Irish Interdisciplinary Journal of Science & Research (IIJSR)
Vol.4, Iss.3, Pages 01-08, July-September 2020

NER Using Machine Learning Approaches

Md Akram Jawed & Dr. Ravinder Kumar

[1]*M.Tech Scholar, Department of ECE, Al-Falah University, Faridabad, Haryana, India.*
[2]*Assistant Professor, Al-Falah University, Faridabad, Haryana, India.*

Article Received: 02 May 2020 Article Accepted: 09 July 2020 Article Published: 02 August 2020

ABSTRACT

Named Entity Recognition (NER) is a subtask of information extraction task that seeks to identify and classify parts in text that consult with predefined classes such as person names, organisations, places, time expressions, quantities and monetary values. This paper gives the detailed of various approaches and techniques that are used for Named Entity Recognition that identifies the named entities from the given knowledge. It explains the different machine learning models that are employed in the Named Entity Recognition. The performances of varied Named Entity Recognition systems that are developed using machine learning models are reviewed. Further, we examine the area and domain of its potential and conventional usage in modern times. We lastly conclude our multifaceted study by mentioning some of the common challenges that are found in the NER technology and suggest some practical solutions for improving the same.

Keywords: *Named Entity Recognition, Data Extraction, Applications, Challenges.*

1. Introduction

The US agency namely Defence Advanced Research Projects Agency (DARPA) convened many Message Understanding Conferences (MUC) to foster the development of new and improved methods of Information Extraction (IE). Given fields for Information Extraction for the first five conferences were, for example, the cause, the agent, the time and place of an event, and the consequences etc. With every next conference, the numeracy of fields kept increasing. At the sixth conference (MUC-6) the task of recognition of named entities and co-reference was added. For named entity all phrases in the text were supposed to be marked as person, location, organization, time or quantity. [1]

According to A. Mansouri et al. [2] "Recognizing and extracting exact name entities, like Persons, Locations and Organizations are very useful to mining information from text. Learning to extract names in natural language text is called Named Entity Recognition (NER) ". Eftimov et al. [3] defines "NER is a process in which a label (class) or semantic category from a predefined set is assigned to the words or phrases known as entity mentions in order to describe the concept". White Paper of Data Community DC [4] defines "the task of Named Entity Recognition and Classification as the identification of named entities in computer readable text via annotation with categorization tags for information extraction." From the above words of varied nature, we can conclude Named Entity Recognition as "important task that extracts certain information usually a distinct word or expression including Time & Date or Percentage numeric from a huge amount of fully unstructured or semi-structured data in order to minimize the efforts and time taken of searching the whole text thoroughly and repeatedly".

2. Machine Learning Approaches

In machine learning approaches the model should be trained exploitation the annotations that are present on the documents that are annotated. Statistical models are utilized by the machine learning approaches that

specialize in automatically recognizing the sequence labelling algorithmic rule or patterns and create choices supported the determined information. In Named Entity Recognition machine learning approaches focuses on extracting entities and classifying them from the given knowledge. Different options are applied to the text and these representations of text are employed by the machine learning models to recognize the options present within the coaching knowledge that helps to automatically generate patterns for characteristic the information that is analogous within the unseen knowledge. Machine Learning approaches are generally classified into three categories:

1) *Supervised Learning*

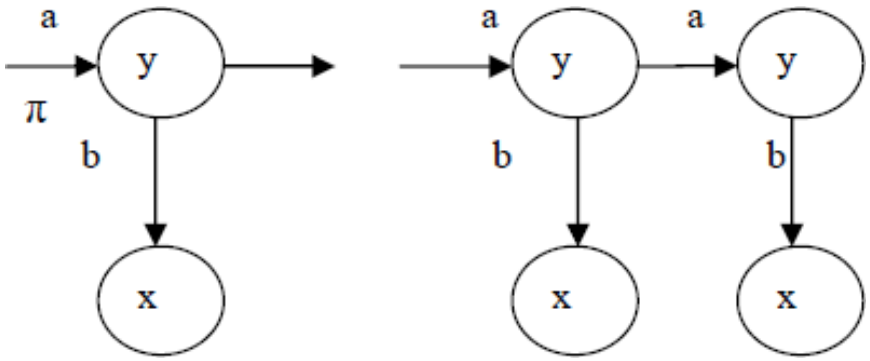
2) *Semi-supervised Learning*

3) *Unsupervised Learning.*

2.1 Supervised Machine Learning: In supervised learning, labelled training data is desired to develop a statistical model because the labelled example benefits the system to take the right decision. It is not conceivable for a supervised learning method to achieve a high-performance without large quantities of training data. In supervised learning, if the quantity of training data is very small it may lead to data sparseness problem. The different supervised learning models are Hidden Markov Model (HMM), Conditional Random Fields (CRF), Support Vector Machine (SVM), Maximum Entropy Model (ME) and Decision Trees.

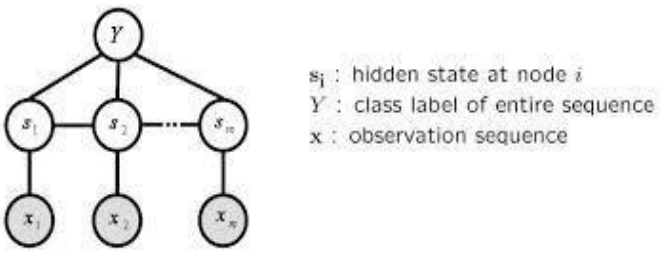
2.1.1 Hidden Markov Model- HMM is one of the earliest supervised models used in Language Processing problems. HMM is a model, analysis of which discloses the stages the result went through before coming in the present form. [5] Daniel M. Bikel, Richard Schwartz, Ralph M. Weischedel (1999) proposes a Markov model namely Identity Finder TM in his Shakespearian Paper. Since HMM model defines the chance for the observation symbols and state transition there's a likelihood related to every transition. Hence, this model will predict the Named entity of the following word once the Named Entity category of the previous word is given [6].Hidden markov model has a model M=(O,Q,A,B,π) where,

A={aij, i, j=1…..N}, B={i(t) by, i=1,……t=1} and O, Q mean a finite set of observation symbol of x and y.[9]



2.1.1 Support Vector Machine: [7] proposed by Corinna Cortes and Vladimir Vapnik (1995) Support Vector Machine is a supervised method for two group classification problems, where an algorithm is put to learn the linear hyperplane which segregates the data into two classes. The hyperplane helps to reason the determined information into positive and the negative category that's present on the alternative sides of the hyperplane. This model computes gap of each vector from the hyperplane that is understood as margin [8].

2.1.2 Conditional Random Field: Conditional Random Field (CRF) is a discriminative probability model. [10] Introduced by John Lafferty et al. (2001).A separate classifier predicts a label for entities solely while not considering the context or the close entities, however a CRF model can take the context into account [11] .Let G = (Y, E) be a graph such that vertex YV is a random variable. Let P(Yv- except Y)= P(Yv-neighbours(Yv)), then Y is a random field. Let X =random variable w.r.t given data sequences to be labelled as Y=random variable over corresponding label sequence. Let G = (V, E) be a graph such that Y = (Yv)vV, so that Y is indexed by the vertex of G. Then (X,Y)is a CRF , when conditioned on X, the random variables Yv obey the HMM Property w.r.t the graph: P(Yv —X,Yw, w v) = P(Yv — X,Yw,w v), where 'w' 'v' means that 'w' and 'v' are neighbours in G.



2.2 Semi Supervised Learning: In semi supervised learning the model is trained employing a bit of tagged knowledge and so the model is tested using another set of unlabeled knowledge. Therefore, once testing the model with unlabeled knowledge the performance of the model is improved by iterating over the predictions that were done by the previous models [12]. for instance, a NER system that is developed to spot the "drug names" are soliciting for some drug names as associate example, and then the NER systems are attempting to seek out these drug names together with the opposite drug names that occur within the similar context. These steps are recurrent many times so the system will establish various drug names.

2.3 Unsupervised Learning: The most disadvantage of supervised learning is that it needs an outsized quantity of annotated knowledge however several languages are not having any annotated knowledge. Therefore, to beat this problem unsupervised technique has been planned in Named Entity Recognition. Cluster is one in every of the everyday approaches to unsupervised learning. In cluster, we can identify the info from the clustered cluster on the premise of comparable context. The most disadvantage of unsupervised learning is that it needs lots of options and it's totally hooked in to the lexical patterns and also the statistics on outsized annotated information [12].

3. Literature Survey

1) N. Kanya and T. Ravi [13], the author has reviewed named entity recognition system that aims to extract significant data from the medicine archives. The author demonstrates the extraction of entities within the datasets like PUBMED and MEDLINE using machine learning algorithms like SVM and CRF. Three corpora (Breast cancer corpus, carcinoma corpus, and Thyroid cancer corpus) were retrieved from the PUBMED corpus, Precision, Recall and f-score are calculated for these corpora.

2) Lin Yao Hong Liu, Yi Liu, Xinxin Li and Muhammad Waqas Anwar [14], the author explains however the feature information generated by the neural networks from the untagged medicine text files are presented as word vectors. This paper presents a medicine named entity recognition that relies on deep neural network. It uses a skip-gram neural network language model for training the word representations with untagged training information.

3) Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat [2], this paper explains named entity recognition strategies like rule based mostly, machine learning based mostly and hybrid Named entity recognition for extracting entities like Persons, locations, and Organizations from the MUC and also the review of those strategies.

Results of this experiment reveal that Hand-made rule NER incorporates a higher exactness and recall for a selected domain, however it still suffers from movableness. Performance of the machine learning based mostly NER depends on the information coaching price. The performance increases after we train the system with the massive quantity of information.

Authors of this paper have planned a brand new Fuzzy Support vector machine for Named entity recognition. Future work of this paper includes implementing completely different models for achieving high results by process an upscale feature set.

4) Shipra Dingare, Malvina Nissim Jenny Finkel, Christopher Manning and Claire Grover [15], the author presents a Named Entity Recognition supported most entropy system for extracting entities present within the medicine text and reviews its performance. This paper demonstrates the extraction of the assorted medicine entities from the Bio artistic and Bio natural language processing dataset of MEDLINE.

5) Appelt et al. [16], proposed a name identification system supported fastidiously handcrafted regular expression referred to as FASTUS. They divided the task into 3 steps: Recognizing Phrases, Recognizing Patterns and Merging incidents.

4. Datasets

Table.1 provides a quick outline of the foremost unremarkably used datasets for major NLP tasks administered through supervised machine learning models, along with corresponding accuracy level of every model.

Table.1 Analysis of different ML techniques used on different dataset

Datasets	Task	Model	Accuracy	Author (s)
ORCHID Corpus	Word Segmentation	Conditional random field	95.79%	Haruechaiyasak*et al.* (2008)
Penn-Tree Bank	POS	Maximum entropy	81.57%	-
MUC-7	NER	Hidden markov model	90.93%	Todorovic*e et al.* (2008)
News corpus (www.jang.com.pk)	POS	Support vector machine	94.15%	Sajjad & Schmid (2009)
Wall Street Journal (WSJ) and Brown Corpus	POS	SVM And Naïve Bays	Not shown	Gillick (2009)

5. Challenges of NER

Automating the text extraction becomes difficult when the source data is written in natural human language. NER, thus, used to make this task easier by extracting them as per the predefined training. Undoubtedly, the success score of NER is more than satisfactory but like any other technology of any domain NER too is not deprived of its shortcomings. Many problems are faced in the applicability of NER, below we are discussing some of them.

(i) Training the System: Supervised Learning techniques are the most common and used one approach in making any NER algorithm. Success stories of well-trained algorithms are at par with human efficiency. But the thing of concern is the availability of such training data. Information varies drastically from one to another. Medical Terminologies are different than Legal and Legal Terms are way different than Economical. Thus, training an NER system which is capable of recognizing wide terms, requires a great deal of expenditure, time and human labor. Apart from that, except English, there is no readymade text corpus for any other language easily available. Without robust training and a huge repository of corpus at disposal, NER tends to fail. Thus, the training aspect is a significant limiting factor in the way of NE

(ii) Vagueness: Natural Languages spoken by humans are a piece of mystery. They can carry various meanings at a time. The same sentence can mean something else when used somewhere else. Humans can communicate without ambiguity because of their metacognitive abilities which the machines lack. This problem too is equally challenging for NER when dealing with natural languages. It can confuse something with its entity and mis-categorize the data or add something to the database which is not desired. In both the cases, its efficiency is hampered.

(iii) Text variations: In writings, people use various method to write the same thing. There are even different spellings for the same word in different literatures. Informal documents include rough shorthand words. For instance, if the entity recognizes 'As Soon As Possible' and encounters ASAP, it can be a Rubik's Cube for NER.

6. Possible Solutions

The problem of Training can be deal with to some extent by using a dual approach of supervised and unsupervised method where possible amount of data is preloaded in the system while a strong algorithm lets it correct itself and learn from mistakes over time. For further improvement, a separate repository can be created where all the data is recorded every time the system makes a mistake, sorted by the frequency of error, which can be analysed later for updating the system.

The language problems are intrinsic which the most complicated one to deal with are apparently. However, machine learning can improve itself over time as it learns the source language by its exhaustive usage.

A similar approach can be taken for the improvement of language-based problems by preloading not just the bare Data Files and the Dictionary of the language but the Grammar Rules, Thesaurus, and Etymology of that language. So the NER system can draw better relationship between the words.

7. Conclusion

New machine learning techniques and approaches are being chalked out by various researchers and academicians every now and then. Seeing the popularity and interest, we can hope to see an improved and error free Named Entity Recognition system in upcoming years. In this paper, we have discussed the meaning of Named Entity Recognition in an extensive and simple language. We then talked about the learning techniques used in this technology and some common methods or models proposed by eminent scholars or researchers with their in-depth analysis. We then discussed some common problems and shortcomings of this technology and categorized them into sequences.

References

[1] Ralph Grishman, Beth Sundheim, "Message Understanding Conference - 6: A Brief History".

[2] Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat "Named Entity Recognition Approaches".

[3] Tome Eftimov, Barbara Koroušić Seljak, Peter Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations", 2017

[4] "An Introduction to Named Entity Recognition in Natural Language Processing" White Paper of Data Community DC online available at (http://www.datacommunitydc.org/blog/2013/04/a-survey-of-stochastic-and-gazetteer-based-approaches-for-named-entity-recognition) [Accessed on 21.09.2018)

[5] Daniel M. Bikel, Richard Schwartz, Ralph M. Weischedel "An Algorithm that Learns What's in a Name", Feb-1999, Vol-34, Issue1-3 P-211-231.

[6] Shrutika Kale, Sharvari Govilkar. "Survey of Named Entity Recognition Techniques for Various Indian Regional Languages", International Journal of Computer Applications (0975–8887) Volume 164 – No 4, April 2017.

[7] Corinna Cortes, Vladimir Vapnik, "Support-Vector Networks" Sep-1995, Vol.20, Issue 3, pp.273 – 297.

[8] Asif Ekbal, Sivaji Bandyopadhyay, "Named Entity Recognition using Support Vector Machine: A Language Independent Approach", World Academy of Science, Engineering and Technology 2010.

[9] Nusrat Jahan, Sudha Morwal and Deepti Chopra, "Named Entity Recognition in Indian Languages Using Gazetteer Method and Hidden Markov Model: A Hybrid Approach," International Journal of Computer Science & Engineering Technology (IJCSET) ISSN: 2229-3345 Vol. 3 No. 12 Dec 2012.

[10] John Lafferty, Andrew McCallum, and Fernando C.N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", June 2001.

[11] Maithilee, L.Patawar, M. A. Potey, "Approaches to Named Entity Recognition: A Survey", International Journal of Innovative Research in Computer and Communication Engineering", December 2015.

[12] Daljit Kaur, Ashish Verma, "Survey on Name Entity Recognition Used Machine Learning Algorithm", International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, 5875-5879.

[13] N.Kanya and T.Ravi, "Named Entity Recognition from Biomedical Text -An Information Extraction Task", ICTACT Journal on Soft Computing, July 2016, ISSN: 2229- 6956.

[14] Lin Yao Hong Liu, Yi Liu, Xinxin Li and Muhammad Waqas Anwar. " Biomedical Named Entity Recognition based on Deep Neutral Network", International Journal of Hybrid Information Technology (2015), pp.279-288.

[15] Shipra Dingare, Malvina Nissim Jenny Finkel, Christopher Manning and Claire Grover, "A system for identifying named entities in biomedical text: how results from two evaluations reflect on both the system and the evaluations", Comparative and Functional Genomics 2005.

[16] Malarkodi C. S, Elisabeth Lex, Sobha Lalitha Devi, "Named Entity Recognition for the Agricultural Domain", Research in Computing Science (2016).