International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

Entity Recognition by Natural Language Processing and Machine Learning

Supriya Shelake[1], Vikas Honmane[2]

[1]Student, Department of Computer Science and Engineering, Walchand College of Engineering Sangli, India
[2]Assistant Professor, Department of Computer Science and Engineering, Walchand College of Engineering, Sangli, India.

Abstract - *Abstract—Named Entity Recognition is a sub-task for Information Extraction.. Data is derived from unstructured and semistructured data by extracting structured information. A subtool that is used to collect targeted information, called Named Entity Recognition (NER). This explains how we handle these and have been adopted to identify, define, and annotate the designated entities by correct tags. In different applications of natural language processing (NLP) such as machine translation, answering questions, summarizing text, extracting information, NER systems are very useful. Various methods and algorithms are used for text summarization. Approaches to machine learning for NER are Conditional Random field (CRF), Support Vector Machine (SVM), Decision Tree. Deep learning approach such as Long Short Term Memory (LSTM). NER is a technique for extracting information that identifies and classifies named entities in text.*

***Key Words*: Named Entity Recognition(NER), Information Extraction, Information Retrieval.**

1. INTRODUCTION

A Named Entity is an object of the real world, such as name of a person, name of place, name of organization, name of product, chemical entities, biomedical entities that can be denoted with proper name or proper noun. Named Entity Recognition is also known as entity extraction categorizes named entities that are present in a text into predefined categories “individuals,” “companies,” “places,” “organizations,” “cities,” “times,” “information terminologies” etc. It strengthens your content with a wealth of semantic information and lets you easily understand every text. Entity Extraction is an entity to shape structured data from unstructured data. Unstructured data ensures that raw data such as social media such as twitter, facebook and instagram are included. Extraction of named entities is separating named entities from unstructured data. Unstructured data includes newspaper articles, various posts, Wikipedia. Entity Recognition is the most important technique for extracting, obtaining information, answering questions, and machine translation.

2. RELATED WORK

There are mainly three NER approaches-linguistic approach, machine learning approach, hybrid approach. Linguistic approach works on handmade rules written by experienced linguists. Previous rule-based NER systems, consisting primarily of lexical grammar, gazette lists and word list triggers. It needs advanced knowledge of the laws of grammar and of other languages. Machine-learning (ML) methods are commonly used in NER. These are simple to train, can be adapted to various domains and languages and are not very expensive. NER can be done with various machine learning approaches like hidden markov model (HMM), conditional random field (CRF), entropy, etc. The hybrid NER method incorporates both a rule-based approach and an approach to machine learning [1]. Author proposed that there are various machine learning algorithms that work for Named Entity Recognition like conditional random field (CRF), hidden markov model (HMM), support vector machine (SVM), Decision tree. The Hidden Markov Model produces a sequence of tokens. It is based upon the property of the Markov chain. The probability of next state occurrence depends on the previous state. Decision trees are most important for prediction and classification but they are computationally expensive for training. Conditional Random field well suited for labelling. Without taking into account neighboring samples, the classifier predicts that the label will be taken into account through CRF. Combine conditional random fields with other features and rules to improve the performance of NER. The Support Vector machine is used for text categorization. SVM is used for identifying hand-written characters. Support Vector Machine is suitable for two classes not for multiple classes[3]. Author proposed many machine learning algorithms for entity extraction. It is critical for working in native languages. Because various languages other than the English language have different grammatical rules [2]. LSTM is also used for entity extraction. LSTM stands for long-short term memory which is Recurrent Neural Network. Recurrent Neural Network is more suitable for sequence data [4]. Bi-directional long short term memory is used to extract entities with conditional random fields. Bi-LSTM-CRF-CNN is a hybrid approach in which CNN layer is aimed at extracting subword information, CRF works well for POS tagging. As a Bi-LSTM input, Bi-LSTM processes the sequence data, output of CNN and word embedding. Conditional Random field employed together with Bidirec-tional LSTM which captures past and future

dependencies. It is suitable for NE tagging [5]. Convolutional Neural Network is a deep learning network used for mining named entities, representing samples with word embedding [6]. The Author has researched three different kinds of deep learning architectures, such as FFN, RNN, and hybrid CNN, and compared the effects of these three models to other Bio-NLP systems involved [7]. Bi-LSTMs at word and character level to determine the state of the art in POS tagging [8].

3. PROPOSED METHODOLOGY

Named Entity Recognition (NER) is a standard NLP problem involving the identification and classification of named entities (people, locations, organisations, etc.) from a piece of text to a predefined list of categories. Recognizing named entities is a specific form of extraction of chunks which uses entity tags along with chunk tags.

3.1 Various approaches to solving NER issues

The NER methods at an early stage were largely based on two-stage rules. The entities are defined and determined during the first stage. The second stage involves the collection and extraction of the entities. By the NER algorithms seem to be better advanced machine learning in NLP the implementation is mathematical models and a broad corpus is trained. These approaches can be classified as supervised in the following categories Unsupervised learning, and semi-supervised learning. Supervised learning requires annotated, large-scale structure to train the model. Supervised approaches to learning with NER are hidden markov model(HMM), maximum entropy (ME), and conditional random field. Semi-supervised learning models on a smaller annotated corpus are trained. Unsupervised learning models typically rely on clustering of vocabulary with other tools like WordNet. The realistic approach to NER can be based either on linguistic grammar or machine learning techniques or a combination of both. Statistical machine learning typically requires a huge, manually noted dataset of training.

3.2 Challenges in Named Entity Recognition

NER, though considered a basic feature of NLP, is questioned by specific dynamics, inherent in every natural language. Few of the challenges are outlined below:

- Ambiguity and Abbreviations: One of the main problems in identifying named entities is language. Recognizing different words that may have different meanings or terms that may be part of numerous sentences. Another critical problem is the classifying words which are similar. Multiple words or phrases written in various forms. Words can be conveniently abbreviated for learning, writing and understanding. Same words can be written in long forms. Another major challenge is words which often require a label to identify.
- Spelling Variations: The vowels (a, e, i o, u) play a very important part in English language. Words that do not make a huge difference in phonetics but make a big difference in the way they are written and their spelling.
- Foreign Words: Words that are not used so much these days, or words that many people don't understand, is another big challenge in this field. Words such as names of people, names of locations etc.

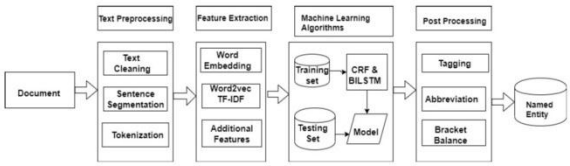

Fig. 1: Methodology

3.3 Data Cleaning

Data from various documents, newspapers, Internet, Social Media, Wikipedia or any articles from various regional languages which is taken as Input. This data is in an unstructured format. So data pre-processing is required for extraction and classification of named entities. NLTK suite of libraries is used for Named Entity Recognition. Spacy library is also used for Named Entity Recognition.

To get structured data from text data preprocessing is required. Data preprocessing involves Sentence Segmentation, tokenization, lemmatization, stopword removal, part-of-speech tagging, TF-IDF.

- Tokenization: Tokenization is a process of breaking up text into sentences and words.
- Sentence Segmentation: Sentence segmentation divides the sentence into words.
- Lemmatization: Lemmatization is the conversion of a word to its root form.
- Stemming: Stemming is reducing related words to a common stem.
- Stopword Removal: Remove commonly used words.

3.4 Chunking and POS Tagging

- Chunking is a process by which phrases are extracted from unstructured text. Chunking works on POS tagging and it uses pos tags as input and as output chunks. Chunking is important in extracting information from texts like locations, personal names, etc. In the construction of the NP, POS tags are used to define chunk grammar. The standard collection of chunk tags such as noun phrase (NP), Verb Phrase (VP), etc. Defines this by defining a common law of normal expression

- POS Tagging: POS Tagging means simply labeling words with the appropriate Part-Of-Speech. The part of speech describes how a word is used in a sentence. Nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions and interjections these are the eight main parts of-speech. POS tagging is a supervised approach to learning which uses features such as the previous word, next word, first letter capitalization, etc. NLTK has a pos tags feature and operates after the tokenization process. In this scenario, the IOB tagging method can be used. B-chunk type is the prefix before a tag means that it is the start of a chunk. I-chunk type is the prefix that shows it is inside a chunk. O is the tag that shows that the token belongs to no chunk.

Table -1: Named Entities with Entity tags

Named Entity	NE Tag
Person	per
Organization	org
Location	loc
Time	tim
Geographical	geo
Geopolitical	gpe
Artifact	art
Natural phenomenon	nat

4. MACHINE LEARNING APPROACHES TO NAMED ENTITY RECOGNITION

For the classification and recognition of named entities, various machine learning techniques such as Hidden Markov Model (HMM), Conditional Random Field (CRF), Decision Tree, Support Vector Machine (SVM) are used.

- CRF based NER framework: Conditional random fields are a group of models that are best suited to predict contextual tasks. For labeling, Conditional Random Field is used. It is typically used for sequence labeling or parsing information, for example, processing of language and CRFs Named Entity Recognition for the POS labeling. For named object recognition activities, CRFs function well. For CRFs, characteristics can be used. For starters, on the lookout i.e. capitalization, attachments. CRFs are used to forecast sequences.

Denote x as the sequence of input states, i.e. the words of a sentence

$$x = (x1, \ldots\ldots\ldots, xm)$$

y as the output states, i.e. the named entity tags.

$$y = (y1, \ldots\ldots\ldots, ym)$$

For a conditional random field, we model a conditional probability

$$\rho(y1, \ldots\ldots, ym | x1, \ldots\ldots xm)$$

Define this by feature map

$$\Phi(x1, \ldots\ldots, xm, y1, \ldots\ldots ym) \epsilon Rd$$

that maps an entire sequence of inputs x together with entire sequence y to some d-dimensional feature vector. Then model the probability with the parameter vector like a log-linear model. This penalizes the model complexity and is known as regularization.

$$\omega \epsilon Rd$$

4.1 Training CRF

L-BFGS is an optimization algorithm. It uses a limited amount of computer memory. L-BFGS stores only a few vectors that represent the approximation implicitly. It is the default algorithm with Elastic Net (L1+L2) regularization. The method of regularization is the introduction of details to avoid overfitting. Training Set contains known results and the model learns from these data to generalize them to other data. The estimation of the output of test system results.

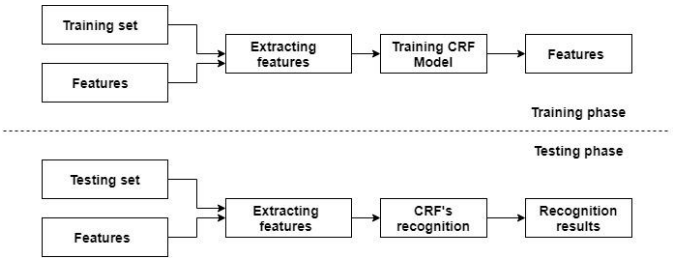

Fig. 2: CRF Model

4.2 Bi-directional-LSTM-CRF Model

The bi-directional LSTM is a combination of two LSTMs, one running forward from "right to left" and the other running backward from "left to right." Bidirectional long short term memory is used for the recognition of entities.

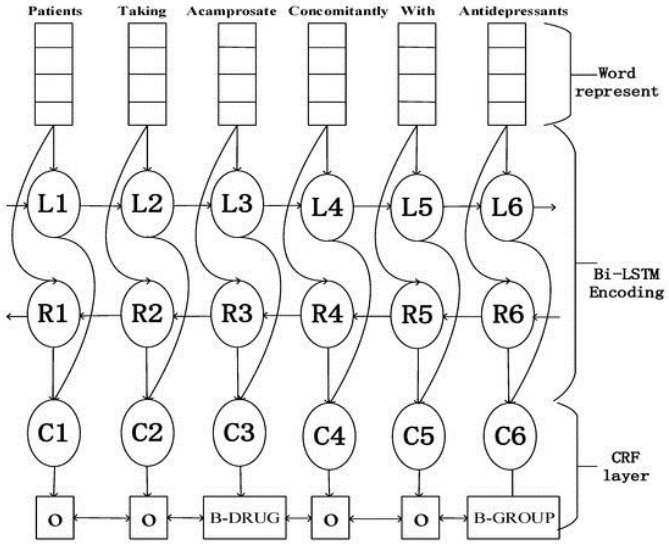

Fig. 3: Bi-LSTM-CRF Model

The two layers of LSTM are forward and backward layers. For capturing past dependencies forward layer is needed and the backward lstm layer is another layer storing future dependency. Entity Recognition is the most important technique for extracting, obtaining information, question answering, machine translation.

Character level vector concatenated as a word presentation with word embedding. Put it on the bi-directional LSTM first and the bidirectional LSTM is loaded into the CRF for abel decoding.

5. PERFORMANCE METRICS

F1-score is used to measure performance metrics. An average of precision and recall is F1-score. Precision indicates how many of the specified elements are significant. Recall indicates how many of the items we found are important. % of selected items that are correct is precision. % of correct items that are selected is recall.

$$f1 - score = 2 * precision * recall \; / \; precision + recall$$

We will use precision, recall and f1-score metrics to measure the model's efficiency as the accuracy is not a reasonable metric for this dataset since we do not have equal number of data points in every class. The below table shows the class wise f1-score of entities.

Table -2: Performance Metrics of NE tags

Performance Metrics

NE tag	Precision	Recall	f1- score
B-per	0.84	0.81	0.83
B-org	0.72	0.74	0.73
B-Geo	0.85	0.89	0.87
B-Gpe	0.97	0.93	0.95
I-per	0.85	0.87	0.86
I-org	0.72	0.81	0.76
I-Geo	0.79	0.78	0.79
I-Gpe	0.93	0.55	0.69

6. CONCLUSIONS

In this study, different language resources should be developed for performance on domain knowledge. Natural language processing is used for cleaning the text using Nltk suite of libraries. In the next step, we will pick more reliable data, create a more complete stopword list and incorporate machine learning and deep learning methods. We proposed a Bidirectional LSTM and CRF model for entity recognition. Lack of an annotated corpus is an obstacle to research. Annotated corpus is easily used for training. The experimental results indicate that the conditional random field is an excellent form of statistical machine learning that has excellent results and recognition of the designated entity.

ACKNOWLEDGEMENT

We are very thankful to all those who have contributed to carry out this work.

REFERENCES

[1] Peng Sun, Xuezhen Yang, Xiaobing Zhao and Zhijuan Wang. "An Overview of Named Entity Recognition," International Conference on Asian Language Processing 2018.

[2] Qianjun Shuai, Runze Wang, Libiao Jin*,Long Pang "Research on Gender Recognition of Names Based on Machine Learning Algorithms,"10th International Conference on Intelligent Human- Machine Systems and Cybernetics 2018.

[3] Deepti Chopra, Nusrat Jahan, Sudha Morwal, "Hindi Named Entity Recognition Aggregating Rule based Heuristic and Hidden Markov Model ," International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.6, November 2012.

[4] Devin Hoesen, Prosa Solusi, Cerdas Bandung, Ayu Purwarianti, "Investigating Bi-LSTM and CRF with POS Tag Embedding for Indonesian Named Entity Tagger," International Conference on Asian Language Processing (IALP) 2018.

[5] Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura and Tomoko Ohkuma, "Character-based Bidirectional

LSTM-CRF with words and characters for Japanese Named Entity Recognition,"Proceedings of the First Workshop on Subword and Character Level Models in NLP.

[6] Xishuang Dong, Lijun Qian, Qiubin Yu, Jinfeng Yang, "A Multiclass Classification Method Based on Deep Learning for Named Entity Recognition in Electronic Medical Records,"IEEE 2016.

[7] Robert Phan, Thoai Man Luu, Rachel Davey, Girija Chetty,"Deep Learning Based Biomedical NER Framework,"IEEE Symposium Series on Computational Intelligence SSCI 2018.

[8] Jason P.C. Chiu, Eric Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs,"IEEE Transactions of the Association for Computational Linguistics, vol. 4, pp. 357370, 2016

BIOGRAPHIES

Ms. Supriya Shelake
P.G Student
Department of Computer Science Engineering
Walchand College of Engineering, Sangli.