PROJECT REPORT FOR CTNLP

ENTITY EXTRACTION SYSTEM

DATE: 05-10-17 *NAME: Remmiya Devi G*

INTRODUCTION

Social media is considered to be a vibrant area where millions of individuals interact and share their views. Processing social media text in Indian languages is a challenging task, as it is a well-known fact that Indian languages are morphologically rich in structure. On transferring such an unstructured text into a consistent format, the data is exposed to feature extraction method. In the huge corpora, information units i.e. entities hold the basic idea of the content. The main aim of the project is to recognise and extract the named entities in the social media Twitter text. The proposed system relies on the proficient count and prediction based word embedding models to extract the features for the words in the dataset. The proposed work makes use of text data from the Twitter resource in the Tamil language. In order to enhance the performance of the system, tri-gram features are extracted from the word embedding vectors. This project includes two experimentations: [i] Implementation and comparison of performance between word2vec and wang2vec systems with regard to tri-gram and 5-gram and [ii] Increasing training data size and implementation of word2vec, glove and word-glove systems with regard to tri-gram embedding. Hence, systems are trained using N-gram embedding features and named entity tags. Implementation of the system is using machine learning classifier, Support Vector Machine (SVM). The implemented system is integrated and an Graphical User Interface has been developed for the corresponding system.

OBJECTIVE OF THIS PROJECT

The main aim of the project is focussed on the primary task of Natural language processing, Entity extraction. The training data is retrieved from the FIRE2015 task. In order to improve the system, the additional dataset is collected from Twitter. Tagging of unlabelled data is done, to increase the training data of the system. The project includes two experimentations. [i] System implementation using features from word2vec and wang2vec embedding models. [ii] System implementation with additional training data, using features from word2vec and glove embedding models. Implementation and evaluation of the system are performed using the machine learning classifier, Support Vector Machine (SVM).

TERMINOLOGIES

(a) *Bio-formatting:* Tagging the beginning and end of each named entity (B - beginning, I - Inside & O- Outside). Words other than entities are tagged as O. If the name of a person has more than one token, first token is marked as B-PERSON, second token is marked as I-PERSON.

(b) *Word embedding models:* Word embedding models are the vector representation of words used for training system. It is a neural network based model used for implementing machine learning algorithms

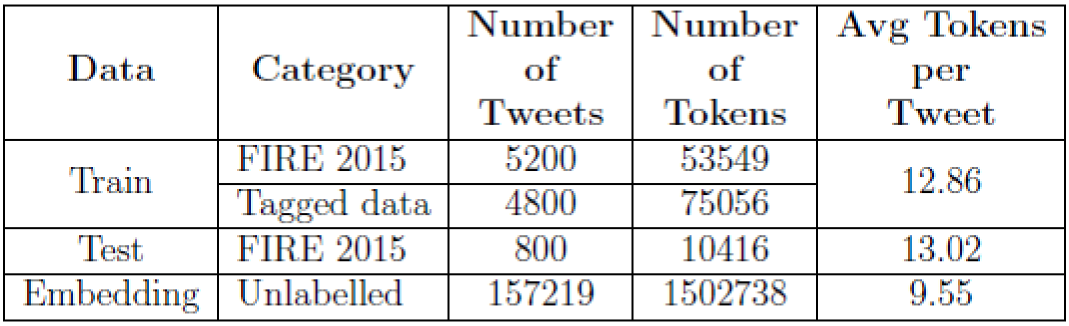
(c) *Glove Embedding:* Popularly known as count based model. It retrieves the vector representation of each word from the frequency of co-occurrences of context words.

DATASET DESCRIPTION

Social media text in the Tamil language from the Twitter platform is utilised in the proposed work. A part of the dataset is from FIRE2015 task and the remaining unlabelled data is extracted from the Twitter resource for the proposed system. Statistics of the dataset used in this project are tabulated in Table 1. Training data includes 10,000 tweets with 1,28,605 tokens, out of which 53,549 tokens are annotated data retrieved from the FIRE2015 dataset and remaining 75,056 tokens are manually tagged.

In the training dataset, there are about 23 named entities represented in BIO format. Among 23 entities in the whole corpus, based on the frequency of occurrence 4 major entities are taken into account for detailed performance analysis.

Table 1. Statistics of Dataset used in Proposed system



Data	Category	Number of Tweets	Number of Tokens	Avg Tokens per Tweet
Train	FIRE 2015	5200	53549	12.86
	Tagged data	4800	75056	
Test	FIRE 2015	800	10416	13.02
Embedding	Unlabelled	157219	1502738	9.55

METHODOLOGY

The proposed system uses Structures Skip-gram based word2vec model for generating word vector representation of words in the dataset. The proposed system 2 uses Co-occurrence matrix based vector representation from glove embedding. In order to enrich the word embedding, Tri-gram embedding of words was generated using the vector representation from Structured Skip-gram (SSG) model and glove embedding model. To improvise this feature set, with the existing trigram feature set, 5-gram feature set was generated that includes features of two preceding and two succeeding words for a target word. The system deals with social media text, specifically Twitter data. The dataset includes raw text, annotation and additional data. Entity tag and corresponding Named Entity is extracted from the annotated data that can be used while training the system. Training data and annotation data is subjected to pre-processing so that each word in train data holds its corresponding tag. Word embedding model generally requires additional dataset apart from training data.

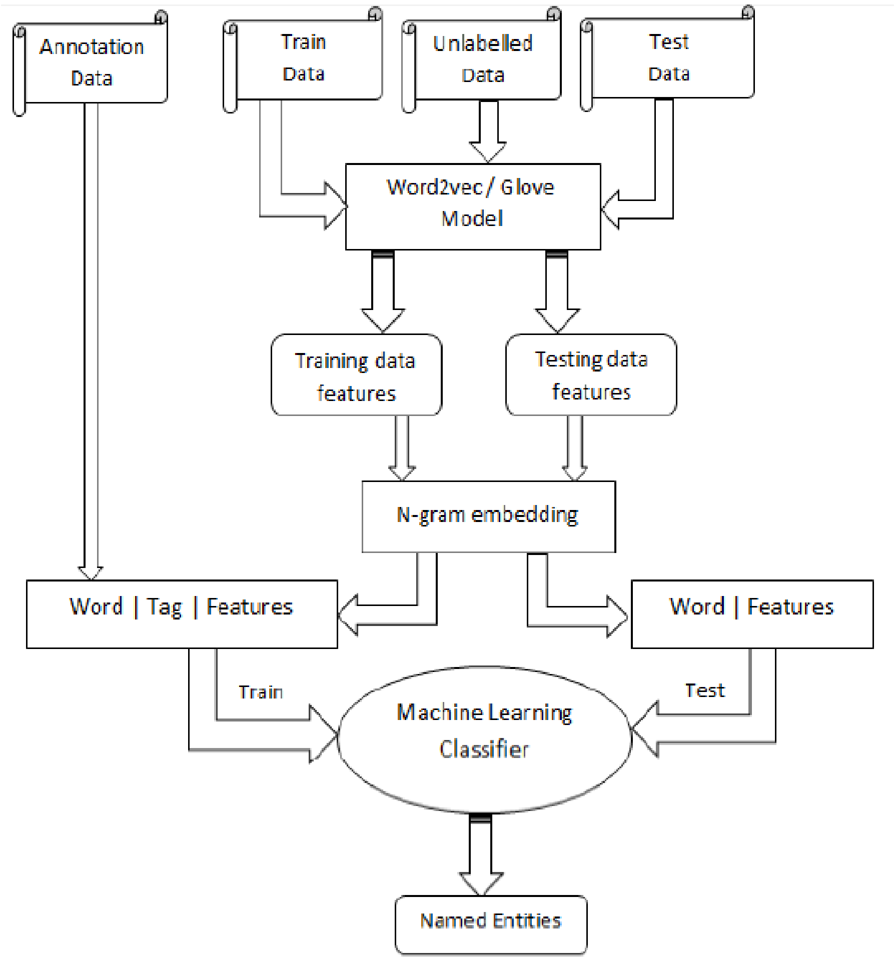
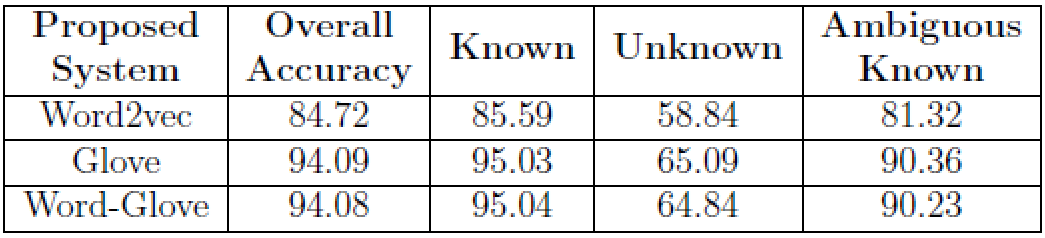


Figure 1. Methodology of proposed system

Structured Skip-gram model is the word embedding model used in the proposed work. Input for SSG model is in the form of sentences; this is because semantic information can be retrieved only from the sentences, not words. The output of word embedding model is the vectors for vocabulary words. The size of word vectors to be retrieved is set as per the need of the task. In this work, during training of SSG model, the size of vectors has been set as 100. With these vectors, Tri-gram embedding of words in train data is formed. Now, the size of the word vectors will be 300 (N*100, since tri-gram N = 3). Trigram embedding model is utilised to extract 5-gram embedding features for each word in the training data. The proposed work includes two systems, Tri-gram embedding and 5-gram embedding using structured skip-gram model from word2vec and glove embedding model. These two systems were implemented using a well-known machine learning based classifier, SVMLight. The proposed work is a novel method for entity extraction in Indian languages. System workflow is described through the schematic sketch in Fig. 1. The dataset extracted from the Twitter platform is exposed to pre-processing steps such as Tokenization, Hyperlinks removal, and BIO-format conversion. The dataset includes several meaningless symbols and tokens. Without proper pre-processing, it will create ambiguity at the time of feature extraction. BIO format conversion for Named entity recognition (NER) tasks is a traditional method in NLP. The named entity tags such as PERSON, LOCATION is converted into B-PERSON, I-PERSON and B-LOCATION and I-LOCATION, where B stands for the beginning tag and I stands for inside tag. The tokens other than named entities in the training data are tagged as O, which stands for outside tag. Based on these pre-defined categories of entity tags, manual tagging of part of the unlabelled dataset is done in this work.

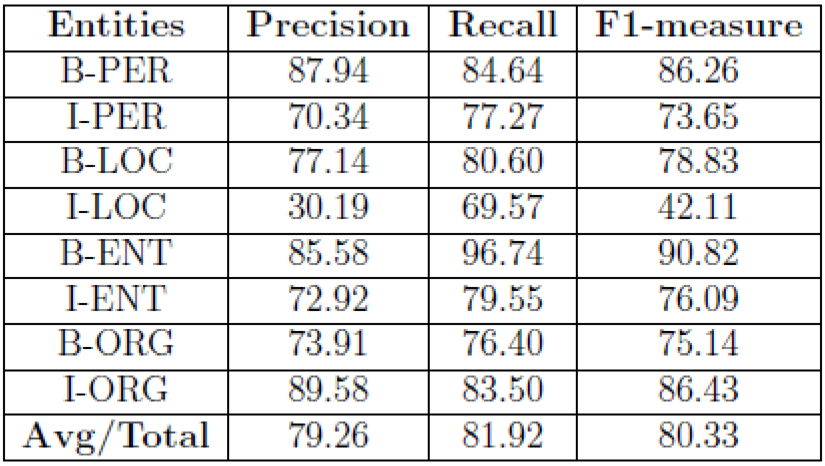
EXPERIMENT RESULTS

Table 3. Cross Validation results for the systems based on Word2vec, Glove and Word-Glove features



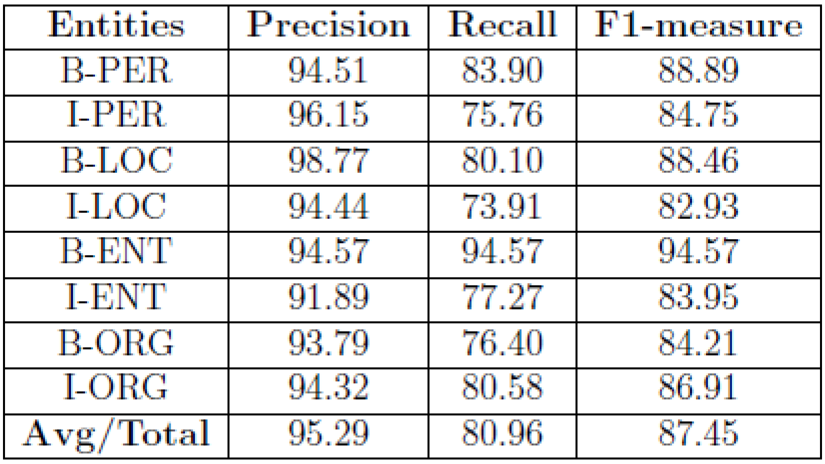
Proposed System	Overall Accuracy	Known	Unknown	Ambiguous Known
Word2vec	84.72	85.59	58.84	81.32
Glove	94.09	95.03	65.09	90.36
Word-Glove	94.08	95.04	64.84	90.23

Table 4. Performance Evaluation of system using word2vec embedding features



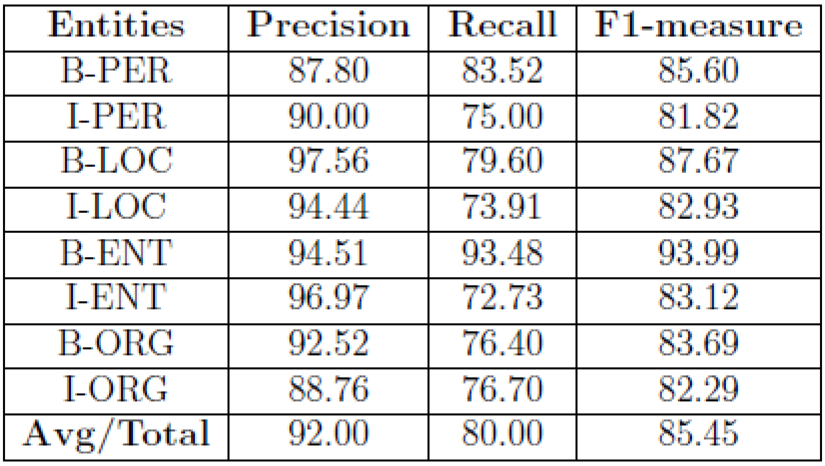
Entities	Precision	Recall	F1-measure
B-PER	87.94	84.64	86.26
I-PER	70.34	77.27	73.65
B-LOC	77.14	80.60	78.83
I-LOC	30.19	69.57	42.11
B-ENT	85.58	96.74	90.82
I-ENT	72.92	79.55	76.09
B-ORG	73.91	76.40	75.14
I-ORG	89.58	83.50	86.43
Avg/Total	79.26	81.92	80.33

Table 4. Performance Evaluation of system using Glove embedding features



Entities	Precision	Recall	F1-measure
B-PER	94.51	83.90	88.89
I-PER	96.15	75.76	84.75
B-LOC	98.77	80.10	88.46
I-LOC	94.44	73.91	82.93
B-ENT	94.57	94.57	94.57
I-ENT	91.89	77.27	83.95
B-ORG	93.79	76.40	84.21
I-ORG	94.32	80.58	86.91
Avg/Total	95.29	80.96	87.45

Table 4. Performance Evaluation of system using Word-Glove embedding features



Entities	Precision	Recall	F1-measure
B-PER	87.80	83.52	85.60
I-PER	90.00	75.00	81.82
B-LOC	97.56	79.60	87.67
I-LOC	94.44	73.91	82.93
B-ENT	94.51	93.48	93.99
I-ENT	96.97	72.73	83.12
B-ORG	92.52	76.40	83.69
I-ORG	88.76	76.70	82.29
Avg/Total	92.00	80.00	85.45

CONCLUSION

In this paper, the word features are focused on the most trending embedding models such as word2vec and glove. Feature enrichment is achieved by performing N-gram embedding. The results clearly depict the fact that tri-gram features outperform uni-gram features. Better clarity in tagging of entities during training can disambiguate the minor differences. The proposed system has proven its novelty through the joint embedding features and implemented using machine learning classifier. It can be seen that glove embedding vectors capture the semantic meaning of words better than word2vec features.