# AMRITA - CEN@NEEL : Identification and Linking of Twitter Entities

Barathi Ganesh H B, Abinaya N, Anand Kumar M, Vinayakumar R, Soman K P
Centre for Excellence in Computational Engineering and Networking
Amrita Vishwa Vidyapeetham, Coimbatore, India
barathiganesh.hb@gmail.com, abi9106@gmail.com, m_anandkumar@cb.amrita.edu
vinayakumarr77@gmail.com, kp_soman@amrita.edu

## ABSTRACT

A short text gets updated every now and then. With the global upswing of such micro posts, the need to retrieve information from them also seems to be incumbent. This work focuses on the knowledge extraction from the micro posts by having entity as evidence. Here the extracted entities are then linked to their relevant DBpedia source by featurization, Part Of Speech (POS) tagging, Named Entity Recognition (NER) and Word Sense Disambiguation (WSD). This short paper encompasses its contribution to #Micropost2015 - NEEL task by experimenting existing Machine Learning (ML) algorithms.

## Keywords

CRF, Micro posts, NER

## 1. INTRODUCTION

Micro posts are a pool of knowledge with scope in business analytics, public consensus, opinion mining, sentimental analysis and author profiling and thus indispensable for Natural Language Processing (NLP) researchers. People use short forms and special symbols for easily conveying their message due to the limited size of micro posts which has eventually built complexity for traditional NLP tools [3]. Though there are number of tools, most of them rely on least ML algorithms which are effective for long texts than short texts. Thus by providing suffcient features to these algorithms the objective can be achieved. We experimented the NEEL task with the available NLP tools to evaluate their effect on entity recognition by providing special features available in tweets.

## 2. SELECTION OF ALGORITHMS

### 2.1 Tokenization

Tokenizing becomes highly challenging in micro posts due to the absence of lexical richness. It includes special symbols (:-), #, @user), abbreviations, short words (lol, omg), misspelled words, repeated punctuations and unstructured words (goooood nightttt, helloooo). Hence these micro posts were fed to the dedicated twitter tokenizer which accounts language identification, a lookup dictionary for list of names, spelling correction and special symbols [4][5] for effective tokenization.

### 2.2 POS Tagger

Due to the conversional nature of micro blogs with non-syntactic structure it becomes difficult in utilizing general algorithms with traditional POS tags in Penn Treebank and Wall Street Journal Corpus [6]. O'Conner et al. used 25 POS tagset which includes dedicated tags (@user, hash tag, G, URL, etc.) for twitter and reports 90% accuracy on POS tagging [7]. The ability of resolving independent assumptions and overcoming biasing problems make CRF as promised supervised algorithm for sequence labeling applications [8]. TwitIE tagger: which utilizes CRF to build the POS tagging model was thus used.

### 2.3 Named Entity Recognizer

CRF and SVM produced promising outcome for sequence labeling task which prompted us to use the same for our experiment. Long range dependency of the CRF can also solve Word Sense Disambiguation (WSD) problem over other graphical models by avoiding label and casual biasing during learning phase. Both CRF and SVM allow us to utilize the complicated feature without modeling any dependency between them. SVM is also well suited for sequence labeling task since learning can be enhanced by incorporating cost models [9]. These advantages provide flexibility in building expressive models with CRF suite and MALLET tools [10][11].

## 3. EXPERIMENTS AND OBSERVATION

The experiment is conducted on i7 processor with 8GB RAM and the flow of experiment is shown in Figure 1. The training dataset consists of 3498 tweets with the unique tweet id. These tweets have 4016 entities with 7 unique tags namely Character, Event, Location, Organization, Person, Product and Thing [1][2]. POS tag for the NER is obtained from TwitIE tagger after tokenization which takes care of the nature of micro posts and provides an outcome desired by the POS tagger model. The tags are mapped to BIO Tagging of named entities. Considering the entity as a phrase, token at the beginning of the phrase is tagged as 'B-(original tag)' and the token inside the phrase is tagged as 'I-(original tag)'. Feature vector constructed with POS tag and addi-

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

tional 34 features like root word, word shapes, prefix and suffix of length 1 to 4, length of the token, start and end of the sentence, binary features - whether the word contains uppercase, lower case, special symbols, punctuations, first letter capitalization, combination of alphabet with digits, punctuations and symbols, token of length 2 and 4 , etc.
After constructing the feature vector for individual tokens in the training set and by keeping bi-directional window of size 5, the nearby token's feature statistics are also observed to help the WSD. The final windowed training sets are passed to the CRF and SVM algorithms to produce the NER model. The development data has 500 tweets along with their id and 790 entities [1][2]. The development data is also tokenized, tagged and feature extracted as the training data for testing and tuning the model. The developed model performance is evaluated by 10- fold cross validation of training set and validated against the development data. The accuracy is computed as ratio of total number of correctly identified entities to the total number of entities and tabulated in Table 1.

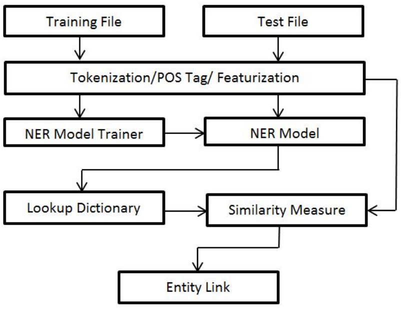

**Figure 1: Overall Model Structure**

$$Accuracy = \frac{\sum correctly\ identified\ entities}{total\ entities} \times 100 \quad (1)$$

MALLET incorporates O-LBFGS which is well suited for log-linear models but shows reduced performance when compared to CRFsuite which engulfs LBFGS for optimization [12][13]. SVM's low performance can be improved by increasing the number of features which will not introduce any over fitting and sparse matrix problem [9].
The final entity linking part is done by utilizing lookup dictionary (DBpedia 2014) and sentence similarity. The entity's tokens are given to the look up dictionary which results in few related links. The final link assigned to the entity is based on maximum similarity score between related links and proper nouns in the test tweet. Similarity score is computed by performing dot product between unigram vectors of proper nouns in the test tweet and the unigram vectors of related links from lookup dictionary. Entity without related links is assigned as NIL.

## 4. DISCUSSION

This experimentation is about sequence labeling for entity identification from micro posts and extended with DBpedia resource linking. By observing Table 1, it is clear that CRF shows great performance and paves way for building a smart NER model for streaming data application. Even though CRF seems to be reliable, it is dependent on the feature that has direct relation with NER accuracy. The utilized TwitIE tagger shows promising performance in both the tokenization and POS tagging phases. The special 34 features extracted from the tweets improves efficacy by nearing 13% greater than the model with absence of special features. At linking part, this work is limited using dot product similarity which could be improved by including semantic similarity.

**Table 1: Observations**

| Tools | 10 Fold-Cross Validation | Development Data | Time (mins) |
|---|---|---|---|
| Mallet | 84.9 | 82.4 | 168.31 |
| SVM | 79.8 | 76.3 | 20.15 |
| CRFSuite | 88.9 | 85.2 | 4.12 |

## 5. REFERENCES

[1] Rizzo, Giuseppe and Cano Basave, Amparo Elizabeth and Pereira, Bianca and Varga, Andrea, *Making Sense of Microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge.*, In 5th Workshop on Making Sense of Microposts (#Microposts2015), pp. 44–53, 2015.

[2] Matthew Rowe and Milan Stankovic and Aba-Sah Dadzie, *Proceedings, 5th Workshop on Making Sense of Microposts (#Microposts2015): Big things come in small packages, Florence, Italy, 18th of May 2015*, 2015.

[3] Dlugolinsky S, Marek Ciglan and M Laclavik, *Evaluation of named entity recognition tools on microposts*, INES, 2013 , pp. 197-202. IEEE, 2013.

[4] Bontcheva K, Derczynski L, Funk A, Greenwood M A, Maynard D, and Aswani N, *TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text*, In RANLP, pp. 83-90, 2013, September.

[5] Brendan O'Connor, Michel Krieger and David Ahn, *TweetMotif: Exploratory Search and Topic Summarization for Twitter*, ICWSM, 2010.

[6] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller and Justin Martineau, *Annotating named entities in Twitter data with crowdsourcing*, 2010.

[7] Kevin Gimpel, et al, *Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments*, HLT'11, 2011.

[8] John Lafferty,Andrew McCallum and Fernando Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, 2001.

[9] Chun-Nam John Yu, Thorsten Joachims, Ron Elber and Jaroslaw Pillardy, *Support vector training of protein alignment models*, in Research in Computational Molecular Biology, 2007.

[10] Naoaki Okazaki, *CRFsuite: a fast implementation of Conditional Random Fields (CRFs)*, 2007.

[11] McCallum and Andrew Kachites, *MALLET: A Machine Learning for Language Toolkit*, http://mallet.cs.umass.edu, 2002.

[12] Galen Andrew and Jianfeng Gao, *Scalable Training of L1-Regularized Log-Linear Models*, ICML, 2007.

[13] Jorge Nocedal, *Updating Quasi-Newton Matrices with Limited Storage*, Mathematics of Computation, Volume 35, Number151, pp:773-782, 1980.