Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881; IMPACT FACTOR :8.017(2022); IC VALUE:5.16; ISI VALUE:2.286
Peer Reviewed and Refereed Journal: VOLUME:11, ISSUE:11(5), November: 2022
Online Copy of Article Publication Available (2022 Issues)
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: 2nd November 2022
Publication Date:10th December 2022
Publisher: Sucharitha Publication, India
Digital Certificate of Publication: www.ijmer.in/pdf/e-CertificateofPublication-IJMER.pdf

DOI: http://ijmer.in.doi./2022/11.11.83
www.ijmer.in

LSTM (LONG SHORT-TERM MEMORY) BASED NAME ENTITY RECOGNITION

[1]Madhu Jain and [2]Preeti Sodhi
[1]M. Tech, Scholar and [2]Assistant Professor
[1&2]Universal Group of Institutions
Lalru, Punjab, India

Abstract
A named entity refers to anything that has to do with a name. Named entity recognition is a method of identifying and classifying people by their names. NER is a key subtask of Information Extraction. In a number of domains of knowledge and science, information extraction, question answering, machine translation, automatic document indexing, cross-lingual information retrieval, text summarization, and other NER applications can be found and observed. We use the deep learning approach LSTM to build Named-Entity Recognition (NER) for Indian Language in this study.

Keywords: NLP, name entity recognition, deep learning, LSTM etc.

Introduction

Natural Language Processing (NLP) is a computerised method to text analysis that is founded on a set of theories as well as a set of technology. We detect and classify proper names in text into specified categories in Named Entity Recognition, a subtask of Information Extraction. The following are possible classifications for the named entities:

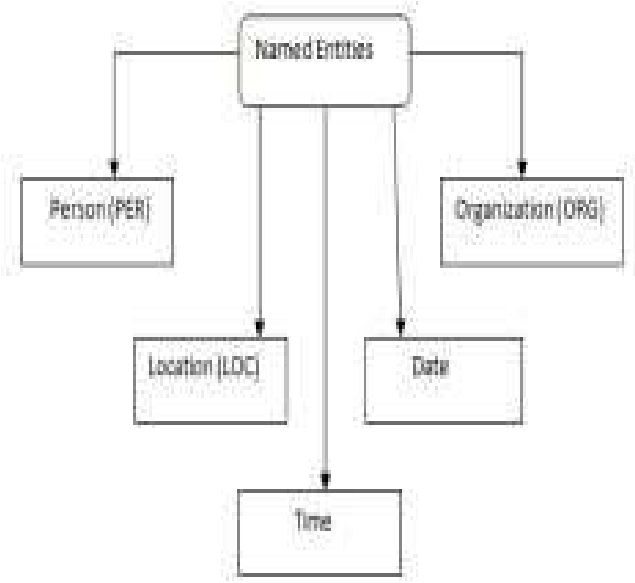


Fig 1: Name entity classification

Many natural language processing tasks, such as machine translation, more accurate internet search engines, automatic document indexing, automatic question-answering information retrieval, and so on, rely on NER. For these purposes, a precise NER system is required.

In the previous few decades, significant progress has been achieved in the field of named entity recognition (NER) for text documents, with outstanding results. NER is often approached as a two-step process including the identification of relevant words and their classification. The first step is to recognise acceptable nouns in the text, and the second is to classify these proper nouns into one of several categories, such as person name, organisation name, geographical name, and others. The fundamental issue with NER is determining how to tag words and which tags should be allocated to entities like as people, places, and things. When there are

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881; IMPACT FACTOR :8.017(2022); IC VALUE:5.16; ISI VALUE:2.286
Peer Reviewed and Refereed Journal: VOLUME:11, ISSUE:11(5), November: 2022
Online Copy of Article Publication Available (2022 Issues)
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: 2nd November 2022
Publication Date:10th December 2022
Publisher: Sucharitha Publication, India
Digital Certificate of Publication: www.ijmer.in/pdf/e-CertificateofPublication-IJMER.pdf

DOI: http://ijmer.in.doi./2022/11.11.83
www.ijmer.in

ambiguities in a document, we must take them into consideration. Sometimes there are ambiguities in the document, and we must deal with them in order to provide the right tag.

Applications of NER

NER is used in the majority of NLP applications. A handful of its applications are highlighted in the following list.

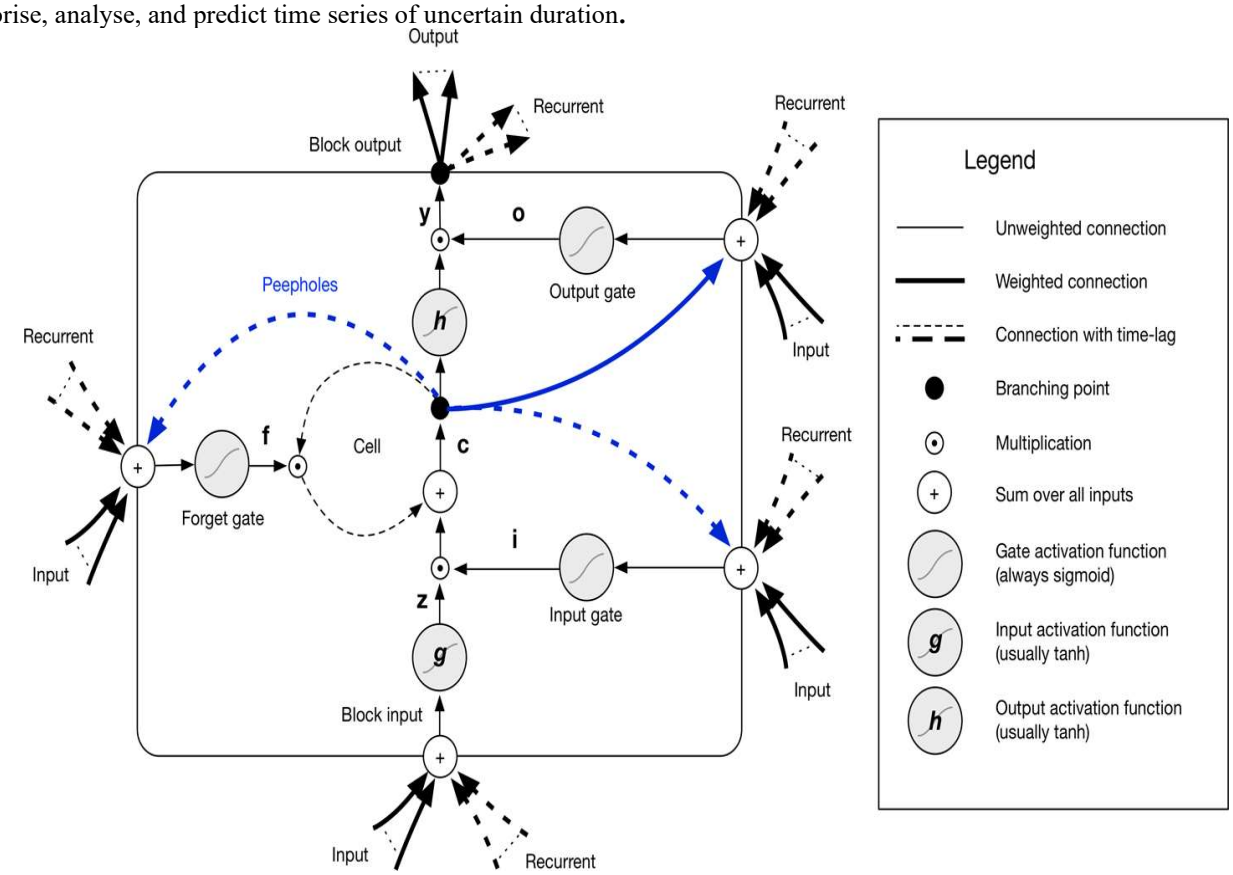
- NER is extremely beneficial to search engines. NER aids in the organisation of textual data, and structured data aids in the effective indexing and retrieval of texts for search.
- NER also has applications in machine translation. The majority of the time, entities classified as Named Entities are transliterated rather than translated.
- If the reader could see the listed entities before reading an article, they would be able to obtain a good notion of what the piece is about.
- Automatic book indexing: Named Entities make up the majority of the words indexed in a book's back index.
- Used in the biomedical field to identify proteins, medications, and disorders, among other things. NE Because it gives structure to raw data, tagger is frequently a sub-task in most information extraction processes.

LSTM (Long Short Term Memory) based Name Entity Recognition

Long short-term memory (LSTM) is a deep learning architecture that uses an artificial recurrent neural network (RNN). Sepp Hochreiter and Jurgen Schmidthuber proposed it in 1997. LSTM has feedback connections, unlike traditional feed-forward neural networks. It can handle not only single data points (like photos), but also complete data streams (such as speech or video).

LSTM can be used for tasks like un-segmented, linked handwriting recognition or speech recognition, for example.

A cell, an input gate, an output gate, and a forget gate make up a standard LSTM unit. Three gates control the flow of information into and out of the cell, and the cell remembers values across arbitrary time intervals. The LSTM algorithm is well-suited to categorise, analyse, and predict time series of uncertain duration**.**



Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881; IMPACT FACTOR :8.017(2022); IC VALUE:5.16; ISI VALUE:2.286
Peer Reviewed and Refereed Journal: VOLUME:11, ISSUE:11(5), November: 2022
Online Copy of Article Publication Available (2022 Issues)
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: 2nd November 2022
Publication Date:10th December 2022
Publisher: Sucharitha Publication, India
Digital Certificate of Publication: www.ijmer.in/pdf/e-CertificateofPublication-IJMER.pdf

DOI: http://ijmer.in.doi./2022/11.11.83
www.ijmer.in

$$
\begin{aligned}
\mathbf{i}_t &= \sigma(\boldsymbol{W}_i\mathbf{h}_{t-1} + \boldsymbol{U}_i\mathbf{x}_t + \boldsymbol{b}_i) \\
\mathbf{f}_t &= \sigma(\boldsymbol{W}_f\mathbf{h}_{t-1} + \boldsymbol{U}_f\mathbf{x}_t + \boldsymbol{b}_f) \\
\tilde{\mathbf{c}}_t &= \tanh(\boldsymbol{W}_c\mathbf{h}_{t-1} + \boldsymbol{U}_c\mathbf{x}_t + \boldsymbol{b}_c) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \\
\mathbf{o}_t &= \sigma(\boldsymbol{W}_o\mathbf{h}_{t-1} + \boldsymbol{U}_o\mathbf{x}_t + \boldsymbol{b}_o) \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t)
\end{aligned}
$$

Bidirectional-LSTM

Bidirectional-LSTM It is advantageous to have access to both the past (left) and future (right) for many sequence labelling jobs. To collect past and future information, the essential idea is to show each sequence forwards and backwards to two independent concealed states. We found the optimal model by setting the number of LSTM units to 100 and the dropout and recurring dropout rates to 0.5.

Proposed Work

Building and training a bidirectional LSTM neural network model to recognise named items in text data using the Keras API and TensorFlow as the backend. People, places, and organisations can all be identified using named entity recognition models. Named entity recognition is a valuable pre-processing step for many downstream natural language processing applications such as machine translation, question answering, and text summarization, as well as a standalone tool for information extraction.

Implementation Steps

Import Modules
Load and explore the NER Dataset
Retrieve Sentences and Corresponding Tags
Define Mappings between Sentences and Tags
Padding Input Sentences and Creating Train/Test Splits
Build and Compile a Bidirectional LSTM Model
Train the Model
Evaluate Named Entity Recognition Model

Result and Discussion

1. First import all the modules

```
%matplotlib inline
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
np.random.seed(0)
plt.style.use("ggplot")

import tensorflow as tf
print('Tensorflow version:', tf.__version__)
print('GPU detected:', tf.config.list_physical_devices('GPU'))
```

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881; IMPACT FACTOR :8.017(2022); IC VALUE:5.16; ISI VALUE:2.286
Peer Reviewed and Refereed Journal: VOLUME:11, ISSUE:11(5), November: 2022
Online Copy of Article Publication Available (2022 Issues)
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: 2nd November 2022
Publication Date:10th December 2022
Publisher: Sucharitha Publication, India
Digital Certificate of Publication: www.ijmer.in/pdf/e-CertificateofPublication-IJMER.pdf

DOI: http://ijmer.in.doi./2022/11.11.83
www.ijmer.in

2. Load Dataset

	Sentence #	Word	POS	Tag
0	Sentence: 1	Thousands	NNS	O
1	Sentence: 1	of	IN	O
2	Sentence: 1	demonstrators	NNS	O
3	Sentence: 1	have	VBP	O
4	Sentence: 1	marched	VBN	O
5	Sentence: 1	through	IN	O
6	Sentence: 1	London	NNP	B-geo
7	Sentence: 1	to	TO	O
8	Sentence: 1	protest	VB	O
9	Sentence: 1	the	DT	O
10	Sentence: 1	war	NN	O
11	Sentence: 1	in	IN	O
12	Sentence: 1	Iraq	NNP	B-geo
13	Sentence: 1	and	CC	O
14	Sentence: 1	demand	VB	O

3. Find Sentences and corresponding tags

```
class SentenceGetter(object):
    def __init__(self, data):
        self.n_sent = 1
        self.data = data
        self.empty = False
        agg_func = lambda s: [(w, p, t) for w, p, t in zip(s["Word"].values.tolist(),
                                                           s["POS"].values.tolist(),
                                                           s["Tag"].values.tolist())]
        self.grouped = self.data.groupby("Sentence #").apply(agg_func)
        self.sentences = [s for s in self.grouped]

    def get_next(self):
        try:
            s = self.grouped["Sentence: {}".format(self.n_sent)]
            self.n_sent += 1
            return s
        except:
            return None
```

```
getter = SentenceGetter(data)
sentences = getter.sentences
```

```
sentences[0]
```

```
[('Thousands', 'NNS', 'O'),
 ('of', 'IN', 'O'),
 ('demonstrators', 'NNS', 'O'),
 ('have', 'VBP', 'O'),
 ('marched', 'VBN', 'O'),
 ('through', 'IN', 'O'),
 ('London', 'NNP', 'B-geo'),
 ('to', 'TO', 'O'),
 ('protest', 'VB', 'O'),
 ('the', 'DT', 'O'),
```

Cover Page

DOI: http://ijmer.in.doi./2022/11.11.83
www.ijmer.in

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881; IMPACT FACTOR :8.017(2022); IC VALUE:5.16; ISI VALUE:2.286
Peer Reviewed and Refereed Journal: VOLUME:11, ISSUE:11(5), November: 2022
Online Copy of Article Publication Available (2022 Issues)
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: 2nd November 2022
Publication Date:10th December 2022
Publisher: Sucharitha Publication, India
Digital Certificate of Publication: www.ijmer.in/pdf/e-CertificateofPublication-IJMER.pdf

4. Perform mapping between tags and sentences

```
word2idx = {w: i + 1 for i, w in enumerate(words)}
tag2idx = {t: i for i, t in enumerate(tags)}
```

```
word2idx
```

```
{'Bychkova': 1,
 'victorious': 2,
 'struggled': 3,
 'ex-President': 4,
 'Rehnquist': 5,
 'Uganda': 6,
 're-tried': 7,
 'Welfare': 8,
 'intent': 9,
 'stance': 10,
 'Depending': 11,
 'pro-rebel': 12,
 'Catalonians': 13,
 'stalls': 14,
 'gored': 15,
 'jeopardy': 16,
 'epitomizes': 17,
 'Bernie': 18,
 'Price': 19,
 'sets': 20,
 'Cayes': 21,
 'however': 22,
 "'T": 23,
 'immorality': 24,
 'climatic': 25,
 'touch': 26,
 'vivid': 27,
 'Shaffi': 28,
```

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881; IMPACT FACTOR :8.017(2022); IC VALUE:5.16; ISI VALUE:2.286
Peer Reviewed and Refereed Journal: VOLUME:11, ISSUE:11(5), November: 2022
Online Copy of Article Publication Available (2022 Issues)
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: 2nd November 2022
Publication Date:10th December 2022
Publisher: Sucharitha Publication, India
Digital Certificate of Publication: www.ijmer.in/pdf/e-CertificateofPublication-IJMER.pdf

DOI: http://ijmer.in.doi./2022/11.11.83
www.ijmer.in

5. Implement LSTM model

```
from tensorflow.keras import Model, Input
from tensorflow.keras.layers import LSTM, Embedding, Dense
from tensorflow.keras.layers import TimeDistributed, SpatialDropout1D, Bidirectional
```

```
input_word = Input(shape=(max_len,))
model = Embedding(input_dim=num_words, output_dim=50, input_length=max_len)(input_word)
model = SpatialDropout1D(0.1)(model)
model = Bidirectional(LSTM(units=100, return_sequences=True, recurrent_dropout=0.1))(model)
out = TimeDistributed(Dense(num_tags, activation="softmax"))(model)
model = Model(input_word, out)
model.summary()
```

```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 50)]              0
_________________________________________________________________
embedding (Embedding)        (None, 50, 50)            1758950
_________________________________________________________________
spatial_dropout1d (SpatialDr (None, 50, 50)            0
_________________________________________________________________
bidirectional (Bidirectional (None, 50, 200)           120800
_________________________________________________________________
time_distributed (TimeDistri (None, 50, 17)            3417
=================================================================
Total params: 1,883,167
Trainable params: 1,883,167
Non-trainable params: 0
_________________________________________________________________
```

```
model.compile(optimizer="adam",
              loss="sparse_categorical_crossentropy",
              metrics=["accuracy"])
```

6. NER prediction on the basis of loss and accuracy

```
model.evaluate(x_test, y_test)
```

```
9592/9592 [==============================] - 12s 1ms/sample - loss: 0.0484 - accuracy: 0.9856
[0.04835232572507321, 0.9855588]
```

```
i = np.random.randint(0, x_test.shape[0]) #659
p = model.predict(np.array([x_test[i]]))
p = np.argmax(p, axis=-1)
y_true = y_test[i]
print("{:15}{:5}\t {}\n".format("Word", "True", "Pred"))
print("-" *30)
for w, true, pred in zip(x_test[i], y_true, p[0]):
    print("{:15}{}\t{}".format(words[w-1], tags[true], tags[pred]))
```

```
Word           True      Pred

------------------------------
The            O         O
United         B-org     B-org
Nations        I-org     I-org
has            O         O
been           O         O
under          O         O
fire           O         O
for            O         O
failing        O         O
to             O         O
stop           O         O
ongoing        O         O
ethnic         O         O
violence       O         O
```


Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881; IMPACT FACTOR :8.017(2022); IC VALUE:5.16; ISI VALUE:2.286
Peer Reviewed and Refereed Journal: VOLUME:11, ISSUE:11(5), November: 2022
Online Copy of Article Publication Available (2022 Issues)
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: 2nd November 2022
Publication Date:10th December 2022
Publisher: Sucharitha Publication, India
Digital Certificate of Publication: www.ijmer.in/pdf/e-CertificateofPublication-IJMER.pdf

DOI: http://ijmer.in.doi./2022/11.11.83
www.ijmer.in

Conclusion and future work

The technology of named entity recognition (NER) is frequently used for extracting meaningful information from unstructured natural language document sets. NER is one of the most important phases in Natural Language Processing (NLP) for text analysis, and it's utilised in both online and standalone applications.

Deep learning can yield good outcomes, according to the results of the experiment. In the English language, the best model for named-entity recognition is the LSTM.

We plan to test our technique on a variety of languages in the future. To develop a hybrid algorithm, mix this method with other neural networks.

References

1. Shinzato, Keiji, et al. "Constructing dictionaries for named entity recognition on specific domains from the Web." Web Content Mining with Human Language Technologies Workshop on the 5th International Semantic Web. 2006
2. Hideki Isozaki. 2001. “Japanese named entity recognition based on a simple rule generator and decision tree learning” in the proceedings of the Association for Computational Linguistics, pages 306-313. India.
3. J. Kim, I. Kang, K. Choi, “Unsupervised Named Entity Classification Models and their Ensembles”, in the proceedings of the 19th International Conference on Computational Linguistics, 2002.
4. John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data” in the proceedings of International Conference on Machine Learning, pages 282-289, Williams College, Williamstown, MA, USA.
5. Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla and Dipti Misra Sharma. 2008.,”Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition” in the proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 25-32, Hyderabad, India.
6. Kumar N. and Bhattacharyya Pushpak. 2006. “Named Entity Recognition in Hindi using MEMM” in the proceedings of Technical Report, IIT Bombay, India.
7. Mandeep Singh Gill, Gurpreet Singh Lehal and Shiv Sharma Joshi, 2009. “Parts-of-Speech Tagging for Grammar Checking of Punjabi” in the Linguistics Journal Volume 4 Issue 1, pages 6-22
8. Goyal, Archana & Gupta, Vishal & Kumar, Manish. (2018). Recent Named Entity Recognition and Classification techniques: A systematic review. Computer Science Review. 29. 21-43. 10.1016/j.cosrev.2018.06.001.
9. Albared, Mohammed & Gallofré Ocaña, Marc & Ghareb, Abdullah & Al-Moslmi, Tareq. (2019). Recent Progress of Named Entity Recognition over the Most Popular Datasets. 1-9. 10.1109/ICOICE48418.2019.9035170.
10. Perera, Nadeesha & Dehmer, Matthias & Emmert-Streib, Frank. (2020). Named Entity Recognition and Relation Detection for Biomedical Information Extraction. Frontiers in Cell and Developmental Biology. 8. 10.3389/fcell.2020.00673.
11. li, Jing & Sun, Aixin & Han, Ray & Li, Chenliang. (2020). A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering. PP. 1-1. 10.1109/TKDE.2020.2981314.
12. Nasar, Zara & Jaffry, Syed Waqar & Malik, Muhammad. (2021). Named Entity Recognition and Relation Extraction: State of the Art. ACM Computing Surveys. 54. 10.1145/3445965.