A Machine Learning Approach for Cross Script Named Entity Recognition -A REVIEW

Anushka Singh
M.Tech Scholar
Department of Computer Science

Ruchi Bhargava
Assistant Professor
Department of Computer Science

Radharaman Institute of Technology and Science
Bhopal, MP

Abstract

Named Entity Recognition (NER) is an important subtask of information extraction. It recognizes and classifies multiword expressions with particular meaning, e.g. persons, locations, organizations etc. Most of the time, these expressions carry the core information of the text. This information can be utilized for better structuring of documents, filtering of essential texts. It can be used as an input for other natural language processing (NLP) tasks like question answering, summarization or machine translation.

There are two fundamental issues of current NER framework. The first issue is the necessity to calibrate the system each new language or domain. There is extensive fall in the quality of the output, when a framework is intended for one space is utilized for another one. Transition from one language to another language is even more complicated. Second issue is the absence of external and semantic knowledge, which is significant for individuals to perceive names in texts such as internet forum posts.

This paper reports about the development of NER framework for Wikipedia dataset crawled based on Cross Script coarse NE Indian context (list of person, location, organization and miscellaneous) by using various machine learning algorithms like Naïve Bayes Classifier, Support Vector Machine, Random Forest Classifier and Conditional Random Filed. The framework uses various types of features that are helpful in predicting different named entities (NEs). The set of features used for this work includes language dependent as well as language independent components. We have built the dataset of 2916 course NEs from Wikipedia page for Cross script Roman Hindi labeled with a label set of four diverse NE classes. We accounted just the labels that signify Person names, Location names, Organization names and Miscellaneous. The framework has been tested with the course token sets of 584 NEs. The performance is evaluated in terms of F1-measaure and accuracy. The F1-measure for Person name, Location name, Organization name is observed as 0.75 using Naïve Bayes Classifier, 0.76 using Support vector Machine Classifier, 0.78 using Random Forest and 0.85 using Conditional Random Filed Classifier. The accuracy for Person name, Location name, Organization name is observed as 78% using Naïve Bayes Classifier, 80% using Support vector Machine Classifier, 81% using Random Forest and 87% using Conditional Random Filed Classifier. By this, we conclude that the Conditional Random Field gives the best F1-measure and accuracy on Wikipedia NER dataset.

Keywords— Named Entity Recognition, Natural language processing, Machine Classifier, Naïve Bayes Classifier, Random Filed Classifier, Cross Script coarse.

I. INTRODUCTION

NER is generally used to identify information units like name, including person name, location names and organization name, numeric expressions like time, date, percent and money expressions for different Information Extraction (IE) and various Natural Language Processing tasks. These entities in text were distinguished and recognized as one of the important subtasks of information extraction and was termed as Named Entity Recognition and Classification (NER). In spite of the fact that these sounds clear that exceptional case emerge to entail complex guidelines, e.g. when is the times of India an organization and when it is an artifact? When is the white house a location and when an organization? Are the branch workplaces of a bank an organization? Is a piece of clothing manufacturing plant an organization or a location? Is phone number a

numeric expression or a location? Is a street name a location? Is early in the day a time? Keeping these things in mind the end goal is to accomplish the consistency of human annotator.

By taking an un-annotated block of document various research on NER framework has been conducted, e.g. "Mr. Donald Trump won the U.S. presidential election on 8th November 2016 " and creating an annotated block of document- " <PERS> Mr. Donald Trump </PERS> won the <LOC> U.S </LOC> presidential election on <TIME> 8th November 2016 </TIME>".

Cross-script Named Entity Recognition (CSNER) is a subfield of information dealing with recognition of named entities written in a language different from it's original language.
Here in this work, we have built Cross script Roman Hindi named entity dataset which is crawled from Wikipedia page based on Indian context and model is proposed to find the named entities like person, location, organization and miscellaneous.
There are majority of applications in natural language processing for Named Entity Recognition (NER). Some of the applications are as follows:
NER is exceptionally convenient for search engines. Named Entity Recognition helps in organizing structured and textual information which assist in productive ordering and retrieval of documents for searching.
It is imperative to recognize if entity is named entity or not in Cross-lingual Information Access Retrieval (CLIR). If entity is NE then token is transliterated rather than translating.
The news aggregation stage is fueled by NER. The information can be examined by using NEs, such as organizing the prevalence of entities over time. However, the enhancement to conventional news aggregation conducted by named entities is the way they associate between things and people.
In machine translation, NER notices it's application. Normally, entities identified as NEs are transliterated rather than getting translated.
If the reader could demonstrate the NEs before reading an article. The reader would get a reasonable thought regarding the contents of the article.
NER is considered very useful in unmanned indexing of books. Majority of the words recorded in the back index of a book are NEs.
NER is useful in identifying NEs like proteins, medicines and diseases etc in biomedical domain.
Named entity tagger is normally a sub-task in large portion of the information extraction tasks because NEs adds structure to raw information.

II. OVERVIEW OF WORK

Support Vector Machine (SVM) is a supervised machine learning algorithm which is used for regression or classification type problem. SVM is generally used for classification type of problems. In this algorithm, each data item is plotted as an n-dimensional space where the value n is considered as the number of features. The classification is performed by drawing the hyper plane which differentiates the classes well. SVM are basically the coordinate of separate observation.
It is simple to have linear hyper-plane between two classes in SVM. The Kernel function takes low dimensional input space to transform into high dimensional space. In other words we say that it converts non separable problems into separable problem. For non-linear separation problem generally Kernel is used. Data transformation is done to discover the process that separates the data based on classes.

Various machine learning algorithms are applied on Wikipedia dataset and the performance is evaluated using 5 fold cross validation test then the accuracy is observed by using confusion matrix. The accuracy obtained by different classifier is shown is fig.6.7. These classifiers obtain the accuracy as 78% using Naïve Bayes (NB) Classifier, 80% using Support Vector Machine (SVM), 82% using Random Forest (RF) and 87% using Conditional Random Field (CRF). Hence by this chart we conclude that the Conditional Random Filed gives the best accuracy on Wikipedia dataset.

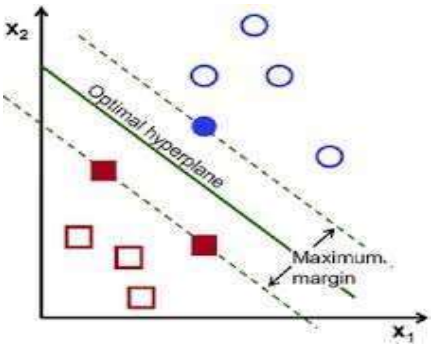

Fig-i: Support Vector Machine Classifier

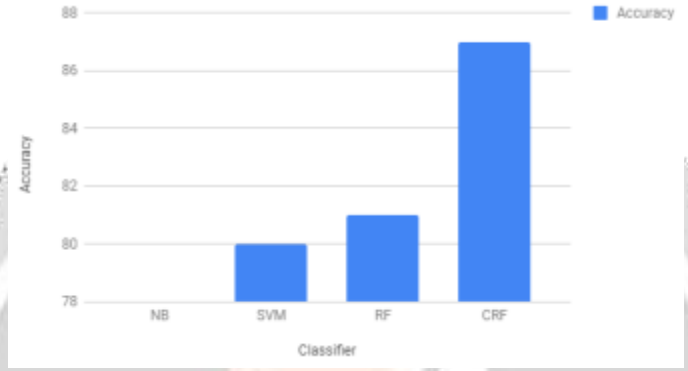

Fig-ii: Accuracy Vs Classifier

III. LITERATURE REVIEW

Literature based on the modelling of multi-storey building using floating column and transfer beam under seismic behaviour. From the detailed literature review, inference is studied.

Aravind Krishnan (2021) propose a novel approach for rapid pro- totyping of named entity recognisers through the development of semi-automatically anno- tated data sets. We demonstrate the proposed pipeline on two under-resourced agglutinating languages: the Dravidian language Malayalam and the Bantu language isiZulu. Our approach is weakly supervised and bootstraps training data from Wikipedia and Google Knowledge Graph. Moreover, our approach is relatively language independent and can consequently be ported quickly (and hence cost-effectively) from one language to another, requiring only minor language-specific tailorin**g.**

Bhargava et al (2016) describe about hybrid approach for code mixing NER in Indian Languages [4]. This framework uses hybrid strategy of a dictionary with supervised classification approach for figuring out entities in Code Mix Text of Indian Languages like Hindi- English and Tamil-English. Dataset contains 2700 Hindi-English tweets and 3200 Tamil-English tweets. There were 22 NEs present in the corpus. A word level NER framework is intended to recognize NEs in a tweet. This technique includes the pipelined approach for recognizing class of NEs. This pipelined approach has been partitioned into four stages: Pre-processing, Number Based NER, Gazetteer List Based NER and Tree Based NE Identifier. This paper attained the most astounding Precision, Recall and F1-measure as 58.84, 35.32 and 44.14 on English-Hindi language pair. . For Tamil-English language pair the Precision, Recall and F1-measure is attained as 58.71, 12.21 and 20.22. The proposed framework stood fifth among the Hindi-English Systems and fourth in the case of Tamil-English.

Liu et al (2016) describe to combine a K-Nearest Neighbor classifier with a linear Conditional Random Field model to demonstrate a semi-supervised learning composition for NER framework [7]. This paper proposes a NER framework to address challenges in recognizing named entities in tweets. To conduct word level classification, a KNN based classifier is embraced. NER systems use pre-labeled results together with other traditional features and this is passed into a linear CRF model, which leads the fine grained tweets level NER. These models are repeatedly retrained with enlarged training set into which high confidently labeled tweets are included. This is a hybrid framework of KNN & CRF model under a semi-supervised learning framework which

separates this model from the existing framework. There are 30 gazetteers used by this framework, which covers common names, countries, locations, temporal expressions and so on. The dataset is prepared by manual annotation of 12,245 tweets as the test. Result observed by this framework demonstrates that this model beats the baselines framework. The result is attained in terms of Precision: 81.6% Recall: 78.8% and F1-measur 80.2% with KNN classifier & the Precision: 82.6%, Recall: 74.8% and F measure: 78.5% without KNN classifier.

Gupta et al (2016) describe about the hybrid approach from code mixed language pairs for entity extraction such as English-Tamil & English-Hindi [1]. The hand-crafted feature set is used by the outcome of classifier. The dataset was prepared by crawling English-Hindi and English-Tamil language mixed tweets from twitter. There were total of 22 entities in training dataset where majority of the entities belong to `Entertainment', `Person' `Location' and `Organization'. Dataset contains 2700 English-Hindi tweets and 2183 English-Tamil tweets. For NER exhaustive set of features are used. These features are portrayed as Context word, Character n-gram, Word normalization, Prefix and Suffix, Word Class Feature, Word Position, Number of Upper case Characters, Test Word Probability and Binary Features. Tokenization and Token Encoding were executed as a major aspect of pre-processing. For labeling the sequence of token CRF classifier is used. After labeling obtained from CRF classifier, the rule and dictionary based post-processing was performed. This paper achieved highest Precision of 81.15% and f-measure of 62.17% on English-Hindi mixed language pair among all the submitted system. For Tamil-English language pair Precision, Recall and F-score is achieved as 79.92%, 30.47% and 44.12%. This framework achieves the best result among the frameworks for code mixed English-Tamil language pair participated in the CMEE-IL task.

Srivastava et al (2011) describe about the hybrid architecture of machine learning & rule based approach to identify NEs [10]. Various machine learning statistical approaches like MaxEnt, CRF and Rule based approach have been experimented on linguistic rules. In overcoming the restrictions of statistical models, linguistic approach plays a vital role for rich language like Hindi. The proposed framework uses voting method additionally to enhance the result of NER. For this work, dataset is obtained from IJCNLP08 website and SSF format is used in annotated Hindi corpus. The framework is trained on the training dataset of 10, 50, 100 and 150 files and tested on 10 files repeatedly for 10 rounds. The result is evaluated by CRF as Precision: 74.28%, Recall: 33.37% and F1-measure: 46.43% using 10 fold cross validation test.

Ekbal and Bandyopadhyay (2010) describe about the development of NER framework for Bengali & Hindi using SVM [3]. This framework uses contextual information of the entities with the variety of features that are helpful in identifying NEs. The dataset contains labeled annotated corpora of 122,467 tokens for Bengali and 502,974 tokens for Hindi with 12 NE classes. This framework uses unsupervised algorithm to induce the lexical context patterns from the part of unlabeled Bengali news corpus. The features are used as lexical patterns to enhance the framework performance. The NER framework is tested with test set of 35K tokens for Bengali and 60K tokens for Hindi. The evaluation result is observed as Precision: 80.12% Recall: 88.61% and F-score: 84.15% for Bengali and Precision: 74.34%, Recall: 80.23%, and F-score: 77.17% for Hindi

Ekbal et al (2008) describes the development of NER framework for Bengali using the statistical CRFs [9]. This framework uses contextual information of the tokens with the variety of features that are useful in identifying NE classes. The dataset is prepared from the leading Bengali newspaper by tagging NE Bengali news corpus. This framework is trained with 150K tokens with a NE tag set of 17 tags. The experimental result is evaluated with average Recall: 93.8%, Precision: 87.8% and F-Score: 90.7% using 10-fold cross validation test.

Nayan et al (2008) describe about recognition of NEs for Indian languages [5]. In this framework, various languages based on their similar phonetic matching strategy were used to match the strings. This model uses language independent approach and requires set of rules suitable for language. Firstly, the two tokens to be coordinated must be entitled in a ordinary script. Hence, they confront two decisions. It must change two tokens into some usual intermediate demonstration or transliterate the names written in Indian language to English & then it finds phonetic identity. The designed framework comprises of following module: Crawler, Parser, Phonetic Matcher, Transliteration Rules, and Baseline Task. The framework is tested with dataset which contains both English & Hindi language. The web crawler is used to crawl named entity list of both English & Hindi languages hence the idea of similar dataset is embedded. The evaluation result is observed for English

corpus in terms of Precision, Recall as 81.40% and 81.39%. The framework is tested on 1000 sentences and the result for Hindi is observed as: Precision 80.2% for all named entities, Recall 47.4% for person entities, Recall 42.9% for organization entities and Recall 74.6% for location entities.

IV. PROS AND CONS OF NAÏVE BAYES CLASSIFIER

Pros

- NB classifier is considered simple and quick in predicting the output class of test dataset. This classifier achieves good result in multiclass prediction too.
- NB classifier performs better compared to other model like logistic regression in case of independence assumption and less training data. This classifier performs well with categorical input varibale compared to numerical variable.

Cons

- In test dataset if the categorical variable has a category, which wasn't observed in training dataset then classifier will be assigned with zero probability hence it will not make prediction. This is termed as Zero frequency. This problem is resolved by using Smoothing technique. Laplace estimation is one of the smoothing techniques.
- NB classifier is considered as a bad estimator hence the output of probability from predicted probability is not taken too seriously.
- NB classifier is the hypothesis of independent predictors. It is not possible to discover the predictors set which are totally independent.

V . OBJECTIVE OF THE WORK

The salient objectives of the study have been

identified as follows:

The objective of the cross-script Named Entity Recognition (CSNER) is a subfield of information dealing with recognition of named entities. These named entities are classified into three parts by MUC-7. Entity (ENAME): person, location, organisation. Time Expression (TIME): time, date. Numeric Expression (NUME): percent, money. The dataset was prepared by crawling English-Hindi and English-Tamil language mixed tweets from twitter. In future we would like to build a large dataset and train the model by using a deep learning system

REFERENCES

[1]. Deepak Gupta, Asif Ekbal, Pushpak Bhattacharyya, "A Hybrid Approach for Entity Extraction in Code-Mixed Social Media Data", Fire 2016 shared task, 2016.

[2]. Asif Ekbal, Sivaji Bandyopadhyay, A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi, LiLT Volume 2, Issue 1, November 2009.

[3]. Asif Ekbal, Sivaji Bandyopadhyay, Bengali Named Entity Recognition using Support Vector Machine, World Academy of Science, Engineering and Technology 39- 2010.

[4]. Rupal Bhargava, Bapiraju Vamsi Tadikonda, Yashvardhan Sharma, "Named Entity Recognition for Code Mixing in Indian Languages using Hybrid Approach", Facilities 23, 10 (2016).

[5]. Animesh Nayan, B. Ravi Kiran Rao, Pawandeep Singh, Sudip Sanyal and Ratna Sanyal, "Named Entity Recognition for Indian Languages", Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, Hyderabad, India. Asian Federation of Natural Language Processing, January 2008.

[6]. Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay, "Language Independent Named Entity Recognition in Indian Languages", In Proceedings of IJCNLP workshop on NERSSEAL 2008.

[7]. Xiaohua Liu, Shaodian Zhang, Furu Wei, Ming Zhou, "Recognizing Named Entities in Tweets", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 359–367, Portland, Oregon, 2011 Association for Computational Linguistics, June 19-24, 2011.

[8]. Asif Ekbal, Sivaji Bandyopadhyay, "Bengali Named Entity Recognition using Support Vector Machine", Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 51–58, Hyderabad, India, 2008 Asian Federation of Natural Language Processing, January 2008.

[9]. Asif Ekbal, Rejwanul Haque, Sivaji Bandyopadhyay, "Named Entity Recognition in Bengali: A Conditional Random Field Approach", In Proc. of 3rd IJCNLP, 51-55, 2008.

[10]. Shilpi Srivastava, Mukund Sanglikar and D.C Kothari, "Named Entity Recognition System for Hindi Language: A Hybrid Approach", International Journal of Computational Linguistics (IJCL), Volume (2) : Issue (1) : 2011.

[11]. Anil Kumar Singh, "Named Entity Recognition for South and South East Asian Languages: Taking Stock", p. 5-7, In IJCNLP 2008.

[12]. Lafferty, J., McCallum, A., Pereira, F., "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", In: Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, p. 1-5, 2001.

[13]. Sriparna Saha, Asif Ekbal, "Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition", Data & Knowledge Engineering, Volume 85, Pages 15-39, May, 2013.

[14] A. Borthwick, "Maximum Entropy Approach to Named Entity Recognition". PhD thesis, New York University, 1999.

[15]. T. Mandl, C. Womser-Hacker, "The effect of named entities on effectiveness in cross-language information retrieval evaluation", In: Proceedings of the 2005 ACM Symposium on Applied Computing (SAC 2005), pp. 1059–1064, 2005.

[16]. A. McCallum, W. Li, "Early results for named entity recognition with conditional random fields, feature induction and Web-enhanced lexicons", In: Proceedings of CoNLL, Canada, pp. 188–191, 2003

[17]. Sudeshna Sarkar, Sujan Saha and Prthasarthi Ghosh, "Named Entity Recognition for Hindi", In Microsoft Research India Summer School talk, p. 21-30, May 2007.

[18]. W. Li, A. McCallum, "Rapid development of Hindi named entity recognition using conditional random fields and feature induction", ACM Transactions on Asian Languages Information Processing 2, 290–294, http://dx.doi.org/10.1145/979872.979879, 2004.

[19]. R. Florian, A. Ittycheriah, H. Jing, T. Zhang, "Named entity recognition through classifier combination", In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, 2003.

[20]. E.F. Tjong Kim Sang, F. De Meulder, "Introduction to the Conll-2003 shared task: language independent named entity recognition", In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp. 142–147. 2003.

[21]. Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla and Dipti Misra Sharma, "Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition", Workshop on NER for South and South East Asian Languages, IJCNLP 2008.

[22]. Burger, John D., John C. Henderson, and T. Morgan, "Statistical Named Entity Recognizer Adaption". In Proceedings of the CoNLL Workshop, pages 163–166. Taipei, Taiwan 2002.

[23]. Cucerzon, S. and David Yarowsky, "Language independent named entity recognition combining morphological and contextual evidence". In Proceedings of the 1991 Joint SIGDAT conference on EMNLP and VLC. Washington, D.C.1991.

[24]. Kumar, N. and Pushpak Bhattacharyya, "Named entity recognition in Hindi using memm". Technical report, IIT Bombay, India, 2006.

[25]. Kumar, P. Praveen and V. Ravi Kiran, "A Hybrid Named Entity Recognition System for South Asian Languages", In Proceedings of the IJCNLP08 Workshop on NER for South and South Asian Languages, pages 83-88, 2008.

[26]. Sujan Kumar Saha , Sanjay Chatterji, Sandipan Dandapat, Sudeshna Sarkar, Pabitra Mitra, "A Hybrid Approach for Named Entity Recognition in Indian Languages", Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 17–24, 2008.

[27]. P. Shishtla, P. Pingali , V. Varma, "A Character ngram Based Approach for Improved Recall in Indian Language NER", In Proceedings of IJCNLP Workshop on NER for South and South East Asian Languages, 2008.

[28]. Li Wei and McCallum Andrew, “Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction”, In ACM Transactions on Computational Logic, 2004.

[29]. Babych, Bogdan, A. Hartley, “Improving machine translation quality with automatic named entity recognition”, In Proceedings of EAMT/EACL 2003 Workshop on MT and other language technology tools, 1-8, Hungary.

[30]. Prasad Pingli et al. “A Hybrid Approach for Named Entity Recognition in Indian Languages”. IJCNLP, 2008.

[31]. Shilpi Srivastava, Siby Abraham, Mukund Sanglikar: “Hybrid Approach for Recognizing Hindi Named Entity”, Proceedings of the International Conference on Managing Next Generation Software Applications - 2008 (MNGSA 2008), Coimbatore, India, 5th- 6th December 2008.

[32]. Shilpi Srivastava, Siby Abraham, Mukund Sanglikar, D C Kothari: “Role of Ensemble Learning in Identifying Hindi Names”, International Journal of Computer Science and Applications, ISSN No. 0974-0767.

[33]. Hanna M. Wallach, "Conditional Random Fields: An Introduction”, Technical Report, University of Pennsylvania. 4-5, 2004