NAMED ENTITY RECOGNIZATION

A Project Report

Submitted in the partial fulfillment of the requirements for the award of the degree of

# Bachelor of Technology in

Department of Computer Science and Engineering

By

Manichand (2010030455)

Dileep Reddy (2010030416)

Harsha Vardhan (2010030173)

Prakash Raj (2010030533)

UNDER THE SUPERVISION OF

DR. ARPITA



Department of Computer Science and Engineering

K L University Hyderabad,

Aziz Nagar, Moinabad Road, Hyderabad – 500 075, Telangana, India.

March

**Declaration**

The Social Internship Report entitled “**NAMED ENTITY RECOGNIZATION**” is a record of bonafide work of Manichand (2010030455), Dileep Reddy(20100030415), Harsha (2010030173), Prakash raj(2010030533), submitted in partial fulfillment for the award of B.Tech in the Department of Computer Science and Engineering to the K L University, Hyderabad. The results embodied in this report have not been copied from any other Departments/ University/ Institute.

<Signature of the Students>

**Certificate**

This is to certify that the Social Internship Report entitled “**NAMED ENTITY RECONIZATION**” is being Manichand(2010030455), Dileepreddy(20100030415), Harsha(2010030173), Prakashraj(2010030533), submitted in partial fulfillment for the award of B. Tech in CSE to the K L University, Hyderabad is a record of bonafide work carried out under our guidance and supervision.

The results embodied in this report have not been copied from any other departments/ University/Institute.

## Signature of the Supervisor

## Signature of the HOD Signature of the External Examiner

**ACKNOWLEDGEMENT**

First and foremost, we thank the lord almighty for all his grace & mercy showered upon us, for completing this project successfully.

We take grateful opportunity to thank our beloved Founder and Chairman who has given constant encouragement during our course and motivated us to do this project. We are grateful to our Principal **Dr. L. Koteswara Rao** who has been constantly bearing the torch for all the curricular activities undertaken by us.

We pay our grateful acknowledgement & sincere thanks to our Head of the Department

**Dr. Chiranjeevi Manike** for his exemplary guidance, monitoring, and constant encouragement throughout the course of the project. We thank **Dr. ARPITA** of our department who has supported throughout this project holding a position of supervisor.

We whole heartedly thank all the teaching and non-teaching staff of our department without whom we won’t have made this project a reality. We would like to extend our sincere thanks especially to our parent, our family members and friends who have supported us to make this project a grand success.

**TABLE CONTENTS**

1. **ABSTRACT**
2. **INTRODUCTION**
3. **PROBLEM STATEMENT**
4. **LITERATURE SURVEY**
5. **DATASET COLLECTION**
6. **FLOW DIAGRAM**
7. **PREPROCESSING FLOWCHAT**
8. **PREPROCESSING TECHNIQUE CODES**
9. **PROPOSED MODEL/TECHNIQUE**

**10. CONCLUTION**

**11.REFERENCES**

**ABSTRACT :**

With the increase in availability of data, extraction of useful information from this data has become most important activity across all domains. When the data is available as documents written in natural language, information extraction becomes more challenging. Named Entity recognition (NER) is a technique used extensively for automatic extraction of useful information from unstructured natural language document collections. Used both, for web applications as well as stand-alone systems, NER is considered as one of the major step in Natural Language Processing (NLP) for analysis of text. This paper discusses basics of NER, various algorithms used for NER and major applications and challenges in the field of NER.

Keywords: Named Entity Recognition, Information Extraction, Natural Language Processing, Part of Speech.

**INTRODUCTION :**

Named Entity Recognition (NER) is a natural language processing task that attempts to find named entities (persons, locations, organizations, etc.) within a text. Since the tag of one word depends on neighboring words (and their tags), the problem is modeled as a sequence labeling task, wherein a sentence is considered as an input, and the goal is to produce the tags of all words jointly, so as to maximize accuracy**.**

**PROBLEM STATEMENT :-**

Named Entity Recognition (NER) is a natural language processing task that attempts to find named entities (persons, locations, organizations, etc.) within a text. Since the tag of one word depends on neighboring words (and their tags), the problem is modeled as a sequence labeling task, wherein a sentence is considered as an input, and the goal is to produce the tags of all words jointly, so as to maximize accuracy.

**LITERATURE SURVEY:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TITTLE | AUTHOUR | PUBLISH YEAR | PROS | CONS |
| Neural Architectures for Named Entity Recognition | Guillaume Lample,Miguel Ballesteros, Sandeep Subramanian ,Kazuya Kawakami Chris Dyer | 2016 | automatically categorizing the articles in defined hierarchies and enable smooth content  discovery | Ambiguity and Abbreviations - In identifying named entities is language.Recognizing words which can have multiple meanings or words that can be a part  of different sentences |
| Named Entity Recognition with Bidirectional LSTM-CNNs | Jason P.C. Chiu, Eric Nichols | 2016 | automatically scan entire articles and reveal which are the major people, organizations,  and places discussed in them. | Same words can be written in long forms.Words which will sometimes require some label for identification. |
| A Unified MRC Framework for Named Entity Recognition | Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, andJiwei Li | 2020 | used in information extraction to identify and segment the named entities and classify or categorize them under various predefined classes | Words which are not used very frequentlythese days, or words that are not heard by a lot of people |
| TENER: Adapting Transformer Encoder for Named Entity Recognition | Hang Yan, Bocao Deng, Xiaonan Li, Xipeng Qiu | 2019 | helps you easily identify the key elements in a text, like names of people, places,  brands, monetary values, and more | the target words are mainly proper nouns or unregistered words. |
| Few-shot Learning for Named Entity Recognition in  Medical Text | Maximilian Hofer, Andrey Kormilitzin,  Paul Goldberg,  Alejo Nevado-Holgado | 2019 | Extracting the main entities in a text helps sort unstructured data and detect important information, which is crucial if you have to deal with large datasets. | It is very hard to collect to all datasets. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Named Entity Recognition using Hidden Markov  Model (HMM) | Sudha Morwal  , Nusrat Jahan  and Deepti Chopra | 2020 | Mostly the researcher uses hybrid NER system which take advantages of both rule-based and statistical approaches so that the performance of NER system can be improved. | The sequence labeling, in addition to having a relationship with individual words, also relates to such aspects as the observed sequence length, word context and others. |
| Using Stanford NER and Illinois NER to Detect Malay Named Entity Recognition | S. Sulaiman, R. Abdul Wahid, S. Sarkawi, and N. Omar | 2021 | These documents consist different types of categories such as news, business, education and articles. This experiment was conducted to see whether the English NER can be used to tag Malay Named Entity. | Due to limitations in accessing a Malay tagged corpus, we used online newspaper reports which were selected randomly as our testing data. |
| A Survey on Recent Advances in Named Entity Recognition from Deep  Learning models | Vikas Yadav  Steven Bethard | 2020 | The survey of models for named entity recognition, covering both classic feature-engineered machine learning models, and modern feature-inferring neural network models has yielded several important insights. | Words which are not used very frequentlythese days, or words that are not heard by a lot of people |
| Telugu named entity recognition | SaiKiranmai Gorla, Sai Sharan Tangeda, Lalita Bhanu Murthy Neti & Aruna Malapati | 2021 | It will recongise telugu words | the target words are mainly proper nouns or unregistered words. |
| Iterative Named Entity Recognition with Conditional Random Fields | Ana Alves-Pinto ,Christoph Demus ,Michael Spranger ,Dirk Labudde and Eleanor Hobley | 2021 | automatically categorizing the articles in defined hierarchies and enable smooth content  discovery | Same words can be written in long forms.Words which will sometimes require some label for identification. |

The first comprehensive NER survey was Nadeau and Sekine (2007), which covered a variety of supervised, semi-supervised and unsupervised NER systems, highlighted common features used by NER systems during that time, and explained NER evaluation metrics that are still in use today. Sharnagat (2014) presented a more recent NER survey that also included supervised, semi-supervised, and unsupervised NER systems, and included a few introductory neural network NER systems. There have also been surveys focused on NER systems for specific domains and languages, including biomedical NER,(Leaman and Gonzalez, 2008), Chinese clinical NER (Lei et al., 2013), Arabic NER (Shaalan, 2020; Etaiwi et al., 2019), and NER for Indian languages (Patil et al., 2021).The existing surveys primarily cover feature-engineered machine learning models (including supervised, semi-supervised, and unsupervised systems), and mostly focus on a single language or a single domain. There is not yet, to our knowledge, a comprehensive survey of modern neural network NER systems, nor is there a survey that compares feature engineered and neural network systems in both multi-lingual (CoNLL 2002 and CoNLL 2003) and multi-domain (e.g., news and medical) settings.

**DATASET COLLECTION:**

|  |  |  |
| --- | --- | --- |
| Data set | charcterstics | technique |
| NCBI-disease | consists of 22,137 sentences totally and is split into 14,987, 3,466 and 3,684 sen- tences for the training It includes annotations for 4 types of named entities: PERSON, LOCATION, ORGANIZATION and MISC | BERT, Bi-LSTM, CNN |
| Ontonotes v5 | made up of 1,745 K English, data.  full list of 18 categories | Transformer, LSTM,MLMET |
| CoNLL 2003 | It has 18 entities  It has 14,987 testing sets | knowledge distillation, GCN,  Bender |

**FLOWDIAGRAM:**

Feature Extraction/

Information extraction

Lebelling

Pre - processing

Document collection

Data training Data test

NER Algorithms

Feature Extraction/

Information extraction

Results

**PREPROCESSING FLOWCHART:**

**TOKENIZATION**

**STOPWORD REMOVAL**

**STEMMING**

**POS-TAGGING**

Evaluate

**LEMMATIZATION**

**PREPROCESSING TECHNIQUE CODES:**

**TOKENIZATION:**

*import* nltk

sentence\_data = "The First sentence is about Python. The Second: about Django. You can learn Python,Django and Data Ananlysis here. "

nltk\_tokens = nltk.sent\_tokenize(sentence\_data)

print (nltk\_tokens)

**STOPWORD REMOVAL:**

*import* nltk

*from* nltk.corpus *import* stopwords

print(stopwords.words('english'))

**STEMMING:**

*from* nltk.stem *import* PorterStemmer

*from* nltk.tokenize *import* sent\_tokenize, word\_tokenize

ps = PorterStemmer( )

text\_example = "your text goes here"

words = word\_tokenize (text\_example)

*for* w *in* words :

 print(ps.stem(w))

**LEMMATIZATION:**

*import* nltk

*from* nltk.stem *import*   WordNetLemmatizer

wordnet\_lemmatizer = WordNetLemmatizer()

text = "studies studying cries cry"

tokenization = nltk.word\_tokenize(text)

*for* w *in* tokenization:

    print("Lemma for {} is {}".format(w, wordnet\_lemmatizer.lemmatize(w)))

**POS-TAGGING:**

*import* nltk

*from* nltk *import* word\_tokenize

sentence = "I am going to school"

print (nltk.pos\_tag(word\_tokenize(sentence)))

**PROPOSED TECHNIQUES / MODEL: -**

LSTM networks are well-suited to classifying, processing and making predictions based on time series data.

The first part chooses whether the information coming from the previous timestamp is to be remembered or is irrelevant and can be forgotten. In the second part, the cell tries to learn new information from the input to this cell. At last, in the third part, the cell passes the updated information from the current timestamp to the next timestamp.

These three parts of an LSTM cell are known as gates. The first part is called Forget gate, the second part is known as the Input gate and the last one is the Output gate.

Diagram

Description automatically generated

**Conclusion:**

Named Entity Recognition, a sub process of natural language processing, plays very important role in automated information extraction NER is suited to any situation in which a high-level overview of a large quantity of text is helpful. With NER, you can, at a glance, understand the subject or theme of a body of text and quickly group texts based on their relevancy or similarity.

**References:-**

1. Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. "Neural architectures for named entity recognition." arXiv preprint arXiv:1603.01360 (2016).
2. Chiu, Jason PC, and Eric Nichols. "Named entity recognition with bidirectional LSTM-CNNs." Transactions of the association for computational linguistics 4 (2016): 357-370.
3. Li, Xiaoya, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. "A unified MRC framework for named entity recognition." arXiv preprint arXiv:1910.11476 (2019).
4. Yan, Hang, Bocao Deng, Xiaonan Li, and Xipeng Qiu. "TENER: adapting transformer encoder for named entity recognition." arXiv preprint arXiv:1911.04474 (2019).
5. Hofer, Maximilian, Andrey Kormilitzin, Paul Goldberg, and Alejo Nevado-Holgado. "Few-shot learning for named entity recognition in medical text." arXiv preprint arXiv:1811.05468 (2020).
6. Morwal, Sudha, Nusrat Jahan, and Deepti Chopra. "Named entity recognition using hidden Markov model (HMM)." International Journal on Natural Language Computing (IJNLC) Vol 1 (2022).
7. Sulaiman, S., R. Wahid, S. Sarkawi, and N. Omar. "Using stanford NER and Illinois NER to detect malay named entity recognition." Int. J. Comput. Theory Eng 9, no. 2 (2019): 147-150.
8. Yadav, Vikas, and Steven Bethard. "A survey on recent advances in named entity recognition from deep learning models." arXiv preprint arXiv:1910.11470 (2019).
9. Gorla, SaiKiranmai, Sai Sharan Tangeda, Lalita Bhanu Murthy Neti, and Aruna Malapati. "Telugu named entity recognition using bert." International Journal of Data Science and Analytics (2022): 1-14.
10. Alves-Pinto, Ana, Christoph Demus, Michael Spranger, Dirk Labudde, and Eleanor Hobley. "Iterative Named Entity Recognition with Conditional Random Fields." Applied Sciences 12, no. 1 (2021): 330.

**BERT:**

1. Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36, no. 4 (2020): 1234-1240.

**Transformer: -**

1. Xu, Lu, Zhanming Jie, Wei Lu, and Lidong Bing. "Better Feature Integration for Named Entity Recognition." *arXiv preprint arXiv:2104.05316* (2021).

**knowledge distillation:-**

1. Zhou, Wenxuan, and Muhao Chen. "Learning from Noisy Labels for Entity-Centric Information Extraction." *arXiv preprint arXiv:2104.08656* (2021).

**Bi-LSTM:-**

1. Jang, Beakcheol, Myeonghwi Kim, Gaspard Harerimana, Sang-ug Kang, and Jong Wook Kim. "Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism." Applied Sciences 10, no. 17 (2020): 5841.

**CNN:-**

1. Zhu, Qile, Xiaolin Li, Ana Conesa, and Cécile Pereira. "GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text." Bioinformatics 34, no. 9 (2018): 1547-1554.

**LSTM:-**

1. Cho, Minsoo, Jihwan Ha, Chihyun Park, and Sanghyun Park. "Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition." Journal of biomedical informatics 103 (2020): 103381.

**MLMET:-**

|  |
| --- |
|  |
|  |

1. Dai, Hongliang, Yangqiu Song, and Haixun Wang. "Ultra-Fine Entity Typing with Weak Supervision from a Masked Language Model." arXiv preprint arXiv:2106.04098 (2021).

GCN :-

1. Seti, Xieraili, Aishan Wumaier, Turgen Yibulayin, Diliyaer Paerhati, Lulu Wang, and Alimu Saimaiti. "Named-entity recognition in sports field based on a character-level graph convolutional network." Information 11, no. 1 (2020): 30.

**Bender:-**

1. Li, Jing, Aixin Sun, Jianglei Han, and Chenliang Li. "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering (2020).