Named Entity Recognition: An Introduction

Irfan Ahmad Shiekh

Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri 185234, J&K, India.

Abstract:

Named entity recognition is a process and study of identification of entities that are proper nouns and classifying them to their appropriate predefined class, also called as tag. Named entity recognition is also called as entity chunking, entity identification and entity extraction. It is a sub-task of information extraction, where structured text is extracted from unstructured text. Popular applications of NER are machine translation, text mining, data classification, question answering system. This paper presents a quick introduction on different aspects of Named Entity Recognition like purpose of NER, challenges in NER, app- roaches in NER, working of NER, implementation and applications of NER.

Keywords: Named Entity, Named Entity Recognition, Machine Translation, Working of NER

Introduction:

A named entity is aword form that detects the elements having same properties from a group of elements. It is called as a rigid pointer oran atomic element ormember of the semantic class which mayvary depending upon the domain of interest. For example — inBiomedicine province, entities of interest are geneand gene products; in general province, person, location, organization, number, date, time, etc are important entities; in the homeopathic province,medicine names and disease names are recognized as entities.Named Entity Recognition (NER) is a form of Natural Language Processing (NLP) that involves detecting and sorting named entities in a given text into predefined categories similar as person names, locations, organizations, time expressions, quantities, monetary values, and so on. The main aim of NER is to automatically removing useful information from formless text by recognizing and classifying named entitiesdirectly.For example, consider the following sentence: "Khan works at Google in California." NER would identify "Khan" as a person, "Google" as an organization, and "California" as a location. This process of identifying and categorizing named entities is useful for many applications such as information extraction, text summarization, question answering, sentiment analysis, and more.Named Entity Recognition (NER) is a subfield[1] of Natural Language Processing (NLP) that involves recognizing and extracting

named entities (such as people, organizations, locations, and other entities) from unstructured text. Natural Language Processing is the capability of a computer program to recognize human language as it is said and composed, i.e, referred to as a natural language. Natural Language has resided for more than 50 years and has roots in the area of linguistics. Language processing appears in several stairs in which Named Entity Recognition is one of the eminent forms. In recent years, there have been several directions and advancements in the field of NER, including: Pre-trained language models, cross-lingual NER, and multi-task learning. Over-all,the area of NER is rapidly advancing, with new techniques and models being developed to improve accuracy, efficiency, and language.In simple words, Named Entity Recognition is the method of finding thenamed entities from the text like a name of person, organization, location, and so on.

Purpose Of Named Entity Recognition:

Natural Language Processing is the capability of a computer program to recognize human language as it is said and composed,i.e. referred to as a natural language. Natural Language has resided for more than 50 years and has roots in the area of linguistics. Language processing appears in several stairs in which Named Entity Recognition is one of the eminent forms. Named entity recognition is a technique in Natural Language Processing that can scan naturally whole text and eliminates some basic entities and groped them into the given forms, such as places, organizations, persons, product names,etc. In simple words we can say, Named Entity Recognition is the method of finding the named entities from the text. Any word which shows the name of a person, organization, location, and so on, is a named entity.

Challenges In Named Entity Recognition:

Named Entity Recognition is supposed to be a introductory function of Natural Language Processing. There are many sided complications that are transferred in any natural language. Named Entity Recognition is also challenged by different complications. A many of the challenges[2] are described below:

1. **Ambiguity and Abbreviation:**Language is one of the significant challenges in recognizing the named entities. Recognizing words that can have different meanings and that can become part of various sentences. Distributing similar words from documents is another challenge. Words can be condensed for the simplicity of writing and

understanding. Words that will sometimes bear any label for recognition are another challenge.

2. **Spelling variation:**In English Language, vowels (a, e, i, o, u) play a leading role. Words that do not make a difference in phonetics but create a difference in writing and spelling.
3. **Foreign words**:Another challenge to Named Entity Recognition is: Many words that aren't frequently used these days, or a lot of people are not aware of some words like person names, location names, etc.

Approaches Of Named Entity Recognition:

There are primarily two approaches that are employed NER. [3]These include:

Rule based approach and Machine learning based approach[4].

Rule Based Approach

Rule based approach is also known as handcrafted approach. It's of two types;

1. **List lookup approach:**In this approach, gazetteers are used that consists of different list of named entity classesand a simple look of operation is performed to conclude whether a word is a named entity or not. However, also a named if a particular word is set up in a named entity class. Entity label is distributed to that word according to the named entity class in which it is set up. This methodology is easy and fast. The disadvantage of this approach is that it cannot overcome the problem of doubtfulness.
2. **Linguistic approach**:In this approach, a linguistic, who has an in depth knowledge about the alphabet of specific language constructs some rules, so that the named entities can be detected as well as classified fluently. [5]The rules that are constructed are language independent and cannot be used to identify named entities in some other language.

Machine Learning Approach:

This method is also known as automated method or statistical method. Machine learning method is more efficiently and constantly used as compared to the Rule based approach.

1. **Hidden markov model (HMM):**HMM is a statistical-based approach in which states are hidden or unobserved. It is based on the Markov Chain Property i.e. the probability of circumstance of the coming state is dependent on the just formal state. HMM is easy to apply. The disadvantage of this approach is that it requires lot of training in order to get better results and it can be used for large dependencies[6].

2. **Maximum entropy markov model (MEMM):**It combines the conception of HMM and MEM. While preparing, this model makes sure that the unknown values in a Markov Chain are connected and are not temporarily autonomous of each other. The large reliance problem of HMM is resolved by this model. Also, it has advanced recall and precision as compared to HMM. The disadvantage of this approach is the marker(label) base problem. The chances of transition from a particular state must add to one. MEMM favors those states through which less number of transitions occur.
3. **Conditional random field (CRF):**It is graphical undirected model unlike other classifiers, it also takes into consideration the environment information or the neighboring samples. It's known as Random Field since it computes the tentative probability on the following guest given the present guest values. This approach has benefits same as that of MEMM. Also, it resolves the marker(label) bias problem faced by MEMM[5].
4. **Support vector machine (SVM)**:SVM is a supervised statistical approach. The main ideal of this approach is to find whether a specific vector belongs to a particular target class or not. [7]In this approach, the training as well as testing data relates to the single size vector space. The main advantage of this approach is that it gives high delicacy for the document categorization problem[8].
5. **Decision tree:**It is a well known methodology that's used to eliminate and classify named entities in a given corpus. In this approach, some recognition rules are applied to the untagged training corpus so that named entities are recaptured. Now, we match these named entities attend with the factual answer crucial handed by the humans. However, also it is if the named entity is same as the answer key. Referred to as the positive illustration differently it's known as negative illustration. [9]A decision tree is make that classifies the named entities in the testing document. [10]The leaf node of decision tree shows the resultant value of test.

Working Of Named Entity Recognition:

Naturally, we can identify named entities like people, values, locations, etc from a proper document. For example: Assume the following sentence:

Sentence: Anil Ambani, the CEO of Jio company, is walking in the streets of Mumbai.

Three types of named entities are identified in the above sentence that is:

Table 1:

CATEGORY	ENTITIES
Person	Anil Ambani
Company	Jio
Location	Mumbai

With the help of computers, we can also identify named entities. So, for that, we need to help them in recognizing the entities first and then they can differentiate them. For achieving this NLP and ML are used.

Natural Language Processing: NLP is a discipline of linguistics and AI that study the correlations between computer and human language. This helps the system to perceive the rules and language and also helps in making intelligent machines that can smoothly extract meaning from the text and speech.

Machine Learning: This may be defined as the beach of AI, which constructs a system that learns from the data. it helps systems learn and enhances over time.

To figure out, what an entity is the NER system has to be capable to identify a word or string of words that make an entity (e.g. Mumbai) and determine which entity form it belongs to.In short, we can say that the heart of the NER system is a two-step process;

1. **Detect an entity (named):** For the first step NER system is identifying an entity from the given text. An entity may be a word or class of words.

2. **Categorize the entity:** The second step requires the generation of entity forms.

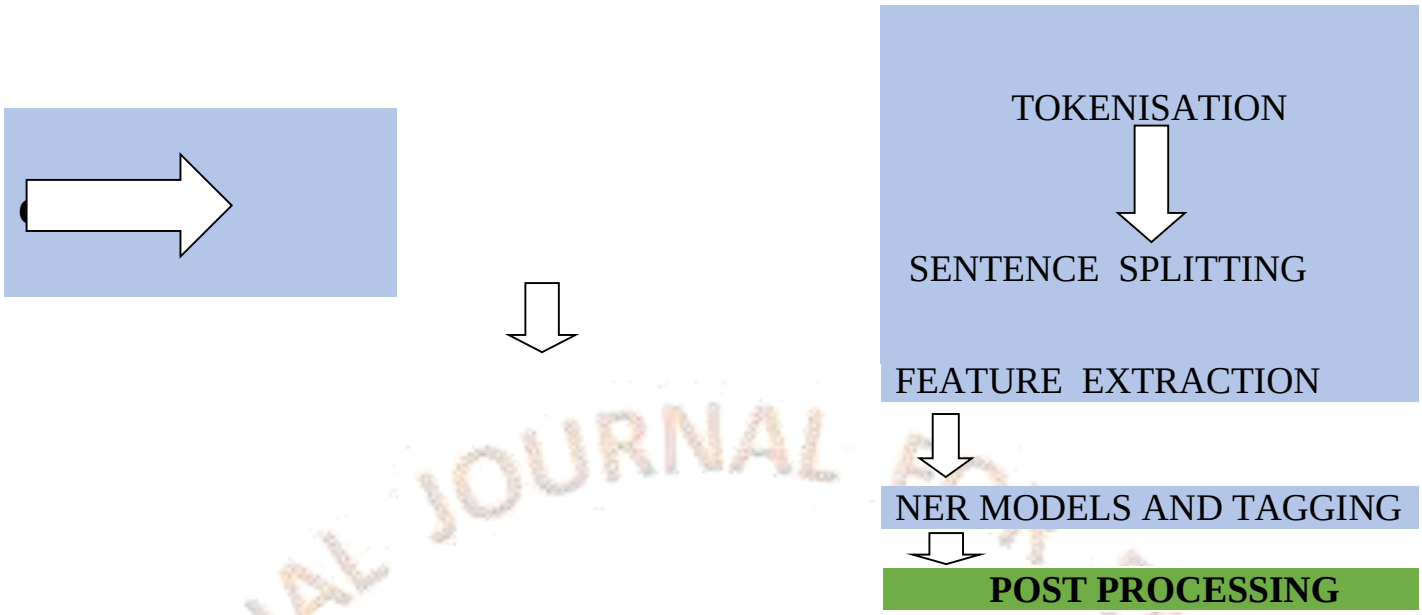

Fig. 1: Shows this is an overview of Named Entity Recognition.

Implementation Of Named entity Recognition:

1. NER Using Natural Language Tool KIT (NLTK)

NLTK is a main platform for making Python programs to work with human language data. It gives easy-to-use interfaces to over 50 corpora resources similar as Word-net, along with a suite of document processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic logic, and wrappers for industrial-strength NLP libraries.

To remove the named entities using NLTK in python, we use the nltk.ne_chunk() function. To achieve named entity recognition with NLTK, we have to execute threestepsthatarebelow:

1. Transfer text into tokens using the function word_tokensize().
2. For every word it finds part of speech tag using the function pos_tag().
3. Transfer the list that holds tuples of words and POS tags to the ne_chunk().

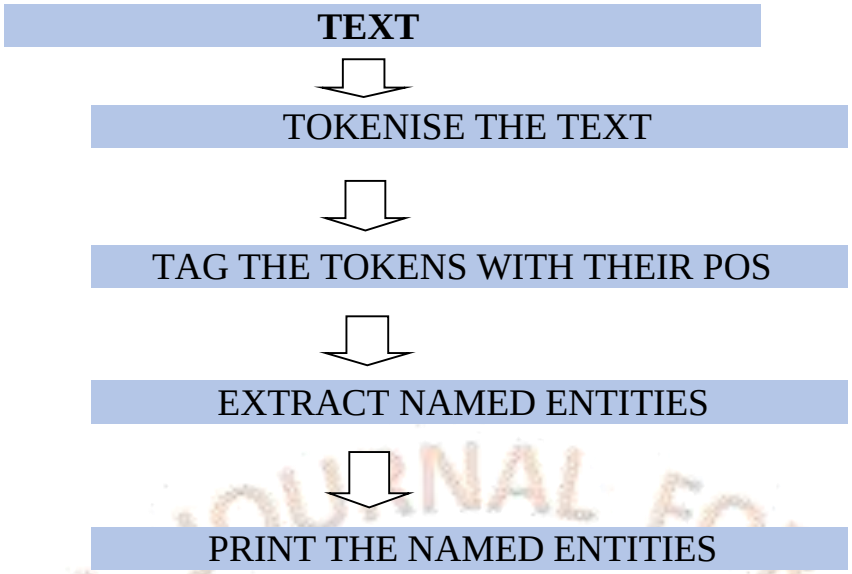

Fig.2: Shows the Named Entity Recognition using NLTK.

2. NER Using SpaCy

Spacy is a popular open-source library for Natural Language Processing in Python. It gives a fast and effective way to perform various NLP tasks, including Named Entity Recognition.

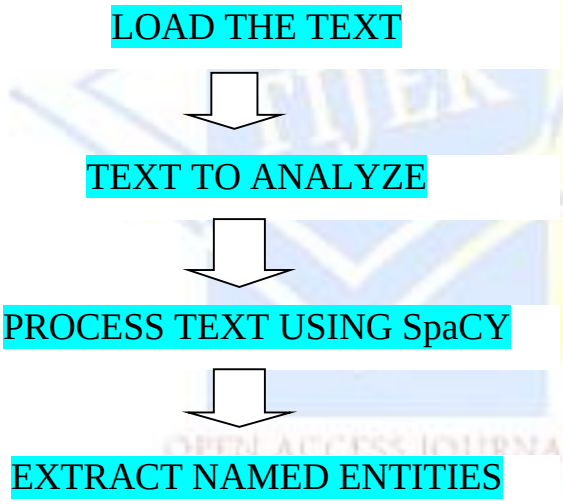

Fig.3: Here is a diagram that shows how NER using Spacy.

SpaCy uses the ent.text and ent.label property. The ent.text property provides the named entity text, and the ent.label property provides the named entity label, such as ORG for organization, GPE for geopolitical entity, and MONEY for monetary value. SpaCy also allows for personalize the named entity recognition pipeline and adding our own entities to the model for good performance.

Performance Metrics:

Performance Metrics is extremely important since it reveals the performance of a NER based system in terms of PRECISION, ACCURACY AND F-MEASURE. The result of a NER system may be nominated as "response" and the interpretation of human as the "answer key". We consider the following terms:

1. **Correct**- if the reaction is similar as the answer key.

2. **Incorrect**- if the response isn't same as the answer key.

3. **Missing**- if answer key is set to be tagged, but response isn't tagged.

4. **Spurious**- if response is said to be tagged but answer key isn't tagged[11].

Hence, precision, recall and f-measure can be defined as follows:

Precision (P)= correct / (correct + incorrect + missing)

Recall(R)=correct / (correct + incorrect + spurious)

F-Measure=(2*P*R) / (P + R)

Applications Of Named Entity Recognition:

NER acts a base for numerous important areas to manage abundant of digital information in the structured form. It acts as a preprocessing tool to break numerous complex activities. Some of them are as below;

1. Information Extraction (IE):

The delicacy of IE system depends upon proper nouns i.e. named entities as they carry important information about the document itself. So NER is considered important step to information extraction. NER is prerequisite for a event origin as well as a relation origin task [12] channels start with fitting named entities to identify relations expressed between entities and ideas and also end with the category of the relation type.

2. Question-Answering (QA):

QA is concerned with structure systems that induce answers to questions asked by humans being in natural language. These systems are classified according to the type of questions asked by operators. One important type of questions is factoid type questions which generally start with wh-word (what, when, which, where, who) and bear answers in a expressions or small statement[13]. So NER System works as a essential element in a question -answering system. The reason behind the employment of NER as an element in a QA System is to find the answer of numerous fact-based questions and these answers are entities that can be recognized by the NER system only. Thus, incorporating the NER in a QA system makes the task of chancing answers to some of the questions much easily.

3. Machine Translation:

Automatic machine translation is the procedure of converting documents or speech from source language to target language by the computer automatically without human intervention. Correct named entities identification is a demanding task in machine translation[14] because proper names need to be faced differently than other type of words. Named entities bear different approaches for conversion due to specific conversion rules that apply to them. Failure in correct identification of named entities not only effects verbal and syntactic structure of the conversion but also immediate and original environment in the document. The quality of machine conversion system can be bettered with the use of automatic NER System.

4. Semantic Search:

Semantic search aims to search for the information and the knowledge on the web by better understanding user intentions and directly answers the queries then traditional search. It brings the capability to prize applicable answers and delivers more individualized results. Origin of named entities and ideas from documents makes semantic search more important and robust[15]. NER task has been extensively used in web search queries as it assists in better understanding their semantics by exploiting the representation of contextual prompts around the named entities. Detecting and assaying the named entities confirming of a search query makes search machines possible for meeting operators search purpose.

6. Other Applications:

NER has numerous other operations in the medical field. For discovery of adverse medicine effect[16], identification of heart complaint threat factors[17], derivation of biomedical entities, etc. NER is demanding in biomedical sphere due to presence of semantically related entities in the data, variations in names of same ideas, common acronyms and condensations etc.

Conclusion:

Named Entity Recognition is an emerging field and is continuously improving due to its major contribution in many natural language applications. The aim of NER is to automatically eliminate relevant information from unstructured text data, which can be helpful for various applications such as information recovery, document classification, sentiment analysis, and machine translation.NER has been broadly studied and has obtained great results in various domains and languages. It is a challenging task due to the ambiguity of natural language, the complexity of named entities, and the diversity of text data. Although the development made in NER, there are still challenges to be addressed, such as handling rare and out-of-vocabulary named entities, improving cross-domain and cross-language generalization, and dealing with noisy and ambiguous text data. In short, NER is an essential task in NLP, and it has become an active research area due to its significance to various real-world applications. Its continued development and improvement will be crucial for advancing the field of NLP and realizing its full capacity.

References:

[1] Dhingra, B., & Dubey, A:"Named Entity Recognition: A Review of NLP Approaches and Techniques". In International Conference on Electronics, Information, and Communication, (2021).

[2] S. Kamath and R. Wagh: "Named Entity Recognition approaches and challenges". In International journal of advanced research in computer and communication engineering, (2017).

[3] Animesh Nayan, B. Ravi Kiran Rao, Pawandeep Singh, Sudip Sanyal and Ratna Sanya: "Named Entity Recognition for Indian Languages".In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages ,(2008).

[4] Kamaldeep Kaur, Vishal Gupta: “Name Entity Recognition for Punjabi Language”. In IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS),(2012).

[5] Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay: “Language Independent Named Entity Recognition in Indian Languages” .In Proceedings of IJCNLP-08 Workshop on NER for South and South East Asian Languages, (2008).

[6] Lawrence R. Rabiner: "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". In Proceedings of the IEEE,(1989).

[7] Asif Ekbal and Sivaji Bandyopadhyay: “Named Entity Recognition using Support Vector Machine: A Language Independent Approach”. InInternational Journal of Electrical and Electronics Engineering, (2010).

[8] Asif Ekbal and Sivaji Bandyopadhyay: “Bengali Named Entity Recognition using Support Vector Machine”.In Proceedings of the IJCNLP -08 workshop on NER for South and South East Asian Languages, (2008).

[9] Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis and Constantine D. Spyropoulos: “Learning Decision Trees for Named-Entity Recognition and Classification”. In EACI Workhop on Machine Learning for Information Extraction, (2000).

[10] Hideki Isozaki: “Japanese Named Entity Recognition based on a Simple Rule Generator and Decision Tree Learning”. Available at:http://acl.ldc.upenn.edu/acl2001/MAIN/ISOZAKI.PDF

[11] Darvinderkaur, Vishal Gupta: “A survey of Named Entity Recognition in English and other Indian Languages”. In IJCSI International Journal of Computer Science Issues,(2010).

[12] S. Riedel and A. McCallum: "Modeling relations and other mentions without labeled text". In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, (2010).

[13] N. Indurkhya, F. J. Damereau (Eds.), Handbook of Natural Language Processing, Second ed. Champan and Hall/CRC, Boca Raton, (2010).

[14] Zixiang Ding, Yu Zhang, and Qun Liu: In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015).

[15] I. Habernal, KM Konopi, SWSNL: “Semantic web search using natural language”. In Expert Syst, Appl., (2013).

[16] F. Zhu, P. Patumcharenpol, et al: “Biomedical text mining and its applications in cancer research”. In J. Biomed. Inform, (2013).

[17] J. Urbain: “Mining heart disease risk factors in clinical text with NER and distributional semantic models”. In J. Biomed. Inform, (2015).