WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Named Entity Recognition for Machine Learning Application

Dr. Ch. Pulla Rao[1], Y. Narmada[2], T. Durga Prasad[3], B. Umesh Chand[4], M. Sanjay Kumar[5]

[1]Head of the Department, [2,3,4,5]UG Students Department of Electronics and communication Engineering
DVR & Dr. HS MIC College of Technology

ABSTRACT

Named entity recognition (NER) is a natural language processing technology used to extract information from unstructured text data such as e-mails, newspapers, blogs, and so on. NER is the process of recognizing nouns such as persons, places, organizations, and so on that are mentioned in the text, sentence, or paragraph string. Many different libraries and natural language processing tools in Java, Python, and C are available for developing the NER system. All of these tools contain pre-trained NER models that can be imported, used, and updated or customized as needed. This paper describes various NLP libraries, such as Python's SpaCy, Apache Open NLP, and TensorFlow. Some of these libraries include a pre-built, customizable NER model. These libraries are compared based on training accuracy, F-score, prediction time, model size, and simplicity of training. All the models use the same training and testing data. When the entire performance of all the models is considered, Python's Spacy provides the highest accuracy and the best result.

I. INTRODUCTION

Identification and classification of entities in text into specified categories, such as names, places, organizations, and so forth, is the problem of named entity recognition (NER), a subtask of natural language processing (NLP). Numerous NLP applications, such as machine translation, information retrieval, and question-answering systems, all depend on NER. The steps in the NER process include analyzing a text, locating the words or phrases that represent entities, and categorizing those words or phrases. Typically, machine learning algorithms that have been trained on substantial volumes of labelled data are used for this.

A critical stage in the task of Natural Language Processing (NLP) is the process of detecting Named Entities (NEs) from a textual material and categorizing them into various conceptual categories (Name, Place, Party, Designation). Named Entity Recognition (NER) is the name of this process. Although information is readily available in this era of the World Wide Web, very little work has been done specifically on NLP and analytics in Indian languages. NER is a key requirement for applications such as information retrieval, machine translation, question answering, text summarization, effective search engines, etc.

NER development for Indian languages continues to be difficult due to the syntactic and semantic complexity of Indian languages and a lack of processing tools. Since Hindi is the official language of India, it has been chosen for this project's NER tool development since it can be more accurate than existing tools thanks to a hybrid NLP ontology, rule-based methodology, and machine learning. Since Gujarati, Bengali, Marathi, and other Indo-Aryan languages do not have ideal NER tools, the current technique can be utilized for these languages as well since they are all members of the Indo-Aryan family. Hindi is one of these languages. NER can be accomplished in one of three ways. utilizing machine learning, ontologies, linguistic rulesets, or a hybrid method combining both. The best methods for forecasting unidentified things are rule-based systems and machine learning, respectively. So, for Indian language, hybrid NER systems are most effective.

For NER, a variety of methods are employed, such as rule-based systems, statistical models, and deep learning-based methods. Rule-based systems employ manually created criteria to identify entities, whereas statistical models use probabilistic algorithms to forecast the possibility that a word or phrase represents an entity. Deep learning-based systems make use of neural networks to discover patterns in the data and recognize items automatically. When working with loud, unclear, or unstructured data, NER might be difficult. But more recent developments in deep learning and machine learning have significantly increased NER accuracy, making it a vital tool in many NLP applications.

II. LITERATURE REVIEW

C. Thielen et al. created a framework to discriminate between words with uppercase letters and those with lowercase letters. According to this paradigm, all named entities are tokens that begin with an uppercase alphabet. When there is ambiguity, words are conjectured. The positioning of entities in ambiguous situations is also considered by the methodology.

A deep neural network-based model for recognising named entities was introduced by Yadav and Bethard et al. The researchers asserted to have used the most up-to-date technology, which is considerably superior to earlier studies based on supervised, unsupervised, and feature engineering approaches and effectively recognises entities.

The proposal was made by R Fenny Syafarian and Rio Yunanto. The goal of this study is to aid researchers in locating and mapping machine learning algorithms using the findings of earlier studies that focused on named-entity recognition. The research method used in this study looks at academic articles about introducing named entities using machine learning. Using Google Scholar, a collection ofarticles from 2018 to 2020 was generated. If machine learning methods have been applied in named-entity recognition research is one of the important research problems that this study needs to address.

III. WORKING OF NER

NLP: Assists machines in understanding the principles of language and in developing intelligent systems that can readily draw meaning from text and speech.

Machine learning: Aids machines in learning and improving over time by utilizing various algorithms and training data.

Any NER model has two steps: i) detect a named entity and ii) categorize the entity. People, places, money, and other elements from a variety of categories are all easily detectable by humans. Computers must first recognize them before classifying them in order to perform the same task. This is accomplished via NLP and machine learning.

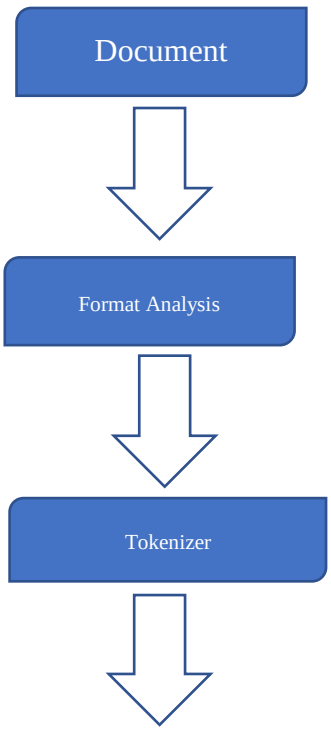
NLP: Aids in the development of intelligent systems that can quickly infer meaning from speech and text by teaching machines how to interpret linguistic rules.

Machine learning: Using a variety of algorithms and training data, machines can learn and get better over time.

Every NER model follows a two-step method to identify named entities and then classify them.

Information Extraction: In the first step of NER, named entities that are mentioned in the document, paragraph, sentence, and texts are found and prepared for extraction of information. Entity extraction is a text analysis technique that automatically extracts certain data from unstructured text and categorizes it in accordance with predetermined categories using Natural Language Processing (NLP). These groups are known entities, which are noun-representing words or phrases. In addition to proper names,this also applies to numerical expressions of time or quantity, such as dates, phone numbers, or monetary amounts. The entire extraction procedure entails tagging speech, identifying sentence boundaries, capitalization rules, and co-reference in the texts that are more crucial to use and find more precise search phrases for.

Searching the Entities: The next step in NER is to look up potential entities to mention in the text. Pseudonyms, numerous sites, and informational web pages are also taken into account while searching for synonyms. The searcher keeps a restricted group of entities in order to maintain the balance with accuracy and recall the correct entity while minimizing the calculation required to recognize such entitie



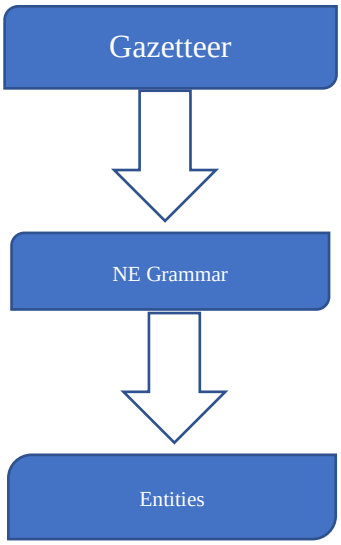


Fig : Working Of NER

1. Document:

These are essentially files or papers that contain data that is used as process input. In this situation, the input could be a phrase, group of words, or paragraph.

2. Format Analysis:

In essence, these are files or papers that contain data used as process input. A sentence, group of words, or chapter could be the input in this situation.

3. Tokenizer:

Tokenization is a technique used in natural language processing to break down phrases and paragraphs into simpler language-assignable elements. Gathering the information (a sentence) and disassembling it into manageable pieces is the first stage in the NLP process.

4. Gazetteer

When used in conjunction with a map or a complete atlas, a gazetteer is a geographical dictionary or directory that is a valuable source of information about places and place names.

METHODOLGY

1. NER Using spaCy

Python and Cython are used by the spaCy software package to carry out sophisticated natural language processing.

This library was created by Matthew Hannibal and Ines Montano, who founded the software business Explosion. It was released under the MIT licence. In contrast to the natural language toolkit (NLTK), which is frequently used for research, SpaCy is intended for the production environment.

Advanced Natural Language Processing (NLP) libraries for Python and Cython are provided by spaCy. One of the best NLP annotation tools available is spaCy. NLP software is being utilised more and more to analyse and analyse data.

Unstructured textual data is generated in large quantities, and it is important to process such data and apply knowledge. To do that, the data must be formatted in a way that computers can interpret. Natural language processing can assist you in doing that.

With the help of ready-to-use pre-trained language-specific models, you can use spaCy to carry out NLP operations such as parsing, tagging, NER, lemmatizer, tok2vec, attribute ruler, and more. There is one multi-language pipeline component and support for 18 different languages.

SpaCy Design for NER: Through sophisticated natural language processing, the spaCy Python package enhances NLP. This programme, which is intended for production use, can be used to create applications that manage and comprehend vast quantities of text.

This method can be used as a preprocessing Zstep for deep learning, or it can be utilized to organize data or recognize natural language. SpaCy has a number of other functions in addition to tokenization, parts-of-speech tagging, text categorization, and named entity recognition. Use of SpaCy for NER tasks is extremely simple.

Despite the fact that we frequently have to alter the data we utilize to meet our business needs, the model works well with any kind of text. A powerful NER system written in Python is used in spaCy to label contiguous groupings of tokens. This model offers a default technique for classifying a variety of names and numbers, including those for individuals, organizations, languages, events, etc.

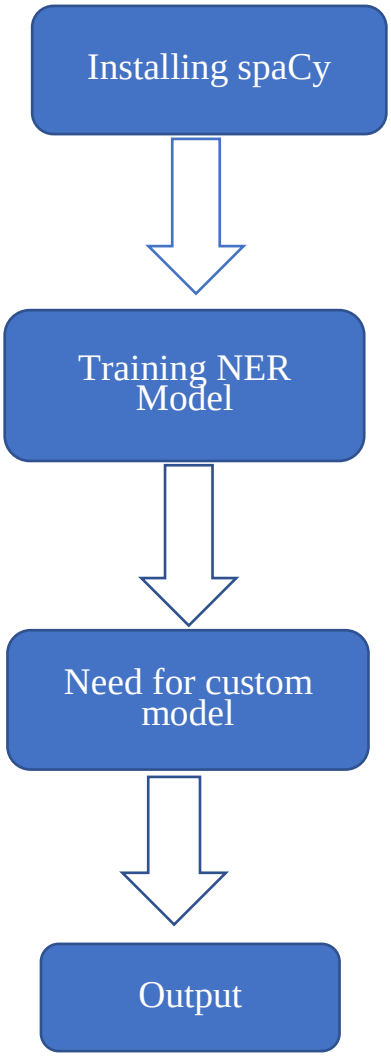
By training the model with more recent examples, SpaCy provides us with a wide range of options for adding more entities. If required, NER can also be altered using any classes.

The MIT license for the English language makes the following four pre-trained spaCy models available.

- en_core_web_sm(12 mb)
- en_core_web_md(43 mb)
- en_core_web_lg(741 mb)

Selecting "NER" as the components and hardware

based on system availability on the Config page will have an impact on the architecture, pretrained weights, and hyperparameter choices. We can also choose to optimize for efficiency (faster inference, smaller model, lower memory consumption) or higher accuracy (possibly larger and slower model). Once finished, click the download button on the bottom right side to access the config file.



Need for custom NER model: SpaCy has a built-in named recognition pipeline. Even if it works effectively, your text may not always be entirely accurate. Depending on the context, a word may be categorized as a person or an organization. Additionally, there are situations when the desired category in the built-in spaCy library may not be available. Identifying named entities (people, locations, organizations, etc.) from a passage of text and categorizing them into a preset set of categories is known as named entity recognition (NER), and it is a common NLP task.

2. Machine Learning Approach To NER

Here, we discuss the supervised machine learning method to NER, which is currently the most popular strategy among researchers and is also the approach we apply in the experiments that follow in this thesis.

The schema for this strategy is shown in Figure. There are two stages: test and training. The text and labels used in the training phase are triples (start index, finish index, type), and they are the only inputs. Preprocessing options for this input include lemmatization, part-of-speech tagging, stemming, and tokenization.

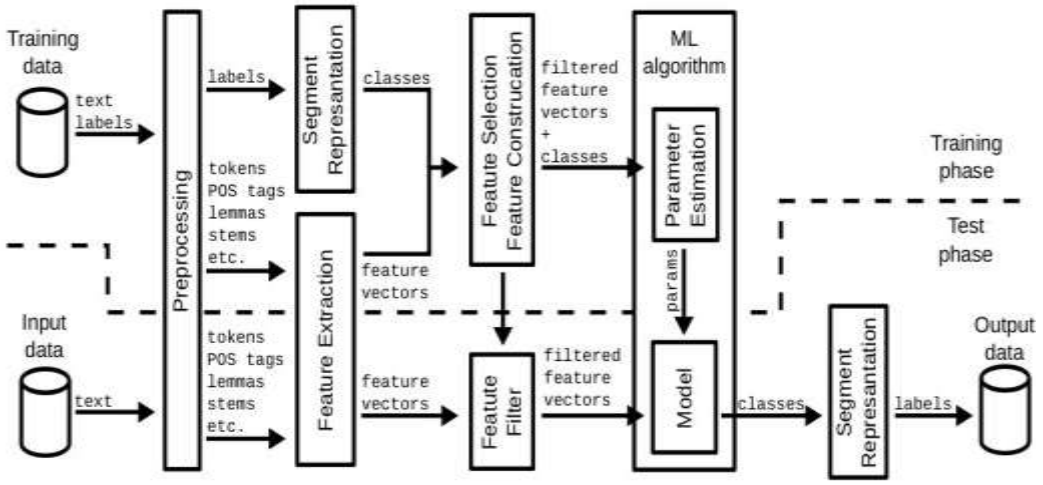


Fig: general schema of supervise machine learning approach to NER

The next step is to format the input so that a machine can understand it. For classification, the token (word) is used as the fundamental unit.

As a result, we generate a vector representation for each word in which each significant aspect of the word and its context is represented by a numerical value. In this thesis, this procedure is referred to as feature extraction (the word is confusing). Each feature vector is given a class based on the labels depending on the selected segment representation.

The process of choosing features and/or building is optional. In this step, the feature vector dimension is reduced without the information being diminished (feature selection), or new features are constructed by combining existing features.

The chosen machine learning algorithm's parameter estimation is the final stage. A model is produced at the conclusion of the training phase, and it is employed in the testing phase.

The same procedures are used in the test phase, but the input data is not labelled. Use of the same characteristics is required. The feature filter retains and uses the feature selection decisions made during the training phase. Based on the knowledge gained from the training data, the model oversees classifying the tokens. The parameters of the model express the experience. Using the selected segment representation, the model's classes are converted to labels.

When opposed to rule-based methods, machine learning techniques are more effective because the system may be trained and applied to a variety of domains. The goal of the NER approach in the ML NER system is to convert the identification problem into a classification problem, which is subsequently solved using statistical models. In theory, the ML system can identify and categorize NEs into specific NE classes like places, people, organizations, etc. Most of the current research in NE uses machine learning, often known as statistical, for all major languages, including Arabic. NE tagging judgements from annotated texts are frequently made using ML algorithms. By searching for patterns and correlations to model, the ML method for language analysis works bottom-up.

Three distinct categories of machine learning (ML) exist: semi-supervised learning, unsupervised learning, and supervised learning. The most widely used Supervised Learning (SL) strategies for Named Entity Recognition are those that treat the NER problem as a classification task and necessitate the availability of sizable annotated datasets. Because the system may be educated and used in other domains, learning methods are more effective than rule-based ones.

COMPARISION OF ALGORITHEM

For applications requiring sequence labelling, such as Named Entity Recognition (NER), two well-liked machine learning techniques are HMM (Hidden Markov Models) and MEMM (Maximum Entropy Markov Models). Despite the fact that both models are Markov models, they are different in how they manage dependencies between input features and output labels. The conditional probability distribution of the output labels given the input features is modelled differently in HMM and MEMM.

The first-order Markov assumption is used to describe the conditional probability distribution in HMM, which means that the probability of the subsequent output label depends solely on the current input feature and the preceding output label. Given the output labels, HMM presupposes that the input features are independent of one another. The model has a number of hidden states that describe the fundamental organization of the data, and the Baum-Welch algorithm is used to train the model's parameters.

On the other hand, MEMM calculates the conditional probability distribution using a maximum entropy model. In contrast to HMM, MEMM calculates the likelihood of the following output label by taking into account all input features. MEMM allows for more complicated dependencies between the input and output by modelling the conditional probability of the output label given the whole sequence of input attributes. The conditional distribution is modelled by MEMM as a collection of entropy-maximizing decision rules with the restriction that the probability of the output label must be consistent with the input attributes.

Overall, the way that HMM and MEMM simulate the conditional probability distribution is the primary distinction between them. While MEMM models the probability of the output label given the entire sequence of input features, allowing for more complex dependencies between the input and output, HMM makes the first-order Markov assumption that the next output label only depends on the current input feature and the previous output label.

IV. APPLICATIONS

Natural language processing (NLP) is a common application, and Spacy is a well-liked Python package for NLP applications that provides effective tools for NER. Here are some examples of how SpaCy's NER is used specifically. The most common application of NER is as a pretreatment tool for additional natural language processing tasks. There are several factors that make it vital to handle NEs carefully when performing machine translation. The lack of NE representation in texts is the first issue. Even in a large corpus, a name can only appear once; this problem cannot be resolved by the context-based patterns (or features) employed by modern machine translation systems.

The sparsity issue can be considerably reduced when NEs of the same type appear in similar contexts.

The system will notify you of the event and provide you with the necessary details. In general, it can be utilized in lecture halls, schools, and seminar rooms to update.

1. Information Extraction

SpaCy's NER is used to extract information from text data for instance it can be used to extract names of people, companies, and locations etc. The information can be used

To build.

2. Named Entity Linking

To connect named entities to external knowledge bases like Wikipedia, one can use SpaCy's NER. This can also help to clarify terms either ambiguous meaning and give more details on terms that were not present in the original text. Applications like chatbots, assistants or intelligent search engines can benefit from this.

3. Text Classification

For text classification based on named entities, spaCy's NER can be employed. It can be used, for instance, to categorize new stories by industry or find social media posts about goods or services. Targeted advertising, competitor analysis, and market research are all possible uses for this data.

ADVANTAGES

1. Efficient search Algorithm
2. Information extraction
3. Sentiment analysis
4. Document classification and summarization
5. Question-answering systems
6. Machine translation
7. Data mining and knowledge discovery
8. Personalization and recommendation system

V. CONCLUSION

Named Entity Recognition has been improving continuously for more than 15 years. The singular use is to extract several types of information (name, date, time, and place) from the text. Over 200 distinct types of objects and more than 20 different languages are also present. Most studies look for specific information on topics like news stories, website information, etc.

I hope that this essay provides a broad overview of techniques for creating NER systems, from manually assigning rules to generating accurate outcomes. However, establishing the regulations takes time as well. Without a large labelled corpus, semi-supervised and unsupervised learning approaches can quickly recognize entities. Supervised learning requires a significant labelled corpus.

You should be aware that choosing the machine learning algorithm and evaluation methods is essential for figuring out how well the system we designed is doing.

ACKNOWLEDGMENT

This work is based on our project from the DVR & Dr. Hs Mic College of Technology. We appreciate Head of the Department, ECE Dr. CH. Pulla Rao Ph.d, cooperation and direction during the study.

REFERENCES

1. Nadeau, David, and Sekine, Satoshi. A Survey of Named Entity Recognition and Classification. Lingvisticae Investigiones 30(1), 2007.
2. Ratinov, Lev, and Roth, Dan. Design Challenges and Misconceptions in Named Entity Recognition. Proceedings of the Workshop on Language in Social Media, 2011.
3. Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. The Stanford CoreNLP Natural Language Processing Toolkit. Communications of the Association for Computing Machinery 56(2), 2013.
4. Jurafsky, Dan, and Martin, James H. Speech and Language Processing. Pearson Education, 2014.
5. What is named entity recognition (NER)? — Definition from WhatIs.com. https://www.techtarget.com/whatis/definition/named-entity-recognition-NER
6. Neural Methods for Event Extraction: https://tel.archives-ouvertes.fr/tel- 01943841/document
7. A Compact Survey on Event Extraction: Approaches and Applications. https://arxiv.org/pdf/2107.02126.pdf
8. C. J. Saju and A. S. Shaja, "A Survey on Efficient Extraction of Named Entities from New Domains Using Big Data Analytics," 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), Tindivanam, 2017, pp. 170- 175