**International Journal of Distributed Artificial Intelligence**
Volume 14 • Issue 1

# Hybrid Model for Named Entity Recognition

Nikhil Chaturvedi, Shri Vaishnav Vidyapeeth Vishwavidyalaya, India*
Jigyasu Dubey, Shri Vaishnav Vidyapeeth Vishwavidyalaya, India

## ABSTRACT

Named entity recognition is an important factor that has a direct and significant impact on the quality of neural sequence labelling. It entails choosing encoding input data to create grammatical and semantic representation vectors. The main goal of this research is to provide a hybrid neural network model for a specific sequence labelling task such as named entity recognition. Three subnetworks are used in this hybrid model to ensure that information at the character, capitalization levels, and word-level contextual representation is fully utilized. The authors used different samples for training and development sets on the CoNLL-2003 dataset to show that the model could compare its performance to that of other state-of-the-art models.

## KEYWORDS

Bi-LSTM, CNN, CRF, Named Entity Recognition, Sequence Labelling

## INTRODUCTION

Sequence labelling is a subset of pattern recognition that involves applying an algorithm to assign a categorical label to each element of a sequence of observed values. Sequence labelling is a broad term that includes tasks like named entity recognition, text chunking and POS tagging.

The goal of NER, also known as entity chunking, is to detect named entities in text that correspond to predefined classes such as entity, time, and numeric, as well as seven subclasses: human name, place of origination, time, date, currency, and percentage. This paper focuses on the CoNLL2003 corpus (Sang & De Meulder, 2003), which was built from a variety of news corpora and contains four distinct elements (person, place, organization, and miscellaneous). In named entity recognition, a word's label is made up of two parts: "X-Y," which denotes the labelled word's position, and "Y," which indicates the applicable taxonomic category. NER is a tough problem in natural language processing that is required by search engines, question-answering systems, and translation systems. When a text has a named entity with a defined meaning, the translation system will typically translate the named entity's constituent terms independently, resulting in inaccurate translation results. If the

DOI: 10.4018/IJDAI.311063 *Corresponding Author

system recognises the thing first, the translation algorithm will have a better idea of how the words are put together and what they mean.

Modern approaches to NER (Sun et al., 2019) use vector embeddings such as Word2Vec1, GloVe (J. Pennington et al, 2014) or FastText2 to exploit the semantic content of words, convolutional neural networks are used to represent character-level properties of named entities, bi-directional LSTM (Hochreiter & Schmidhuber, 1997),(Green & Karras, 2012), (Wang et al., 2020) is used to model word order, and CRF (Godin et al., 2018) is used to model tag sequences probabilistically. As a named entity, another crucial element to examine is word capitalization might sometimes be composed of numerous capitalised components in a sentence. To effectively exploit the semantic, sequential, and character-level components of the NER challenge, we provided a hybrid model comprised of three encoding sub-networks in our proposed study. In contrast to (Huang et al., 2015), used CNN to extract character-level information in our proposed model. (Lample et al., 2016) work also has some similarities to ours. In the above both methods used Bi-LSTM to capture both features character and contextual level representation of words. In real time, they combine pre-trained word embeddings with capitalisation features. For that in our model we use dual subnetworks, Bi-LSTM, and CNN, to collect capitalization (Bodapati et al., 2019) and character-level characteristics separately. To reflect the sentence's rich semantic and grammatical properties, these two sub-networks' outputs are combined with pre-trained word embeddings.

## RELATED WORK

Several researchers have contributed significantly to the field of named entity recognition by inventing numerous ways of extracting sequence tagging from English text.

(Sun et al., 2019) highlight how previous scholars surveyed named entity recognitions in the statistical machine learning period, although NER task has evolved significantly in the last decade. On the one hand, transfer learning, deep learning, knowledge bases, and other methodologies are increasingly being used in NER systems. To demonstrate how these changes have occurred, they present an overview of NER based on 162 articles published at NLP-related conferences between 1996 and 2017.

(Zeng et al., 2018) suggested two-character feature learning models for NER tasks: one for learning local semantic features in word characters and the other for learning position-related information in word characters using concatenations and staking. The presented models are the first to learn character features in NER tasks using CNN and Bi-LSTM. They suggested attention mechanisms in the future to improve model capabilities at the character feature level by learning the combined weights of several character modules. The attention mechanism is used at the word level to model the relationship between tags and words.

(Gridach, 2017) suggested a novel neural network architecture for biological NER in this study used Bi-LSTMs, pretrained word embeddings, character-level embeddings, and CRF in this neural network model. On the two datasets, they did better than the previous NER system and did better than the best state-of-the-art system.

They developed a data augmentation technique for enhancing the robustness of NER models to capitalization errors in (Bodapati et al., 2019). Data augmentation outperforms earlier methods in terms of robustness while maintaining well-formed text performance and enhancing generalisation to noisy text. We saw this across all models, languages, and dataset sizes. Also, for many natural languages, data augmentation is easy to set up and doesn't require any extra language-specific resources.

They proposed using the capitalization and punctuation recovery approach to improve named entity recognition from Vietnamese speech in this study by (Nguyen et al., 2020). They presented the first voice dataset for the Vietnamese language, which laid the framework for studying extracted entities in speech. They also showed that a pre-trained language model for Vietnamese that could be used for NER tasks was effective. This model out formed the state-of-the-art on the VLSP 2018 dataset.