International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 6, September-October 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

A Survey of Named Entity Recognition Techniques

Preeti Sondhi[1], Adil Ayoub[2]

[1]Assistant Professor, [2]M Tech Scholar,
[1,2]Universal Group of Institutions, Lalru, Punjab, India

ABSTRACT

The extraction of relevant information from data has become the most significant activity across all domains as data availability has increased. Information extraction becomes more difficult when data is accessible in the form of documents written in natural language. Named Entity Recognition (NER) is a technique for extracting meaningful information from unstructured natural language document collections that is widely utilised. NER is one of the primary steps in Natural Language Processing (NLP) for text analysis, and it is used for both online applications and stand-alone systems. This paper covers the fundamentals of NER as well as the various NER algorithms.

How to cite this paper: Preeti Sondhi | Adil Ayoub "A Survey of Named Entity Recognition Techniques" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-6, October 2021, pp.760-763, URL: www.ijtsrd.com/papers/ijtsrd47493.pdf

IJTSRD47493

KEYWORDS: *Natural Language Processing, Part of Speech, Named Entity Recognition, Information Extraction*

INTRODUCTION

The process of identifying and classifying all proper nouns in a text document or a sentence into specified classes such as people, places, organisations, dates, addresses, and time expressions is known as named entity recognition. The proper names identified in a text are referred to as Named Entities. A person's name, an organization's name, a location's name, and date and time expressions are examples of identified text. The essential responsibilities of NLP are to make a computer acceptable and to divide these named entities into pre-defined groups. Named Entity Recognition is the name given to this task. Information Extraction is another name for it.

Entity With a Name One of the most important tasks in Natural Language Processing is recognition (NLP). For many years, NER has been a hotbed of research. A named entity is a word or a phrase that clearly identifies a piece of information. One of the most important uses of Natural Language Processing is automatic data extraction. The majority of information extraction technologies take a rigorous approach to extracting data. These approaches are usually tailored to a certain text or language. Words in content are found and arranged into specified classes using NER frameworks. In many domains, data related to NER may be confidential and a technique is required to which it will process any kind of documents without any kind of pre requisite knowledge. One of the major roles of NLP is to generate models that will be useful for human to machine type communication.

PERFORMANCE EVALUATION METRICS ARE:

- Precision (P): Precision is the fraction of the documents retrieved that are relevant to the user's information need.
- Precision (P) = correct answers/answers produced
- Recall (R): Recall is the fraction of the documents that are relevant to the query that are successfully retrieved. Recall (R) = correct answers/total possible correct answers
- F-Measure: The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is

$$\text{F-Measure} = (\beta^2+1)PR/(\ \beta^2R+P)$$

APPROACHES FOR NAME ENTITY RECOGNITION

There are different approaches to name entity recognition. It can be categorized into two broad categories:-

A. Rule based (Linguistic) approaches: - Rule based approaches rely on hand-crafted rules, written by language experts, to recognize and classify NEs. Rule-based approaches may contain Lexicalized grammar, Gazetteer lists, List of triggered words etc. There are two disadvantages for using this approach: first is to developing and maintaining rules and dictionaries is a tedious and costly task. Second these systems cannot be transferred to other languages or domains.

B. Machine learning (Statistical) approaches:- Machine learning approaches rely on statistical models to make predictions about name entities in given text. Large amounts of annotated training data are required for these models to be effective, which can prove costly . There are three main machine learning approaches:-Supervised, Semi-supervised, Unsupervised.

1. Supervised Learning:- Supervised learning approaches build predictive models based on the labelled data and true labels. Some of the supervised machine learning techniques is:

➢ Hidden Markov Model (HMM)
➢ Decision Trees
➢ Maximum Entropy (MaxEnt)
➢ Support Vector Machines (SVM)
➢ Conditional Random Fields (CRFs)

2. Unsupervised Learning: Unsupervised learning approaches don't expect any implicit or structural information about the data they are processing. The typical approach to unsupervised learning is clustering. For example, one can try to collect names from clustered groups based on the similarity of context. There are other methods also, which are unattended. Basically, the techniques based on lexical resources (e.g. WorldNet) calculated on lexical patterns and statistics on a large unannotated corpus.

3. Semi supervised learning: The term semi-supervision or weak supervision is still relatively young. The main SSL technology is called bootstrapping and includes a small measure of control, like a row of seeds, for the beginning of the learning process. In semi-supervised approach, a model is trained on an initial set of labelled data and true labels, then, predictions are made on a separate set of unlabelled data, and then improved models are created iteratively using predictions of previously developed models. For example, a system aimed at "disease names" could prompt the user to give a small number of example names.

RELATED WORK

The Paper titled “Named Entity Recognition for Question Answering” [1] proposes a different approach where named entity recognition is done is in a question answer format. As compared to named entity in the traditional way this paper has a better way of getting results. The methodology gives a better chance to find questions to all answers. The future work of this paper includes multiple labels on Entity Recognition on higher Question Answering systems.

The paper titled “Named Entity Recognition: Exploring Features” [2] focuses on complete features in identifying supervised NER, and various combinations of features and their result on recognizing performance. This paper mainly focuses on variety of features, which are mined from a word being labelled and analysis is done on the same and the effectiveness of a supervised NER system, various individual features and combinations on the effectiveness of named entity recognition is also focused in the paper. The paper aims to extend their work on clustering features and their effective combinations of Named entity recognition.

The Paper titled “A Survey of Named Entity Recognition and Classification”[3]aims at improving the named entity recognition for Indian languages using both supervised and non-supervised methods, and various statistical measures are used for the study of the same and also study of neural network approaches for named entity recognition. The paper made observations like language factor, domain factor; entity factor etc. The paper emphasizes on the suitability of neural networks the field of NER.

The paper titled “Named Entity Recognition for Indian Languages:” A Survey [4] deals with how languages play a vital role in NER, Language being the fundamental goal for communication and helps in enabling machine type communication. This paper tells about how language plays a vigorous role in hearing, talking, speaking etc. The major part of NER is to identify and categories different words in a text format into its subsequent categories like person name, place names, quantities etc. E.g.: Sonia is from Gujarat. This can be identified in two ways Sonia is a person name and Gujarat a place name .The paper came up with 13 noun taggers for entity recognition like person names, location names and organization names ,also used Hidden markov model in supervised learning technique and statistical models with generalized learning method in this paper. The major challenge of this paper is that all Indian languages do not have capitalized forms of nouns and Indian languages are varied when compared to other languages. The paper deals with languages such as Oriya, Punjabi and related Indian languages.

The paper titled “Named Entity: History and Future” [5] discusses about the history of NER future of the same, the problems faced and how these problems could be solved. The paper describes how drastic changes are taking place in this field, changes from tagging of only proper names to tagging a wide variety of words and expressions which humans call it information. The paper demonstrates results after three types of study namely weakly supervised, active learning and unsupervised learning. The weakly learning focuses on extracting relations of entities such as book titles, author names etc, active learning have better outcomes that could be achieved by annotating each data that has been tagged. Unsupervised learning refers to data without label. Entity Recognition plays a vital role as a technology for applications in the field of natural language processing.

The paper titled “Named Entity Recognition using Machine Learning and pattern Selection Rules”[6]discusses about significance of machine learning in NER. The paper has proposed methods and rules namely hybrid method and maximum entropy model. The data used are extracted from tagged data sets.

PROPOSED WORK

With the growing availability of data, extracting valuable information from it has become a top priority in all fields. Information extraction gets more difficult when the data is available as documents written in natural language. NER is a technique for extracting usable information from unstructured natural language document sets that is widely used. NER is one of the primary steps in Natural Language Processing (NLP) for text analysis and is used for both online applications and stand-alone systems.

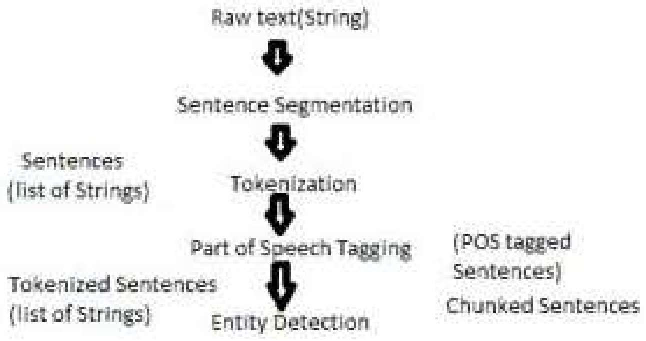
RESEARCH OBJECTIVES

The main goal is to find and classify named items in text into specified categories, such as people's names, organisations' names, locations, events, time expressions, quantities, monetary values, percentages, and so on.

- Detect a named entity
- Categorize the entity

So, first, we must define entity categories such as Name, Location, Event, Organization, and so on, and then feed relevant training data to a NER model. We'll finally educate our NER model to detect and categorise entities by tagging certain examples of words and phrases with their matching entities.

RESEARCH METHDOLOGY



CONCLUSION

Anything that has to do with a name is referred to as a named entity. A suitable sequence of name identification and classification is known as named entity recognition. Information Extraction's key subtask is NER. Information extraction, question answering, machine translation, automatic indexing of documents, cross-lingual information retrieval, text summarization, and other NER applications can be found and observed in a variety of fields of knowledge and science. We give an overview of various ways for identifying Named Entities (NE) in Indian languages in this work.

REFERENCES

[1] Mollá, Diego, Menno Van Zaanen, and Steve Cassidy. "Named entity recognition in question answering of speech data." Proceedings of the Australasian Language Technology Workshop. 2007.

[2] Tkachenko, Maksim, and Andrey Simanovsky. "Named entity recognition: Exploring features." KONVENS. 2012.

[3] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30.1 (2007): 3-26.

[4] Pillai, Anitha S., and L. Sobha. "Named entity recognition for indian languages: A survey." International Journal 3.11 (2013).

[5] Sekine, Satoshi. "Named entity: History and future." Project notes, New York University (2004): 4.

[6] Seon, Choong-Nyoung, et al. "Named Entity Recognition using Machine Learning Methods and Pattern-Selection Rules." NLPRS. 2001.

[7] Kaur, Darvinder, and Vishal Gupta. "A survey of named entity recognition in english and other Indian languages." IJCSI International Journal

of Computer Science Issues 7.6 (2010): 1694-0814.

[8] 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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Kumar N. and Bhattacharyya Pushpak. 2006. “Named Entity Recognition in Hindi using MEMM” in the proceedings of Technical Report, IIT Bombay, India.

[14] 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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.