Automated Named Entity Recognition from Tamil Documents

R. Srinivasan
Department of Computer Science and Engineering
SRM Institute of Science and Technology
Kattankulathur, India
srinirvs89@gmail.com

C.N. Subalalitha
Department of Computer Science and Engineering
SRM Institute of Science and Technology
Kattankulathur, India
subalalitha@gmail.com

***Abstract*—Named Entity Recognition (NER) is a subsequence of words in a document that seeks to detect and classify entities into pre-defined categories such as name of the person, organization and location respectively. The impact on NER is high because a lot of Information Extraction (IE) relations are associated using Named Entities (NEs). This paper presents a pioneering method for extraction of NEs for Tamil using Supervised Learning. This hybrid framework makes use of features that are extracted based on the speciality of the Tamil language NEs. The evaluation has been done by using 1028 number of documents which comprises the standard FIRE corpus and an F-measure of 83.54% has been achieved. A performance comparison with one of the state-of-the-art Tamil NE system has been done and the proposed methodology has achieved better accuracy**.

***Keywords*—Naïve Bayes; Natural Language Processing; Named Entity Recognition; Entity Extraction; Feature Identification**

I. INTRODUCTION

Over the past decade, contents on various media types such as, e-books, websites, blogs, documentaries, emails, chats have increased significantly. According to researcher [12], massive amount of data has been grown every day. Search engines need to process this massive information and provide precise results from the user query. Natural Language Processing (NLP) is an emerging domain of Computer Science related to the area of Human-Computer Interaction. It is the process of extracting meaningful information by processing the human languages. The major applications of NLP are Automatic Summarization, Sentiment Analysis, Navigation Systems, Information Extraction (IE), Question Answering, Word Sense Disambiguation (WSD) and Information Retrieval (IR). The goal of IE is extracting relevant information from unstructured or semi structured documents. The subdomains of IE are Named Entity Recognition (NER), Relationship extraction, Terminology Extraction and Audio Extraction. This paper presents a novel approach for building NER system for Tamil Language. NER in English gives an outstanding result due to the adequate of data and standard dataset also available such as MUC-7[8], CoNLL-2002[14] to evaluate the accuracy.

A. *Tamil Language Challenges*

The task of extracting entities from Tamil language is very challenging due to the following reasons:

Partially free word order: Tamil language has a partially free word order [13]. The meaning of the sentence will remain the same even if the positions of the subject, verb and the object are interchanged. In such cases the Person Name NE and Place NE identification becomes difficult. Sentences without subject, verb or object: In Tamil language, it is possible to construct a sentence without having subject, verb or object. This reduces the number of features that could be extracted to identify the NE.

No capitalization: Capitalization in English first feature that helps in identifying the possible NEs. In Tamil there is no capitalization which induces its own complexity in identifying the NEs.

Polysemic nature: *Polysemy* is the relationship of single word with two or more distinct meanings [5]. Some words are used to refer person name and name of the location. For example, Word Chidambaram refers to name of the person and sometime referred as Location.

Ambiguity: Word refers to different NE's types based on the meanings. For example, Word Periyar (Noun, person name) can also be used as Periyar Street (Location).

Lack of Resources: No standard corpora, Online gazetteer's list, Short words are not available for Tamil language to identify the entities.

Regular Expression, POS Tagger and the Contexts are used as the features. The proposed approach attempts to overcome the above said challenges by using these features in a fixed hierarchical manner. This alleviates the complexities involved in identifying the Tamil NEs.

In sum, the contributions of the proposed approach are twofold.

1) Design of Unique algorithm to extract features using REGEX (Regular Expression) Feature Extraction, Morphological Feature Extraction, Context Feature Extraction in the specified order.
2) Incorporation of word level Naïve Bayes classification to classify the entities.

To the best of our knowledge, this research is the extracting and classifying the NEs in Tamil languages that uses word level Naïve Bayes classification. The rest of the paper is organized as follows:

Section 2 contains Background details about NER, Section 3 discusses Literature Survey, Section 4 demonstrates the proposed approach to identifying and extracting entities from Forum for Information Retrieval (FIRE), Section 5 contains the Results and Discussion obtained and Section 6 describes the Conclusions.

II. BACKGROUND

NER is the process of extracting entities such as person name, organisation name, date, time, money and location. Extracting entities from documents is one of the

978-1-7281-0419-5/19/$31.00 ©2019 IEEE

challenging tasks and it is essential for NLP applications. NER has been done for various languages such as English, Turkish, Spanish, and also for Indian languages. NEs are extracted by using Rule-based approach, Learning-based approach and Hybrid-based approach [16]. Early works on NER mainly focused on Rule based approach. Rule based approach is a methodology consists of hand coded rules to identify the entities with the help of generic resources such as gazetteers or dictionaries. It follows syntactic rules to extract and classify the named entities. It takes a lot of human efforts to generate the hand coded rules. The major drawbacks of rule-based approach are domain specific and time consuming [15,16]. Learning based approaches can be broadly classified as unsupervised, supervised and semi-supervised approaches [4]. Supervised Learning approaches are based on the knowledge of providing labelled training data, annotated by domain experts manually [16]. The labelled data is used to train the model which is further used to identify and classify the named entities to the test data. Semi-supervised approaches use a small amount of labelled training data called seed data. Seed data is mostly used to identify the context features or pattern to each named entity category. The identified pattern is used to extract the entities in the test data. Unsupervised learning approaches are use unlabelled data to classify the named entities. These approaches always rely on statistical techniques, Wordnet, Similarity context between words on a large unannotated corpus [18]. Hybrid Learning approaches that combines Rule-based techniques and Learning based approaches.

III. Literature Survey

Many techniques have been used to obtain the outstanding result in NER due to abundance of data. Researchers are moved towards the supervised approaches such as Support Vector Machine (SVM) [1], Hidden Markov Model (HMM) [3,7,8,10], Conditional Random Field (CRF) [3,6] due to exponential growth on data.

Zhenfei Juet has used a Support Vector Machine (SVM) to recognize the named entity in biomedical abstracts [1]. The author has used a Genia corpus that consists of around 4,60,000 words approximately and obtains a precision of 84.24%. Xia Han et. al. showed that semantic model feature selection combines SVM and K-Nearest Neighbour (KNN) algorithm named as EK-SVM-KNN (Extending K value SVM-KNN) algorithm. It enhanced the quality of NER and made great contribution to the unstructured information process and obtains the F-Measure of 85.6% [9].

Branimir T. Todorovic et. al. has shown the Context HMM is approximately twice faster in both training and classification, and has larger precision (94.88%), recall (93.62%) and F-measure (94.24%) than our implementation of Bikel's HMM [7] named entity recognizer [8]. Branimir used a standard dataset as MUC-7 text corpus and produced the best result compared to Bikel's HMM.

Kishorjit Nongmeikapam et. al. built a Conditional Random Field (CRF) model to extract the NER from Manipuri language. It is difficult to extract NER from Manipuri language which is relatively free word order. The author had used a standard corpus to evaluate the CRF model and achieve F-score of 83.33%.

Dilek Kucuk and Adan Yazici proposed a preliminary version of hybrid named entity recognizer or Turkish language [17]. The author has extracted the feature using Rule-based approach and identify the named entities using Rote learning component. Evaluation dataset include the genres of news text, financial news text, child stories historical text and also extract entities from video dataset. This hybrid approach produced a significant result for Turkish texts and news videos. N. Jeyashenbagavalli et. al. recommended a hybrid approach that merge both Rule-based and the machine learning approach to achieve high F-measure score [10]. Initially, classifier is trained using a small number of labelled sentences with all the features. The process continues until all the unlabelled sentences to be included into the training set. The proposed method can learn features using Rule-based approach and applied to HMM and Expectation-Maximization (EM) model to extract entities from Tamil documents. To the best of our knowledge, Hybrid approach combine Rule based approach with Naïve Bayes (NM) have never been attempted to extract entities in Tamil news documents. There was some research conducted on extracting entities from Tamil documents which includes Hidden Markov Model. Our proposed approach is completely different and tested using Fire corpus and produce an outstanding result.

IV. Proposed Methodology

The proposed NER approach mainly includes Feature Extraction and Entity Identification. Fig. 1 shows the architecture of the proposed NER model which describes the sequence of steps involved in the proposed system.

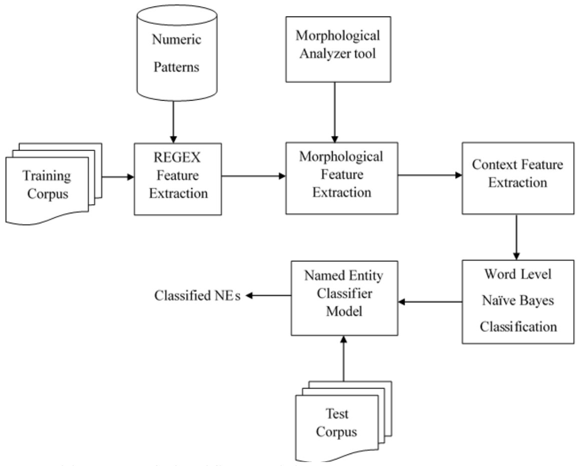
Numeric Patterns
Morphological Analyzer tool
Training Corpus
REGEX Feature Extraction
Morphological Feature Extraction
Context Feature Extraction
Word Level Naïve Bayes Classification
Named Entity Classifier Model
Classified NEs
Test Corpus

Fig. 1 Architecture of classifier model

Training data of 4111 documents containing 73891 words from news articles in Tamil are taken from the Forum for Information Retrieval Evaluation (FIRE). In 2008, FIRE initiated the new challenges in multilingual information access and a series of FIRE are conducted to evaluate the researches in IE. As the FIRE progressed, the research work on IE has focused on various domains, such as NER, Sentimental Analysis, Question Answering and Information Retrieval [11]. The first step is to extract the features by using Regular Expression (REGEX) feature extraction. Later, Morphological Feature Extraction and Context Feature Extraction are used to identify and extract the features. Once the features are extracted, Naïve Bayes algorithm are used to identify the NEs in the FIRE corpus.

The test data which consist of 1028 documents containing 21,316 words approximately.

A. Regex Feature Extraction

Extracting features using Regex were introduced by M. Collins [19]. Regular expression patterns are the one that successfully extract these numerical data effectively. This is the most effective way to extract features of the date and time entities. Since the corpus contains information about news in Tamil which mostly describes date and time entities that are expressed in number. Regex is used to extract features for Date and Time NEs from the corpus which has more than 1,50,000 words. Once the numeric entities are identified, the volume of data gets reduced. It helps to reduce the complexity of the next step. Once the REGEX features are extracted, it should in a separate class. Table 1 describes some of the patterns for numerical entities.

TABLE I FEATURES FOR DATE AND TIME ENTITY

Patterns	Date and Time Entities Identified
[[0-9] [0-9]-[[A-Z] *] - [2] [0-9] [0-9] [0-9]]	12-Nov-2018
[0-9] [0-9] - [0-1] [0-9] – [0-9] [0-9]	12-10-18
[0-9] [0-2] [[:] | [.]] [0-5] [0-9]	11.52

B. Morphological Feature Extraction

Morphological analyser is used to identify the Parts-Of-Speech (POS) and the case markers. Morphological analyser is very essential for rich inflectional languages like, Tamil, Malayalam, Telugu etc. Case markers are given by the Morphological Analyser is used as the morphological features to classify the NEs appropriately. For instance, In the Example 1, it clearly describes the morphological features of the word மதுரையில். Here இல் is the Locative Case aids in classifying the Location NE. Similarly, a list of case markers is shown in the table 2 are used to classify the NEs appropriately. A word need not necessarily have a case marker in it. In most of the time NEs, case markers will not be present. Such cases, the context feature extraction technique is used to classify the NEs which is explained in the next section.

TABLE II CASEMARKERS

Word	Morphological Analyzer	Morphological Features
மதுரையில்	மதுரை< Entity > ய்< Sandhi > இல்< Locative Case>	இல் (Location NE)
சுந்தரத்துக்கு	சுந்தரம்< Entity > அத்து< Oblique > உக்கு< Dative Case >	உக்கு < Dative Case >

C. Context Feature Extraction

Context Features are extracted using the words Co-occurring along with the Entities identified by the Morphological Analyser. The Context Features are extracted using a context window of words surrounding the Entities. It forms a pattern which could fall in the following categories. The pattern consists of three-words which can fall into three types of context windows namely, ($w_{i-1,}$ w_i) or (w_i, w_{i+1}) or ($w_{i-1,}$ w_i, w_{i+1}). The NEs are identified using the features extracted from the context window. The bootstrapping approach makes use of the initially identified features and also the NEs identified using these features as the seed set. This initial seed set is augmented in an iterative fashion. The Yarowsky algorithm is the first bootstrapping algorithm to become widely known in Word Sense Disambiguation [2,4].

In above mentioned example1 ஜெயகுமார் (jeyakumar) is an entity identified with the help of seed features. By applying Bootstrapping technique, we can identify the new feature தாசில்தார் (Tahsildar) with the existing entity ஜெயகுமார் (jeyakumar) as shown in in example 2. The process is repeated until no different NES are learnt.

Example 1

திரைப்படஇயக்குனர்ஜெயகுமார்அவர்கள்,1959-ஆம்வருடம்அக்டோபர்மாதம் 2-ஆம்தேதிபிறந்தார்.									
	W_{i-1}	W_i	W_{i+1}						
Meaning: Film director Jayakumar was born on the 2nd of October 1959									

Example 2

சங்கராபுரம்அருகேமணல்கடத்தியடிராக்டரை தாசில்தார்ஜெயகுமார்பறிமுதல்செய்தார்								
				W_{i-1}	W_i	W_{i+1}		
Meaning: Tasildar Jayakumar confiscated the tractor's sand tuner near Sankarapuram								

D. Feature Classification algorithm

After identifying all the features, it is possible to group into three different feature sets: features found using Regular Expression, features found from POS, and features based on Context Window. Table 3 describes the various features falling under the three main categories namely, Regex Feature Extraction, Morphology Feature Extraction and Context Feature Extraction.

Feature Extraction algorithm is used to classify the features to identify the different types of entities namely, Person, Organization, Location, Date and Time.

```
    Input: Collection of Sentences (Cs)
    Output: Person name (PF) or Organization name (OF) or Location (LF) or Date(D) or Time(T)
    Notations: Regex represents Regular Expression Technique
    Moa represents Morphological Analyzer
    Cw represents Context Window Technique
    N represents Noun
    F1, F2, F3, F4, F5, F6, F7, F8, F9, F10 areFeatures
Step 1: Read S1, S2, S3 . . . Sn where S represents a sentence
Step 2: For W1, W2, W3 . . . Wn
    If W1 follows F1
    classify the Features in either D or T
    Else if W1 as Noun and follows F2 || F4 || F6 ||F8 (ByApplying Moa)
    classify the Features in PF
   Else if W1 follows F3||F5||F7||F9||F10(By Applying Cw)
    classify the Features in either PF or OF or LF
    Else W1 is not an entity
```

Step 3: Continue with Naïve Bayes (For determining Class Probability, Conditional Probability, Predicted class)
Step 4: Exit

TABLE III. FEATURE SET

Type	Features	Feature Name
Regular Expression	F1	Date, Time
Morphological Analyzer	F2	Parts of speech of Individual words
	F3	Preposition
Context Window	F4	Prefixed by Salutation Words (Person name)
	F5	Prefixed by Substring (Person, Organization, Location)
	F6	Suffixed by Salutation Words (Person name)
	F7	Suffixed by Substring (Person, Organization, Location)
	F8	Gazetteers (Person)
	F9	Previous word of size 3
	F10	Next word of size 3

E. NE Classification Using Naive Bayes

Naïve Bayes is a supervised machine learning algorithm based on baye's theorem. Naive Bayes is a probabilistic classifier and assumes that each feature is independent while classifying the NEs. Let the NE classes be represented Ci. The class of an NE(Ci) is identified when a set of features are given, using the posterior probability(ref). This is done by calculating the conditional probability of a particular class Ci signalled by the feature vectors w = (PF, OF, LF, DF, TF).PF represents the person name feature set, where PF= (w1, w2, …., wn) represents independent features belong to person name. Likewise, each entity has a different feature set which mentioned in the above algorithm.

$$P\left(C_i \mid PF\right) = P(C_i|w_1) * \ldots\ldots\ldots\ldots * P(C_i|w_n) \qquad (1)$$

In NB, calculate the individual probability of a feature are shown in Equation 2.

$$P(C_i|w) = P(C_i)*P(w|C_i) \qquad (2)$$

$P(C_i)$ calculates the prior probability are shown in the Equation 3.

$$P(C_i) = \frac{\text{Total no of features belongs to } C_i \text{ in the training set}}{\text{Total no of features in the training set}} \qquad (3)$$

$P(w|C_i)$ computes the conditional probability for the class C_i.

$$P(w|C_i) = \frac{\text{Total count of w belongs to } C_i + 1}{\text{Total no of features belongs to } C_i + |u|} \qquad (4)$$

Where |U| represents the total number of unique features in the training document. The classifier begins with each of the feature set and the results were documented. Each feature set works independently to classify the entity class. The class of a word is finally fixed by identifying the feature that has maximum probability for that class as shown in the Equation 5.

$$C^* = \arg\max_c P(C/w) \qquad (5)$$

V. RESULTS AND DISCUSSION

The evaluation has been done by standard metrics namely Precision, Recall and F-measure which are defined as follows.

Named Entities	Precision	Recall	F-measure
Person	90.04	82.87	86.31
Location	90.29	84.32	87.20
Organization	85.89	81.4	83.58
Date and time	82.03	72.66	77.06
Total	87.06	80.31	83.54

$$\text{Recall} = \frac{\text{number of correctly identified person NE 's}}{\text{total number of person NE's in a document}} \qquad (6)$$

$$\text{Precision} = \frac{\text{number of correctly identified person NE 's}}{\text{total number person NE's retrieved by the system}} \qquad (7)$$

F-measure is the harmonic mean of recall and precision

$$\text{F-measure} = \frac{2*\text{precision} \ *\text{recall}}{\textit{precision} \ +\textit{recall}} \qquad (8)$$

We have tested the performance of the system by using the annotated corpus for NER using GMB [21] (Groningen Meaning Bank) in Kaggle. The GMB corpus consists of 13,54,149 words and 2,08,081 tagged entities along with the POS tags. Entities such as Geographical name, person name, organization, time, artifact, event, Natural Phenomenon and geopolitical name are focused. 80 % of the words are used as training data and 20% of the words are used as test data. The features used by the proposed framework are extracted from the training set. The test data consists of 2,09,303 words and 31992 tagged entities. The test data consist of 6986-person name, 7096 organization, 8850 location 5131 date and time entities. The proposed methodology focuses on five entities namely, person, organization, location, date and time. The performance is shown in Table 4, and the average precision, recall and f-measure achieved our model are 89.52%, recall 84.51% and f-measure 86.94%.

TABLE IV PERFORMANCE OF OUR MODEL IN GMB CORPUS

Named Entities	Precision	Recall	F-measure
Person	91.87	85.41	88.52
Location	86.35	82.33	84.29
Organization	88.71	84.65	86.63
Date and Time	91.14	85.64	88.3
Total	89.52	84.51	86.94

The corpus contains 5,139 documents. Out of 5139 documents ,4111 used as training data and remaining 1028 is used as test data. The data set used by the proposed system comprises of news articles and various other online sources. We have manually annotated this data set, with the above-mentioned GMB style named entity tags leading to a total of 11,539 entities and 80,824 non-entities. The annotated entities consist of 3973 person NEs, 3376 location NEs, 2814 organization NEs and 1376 date and time NEs. The test data begins with the 20 % of the fire corpus. Table 5. shows the performance of our model in FIRE corpus and obtains 87.06% precision, 80.31% recall, and 83.54% F-measure.

TABLE V. PERFORMANCE OF OUR MODEL IN FIRE CORPUS

The existing NER model that used GMB corpus has achieved 84% F-measure [22]. It can be observed that the proposed model achieves a F-measure of 86.94%. This is mainly due to usage of features in a hierarchical manner. Also, the Naïve Bayes algorithm has contributed much in increasing the coverage of feature extraction which in turn has increased the F-measure. The reasons for missed out NEs are the inability of the morphological analyzer in handling nouns and those NEs were surrounded by the context features which were not captured by the proposed model. This should be tackled in future by analyzing the semantics of the sentences.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new technique for named entity recognition for Tamil language. The proposed methodology differs from the state of art techniques by grouping the features and tapping the best out of the features by hierarchical ordering of features. This has increased the efficiency of the proposed technique. The bootstrapping technique has aided in increasing the coverage of features. Since the features are perfectly grouped into three classes, the prediction of the Entity class using the Naïve Bayes algorithm has increased. Since language like Tamil lacks bench mark data set, the proposed technique would be a pointer in creating such data sets in future. This will eventually help in building many other applications for Tamil which demands the recognition of NE.

REFERENCES

[1] Zhenfei Ju, Jian Wang, and Fei Zhu, “Named Entity Recognition from Biomedical Text Using SVM,” in Proc. 5th International Conference on Bioinformatics and Biomedical Engineering., 2011.

[2] Michael Collins, and Yoram Singer, “Unsupervised Models for Named Entity Classification,” in Proc. Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora., 1999.

[3] Downey. D, M. Broadhead, and Etzioni, “Locating Complex Named Entities in Web Text,” in Proc 20th International Joint Conference on Artificial Intelligence., pp. 6-12, 2007.

[4] Thenmalar, S. J. Balaji, and T.V. Geetha, “Semi-supervised Bootstrapping approach for Named Entity Recognition,”International Journal on Natural Language Computing., Vol. 4, no. 5, pp. 01-14, October 2015.

[5] AlBader, and Yousuf. B, “Semantic Innovation and Change in Kuwaiti Arabic: A Study of the Polysemy of Verbs,” Ph.D Thesis, University of Sheffield, 2015.

[6] Kishorjit Nongmeikapam,TontangShangkhunem, Ngariyanbam Mayekleima Chanu,Laishram Newton Singh,BishworjitSalam,Sivaji Bandyopadhyay,“CRF based Named Entity Recognition in Manipuri: A highly agglutinative Indian Language,” National Conference on Emerging Trends and Applications in Computer Science, 2011.

[7] Bikel, D. R.Schwartz, and R.Weischedel,“An Algorithm that Learns What's in a Name,” Machine Learning., Vol. 34,no. 1-3, pp.211-231, February 1999.

[8] Branimir T. Todorovic, Svetozar R. Rancic, Ivica M. Markovic, Edin H. Mulalic, Velimir M. Ilic, “Named Entity Recognition and Classification using Context Hidden Markov Model”, in Proc 9th Symposium on Neural Network Applications in Electrical Engineering.,October 2008.

[9] Xia Han, and Rao Ruonan, “The Method of Medical Named Entity Recognition Based on Semantic Model and Improved SVM-KNN Algorithm”, in Proc 7th International Conference on Semantics, Knowledge and Grids., pp. 21-27, December 2011.

[10] K. G. Srinivasagan,S. Suganthi, and N. Jeyashenbagavalli, “An Automated System for Tamil Named Entity Recognition Using Hybrid Approach,” in Proc International Conference on Intelligent Computing Applications., November 2014.

[11] http://www.isical.ac.in/~fire/2013/index.html

[12] Ashwin Satyanarayana, “Intelligent Sampling for Big Data Using Bootstrap Sampling and Chebyshev Inequality,” in Proc 27th Canadian Conference on Electrical and Computer Engineering., pp. 1-6, September 2014.

[13] Selvam. M, A.M. Natarajan, and R. Thangarajan, “Structural Parsing of Natural Language Text in Tamil Using Phrase Structure Hybrid Language Model,”International Scholarly and Scientific Research & Innovation., Vol. 2 no. 3, pp. 737-743, 2008.

[14] Erik F. Tjong Kim Sang, “Introduction to the CoNLL-2002 shared task: language-independent named entity recognition,” in Proc 6th conference on Natural language learning., Vol20, 2002.

[15] Sudha Morwal, Nusrat Jahan, Deepti Chopra, “Named Entity Recognition using Hidden Markov Model (HMM),” International Journal on Natural Language Computing., Vol. 1 no. 4, pp. 15-23, December 2012.

[16] Archana Goyal, Vishal Gupta, and Manish Kumar, “Recent Named Entity Recognition and Classification techniques: A systematic review,”Computer Science Review., Vol. 29,pp. 21–43, August 2018.

[17] Dilek Kucuk, and Adnan Yazici, “A hybrid named entity recognizer for Turkish,”Expert Systems with Applications., Vol. 39, no. 3, pp. 2733–2742, February 2012.

[18] David Nadeau, and Satoshi Sekine, “A survey of named entity recognition and classification,”Lingvisticae Investigationes., Vol 30, no 1, pp. 3 – 26, August 2007.

[19] Collins Michael, “Ranking Algorithms Named-Entity Extraction: Boosting and the Voted Perceptron”, In Proc. Association for Computational Linguistics.,pp. 489-496, July 2002.

[20] Rayner Alfred, Leow Chin Leong, Chin Kim On, and Patricia Anthony, “Malay named entity recognition based onrule-based approach,” in International Journal of Machine Learning and Computing., Vol. 4, no.3,June 2014.

[21] https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus/home