iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** VI **Month of publication:** June 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AN D ENGINEERING TECHNOLOGY (IJRASET)

Optimized Name Entity Recognition of Machine Translation

Navneet Kaur Aulakh1, Er.Yadwinder Kaur2

1Scholar of Department of Computer Science

CGC, Gharuan Mohali (Punjab)

2Associate Professor Department of Computer Science

CGC, Gharuan Mohali (Punjab)

Abstract—Machine Translation is an important part of Natural Language Processing. Machine Translation refers to using machine to convert one natural language to another language. Machine transliteration for English to Punjabi and Hindi languages pair has been done by using statistical rule based approach. Various statistical rules are constructed with the help of syllabification approach. In this paper Name entity recognition (NER) techniques are explained and how they find name entity from the given input. Translation model check the accuracy of target sentences given the source sentence and decoder maximizes the probability of translated text of target language. We proposed a frame work for machine learning. In which we have make a NER model to find proper entities. In the first section introduction of natural language processing has been described. Second section explains machine translation and its approaches. Next section show how name entity can be recognized. In the next section represents the literature of all research. And at last proposed methodology has been explained.

Index Terms—Natural Language Processing (NLP), Machine Translation (MT), Name Entity Recognition (NER), Different Languages. NER model.

I. INTRODUCTION

Natural Language Processing (NLP) is an area of language research and application that explores how system of computers can be used to understand and manipulate natural language text or speech to do many tasks. The aim of Natural Language Processing researchers is to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the particular tasks. NLP is the field whose aims to convert the human language into the formal representation that is easy to manipulate for the computer. The basic goal of Natural language Processing is to enable a person to communicate with a computer in a language that they use in their everyday life [1].

MACHINE TRANSLATION AND ITS APPROACHES

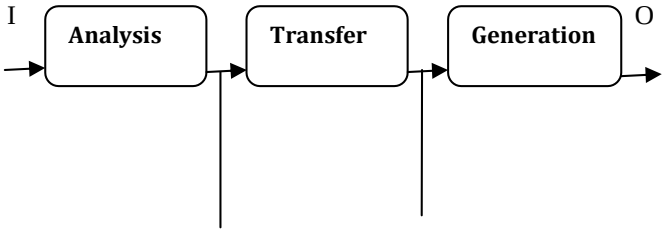
Machine Translation:-Machine translation is defined as the study of planning the systems that can translate one human language into another human language. These systems take input in one natural language and convert it into another human language. The input language is called Source Language and the output language is called Target language. Machine Translation is the sub field of computational linguistic that investigates the use of computer software to translate text or speech from one natural language to another natural language. At the basic level, Machine translation performs simple substitution of word in one natural language for words in another. The literary work is fed to the machine translation system and translation. Machine translation system can breakdown the language barriers by making available work rich sources of literature available to people across the

ISSN: 2321-9653

world.MT also overcomes the technological barrier. In the development process, there are two major goals for machine translation: - (1) Accuracy of translation and (2) Speed [3].

Components of Machine Translation:- Machine translation

Has three main phases- First the system has to analyse the input or source language to create internal representation. Then it manipulates internal representation to transfer a suitable form for target language. Finally it generate the output and target language.



IR on input IR on output

NAME ENTITY RECOGNITION AND APPROACHES

Name Entity Recognition:- Named Entity Recognition is the process of identification and classification of all proper nouns in a given text document or a sentence into pre-defined classes such as persons, locations, organizations, date, address and time expressions. Named Entities are defined as the proper names identified in a text. Identified text may be a person's names, organization's names, location's names, and date and time expressions. To make a computer suitable and divide these named entities into different categories, which are important tasks of NLP. This task is known as Named Entity Recognition. This process is also called Information Extraction [15]. It states-

Ram bought 15 books from Library in 2010. And producing annotated block of text that highlights the names of entities.

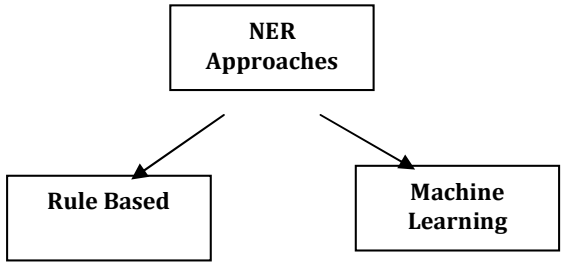
$[Ram]_{Person}$ bought 15 books of $[Library.]_{Organization}$ in $[2010]_{Time}$.

DIFFERENT APPLICATIONS OF NAMED ENTITY RECOGNITION

Name entity recognition is useful in many Natural Language Processing applications, such as- retrieve information is information retrieval, extraction of information, question answering (true or false), parsing, and machine translation (from one language to another). Named entity recognition give the information to users who are looking for person or organization names with fast information [8] .Name entity recognition systems are used in the areas of entity identification in the field of medical science. Earlier, NER systems were used by primarily extraction from journalistic articles and then Automatic Content Extraction (ACE) evaluation also included several types of text styles, such as WebPages and detects text from the speech or any audio.

NAME ENTITY RECOGNITION APPROACHES

Name Entity Recognition has different approaches. It can be divided 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 Name Entities. Rule-based approaches may contain Lexical grammar, Gazetteer lists, List of triggered words [17]. It has two disadvantages of this approach: first is to developing and maintaining rules and dictionaries is a tedious and task is costly. Secondly, it cannot be transferred from one to other languages or domains.

B) Machine learning (Statistical) approaches: - Machine learning approaches rely on statistical models to make predictions about name entities in certain text. It has large amounts of annotated training data are required for these models to be effective, which can prove costly [8]. There are

ISSN: 2321-9653

three main machine learning approaches:-Supervised, Semi-supervised, Unsupervised.

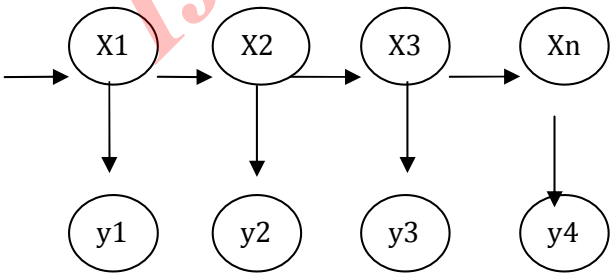
a) Supervised Learning:- SL entails learning a mapping

between a set of input variables X and an output variable Y and applying this mapping to predict the outputs for unseen data. It is the most important methodology in machine learning and it also has an essential importance in the processing of multimedia. SL method builds 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)

1. Hidden Markov Model (HMM)

A Hidden Markov Model is a state machine where the machine enters a state at each time instant and outputs a label from that state. The state that machine is in at any time instant is unknown i.e. hidden and it must be determined by probabilistic methods. HMM is a probabilistic automata based on markov model where a label relates to a state and an observation symbol to a word. State transition and observation symbols are described in probabilistic manner. HMM has a model M=(O,Q,A,B,π) where A={ a_{ij} aalignl ¿¿¡¿ ,i, j=1…..N}, B={ $b_i(y_t)$,i=1,……t=1} and O ,Q mean a finite set of observation symbol of x and y [17].



2. Maximum Entropy (MaxEnt)

Maximum entropy is a conditional probabilistic arrangement model. It has multiple features which are extracted from one word and handle their dependency for the extended term. Maximum entropy is a model for least biased which considers all known facts is the one which maximizes entropy. Each source has a model of exponential that takes the inspection feature as input and distribution over possible next state is a output. The output result labels are related with states. It solves the problem of multiple feature representation and long term dependency issue occurred in HMM. It has increased the recall and greater precision than Hidden Markov Model (HMM). The probability translation leaving any given state must sum to one, so, it is influence towards that states with lower or less outgoing evolutions. The state which has one outgoing state transition will ignore all observations. To overawed Label Bias Problem we can change the state-transition structure or we can start with fully connected model and let the training procedure decide a good structure [19].

3. Conditional Random Fields (CRF)

CRFs are probabilistic framework for labeling and segmenting data. CRFs are a form of Undirected Graphical Models and they model conditional probability distribution Pr(Y|X), where X is vector of random variables representing the input observation to be labeled and Y is a vector of random variables representing labeling of X. The structure of X can be any general graph. CRFs are feature based models which are defined in terms of feature vector F and weight vector W. A feature is a property of X which can take any real value.

Condition Random field (CRF) is a type of discriminative model probability. It has the advantage of maximum entropy instead the label bias problem. CRFs are undirected graphical models and also called random fields which are used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes.

Random field:- Let G = (Y, E) be a graph where each vertex YV is a random variable. Suppose P (Yv-all other Y) = P(Yv-neighbours(Yv)), then Y is a random field. Let X = random variable over data sequences to be labelled Y = random variable over corresponding label sequence. Definition Let G = (V, E) be a graph such that Y = (Yv) vV, so that Y is indexed by the vertices of G. Then (X,Y) is a conditional random field ,

when conditioned on X, the random variables Yv obey the Markov Property with respect to 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.

4. Support Vector Machines (SVM)

A Support Vector Machine for short, is a system which is trained to classify input data into one of two categories.The SVM is based on discriminative approach which is use for positive and negative examples to learn the variance between the dual classes. The Support vector machines are recognized to robustly manage large feature sets and to develop models that maximize their generalizability Take two set of training data for a two-class problem: {(*x*1, *y*1),..... (*xN*, *yN*)}, where xi ε RD is a feature vector of training data of ith sample and yi ε {+1, -1} is the class to which xi related. The main goal is to find a decision function that accurately predicts class y for an input vector x. A non-linear support vector machine classifier states a decision function f (x) = sign (g (x)) assumed value of f(x) is 1for this section. Here, f(x) =+1means x is a member and f(x) =-1 means x is not a member of a definite class. And zi is called support vector and descriptive of training examples, m is the number of support vectors. The computational complexity of g(x) is proportional to the m. SVM and different constants are determined by solving a certain quadratic programming problem, K(*x, zi*) is a kernel that implicitly maps vectors into a higher dimensional space. Typical kernels use dot products (*K*(*x, zi*) = *k*(*x.z*)) .A polynomial kernel of degree d is given by *K*(*x, zi*) = (1+x)d It can use different kernels, and the design of each kernel for a particular application is an important research issue [21].

5. Decision Tree

Decision Tree is a popular and powerful tool for categorizing and calculation. Rules can use for the artificial intelligence and neural network in decision tree. These rules can easily be communicated so that human can well understand and directly use rules in a database access language like SQL so that records failing into a particular classification may be tree. Decision Tree is a classifier in the form of a tree structure where each node represent a leaf node, indicates the value of the output attributes of languages, decision, requires some text to be supported out on a single attribute value with one branch and sub-tree for each possible come the output of the text. This is an approach which acquires knowledge on classification [22].

b) Unsupervised Learning

Unsupervised learning approaches don't expect any implicit or structural information about the data they are processing. Clustering is the typical approach to unsupervised learning. For example, one can try to gather names from clustered groups based on the resemblance of context. Other methods are unattended. Basically, the techniques based on lexical resources (e.g. WorldNet) calculated on lexical patterns and statistics on a large unannotated corpus.

c) Semi supervised Learning

The term semi-supervision or weak supervision is still relatively young. The main SSL technology is called bootstrapping that includes a minor measure of control, for the creation 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 data which is not labelled , and then improved models are created iteratively using predictions of previous progressive models. [16].

II. RELATED WORK

Many researchers have been discussed about Name entity recognition of machine translation. Deepti Bhalla [23] in this name entity comprises two tasks; they can be translated or transliterated. Translation of English language to Punjabi by using statistical rule based approach. Syllabification algorithm is used for translation of entity. In this n-gram probability for syllable calculated. Kamal deep [24] rule based approach is used for addressed the problem of transliterating Punjabi to English language. The proposed transliteration scheme uses grapheme based method to the transliteration problem. Sharma et al. [5] show English-Hindi transliteration by using statistical machine translation in the different notation. This paper WX-notation gives the better result over UTF –notation by English Hindi corpus by using phrase based statistical machine translation. Dhore et al [7] have addressed the problem of MT where give named entity in Hindi using Devanagari script by using conditional random field as a statistical tool of probability. This approach shows machine transliteration of name entities for Hindi-English language using CRF as statistical probability tool. The accuracy of this system is

ISSN: 2321-9653

85.79%. Sweta Kulkarni [15] this paper shows the survey of name entity recognition of languages. In this, they describe the different approaches used for recognition, that measure the the Performance Metrics which is used to evaluate the system of name entity. They consider the existing NER systems for each of the four main South Indian languages: Telugu, Tamil, and Malayalam and evaluate them. Nusrat Jahan [17] in this paper they describe the various approaches used for NER and summery on existing work done in different Indian Languages using different approaches and also describe introduction about Hidden Markov Model (HMM) and the Gazetteer method. They also present several experimental results using Gazetteer method and HMM method is a hybrid approach. Last the paper defines the comparison between these two methods separately and then we combine these two methods so that performance of the system is increased. Ryohei Ageishi [18] combination of statistical with rule based approach is used to recognize name entity in the morphological analysis. HMM is used for tagging the English text of their meaning. Author describes the rule based approach over n consecutive word for the rule extraction. Thoudam Doren Singh [21] there are two different models are used, first using an active learning technique based on the context patterns generated from an unlabelled news corpus and the other based on the well-known Support Vector Machine have been developed. News corpus has been manually annotated with the major name entity tags, namely name of the person, name of location, and Organization's name and to apply SVM. The SVM based system makes use of the different contextual information of the words along with the variety of orthographic word-level features which are helpful in predicting the NE classes. Georgios Paliouras [22] a NERC system assigns semantic tags to phrases that correspond to named entities, such that persons, locations and organisations. This system makes use of two different language resources: a recognition grammar and a lexicon of names that are categorized by the corresponding named-entity types.we evaluated the behaviour of C4.5 on the task of learning decision trees to recognise and classify named entities in text. This approach reduces significantly the effort needed for customising a NERC system to a particular domain. Yunita Sari [25] in this, to extract important facts from unstructured text which later help to populate database entries. Name Entity Recognition is one of the main task needed to develop text mining systems in which it is used to identify and classify entities in the text into predefined categories such as the person's name, organization's name, locations, dates, times, quantity, percentages, etc.

III. PROPOSED METHODOLOGY

The objective of our work is to make a model for name entities that show name of person, name of organisation, location's name and other entities. Then translate it the target language. The system architecture shown in fig. 1 shows the various steps through which our input in a source language has passed and converts it into the target language.

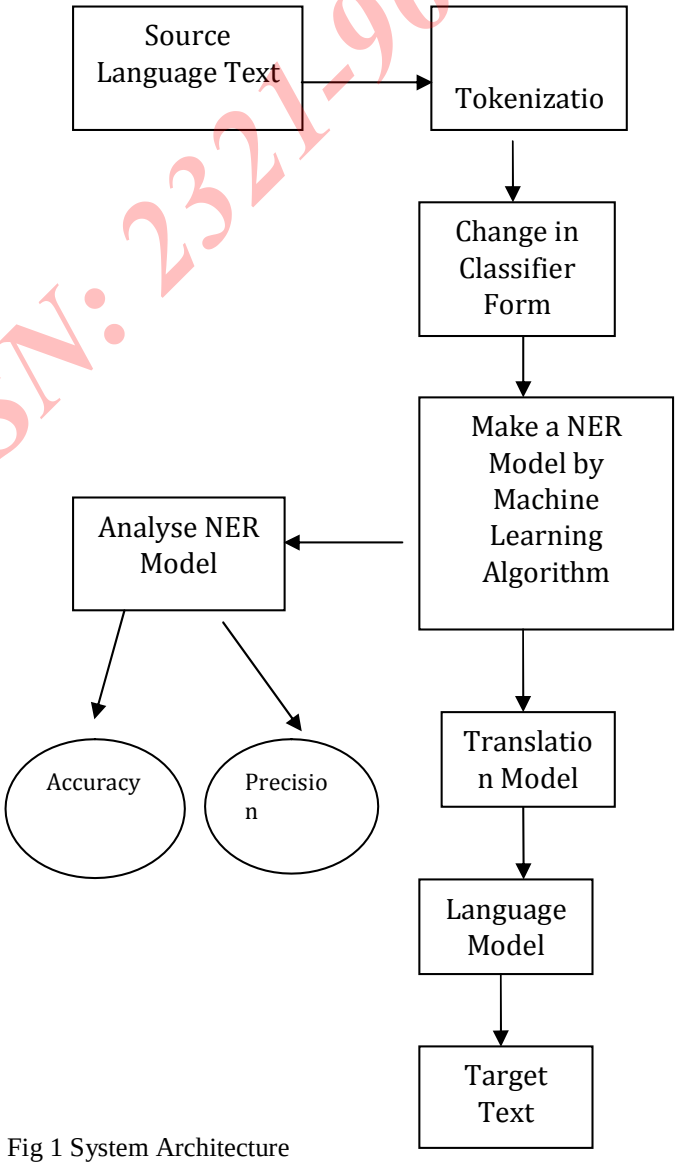


Fig 1 System Architecture

ISSN: 2321-9653

In the first step, it will take an input or source language text that is in English Language. Apply tokenization on that input. Tokenization is the process of breaking down a stream of text or sentence up into different words, number of phrases, symbols, or other significant elements are called tokens. After the tokenization, change it into the classifier form. In this, it will change it into integer form.. Next step is to find the name entities. Make a name entity recognition model by machine learning algorithm. It will show all entities like particular person's name, location and origination of it. This model also evaluates the accuracy and precision of it. Then it sends to the translation model. Translation model directly translate that into the Hindi language. We will use the Google API for the translation model. In language model, it convert to the target language. It will directly translate into the other language. Then it will produce the target text which we want.

CONCLUSION

Machine translation has been an active research sub-field of AI from years. But the challenges faced during translation need to be solved for which more detailed study of various natural languages is required. So still a lot of work is required to develop a completely automatic translation system. Improved Name entity recognition is most important part of machine translation. There are some characters exist in English which are double meaning like you is also written in u. The major inaccuracies in the transliteration are due to poor word selection. In this paper, there have described the recognition system build on statistical techniques. There are many issues left for further improvement. The system itself could be improved. In this investigation, we have discussed how to recognize name entity.Our Frame work enhance the capability of machine translation. In our frame work we classified name and entity which is input of language model. Therefore language mdel perform efficiently.

REFERENCES

[1] Ronan Collobert, Jason Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning"

[2] Harjinder Kaur, Dr. Vijay Laxmi, "A Web Based English to Punjabi MT System for News Headlines," International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

[3] Latha R. Nair, David Peter S., "Machine Translation Systems for Indian Languages," International Journal of Computer Applications (0975 – 8887) Volume 39– No.1, February 2012.

[4] Vishal Gupta, Gurpreet Singh Lehal, "Named Entity Recognition for Punjabi Language Text Summarization," International Journal of Computer Applications (0975 – 8887) Volume 33– No.3, November 2011.

[5] Shubhangi Sharma, Neha Bora and Mitali Halder, " English-Hindi Transliteration using Statistical Machine Translation in different Notation ," International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.

[6] Rejwanul Haque, Sandipan Dandapat, Ankit Kumar Srivastava, Sudip Kumar Naskar and Andy Way, "English—Hindi Transliteration Using Context Informed PB-SMT: the DCU System for NEWS 2009," CNGL, School of Computing Dublin City University, Dublin 9, Ireland.

[7] Yuxiang Jia, Danqing Zhu, Shiwen Yu, "A Noisy Channel Model for Grapheme-based Machine Transliteration," Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 88–91, Suntec, Singapore, 7 August 2009. c 2009 ACL and AFNLP.

[8] Kamal Deep, Dr.Vishal Goyal, "Hybrid Approach for Punjabi to English Transliteration System," International Journal of Computer Applications (0975 – 8887) Volume 28– No.1, August 2011.

[9] Mitali Halder, Anant Dev Tyagi, "English-Hindi Transliteration by applying finite rules to data before training using Statistical Machine Translation," 978-1-4799-2845-3/3/$31.00 ©2013 IEEE.

[10] Deepti Bhalla, Nisheeth joshi, Iti mathur, "Improving the quality of machine translation output using novel name entity translation scheme," 987-1-4673-7/13/$31.00©2013 IEEE.

[11] Darvinder kaur, Vishal Gupta, "A survey of Named Entity Recognition in English and other Indian Languages," IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010.

[12] Kamaljeet Kaur Batra and G S Lehal, "Rule Based Machine Translation of Noun Phrases from Punjabi to English," IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.

[13] Jiyou Jia, "The Generation of Textual Entailment with NLML in an Intelligent Dialogue System for Language Learning CSIEC," 978-1-4244-2780-2/08/$25.00 ©2008 IEEE.

[14] Harjinder Kaur, Dr. Vijay Laxmi, " a survey of machine translation approaches," international journal of science, engineering and technology research (ijsetr) volume 2, issue 3, march 2013.

[15] Malarkodi, C S., Pattabhi, RK Rao and Sobha, Lalitha Devi, "Tamil NER – Coping with Real Time Challenges", Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 23–38, COLING 2012, Mumbai, December 2012.

[16] Yunita Sari, M. Fadzil Hassan, Norshuhani Zamin, "A Hybrid Approach to Semi-Supervised Named Entity Recognition in Health, Safety and Environment Reports,"International Conference on Future Computer and Communication © 2009 IEEE.

[17] 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.

[18] Ryohei Ageishi, Takao Miura, "Name entity recognition based on a hidden markov model in part of speech tagging," 978-1-4244-2624-9/08/$25.00 ©2008 IEEE

[19]https://www.google.co.in/#q=LANGUAGE+INDEPENDENT+NAMED+ENTITY+RECOGNITION.

[20] Brahmaleen K. Sidhu,Arjan Singhand Vishal Goyal, " Identification of Proverbs in Hindi Text Corpus and their Translation into Punjabi,"JOURNAL OF COMPUTER SCIENCE AND ENGINEERING, VOLUME 2, ISSUE 1, JULY 2010.

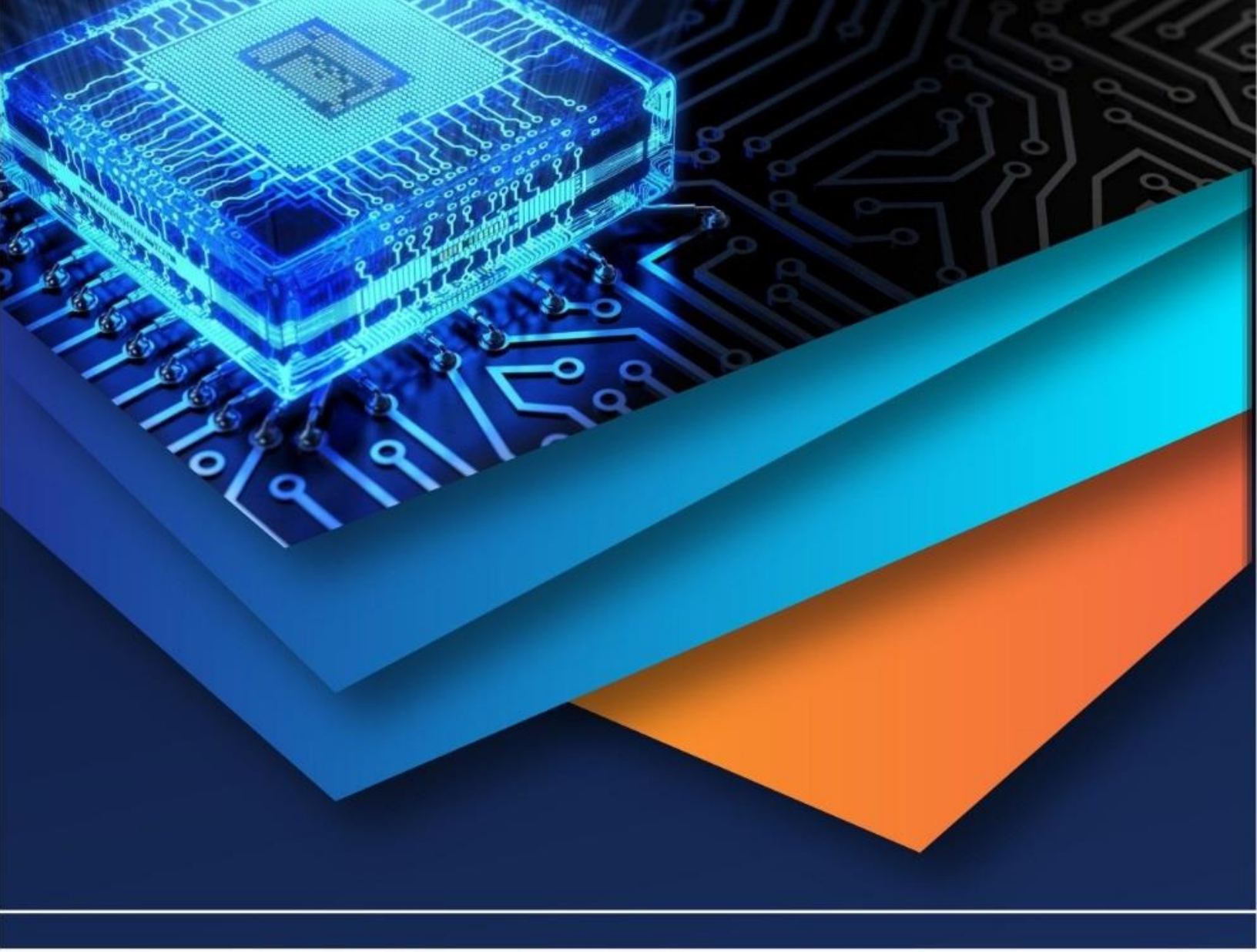
[21] Thoudam Doren Singh, Kishorjit Nongmeikapam, Asif Ekbal and Sivaji Bandyopadhyay, "Named Entity Recognition for Manipuri Using Support Vector Machine," 23rd Pacific Asia Conference on Language, Information and Computation, pages 811–818.

[22] Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis and Constantine D. Spyropoulos, "Learning Decision Trees for Named-Entity Recognition and Classification," Institute of Informatics and Telecommunications, NCSR "Demokritos", 15310.

[23] http://en.wikipedia.org/wiki/Natural_language_processing

[24] Kamal Deep and Vishal Goyal, "DEVELOPMENT OF A PUNJABI TO ENGLISH TRANSLITERATION SYSTEM," International Journal of Computer Science and Communication Vol. 2, No. 2, July-December 2011, pp. 521-526.

[25] Yunita Sari, M. Fadzil Hassan, Norshuhani Zamin, "A Hybrid Approach to Semi-Supervised Named Entity Recognition in Health, Safety and Environment Reports,"International Conference on Future Computer and Communication © 2009 IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY