A NOVEL APPROACH FOR ENTITY EXTRACTION IN CODE MIXED DATA

Ganta Christina

Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India.

Dr. Selvani Deepthi Kavila

Assistant Professor, Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India

Badda Sravani

Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India

Boddu Sri Sankara Avinash

Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India

Abstract:

In the field of Natural Language Processing (NLP), Named Entity Recognition (NER) is one of the major task. The main challenge in this extraction is to extract Entities that lies in the inadequate information available in a tweet. There has been plenty of work done on this domain of entity extraction but it was mainly focused on popular languages such as English. In general extraction of entities from an informal text makes it difficult and for data that is written in two or more languages (code-mixed) makes it more difficult. In this paper the author has proposed the Machine Learning algorithms like Decision tree, and Conditional Random Field (CRF) with efficiencies of 60% and 76% respectively. The dataset was collected from FIRE-2016.

Keywords:

Social media text, Entity extraction, Code-mixed data, CRF, BIO format, Decision Tree.

1. INTRODUCTION:

Multilingual speakers often switch back and forth between languages when speaking or writing. This language interchange involves complex grammar, and the terms “code-switching” and “code-mixing” are used to describe it. An entity in a text is simply a proper noun such as name, city, place, product, organization and so on. Entity extraction in code mixed data is a process to extract the named entities which are present in the given text that is a code mixed data. Here we mainly concentrate on Hindi-English code mixed language. A significant number of researches are done in this field and some of them are Malayalam-English, Tamil-English so now the main target is on Hindi-English code mixed language. CRF and Decision Tree machine learning algorithms are used in entity extraction in code mixed languages.[Banerjee et al, 2017] Proposed formal as well as informal language-specific features to prepare the classification models and employed four machine learning algorithms (Conditional Random Fields, Margin Infused Relaxed Algorithm, Support Vector Machine and Maximum Entropy Markov Model) for the NE recognition (NER) task. [Gupta et al,2016] Proposed a hybrid approch for entity extraction from code mixed language pair English-Tamil. We use a rich linguistic feature set to train Conditional Random Field (CRF) classifier.

This entity extraction has many domains, some of them are :

1. Entity extraction is useful for those who are supposed to use voice to text conversion techniques such as siri, google assistant. This is used for those who wants to understand foreign phrases or sentences, best example is google

translator (converting from one language to another language).

2. Named Entity Recognition can automatically scan entire articles and reveal which are the major people, organizations, and places discussed in them, knowing the relevant tags for each article help in automatically categorizing the articles in defined hierarchies and enable smooth content discovery.

Following is a instance from a Twitter corpus of Hindi-English code-mixed texts also transliterated in English.

NOTE: In the below example, English words are in bold letters and Hindi words are in italics for better understanding.

T1: "*agar #notebandi ke* **time political party** *bhi* **#rti** *ke daayre me aa jati to #sukmath* **#kashmir** *me patthar* **attack** *na hote*"

Translation: "At the time of notebandi (Indian banknote demonetisation) if political party came under RTI's scope then in Kashmir stone attack would not have happen"

However, there is complication in social media data itself. First, the shortness of text in tweets makes it difficult to interpret. Second, as these micro text have more than one language in them, they tend to be less grammatical when compared with text in a single language.

Most of the research has, however been focused on resource rich languages, such as English, German, French and Spanish. However entity extraction and recognition from social media text for Indian languages and Code-Mixed text have been introduced a bit late.

2. RELATED WORK :

In recent years, many works were carried out in the field of processing text on code-mixed data. Vyas Y et al,2014 has worked on English-Hindi language social media content POS (Part-of-Speech) tagging was performed. Barman U et al,2014 has worked on code-mixed data of Bengali, Hindi and English a language identification task was carried out for Facebook data. Jamatia A et al,2018 discussed part-of-speech tagging of the corpora using both a coarse-grained and a fine-grained tag set, and compare their complexity to several other code-mixed corpora based on a Code-Mixing Index. Anupam Jamatia et al, 2015 has worked on POS (Part-of-Speech) tagging for Hindi-English code mixed data of Facebook and Twitter was performed with 90% result. Presented a language and POS tagged Hindi-English dataset of 1,489 tweets (33,010 tokens) that closely resembles the topical mode of communication on Twitter. Kushagra Singh et al,2018 has worked on the dataset is more extensive than any existing code-mixed POS tagged dataset and is rich in Twitter specific tokens such as hashtags and mentions, as well as topical and situational information. Three different methodologies are proposed in this paper for extracting entities from Hindi-English and Tamil-English code-mixed data. BIO-tag formatting is done as a pre-processing step. Extraction of trigram embedding is performed during feature extraction. Remmiya Devi G et al,2016 has developed of the system is carried out using Support Vector Machine-based machine learning classifier. Irshad Ahmad Bhat has Presented a simple feed forward neural network for Named Entity Recognition (NER) that use distributed word representations built using word2vec and no other language specific resources but the unlabeled corpora. Deepak Gupta et al,2016 has worked on the problem of code-mixed entity extraction comprises of two sub-problems, viz. entity extraction and entity classification.

3. PROPOSED SYSTEM:

A. Corpus :

The Hindi-English code mixed data taken for this experiment is the data which is collected from tweets that is collected from last 8 years. This data contains of topics like Politics, Sports, Social etc.. related to India(since the data being processed conatins Hindi). Extensive pre-processing is done to the corpus it self where noisy tweets are removed which only contain hashtags and the data which is either only in English or only in Hindi (Devanagari script) is also removed. So the data which will be further considered is based only on Hindi-English code-mixed data.

B. Preprocessing and Annotation: Named Entity Tagging :

Once the data is taken (corpus) then all the stops words are removed and the given input is tokenized sentence wise.

T1: "agar #notebandi ke time political party bhi #rti ke daayre me aa jati to #sukmath #kashmir me patthar attack na hote"

TOKENS: "agar", "#notebandi", "ke", "time", "political", "party", "bhi", "#rti", "ke", "daayre", "me", "aa", "jati", "to", "#sukmath", "#kashmir", "me", "patthar", "attack", "na" "hote".

Next the data is tagged based on three named entities that are Person, Location and Organisation. These are tagged in BIO format (Beggining, Intermediate , Other) which result in a total of 7 tags they are:

'Per' tag refers to the 'Person' entity which is the name of a Person**. B-Per** Indicates the Beginning of a Person's name. **I-Per** Indicates the intermediate of a Person's name. **'Org'** tag refers to the named entity of Organisation that is given to the names of social, political groups like Congress, Bhartiya Jnata Party (BJP),Hindus, Muslims, social media organizations like Instagram, twitter, whatsapp, etc. and also government institutions like State bank of India(SBI), banks, Swiss banks, etc. **B-Org** Indicates the Beginning of a Organizations's name. **I-Org** Indicates the intermediate of a Organizations's name. '**Loc**' tag refers to the named entity of location that is given to the names of places for eg. 'Visakhapatnam', '#India', 'Bharath', etc. **B-Loc** Indicates the Beginning of a Locations's name. **I-Loc** Indicates the intermediate of a Locations's name and if it don't fall in above 6 it is marked as **other.**

For the example considered the tags identified as follows:

T1 tags : agar/**other** #notebandi/**other** ke/**other** time/**other** political/**B-Org** party/ **I-Org** bhi/**other** #rti/**other** ke/**other** daayre/**other** me/**other** aa/**other** jati/**other** to/**other** #sukmath/**other** #kashmir/ **B-Loc** me/**other** patthar/**other** attack/**other** na/**other** hote/**other**

C. Data statistics:

In the statistics of data after the data is tokenized and Tagged. For each Tag the number of tokens recognised, count is given in table 1.

Table 1. Data Statistics after Tagging

Tag	Count of Tokens
B-Per	795
I-Per	31
B-Org	1528
I-Org	96
B-Loc	2362
I-Loc	571
Total	5383

D. System Architecture

Initially, the test data is given as an input after a series of steps in preprocessing the data set will be undergoing the BIO format conversion. Then for each word its feature will be tagged which undergoes evaluation metrics using CRF and Decision Tree Algorithms.

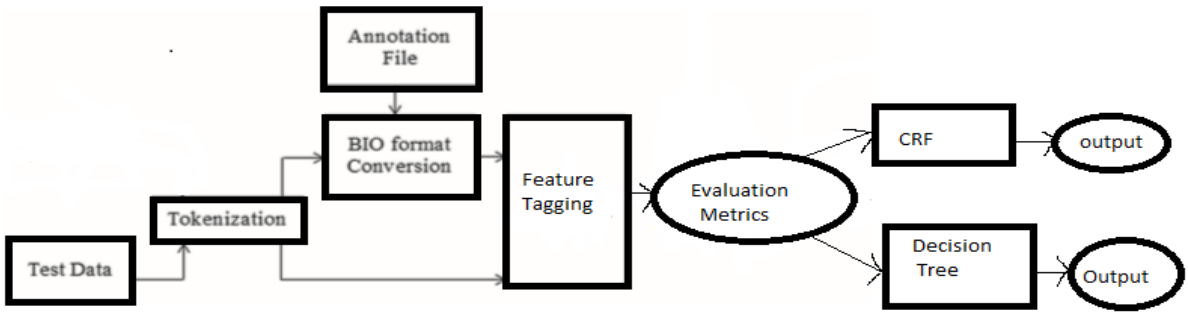

Figure 1. Architecture for Entity extraction in code-mixed data

a) Conditional Random Field (CRF):

CRF is a sophisticated algorithm. It is a class of statistical modelling method often applied in machine learning and used for structured prediction. CRFs fall into the sequence modelling family. Whereas a discrete classifier predicts a label for a single sample without considering "neighbouring" samples, a CRF can take context into account; e.g., the linear chain CRF (which is popular in NLP(Natural Language Processing)) predicts sequences of labels for sequences of input samples.

Below is the formula for CRF where Y is the hidden state (for example, part of speech) and X is the observed variable (in our example this is the entity or other words around it).

$$p(\mathbf{y}|\mathbf{x}) = \underbrace{\frac{1}{Z(\mathbf{x})}}_{\text{Normalization}} \prod_{t=1}^{T} \exp \left\{ \sum_{k=1}^{K} \underbrace{\theta_k}_{\text{Weight}} \underbrace{f_k(y_t, y_{t-1}, \mathbf{x}_t)}_{\text{Feature}} \right\}$$

There are 2 components to the CRF formula:

1. **Normalization**: Notice that there are no probabilities on the right side of the equation where we have the weights and features. However, the output is expected to be a probability and hence there is a need for normalization. The normalization constant Z(x) is a sum of all possible state sequences such that the total becomes 1.
2. **Weights and Features**: This component can be thought of as the logistic regression formula with weights and the corresponding features. The weight estimation is performed by maximum likelihood estimation and the features are defined by the user.

b) Decision Tree:

A decision tree is a tree where each node represents a feature(attribute), each link(branch) represents a decision(rule) and each leaf represents an outcome(categorical or continues value).

There are two measures:

Entropy: Defining a measure commonly used in information theory is called Entropy. It is given as

Entropy(X)= $\sum_{□ \epsilon □} - p(x)\log(p(x))$

Where X is collection of examples

Information Gain: It is a effectiveness of classifying an attribute . For an attribute A information gain can be given as Gain(S, A).

Gain(X, A) = Entropy(X) - $\sum_{□ \epsilon □□□□□□(□)}$(|Xv| / |X|) Entropy(Xv)

Where values(A) is the set of all possible values of A.

4. RESULTS AND PERFORMANCE EVALUATION

Precision:

Precision is the evaluation metrics. The ratio of currently predicted positive observations to the total predicted positive observations will become the precision.

Precision= Tp/(Tp+Fp)

Recall:

Recall is called evaluation metrics and also sensitivity hence the ratio of currently predicted positive observations to all observations in actual class becomes recall.

Recall=Tp/(Tp+Fn)

F1-Score:

It is the weighted average of precision and recall.

F-Score=2*(Recall*Precision)/(Recall+Precision)

Where,

TP- True positive means actual class becomes yes and prediction class is yes.

FP- False positive means actual class becomes no and predicted class is yes.

FN- False negative means actual class is yes and predicted class is no.

Support- Number of words identified in given sentences.

Table 2. Output of CRF model

	Precision	recall	F1-score	Support
B-Loc	0.74	0.56	0.64	795
B-Org	0.76	0.43	0.55	1528
B-Per	0.81	0.57	0.67	2362
I-Loc	0.73	0.26	0.38	31
I-ORG	0.62	0.27	0.38	96
I-Per	0.72	0.42	0.53	571
Other	0.96	0.99	0.98	66760
Micro-avg	0.95	0.95	0.95	72143
Macro avg	0.76	0.50	0.59	72143
Weighted-avg	0.95	0.95	0.95	72143

Table 3. Output of Decision Tree model

	precision	recall	F1-score	Support
B-org	0.93	0.28	0.43	250
B-loc	1.00	0.00	0.01	202
I-per	1.00	0.02	0.04	153
I-loc	0.00	0.00	0.00	10
B-per	0.83	0.41	0.54	645
I-org	0.00	0.00	0.00	23
Other	0.94	1.00	0.97	16653
Micro-avg	0.94	0.94	0.94	18036
Macro-avg	0.6	0.24	0.28	18036
Weighted-avg	0.94	0.94	0.92	18036

From table-2 and table -3 we can draw a comparison that CRF model performs better than the Decision Tree which is shown in table-4

Table 4. Comparison between CRF and Decision tree model

	Precision	recall	F1-score
CRF	0.76	0.50	0.59
Decision Tree	0.6	0.24	0.28

5.CONCLUSION AND FUTURE WORK

In this paper, present code-mixed Hindi-English data which is then BIO tagged for Person, Location and Organisation then it is classified based on CRF and Decision Tree algorithms with efficiencies of 76% and 60% respectively.

As a part of future work the data can be POS(parts of speech) tagged in word level which may give more accurate output. And moreover the data contains limited number of tweets , the tweets considered can be increased . The data considered is Hindi-English in this paper this can be done to other commonly used languages as well and the code-mixed data can be a mix of more than two languages for future work.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their careful reading of this paper and for their helpful comments

REFERENCES:

1. Anupam Jamatia, Björn Gambäck. ” Part-of-Speech Tagging for Code-Mixed English-Hindi Twitter and Facebook Chat Messages”. In: Proceedings of Recent Advances in Natural Language Processing, pages 239–248, Hissar, Bulgaria, (Sep 7–9 2015).

2. Barman U, Das A, Wagner J, Foster J. “ Code mixing: a challenge for language identification in the language of social media.” In: Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014, p. 13 (2014).

3. Banerjee, Somnath, Sudip Kumar Naskar, Paolo Rosso and Sivaji Bandyopadhyay. “Named Entity Recognition on Code-Mixed Cross-Script Social Media Content.” *Computación y Sistemas* 21 (2017).

4. Deepak Gupta, Asif Ekbal, Pushpak Bhattacharyya. “ A Deep Neural Network based Approach for Entity Extraction in Code-Mixed Indian Social Media Text.”(2016).

5. Gupta, D.K., Shweta, Tripathi, S., Ekbal, A., & Bhattacharyya, P.” A Hybrid Approach for Entity Extraction in Code-Mixed Social Media Data. “ In:FIRE(2016).

6. Irshad Ahmad Bhat, Manish Shrivastava ,Riyaz. “Code Mixed Entity Extraction in Indian Languages using Neural Networks” In: FIRE (2016).

7. Jamatia A., Gambäck B., Das A. “ Collecting and Annotating Indian Social Media Code-Mixed Corpora “. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2016, vol 9624. Springer, Cham(2018).

8. Kushagra Singh, Indira Sen, Ponnurangam Kumaraguru. “A Twitter Corpus for Hindi-English Code Mixed POS Tagging.” In: Proceedings of the sixth International Workshop on Natural Language Processing for Social. pages 12–17 (July 2018).

9. Remmiya Devi G., Veena P.V., Anand Kumar M., Soman K.P. "Entity Extraction of Hindi-English and Tamil-English Code-Mixed Social Media Text." In: Majumder P., Mitra M., Mehta P., Sankhavara J. (eds) Text Processing. FIRE 2016, vol 10478. Springer, Cham(2018).

10. Vyas Y, Gella S, Sharma J, Bali K, Choudhury M. " POS tagging of EnglishHindi code-mixed social media content." In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 974–979. Association for Computational Linguistics (2014).