Named Entity Recognition for English Tweets using Random Kitchen Sink Algorithm

Abinaya N[1] Saranya S S[2]
[1,2]Assistant Professor
[1,2]Department of Information Technology
[1,2]Nandha Engineering College

Abstract— The information obtained electronically is vast and difficult for users to access the exact information within permissible time. The various Natural Language Processing (NLP) task has been carried out in the field of Artificial Intelligence (AI) to extract structured information from this large amount of unstructured data. Named Entity Recognition (NER) is one such Information Extraction (IE) system which identifies the elements such as name of person, location, organization, quantities, time expressions etc. and classify into set of pre-defined classes. In this thesis, language independent system is developed using Machine Learning algorithms such as Conditional Random Field (CRF), Support Vector Machine (SVM) and Random Kitchen Sink (RKS). A unique approach has been carried out in implementing Random Kitchen Sink algorithm for Named Entity Recognition. This paper deals with RKS for English tweets. The accuracy for this system is obtained as 82.60%.
Key words: NER, SVM, CRF, RKS, IE, NLP,

I. INTRODUCTION

Named Entity Recognition (NER) includes distinguishing names inside the content as named entities and grouping each such distinguished occurrence into predefined classes [1] [2]. Every word in the text is categorized as named entity or not. Most of the named entities are of nouns and also noun phrase. The named entities can be of name of person, location, organization, time, date, quantity measures etc. Let us consider the following sentence:

Modi [PERSON] visits naxals-affected areas of Chhattisgarh [LOCATION] on 9th May [DATE].

The named entities in this sentence are "Modi" is the name of the individual, "Chhattisgarh" is the name of the place and "9th May" speaks to date. This kind of labeling is done for entities which are not a phrase.

Considering named elements as phrase, the labeling is done as "BIO-tagging". The beginning word of the noun phrase is tagged as "B-tag" and the word inside the phrase is tagged as "I-tag". Consider the sentence:

Narendra [B-PERSON] Modi [I-PERSON] scheduled to visit Naya [B-LOCATION] Raipur [I-LOCATION] today.

In this sentence, "Narendra Modi" is a noun phrase which represents the name of the person and "Naya Raipur" represents the name of the city. "Narendra" is the beginning of the phrase so tagged as "B-PERSON" and "Modi" is continuation of the phrase and tagged as "I-PERSON". Considering the name of the place in the above sentence, "Naya" begins the phrase with the label "B-LOCATION" and "Raipur" continuing the phrase tagged as "I-LOCATION".

Machine Learning in NLP converts the NER problems into Classification problem. These algorithms analyze the patterns and relationship in given example text and develop a set of rules to classify new text [3]. All the machine learning types such as supervised, unsupervised and semi-supervised learning are used for developing NER systems [4]. The NER system developed using machine learning algorithm require large amount of training data in order to identify the similar patterns in the text [5]. This large amount of labelled data is referred as annotated corpora. Example data of NER system is modelled as tokens (i.e word per line) with the label for each token. The label may be type of named entity such as person name, location, date or time [6].

II. RELATED WORKS

Bikel et al., (1998) [7] has developed NER system for English and Spanish language in order to identify and classify names, time and date entities and numerical entities. They demonstrated an experiment stating that training data size of 100,000 words is sufficient to get 90% accuracy. Ralph Grishman [8] developed a rule based NER system in 1995 by considering the dictionaries.

This dictionary composed of name of companies, most commonly used first names, name of all countries, name of major cities etc. Borthwick [5] developed a Machine Learning based system using Maximum Entropy algorithm in 1999. He used 8 dictionaries as gazetteer list. Named Entity Recognition for English has been done using various approaches such as supervised, semi-supervised and unsupervised learning [9]. Collins et al proposed unsupervised model for NER using unlabelled training data. Kim et al [10] also proposed NER system using unsupervised approach which uses small scale dictionary and unlabeled corpus.

A survey has been done in NER systems for Indian languages by Darvinder et al [11] in 2010. The techniques employed for developing NER system are moved from handcrafted rules to machine learning approach [18]. They have concluded that hand-crafted rules provide more accuracy than machine learning approaches. They have categorized the main features that help in identifying the named entities. Gupta and Gurpreet (2011) [21] implemented NER system for Punjabi and applied in text summarization. They used condition based approach with various rules like prefix, suffix, proper name, middle name and last name. The precision, recall and F score are found to be 89.32%, 83.4% and 86.25% respectively.

In 2013, Jisha et al [12] proposed a NER system for Malayalam Language which is a hybrid approach i.e. rule based machine learning. This method is based on probability measures by considering unigram, bigram, trigram and n-grams and provides accuracy of 73.42%. Vijayakrishna and Sobha [1] proposed a domain focused NER system for Tamil using CRF. They used nested tagging for various named entities of tourism domain. They designed a hierarchical

tagset which contains 106 tags. They provided different accuracy for each level as 81.79%, 83.77% and 75.77% for level-1, level-2 and level-3 respectively.

Malarkodi et al (2012) [5] described the challenges faced in developing NER system for Tamil language. They also discussed the way for overcoming these challenges. The automatic NER system was developed using CRF with web data corpus. The total F score of 70.68% was obtained after applying root word rules. An automated system was developed using hybrid approach by Jeyashenbagavalli et al (2014) in [20]. This uses both the rule based approach followed by Hidden Markov Model with E-M algorithm [13]. This method solves the problem of occurrence of single entity in different positions present in same document. The result was observed to be 89.7% as F score [21] [14] [15]. The hybrid three stage NER system was developed in 2008 by Pandian et al [19] and provided accuracy of 72.72% for various entity types.

Rahimi and Recht [16] proposed a powerful tool for classification task using machine learning approach in 2007. Random Kitchen Sink is mainly used for classification and Regression task. It is also used for other kernel methods to implement semi-supervised and unsupervised algorithm [17]. Again in 2009, they chose arbitrary nonlinearities and accomplished the accuracy as greedy algorithm.

III. PROPOSED METHOD

Random kitchen sink algorithm is a machine learning algorithm for classification of nonlinearly separated data set. The conventional nonlinear kernel methods use large nonlinear data set for training the system. It requires large proportion of data points to be stored for classifying new data point. So, space and time requirement is more for classification. Random Kitchen Sink is an alternative for these conventional nonlinear kernel methods. RKS uses only the feature size and does not consider the number of data points for classification.

Random Kitchen Sink algorithm uses GURLS library which is targeted to machine learning, supervised learning to be more specific. It is well suited for large scale machine learning problems especially with multi label problems. Andrea et al implemented GURLS library in Matlab which can handle large matrices.

Random Kitchen Sink uses Radial basis function (RBK) kernel which is a real Gaussian function. The Fourier Transform of any real Gaussian function which is symmetric is also a real Gaussian function.

(3.1) $$F(x_1, x_2) = \left\langle \Phi(x_1), \Phi(y_1) \right\rangle = e^{\frac{-1}{\sigma}|x_1 - y_1|_2^2}$$

$$e^{\frac{-1}{\sigma}|x_1 - y_1|_2^2} = e^{\frac{-1}{\sigma}(x_1 - y_1)^T (x_1 - y_1)}$$

(3.2) $$e^{\frac{-1}{\sigma}(x_1 - y_1)^T (x_1 - y_1)} = e^{\frac{-1}{2}(x_1 - y_1)^T \Sigma^{-1} (x_1 - y_1)}$$

(3.3)

Where

$$\Sigma = \begin{bmatrix} 2\sigma & 0 & \dots & 0 \\ 0 & 2\sigma & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 2\sigma \end{bmatrix}$$

The kernel can be expressed in the form as a Gaussian probability density function. Gaussian density function is the product of n Gaussian functions because the covariance matrix diagonal. The kernel is interpreted as probability density function and associated variables are independent.

The kernel can be expressed in the form as a Gaussian probability density function. Gaussian density function is the product of n Gaussian functions because the covariance matrix diagonal. The kernel is interpreted as probability density function and associated variables are independent.

Let $x_1 - x_2 = z$ then the kernel function is $f(z) = e^{\frac{-1}{2} z^T \Sigma^{-1} z}$. Let $F(\Omega)$ represent Fourier Transform of $f(z)$ which is given as

(3.4) $$F(\Omega) = \frac{1}{2\Pi} \int_{-\infty}^{\infty} f(z) e^{-jz^T \Omega} dz$$

Since $f(z)$ is Gaussian $F(\Omega)$ is again Gaussian. It is a multivariate Gaussian function with variance. $F(\Omega)$ is a Gaussian (multivariate) density function.

(3.5) $$F^{-1}(\Omega) = \left\langle \Phi(x_1), \Phi(y_1) \right\rangle = \int_{-\infty}^{\infty} F(\Omega) e^{jz^T \Omega} dz$$

This can be interpreted as expected value of the quantity $e^{jz^T \Omega}$.

(3.6) $$E(e^{jz^T \Omega}) = \int_{-\infty}^{\infty} F(\Omega) e^{jz^T \Omega} dz$$

The expected value of any function of random variable is obtained by numerous independent samples from the associated probability density function and an average. Since z is n-tuple, Ω is n-tuple and one particular vector Ω_i can be easily generated.

(3.6) $$E(e^{jz^T \Omega}) = \frac{1}{k} \sum_{i=1}^{k} e^{jz^T \Omega_i} = \frac{1}{k} \sum_{i=1}^{k} e^{j(x-y)^T \Omega_i}$$

(3.7) $$\frac{1}{k} \sum_{i=1}^{k} e^{j(x-y)^T \Omega_i} = \frac{1}{k} \sum_{i=1}^{k} e^{jx^T \Omega_i} \overline{e^{jy^T \Omega_i}}$$

IV. EXPERIMENTAL RESULTS

This work focuses on identifying the named entities in the given text. The size of the dataset considered for this work is given in Table 1. The training and test corpus consist

of 1500 and 100 tokens along with their POS tag features respectively. The total number of entities in the corpus is 184. Random Kitchen Sink algorithm does not support data in text format. So, the dataset is converted into binary matrix.

Dataset	No of words	Dimension	No of Tweets
Train	1500	1500 x 5760	125
Test	100	100 x 5760	17

Table 1.size of dataset

Binary Matrix generation is done with the help of scikit learn module in Python with DictVectorizer. Dict objects. The categorical features in the corpus are given as "attribute-value" pairs. Using the scikit learn module in Python, both the training and testing data and their features are stored as dictionary and unique characters are identified. We extract the features around each individual token of a corpus which resulted in wide matrix with most of its values are zero and one is being present at the location where the token is present in dictionary. The features that are absent are not stored in the matrix which represents 0.page.

The list of entities identified by running the algorithm is given in Table II. This tweet consists of mainly 8 tags namely B-PERSON, I-PERSON, B-ORGANIZATION, I-ORGANIZATION, B-PRODUCT, I-PRODUCT, B-LOCATION and I-LOCATION. The accuracy obtained from this system is 82.60%. The accuracy obtained in each tag is separately shown in Table II and a comparison between the identified entities and total numbers of entities present in tweets are shown in Figure 1.

From Table II, the entity such as I-PERSON and I-PRODUCT shows less accuracy when compared to other entities. The total number of entities in tweets is 184. The number of entities identified by Random Kitchen Sink algorithm is 152. The entity I-LOCATION shows higher accuracy.

Entities	No of entities present	No of entities identified	Accuracy (%)
B-PERSON	23	18	78.26
I-PERSON	16	12	75.00
B-ORGANIZATION	29	23	79.31
I-ORGANIZATION	17	15	88.23
B-PRODUCT	27	24	88.88
I-PRODUCT	12	9	75.00
B-LOCATION	36	29	80.55
I-LOCATION	24	22	91.66
Total	184	152	82.60

Table 2.number of entities present and identified

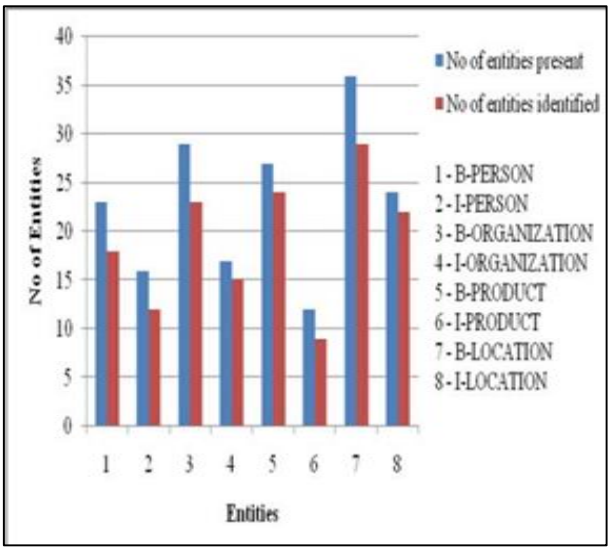

Fig. 1: Comparison of Entities present and identified.

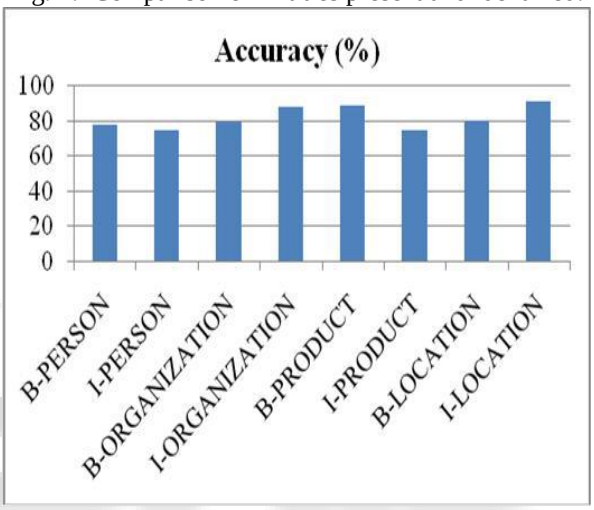

Fig. 2: Separate accuracy for each entity.

V. Conclusion

Random Kitchen Sink algorithm applied for Named Entity Recognition shows almost equal performance with flourishing algorithm named Conditional Random Field (CRF). This algorithm can be modified to get input of text data. This can also be extended to apply for various languages and tweets from various languages. The number of features can be increased in order to improve the accuracy of this system.

References

[1] Vijayakrishna, R., Devi, S. L. (2008, January). Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields. In IJCNLP (pp. 59-66).
[2] Nadeau, D., Sekine. S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.
[3] Nayan. A., Rao, B.R.K., Singh, P., Sanyal, R. (2008, January). Named Entity Recognition for Indian Languages. In IJCNLP (pp. 97-104).
[4] Farkas, R., Szarvas, G., Kocsor, A. (2006). Named Entity Recognition for Hungarian Using Various Machine Learning Algorithms. Acta Cybern., 17(3), 633-646.

[5] Malarkodi, C. S., Pattabhi, R. K., Sobha, L. D. Tamil NER– Coping with Real Time Challenges. In 24th International Conference on Computational Linguistics(p. 23).
[6] Talukdar, G., Borah, P. P., Baruah, A. (2014). A Survey of Named Entity Recognition in Assamese and other Indian Languages. arXiv preprint arXiv:1407.2918.
[7] Bikel, D. M., Schwartz, R., Weischedel, R. M. (1999). An algorithm that learns what's in a name. Machine learning, 34(1-3), 211-231.
[8] Grishman, R. (1995, November). The NYU System for MUC-6 or Where's the Syntax?. In Proceedings of the 6th conference on Message understanding (pp. 167-175). Association for Computational Linguistics.
[9] Saha, S. K., Chatterji, S., Dandapat, S., Sarkar, S., Mitra, P. (2008, January). A hybrid approach for named entity recognition in indian languages. In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages (pp. 17-24).
[10] Morwal, S., Jahan, N., Chopra, D. (2012). Named entity recognition using hidden Markov model (HMM). International Journal on Natural Language Computing (IJNLC), 1(4).
[11] Nadeau, D. (2007). Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision.
[12] Dey, A., Paul, A., Purkayastha, B. S. Named Entity Recognition for Nepali language: A Semi Hybrid Approach.
[13] Borthwick, A. (1999). A maximum entropy approach to named entity recognition (Doctoral dissertation, New York University).
[14] Jahan, N., Morwal, S., Chopra, D. (2012). Named entity recognition in indian languages using gazetteer method and hidden markov model: A hybrid approach. IJCSET, March.
[15] Kaur, D., Gupta, V. (2010). A survey of named entity recognition in english and other indian languages. The Proceedings of the IJCSI, 239-245.
[16] Rahimi, A., Recht, B. (2007). Random features for large-scale kernel machines. In Advances in neural information processing systems (pp. 1177-1184).
[17] Rahimi, A., Recht, B. (2009). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In Advances in neural information processing systems (pp. 1313-1320).
[18] Singh, T. D., Nongmeikapam, K., Ekbal, A., Bandyopadhyay, S. (2009). Named Entity Recognition for Manipuri Using Support Vector Machine. InPACLIC (pp. 811-818).
[19] Pandian, S., Pavithra, K. A., Geetha, T. (2008). Hybrid three- stage named entity recognizer for tamil. INFOS.
[20] Srinivasagan, K. G., Suganthi, S., Jeyashenbagavalli, N. (2014, March). An Automated System for Tamil Named Entity Recognition Using Hybrid Approach. In Intelligent Computing Applications (ICICA), 2014 International Conference on (pp. 435-439). IEEE.
[21] Gupta, V., Lehal, G. S. (2011). Named Entity Recognition for Punjabi Language Text Summarization. International Journal of Computer Applications,33(3), 28-32.

