NER from Tweets: SRI-JU System @MSM 2013

Amitava Das[1], Utsab Burman[2], Balamurali A R[3] and Sivaji Bandyopadhyay[4]

[1&3]Samsung Research India,
Bangalore, India.

[2&4]Department of Computer Science &
Engineering, Jadavpur University
Kolkata, India

{amitava.santu[1],utsab.barman.ju[2],balamurali.ar[3]}@gmail.com, sivaji_cse_ju@yahoo.com

Abstract. Now a day Twitter has become an interesting source of experiment for different NLP experiments like entity extraction, user opinion analysis and more. Due to the noisy nature of user generated content it is hard to run standard NLP tools to obtain a better result. The task of named entity extraction from tweets is one of them. Traditional NER approaches on tweets do not perform well. Tweets are usually informal in nature and short (up to 140 characters). They often contain grammatical errors, misspellings, and unreliable capitalization. These unreliable linguistic features cause traditional methods to perform poorly on tweets. This article reports the author's participation in the Concept Extraction Challenge, Making Sense of micro posts (#MSM2013). Three different systems runs have been submitted. The first run is the baseline, second run is with capitalization and syntactic feature and the last run is with dictionary features. The last run yielded than all other. The accuracy of the final run has been checked is 79.57 (precision), 71.00 (recall) and 74.79 (f-measure) respectively.

1 Introduction

Micro posts are the new form of communication in the web. Posts from different social networking sites and micro blogs reflect the present social, political and other events through user's text. Due to the limitation of message length (140 characters) and the noise of user generated content it is difficult to extract the concepts from them.

The different forms of user gen-erated noise makes Twitter text extreme noisy for standard NLP tasks. Such as -

a. Abbreviations and short forms of phonetic spelling (Examples: nite - "night", sayin -"saying"), inclusion of letter/number such as gr8-"great".

b. Acronyms (Examples: lol-"laugh out loud", iirc-"if I re-member correctly" etc).

c. Typing error/ misspelling in tweets. Examples: wouls-"would", ridiculous-"ridiculous".

d. Punctuation omission/error. (Examples: im -"I'm", dont-"don't").

e. Non-dictionary slang in tweets. This category includes word sense disambiguation (WSD) problems caused by slang uses of standard words, e.g. that was well mint ("that was very good"). It also includes specific cultural reference or group-memes.

· **#MSM2013 Workshop Concept Extraction Challenge Proceedings** ·
available online as CEUR Vol-1019, at: http://ceur-ws.org/Vol-1019
Making Sense of Microposts Workshop @ WWW'13, May 13th 2013, Rio de Janeiro, Brazil

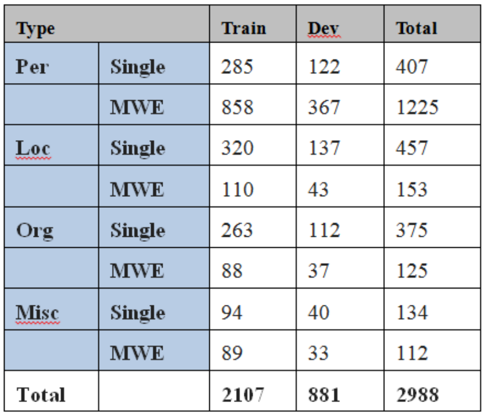
f. User's wordplay in tweets. This includes phonetic spelling and intentional misspelling for verbal effect e.g. that was soooooo great (“that was so great”).

g. Censor avoidance. This includes use of numbers or punctuation to disguise vulgarities, e.g. sh1t, f***, etc.

h. Presence of emoticons. While often recognized by a human reader, emoticons are not usually understood in NLP tasks such as Machine Translation and Information Retrieval. Examples: :) (Smiling face), <3 (heart).

2 Data

Table 1. NE Distribution of Training and Development Set



Type		Train	Dev	Total
Per	Single	285	122	407
	MWE	858	367	1225
Loc	Single	320	137	457
	MWE	110	43	153
Org	Single	263	112	375
	MWE	88	37	125
Misc	Single	94	40	134
	MWE	89	33	112
Total		2107	881	2988

The work has been done on MSM-2013 dataset. The datasets were available in 2 subsets as training and test datasets. No development set has been provided therefore the training data was divided into 2 further subsets (in 70%-30% ratio). The name entities are considered as two types - single word NE and multiword NE. The division of the available training data was made based on the presence of 4 different types of name entities with each type single and multiword. The statistics of the above process is elaborated in Table 1.

3 Experiment

Three different runs have been submitted. This is a CRF based system and the features are described below. Yamcha toolkit has been used for CRF implementation.

3.1 Baseline

Our baseline system incorporates the part of speech tags, stemmed tokens to train the baseline classifier. For POS tags of a micro post, we used CMU-POS tagger tool[1] which is specialized for tweets.

3.2 Capitalization

Capitalization of tokens is one of the key features to recognize the name entities in micro posts. It has been used as a binary feature in the classifier.

3.3 Predicate Rules

Generally the position of a name entity in a sentence is always close to the positions of functional words. For example in, of, near and etc. N-grams rules have been developed and used to train the classifier.

3.4 Out of Vocabulary Words

Most of the name entities are not the dictionary words. We used Samsad[2] & NICTA dictionary[3] in the experiment.

3.5 Gazetteers

For Location and MISC types two separate lists has been augmented. The LOC type consists of 220 country names and 100 popular city names. The MISC type has 110 NEs of different types. Mostly the error case in the Dev set.

We have experimented with series of features. Tweets are extremely noisy and therefore a concise set of named entity clue is very hard to finalize. Indeed person and organization categories are relatively naïve but location and miscellaneous category are very hard for a classifier.

4 Performance

The performance results on the Dev set is been reported in the Table 2. It should be noted the actual result on the test is yet to be evaluated by the organizer of MSM.

[1] http://www.ark.cs.cmu.edu/TweetNLP/

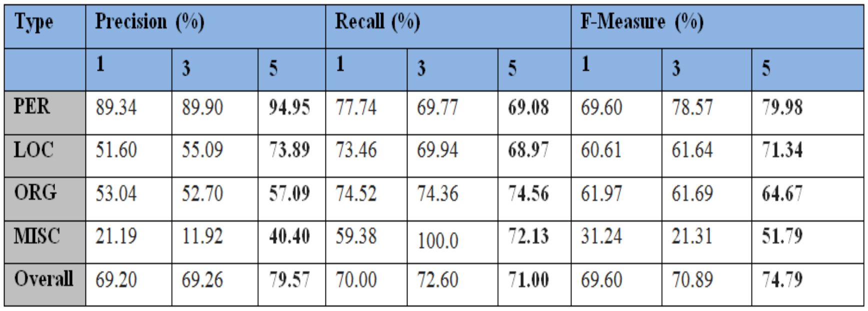
[2] http://dsal.uchicago.edu/dictionaries/biswas-bengali/

[3] http://www.csse.unimelb.edu.au/~tim/etc/emnlp2012-lexnorm.tgz

We run multiple iterations to reach the final accuracy. Broadly they could be categorized in 5 genres, as reported below. Among those iterations 3 best runs (1, 3 and 5) have been submitted. The details of the features used in each runs are as below and the scores are elaborated in Table 2.

1) Baseline: POS + Stem
2) 1 + Capitalization: Capitalization feature
3) 2 + N-Grams FW Predicates: in, of, or features
4) 3 + OOV
5) 4+Gazetters: LOC Dict + MISC Dict

Table 2. Experiment Results on Development Set



Type	Precision (%)			Recall (%)			F-Measure (%)		
	1	3	5	1	3	5	1	3	5
PER	89.34	89.90	**94.95**	77.74	69.77	**69.08**	69.60	78.57	**79.98**
LOC	51.60	55.09	**73.89**	73.46	69.94	**68.97**	60.61	61.64	**71.34**
ORG	53.04	52.70	**57.09**	74.52	74.36	**74.56**	61.97	61.69	**64.67**
MISC	21.19	11.92	**40.40**	59.38	100.0	**72.13**	31.24	21.31	**51.79**
Overall	69.20	69.26	**79.57**	70.00	72.60	**71.00**	69.60	70.89	**74.79**

5 Conclusion

In this paper we present a novel method for identification and classification of name entities based on the features. Though classifying named entities from twitter data is hard because of the noise and non-grammatical nature.

In this article we report our scores based on dev. set, we will incorporate the evaluation scores of #MSM2013 to support our evaluation framework.

Form the features that took part in our experiments, the gazetteer list, used in our experiment is small. We will try to include more in future.

We have observed that a-few Structural information can help to increase the results. For example - URL, Mention and Hash Tag. Our exploration is to find out more viable features that help to understand the semantics of micro post.

References

1. Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-

of-speech tagging for twitter: Annotation, features, and experiments. CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2010.

2. Ritter, Alan, Sam Clark, and Oren Etzioni. "Named entity recognition in tweets: an experimental study." In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524-1534. Association for Computational Linguistics, 2011.
3. Finin, Tim, et al. "Annotating named entities in Twitter data with crowdsourcing." Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Association for Computational Linguistics, 2010.
4. Han, Bo, and Timothy Baldwin. "Lexical normalisation of short text messages: Makn sens a# twitter." In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 368-378. 2011.