AMRITA_CEN-NLP@FIRE 2015:CRF based Named Entity Extraction for Twitter Microposts

Sanjay S.P
Centre for Excellence in Computational Engineering and Networking, Amrita Vishwa Vidyapeetham
Ettimadai, Coimbatore. India
sanjay.poongs@gmail.com

Anand Kumar M
Centre for Excellence in Computational Engineering and Networking, Amrita Vishwa Vidyapeetham
Ettimadai, Coimbatore. India
m_anandkumar@cb.amrita.edu

Soman KP
Centre for Excellence in Computational Engineering and Networking, Amrita Vishwa Vidyapeetham
Ettimadai, Coimbatore. India
kp_soman@amrita.edu

1 ABSTRACT

This proposed method implements the Named Entity Recognition (NER) for four dialects Such as English, Tamil, Malayalam, and Hindi. The results obtained from this work are submitted to a research evaluation workshop Forum for Information Retrieval and Evaluation (FIRE 2015). It is single-layered problem which is divided into multi- layered this step is called pre-processing; it has three levels of named entity tags which are referred as BIO format. This format is trained using Condition Random field(CRF) are used for implementing in NER system , the results obtained are grouped back to single-label or single-tagged referred as Format converting. In FIRE 2015, we developed English, Tamil, Malayalam, and Hindi NER system using CRF. The FIRE estimated the average precision for all the four languages.

CCS Concepts

• **Theory of computation~Conditional random feild** • **Computing methodologies~Natural language processing** • *Information systems~Information extraction* • *Human-centered computing~Social tagging systems*

Keywords

Named Entity Recognition (NER); Natural Language Processing (NLP); Conditional Random Fields (CRF). Entity Extraction from Social Media Text -Indian Languages (ESM-IL);

2 INTRODUCTION

Named-entity recognition (NER) is known as entity chunking, entity identification and entity extraction. It is an information extraction that find and locate , classify elements in text documents into defined categories such as organizations, the names of persons, locations, quantities, , expressions of times, monetary values, percentages, etc. That seeks to locate and classify elements in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

The Tweets are the general user data which the user use to communicate with others. The Tweets contains all the named entity like Person, organization, location, money, data, time, etc. the entity recognition is little difficult to the normal entity extraction due to user typed data which has no pre format or it may contains many short forms and mixed data .The NER is used in d IT sectors, tweets and conversation monitoring etc.The given files are converted into BIO format for the training the data .no preprocess where done or no data is modified in the languages. After the BIO format conversion the needed features has been extracted along with the pos tagged words which is given to the CRF++ for training and testing the data.The remaining discussion in this paper are , 3 Task descripition,4 system overview 5 conclusion ,6 acknowledgement .

3 TASK DESCRIPTION

The task provided to us is challenged with 2 types of data set .The first file contains the TWITTER data, and the second file contains the ANNOTATION file which has the information of the tag ,index, and length of the Twitter data's.so the given data is first preprocessed into BIO format and then extra features are added to it and then trained using CRF. This work is based on Conditional Random Field (CRF). It is used for developing NER system based on Machine learning approach. It is a customizable tool in which Feature sets can be redefined and fast training is possible. The converted BIO format is used for training the CRF. And output results are generated. The BIO format was like:

Table 1

Data	**BIO-tag**
@aajtak Mr.BAssi	O
…	O
…	O
Delhi	B-ORGANIZATION
Govt.may	I-ORGANIZATION
involve	O

The given data's are like:

623056949634994177 1945618028 @aajtak Mr.BAssi …. Delhi Govt.may involve

The first 2 numbers are Twitter id and user id which was mapped with an Annotation file which was in the format like:

Twitter id:623056949634994177 Userid:1945618028 NETAG:ORGANIZATION NE:Delhi Govt.may Index:105 Length:14

3.1 TRAINING DATA SET

The challenge in this task provided with 2 types of data. They provided data's for four language English, Tamil, Hindi, and Malayalam. The data provided is converted into BIO format and then it is trained using the CRF. The number of sentences for the training and testing data are given in the table below.

Table 2

Language	English	Tamil	Malayalam	Hindi
Train Data	5941	6000	8426	7983
Test Data	9595	8222	4121	10752

The data for 4 language are taken and then trained .the training set include feature files unigram feature and bigram feature. The languages for which these features are obtained are given bellow.

4 SYSTEM OVERVIEW

The training data with extracted features are then given to CRF++.The template file is the information for extracting the features .the CRF trains the data according to the template file and produce the model files. There are 2 model files for which the template files are altered for Unigram and Bigram features. The extracted feature file along with the NE tagged data is now trained using CRF++

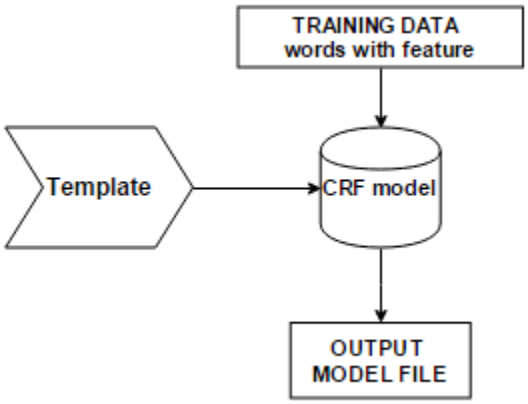

Figure 1

The extracted model files are then used for testing.

4.1 CRF MODEL BASED NER

In this task CRF++ is used for training and testing. The extracted features are trained using CRF. The template file which contains all the information to extract the feature. Each sentence is separated by an empty line. The CRF will generated a 2 model files ,the first model files has only unigram features and the second model files has Unigram and Bigram features. The languages for which unigram and bigram feature. In the example data flow diagram (4.1.1) the words w1, w2, w3, w4 are given to the feature extraction unit where all the binary features and the pos tag features, culture, length, position features are extracted and added with the BIO format. File which is then given to the CRF++ along with the CRF model file which has the trained data file. The CRF will return the output of the tagged file in the BIO format. The format conversion block will convert the file back to the ANOTATION format for the evaluation.

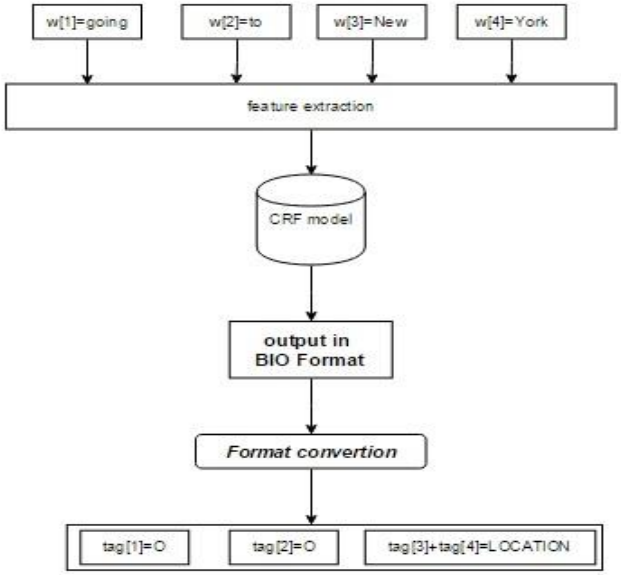

Figure 2

4.2 Features

Context words:

The previous word and the next word is considered for training the data.

POS tag:

The training and testing data are POS tagged with the tagger tools. Twitter POS tagger does not exist for other language than English, so we used the standard POS taggers except for Tamil. The Twitter POS tagger by Gimbel [7] is used for English language .Malayalam POS tags are retrieved from the Malayalam POS tagger. NLTK Hindi POS tagger is used for tagging the Hindi tweets. The pos tagged data plays an important role as they improve the accuracy.

Prefix and suffix:

The prefix suffix features will check the previous and next letter. The 2, 3, 4 are the count of the letters which they check before and after which is added for all the 4 language.

Clusters:

The clusters are taken only for the English language, the brown cluster is used for the English. There are no cluster tool or the Indian languages so this feature is not taken for other languages.

The linguistic Feature: the extracted features for the 4 languages are given below

Table 3

Features	English	Hindi	Tamil	Malayalam
Context words: The Previous and the next word	✓	✓	✓	✓
Pos tag: The part of speech tag for the current word	✓	✓	✓	✓
Prefix and suffix : The prefix suffix of length 3,4,5 are taken	✓	✓	✓	✓
Clusters : using brown cluster	✓	X	X	X
Shape feature	X	X	✓	X
Length : the word length as a feature	✓	✓	✓	✓
Position: position of the word as a feature	✓	✓	✓	✓

The binary features for the languages are given bellow.

Table 4

Binary Features	English	Hindi	Tamil	Malayalam
Contains number	✓	✓	✓	✓
Capitalization	✓	X	X	X
Contains Dot	✓	✓	✓	✓
ends with Comma	✓	✓	✓	✓
Ends with !	✓	✓	✓	✓
Ends with ?	✓	✓	✓	✓
Contains #	✓	✓	✓	✓
Contains ‘s	✓	X	X	X

The extracted features are combined with the BIO file and then tested.

Binary features:

In this binary features the values will be either 1 or 0. The feature is 1 if there exist a (.,!? #). This features are called binary features and for English capitalization and‘s is also taken as a binary features.

4.3 SYSTEM EVALUATION

Approximate match metric is used for evaluating partial correctness of the named entity. The right boundary should match. The named entity tag should be same as the gold standard tag. The tags that are perfectly matched are given weightage of 1 and partially matched tags are given weightage of 0.5. Among 10 Named Entities identified by the system, if 4 are perfectly identified and 5 are partially identified then approximate match = ((4∗1) + (5∗0.5))/10 = 0.65.

4.4 Runs

Table 5

In this task we have submitted 2runs.

Language	Unigram feature only(Run 1)	Unigram and Bigram feature(Run 2)
Hindi	✓	X
English	✓	✓
Tamil	✓	✓
Malayalam	✓	X

Run1: Hindi, English, Tamil and Malayalam runs with only unigram features are trained in CRF and tested.

Run2: English and Tamil files with unigram and bigram features are trained and tested.

4.5 Results

Language	Hindi			Tamil			Malayalam			English		
Runs	P	R	F	P	R	F	P	R	F	P	R	F
Run1	74.65	5.26	9.83	70.11	19.81	30.89	60.05	39.94	47.97	46.78	24.90	32.50
Run2	-	-	-	54.87	18.91	28.13	-	-	-	46.88	25.64	33.15

5 CONCLUSION

In this paper we briefly discussed about the NER for twitter data.we used CRF++ for the tagging of the data. The extended features has been discussed and table for all the linguistic features and Binary features has been briefly explained. The tagged data has been identified .since Twitter data is huge so we are in the need for Entity extraction for various purposes.

The future work will be based on added more rich features like clustering the data for all the Indian languages. We need to perform an error analysis so we could improve the effectiveness of the data.

6 AKNOWLEDGEMENT

We would like to thank Forum of Information Retrieval and Evaluation (FIRE 2015) organizers for organizing a wonderful research evaluation workshop and giving opportunities for researchers to present their work on Natural Language Processing (NLP). We also like to thank Computational Linguistics Research Group (CLRG), AU-KBC Research Centre, for organizing the Entity Extraction from Social Media Text Indian Languages (ESM-IL) Task.

REFERENCE

[1] Abinaya.N, Neethu John, M. Anand Kumar and Dr.K.P. P Soman - Amrita University.AMRITA@FIRE-2014: Named Entity Recognition for Indian Languages*FIRE 2014*.

[2] P Gupta, Kalika Bali, R E Banchs, M Choudhury, P Rosso.Query Expansion for Mixed-Script Information Retrieval*. In Processing's of the 37th international ACM SIGIR conference on Research & development in information retrieval2014.*

[3] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In Proc. of ICML, pp.282–289, 2001

[4]J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In Proc. of ICML, pp.282-289, 2001

[5]Karen Stepanyan , George Gkotsis , Vangelis Banos , Alexandra I. Cristea , Mike Joy, A hybrid approach for spotting, disambiguating and annotating places in user-generated text, Proceedings of the 22nd international conference on World Wide Web companion, May 13-17, 2013, Rio de Janeiro, Brazil

[6] Tuan Tran , Mihai Georgescu , Xiaofei Zhu , Nattiya Kanhabua, Analysing the duration of trending topics in Twitter using wikipedia,

Proceedings of the 2014 ACM conference on Web science, June 23-26, 2014, Bloomington, Indiana, USA

[7] Gimpel, Kevin, et al. "Part-of-speech tagging for twitter: Annotation, features, and experiments." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011.

[8] Kalika Bali, Yogarshi Vyas,Monojit Choudhury– Microsoft India and University of Maryland.POS Tagging of English-Hindi Code-Mixed Social Media Content.Proceedings of the 2014 EMNLP pages 974–979, October 25-29 (2014).