IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-8, Special Issue-1, Feb 2020 E-ISSN: 2347-2693

Supervised Machine Learning approach for Extracting Named Entities from Hindi-English Mixed Social Media Text

Suparna Arya[1*], Amit Majumder[2], Sulabh Majumder[3], Aparajita kundu[4], Nuzhat shamim[5], Ira Nath[6]

[1,2,3,4,5,6]Dept. of Computer Science and Engineering, JIS College of Engineering, West Bengal, India

[*]*Corresponding Author: suparnaarya1997@gmail.com*

Available online at: www.ijcseonline.org

***Abstract*—** Named Entity Recognition (NER) is a task of identifying named entities from text written in Natural Language. In this task, a string of text in the form of sentence or paragraph is accepted as input and relevant nouns like names of people, places, organizations etc. that are mentioned in that string are identified. This task belongs Information Extraction of the field of Natural Language Processing (NLP). Significant amount of work has been carried out on named entities recognition, but most of the researches have been done for resource-rich languages and domains. It is a challenging task for an informal text and code-mixed text which complicates the process with its unstructured and incomplete information. In this paper, we propose a method of extracting named entities from code-mixed data with different machine learning based algorithms using content and contextual features extracted from code-mixed data.

***Keywords*—** Named Entity, Machine Learning, Support Vector Machine, Decision Tree, K-Nearest Neighbour

I. INTRODUCTION

It has been seen that news and publishing houses are generating huge amount of online contents in regular basis. This huge amount of information needs to be managed correctly to get the most use of each article. Named Entity Extraction system can automatically extract the most useful information related to people, organizations, and places which are discussed in the text.

Multilingual speakers often switch back and forth between languages when they speak or write and it involves code-mixing from different languages. It has been seen that there are some complications in social media data itself. The shortness of micro-blogs makes them hard to interpret. Ambiguity is another major issue in code-mixed data. Micro-texts exhibit much more language variation and tend to be less grammatical than longer posts. Extracting useful information like named entities is going to be one difficult task. In this work, we use machine learning based approach with some relevant features to extract named entities from code-mixed data.

II. BACKGROUND AND RELATED WORK

Bali et al [1]. performed analysis of data from Face-book posts generated by English-Hindi bilingual users. Analysis depicted that significant amount of code-mixing was present in the posts. Vyas et al. [1,2] formalized the problem, created a POS tag annotated Hindi-English code-mixed corpus and reported the challenges and problems in the Hindi-English code-mixed text. They also performed experiments on language identification, transliteration, normalization and POS tagging of the Dataset. Sharma et al. [3] addressed the problem of shallow parsing of Hindi-English code-mixed social media text and developed a system for Hindi-English code-mixed text that can identify the language of the words, normalize them to their standard forms, assign them their POS tag and segment into chunks.

In Named Entity Recognition there has been significant research done so far in English and other resource rich languages Morwal et al. [4], but same cannot be said for code-mixed text due to lack of structured resources in this domain. Significant work was carried out on bengali data and code-mixed data for named entity recognition by Ekbal et al [6,7]. One hybrid model for NER on Hindi-English and Tamil-English code-mixed dataset was proposed by Bhargava et al. [5]. Bhat et al [8]. proposed a neural network architecture for NER on Hindi-English code-mixed Dataset. Named Entity Recognition for Hindi-English Code-Mixed Social Media Text was also addressed by Singh et al [9]. In the work of Named Entity Recognition in Tweets [10], it has been seen that the performance of the existing named extraction systems is not so good. There is a need of Named Entity Extraction system which is tuned on Tweets dataset.

III. PROPOSED APPROACH

Named entity extraction is one task which belongs to Information Extraction (IE) under NLP. We use machine learning approach by considering some useful features retrieved from code-mixed twitter dataset. For machine learning purpose we use Support Vector Machine, Decision Tree and K-Nearest Neighbour algorithm. All these algorithms are used for classifying data. The steps of our proposed approach are given below.

1. Collect raw text data from code-mixed twitter social media.
2. Segment the dataset into sentences.
3. Tokenize the sentences to find out tokens from the sentences.
4. Extract some useful features from the tokenized sentences.
5. Use machine learning algorithms to classify the tokens into predefined named entity types.
6. Measure accuracy of the system as a performance measurement in terms of recall, precision and f-score.

IV. FEATURES

As our dataset contains mixed data, finding syntax features is difficult. Therefore, we used content and contextual features only to perform named entity extraction. The following features have been used in our experiment.

- **Word:** This feature is defined as the current word in original form.
- **Word_lower :** The current word in lower-case form. For example, if the current word is "Played", its actual feature value will be "played"
- **Word_1stUpper:** Boolean feature. It is true if the first character of the current word is in upper-case. For example, if the current word is "Played", then its feature value will be 'True' as the word starts with capital character.
- **Word_isAlpha:** Boolean feature. It is true if the current word contains all alphabetic characters.
- **Word_isdigit :** Boolean feature. It is true if the current word contains all numeric characters.
- **Word_isupper :** Boolean feature. It is true if the current word is in upper-case.
- **Word_startsWith# :** Boolean feature. It is true if the current word starts with '#' character.
- **Word_startsWith@:** Boolean feature. It is true if the current word starts with '@' character.
- **Word$_{-1}$:** Previous word of the current word.
- **Word$_{-1}$_1stUpper:** Boolean feature. It is true if the first character of the previous word is in upper-case.
- **Word$_{-1}$_isAlpha:** Boolean feature. It is true if the previous word contains all alphabetic characters.
- **Word$_{-1}$_isdigit:** Boolean feature. It is true if the previous word contains all numeric characters.
- **Word$_{-1}$_isupper:** Boolean feature. It is true if the previous word is in upper-case.
- **Word$_{-1}$_lower:** Boolean feature. It is true if the previous word is in lower-case.
- **Word$_{-1}$_startsWith#:** Boolean feature. It is true if the previous word starts with '#' character.
- **Word$_{-1}$_startsWith@:** Boolean feature. It is true if the previous word starts with '@' character.
- **Word$_{+1}$:** Next word of the current word.
- **Word$_{+1}$_1stUpper:** Boolean feature. It is true if the first character of the next word is in upper-case.
- **Word$_{+1}$_isAlpha:** Boolean feature. It is true if the next word contains all alphabetic characters.
- **Word$_{+1}$_isdigit:** Boolean feature. It is true if the next word contains all numeric characters.
- **Word$_{+1}$_isupper:** Boolean feature. It is true if the next word is in upper-case.
- **Word$_{+1}$_lower:** Boolean feature. It is true if the next word is in lower-case.
- **Word$_{+1}$_startsWith#:** Boolean feature. It is true if the next word starts with '#' character.
- **Word$_{+1}$_startsWith@:** Boolean feature. It is true if the next word starts with '@' character
- **N-gram:** Character level n-gram features consisting of sub-words of length equal to n number of characters. We use 1-gram, 2-gram and 3-gram features.
- **BOS:** Boolean feature. It is true if the current word is the first word of the sentence.
- **EOS:** Boolean feature. It is true if the current word is the last word of the sentence.

V. DATASET AND EXPERIMENTAL RESULTS

The dataset that we used in our experiment are from Hindi-English code-mixed tweets on topics like politics, social events, sports, etc. from the Indian subcontinent perspective. The tags are labelled with 'Person', 'Organization', 'Location' using the BIO standard. Statistics of dataset is given in Table-1.

Table 1: Dataset statistics

Entity Type	Count
B-Per	2138
Other	63499
B-Org	1432
I-Org	90
B-Loc	762
I-Loc	31
I-Per	554
Total	68506

We perform machine learning approach using Support Vector Machine (SVM), K-Nearest Neighbour and Decision

Tree algorithms to extract named entities from code-mixed text data. The experimental results have been shown in Table-2, Table-3 and Table-4.

Table 2: Result by SVM

Named_entity type	precision	recall	f1-score
B-Loc	0.68	0.65	0.67
B-Org	0.69	0.59	0.63
B-Per	0.75	0.63	0.69
I-Loc	0.73	0.26	0.38
I-Org	0.54	0.24	0.34
I-Per	0.69	0.48	0.57
Other	0.97	0.98	0.98

Table 3: Result by K-Nearest Neighbour

Named_entity type	precision	recall	f1-score
B-Loc	0.58	0.67	0.63
B-Org	0.65	0.57	0.61
B-Per	0.77	0.54	0.64
I-Loc	0.57	0.26	0.36
I-Org	0.39	0.08	0.13
I-Per	0.61	0.35	0.45
Other	0.97	0.98	0.98

Table 4: Result by Decision Tree

Named_entity type	precision	recall	f1-score
B-Loc	0.65	0.61	0.63
B-Org	0.67	0.55	0.60
B-Per	0.70	0.63	0.67
I-Loc	0.30	0.26	0.28
I-Org	0.23	0.11	0.15
I-Per	0.63	0.43	0.51
Other	0.97	0.98	0.98

VI. RESULT ANALYSIS AND COMPARISON

We compare the results extracted by the three classifiers and find that the classifier generated by SVM algorithm is best for all types of entity labels (i.e. B-Loc, B-Org, B-Per, I-Loc, I-Org and I-Per). We also compare our experimental results with the existing named entity extraction system [9] which is implemented on the same dataset using CRF (Conditional Random Field) model and we find that our SVM based is better than the existing CRF-based models (based on f1-score value) for all the entity labels B-Loc (+2%), B-Org (+19%), B-Per (+2%), I-Loc (+4%) and I-Per (+2%) except I-Org (-5%).

VII. CONCLUSION

In this experiment we have applied three machine learning algorithms like SVM, Decision Tree and K-Nearest Neighbour algorithms using content and contextual features and we find that SVM provides result better than other algorithms. In our future work we will try to use more efficient features and to use more machine learning algorithms to check whether performance of the system increases or not. In this work, we have not applied deep learning approach. In future we would like to use deep learning approach to extract named entities from code-mixed text data.

REFERENCES

[1] Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "i am borrowing ya mix-ing?" an analysis of english-hindi code mixing in facebook. In Proceedings of the First Workshop on Computational Approaches to Code Switching, pages 116–126.

[2] Yogarshi Vyas, Spandana Gella, Jatin Sharma, Ka-lika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media con-tent. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 974–979.

[3] Arnav Sharma, Sakshi Gupta, Raveesh Motlani, Piyush Bansal, Manish Srivastava, Radhika Mamidi, and Dipti M Sharma. 2016. Shallow parsing pipeline for hindi-english code-mixed social media text. arXiv preprint arXiv:1604.03136.

[4] Sudha Morwal, Nusrat Jahan, and Deepti Chopra. 2012. Named entity recognition using hidden markov model (hmm). International Journal on Natural Language Computing (IJNLC), 1(4):15–23.

[5] Rupal Bhargava, Yashvardhan Sharma, and Shubham Sharma. 2016a. Sentiment analysis for mixed script indic sentences. In Advances in Computing, Com-munications and Informatics (ICACCI), 2016 Inter-national Conference on, pages 524–529. IEEE.

[6] Asif Ekbal and Sivaji Bandyopadhyay. 2008. Bengali named entity recognition using support vector machine. In Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages.

[7] Deepak Gupta, Shubham Tripathi, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A hybrid approach for entity extraction in code-mixed social media data. MONEY, 25:66.

[8] Irshad Ahmad Bhat, Manish Shrivastava, and Riyaz Ahmad Bhat. 2016. Code mixed entity extraction in indian languages using neural networks. In FIRE (Working Notes), pages 296–297.

[9] Vinay Singh, Deepanshu Vijay, Syed S. Akhtar, Manish Shrivastava. Named Entity Recognition for Hindi-English Code-Mixed Social Media Text. In Proceedings of the Seventh Named Entities Workshop, pages 27–35, Melbourne, Australia, July 20, 2018, Association for Computational Linguistics

[10] Alan Ritter, Sam Clark, Mausam, Oren Etzioni; Named Entity Recognition in Tweets: An Experimental Study; in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, July, Year:2011, Address:,Edinburgh, Scotland, UK.

.

Authors Profile

Suparna Arya is currently doing his B.Tech from Computer science and Engineering department, JIS College of Engineering, West Bengal, India. She has interest in Machine Learning, Information Extraction, Natural Language Processing.

Amit Majumder has received B.Tech degree in Computer Science and Engineering from Kalyani Govt. Engineering College, Kalyani, West Bengal, India, and ME degree in Computer Science and Engineering from Jadavpur University, Kolkata, West Bengal, India. Currently he is working as Assistant Professor in CSE department at JIS College of Engineering, Kalyani, West Bengal, India. He has interest in areas of Artificial Intelleigence, Soft Computing, Natural Language Processing, Machine Learning, Information Extraction. Currently he is doing his work on extracting bio-molecular event using Bio-NLP technique.

Sulabh Majumder is currently doing his B.Tech from Computer science and Engineering department, JIS College of Engineering, West Bengal, India. She has interest in Machine Learning, Information Extraction, Natural Language Processing.

Aparajita kundu is currently doing his B.Tech from Computer science and Engineering department, JIS College of Engineering, West Bengal, India. She has interest in Machine Learning, Information Extraction, Natural Language Processing.

Nuzhat shamim is currently doing his B.Tech from Computer science and Engineering department, JIS College of Engineering, West Bengal, India. She has interest in Machine Learning, Information Extraction, Natural Language Processing.

Mrs Ira Nath is presently working as an Assistant Professor in the Department of Computer Science and Engineering of JIS College of Engineering, India. She received the Master of Technology (M.Tech.) degree in Software Engineering from the Maulana Abul Kalam Azad University of Technology, India formerly West Bengal University of Technology, India in 2008. She also received the degree of Bachelor of Technology(B.Tech.) in Computer Science and Engineering from the sameuniversity in 2005. She is presently pursuing her Ph.D in Computer Science & Technology at Indian Institute of Engineering Science and Technology (IIEST), Shibpur, India. Her research interests include Network Security regenerator placement, survivability and routing and wavelength assignment in translucent WDM optical Networks