 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://doi.org/10.1155/2021/6696064.
15) Muralikrishna H, Sapra P, Jain A, Dinesh DA. Spoken Language Identification Using Bidirectional LSTM Based LID Sequential Senones. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2019;p. 320–326. Available from: https://doi:10.1109/ASRU46091.2019.9003947.