Named Entity Extraction From the Colloquial Setting of Twitter

Cassaundra Doerhmann
University of Colorado at Colorado Springs,
Colorado

***Abstract*—This paper suggests a study of Named Entity Recognition (NER) as it applies to Twitter and strategies that can be used to make NER systems more successful in colloquial settings such as Twitter. Named Entities are named nouns which fall into the catagories following: People, Locations, and Organizations. The strategies explored are useing a text normalizer to shape the text into a format that NER programs can recognize and cross checking classifiers to increase the precision of NER tools.**

I. Introduction

Named Entity Recognition is widely used in many different kinds of natural language processing tasks. Named Entity Recognition (NER) is the process of extracting and organizing the names of people, places, and organizations into groups based on commonality [4]. In the area of natural language proccessing many research projects need to identitfy the named entities in order to extract information and relations from texts. Therefore, this process of identifying named entities is necessary for research in the area of natural language proccessing. These named entities can be found by grammar and capitalization patterns but is improved when machine learning is implemented to capture different phrasing of these sentences. For example terms such as "graduated from", "worked at", and "studied at" all suggest that there is a high probability of the prior word being a name [8]. Furthermore, the personal titles such as "Mr.", "Dr.", and "President" suggest that a named entity follows immediately [2].

TABLE I
Example Sentences

President Obama spoke to the troops today.
Mary graduated from Brown.
John vacationed in Spain for the summer.

TABLE II
Classification of Named Entities

Person	Organization	Location
Obama	-	-
Mary	Brown	-
John	-	Spain

Interest in Named Entity Recognition is growing rapidly because of its overwhelming relation to information extraction and to AI strategies. In this project, we attempt to identify and categorize the named entities from Twitter posts.

Twitter posts, also called tweets, have a maximum length of 140 characters, and users often, due to the limited length, forsake grammar and capitalization rules, replacing grammatically correct phrases for slang. Twitter is a huge source of information and therefore it is necessary to discover a way to make these tweets understandable and extract the named entities which are mentioned in these Twitter posts. This will greatly improve the ability of natural language tools to accurately execute the tasks that they are designed to perform. This will allow for better research in the area of natual language processing, especially when in relation to non-normal texts such as blogs, twitter posts, and other texts not written in standard English. Named entity recognition has been used in Natural Language Proccessing (NLP) before, but the focus of these projects has been primarily on texts which are written in standadrd English.

The task of NLP becomes much harder when the text is not written in standard English. Thus it is necessary to have a new system of Named Entity Recognition which takes into account the non-standard laguage used in these colloquial sources.

II. Motivation

Because of the popularity of social networking in today's society, Twitter is quickly becoming the fastest updated source of news in our world today. Not only is it so quickly updated, but the amount of data held within these tweets can be used as a huge resource. With an average of 200 million tweets a day, Twitter often reports the major happenings of the world before the news has the time to broadcast. Because of the immense amount of information contained within these tweets, they, with a little analysis, could be used for many different research projects. However, analysis of this data could be much faster and more accurate if the text was first normalized and the named entities were identified.

The normalization and named entity recognition of Twitter posts could be beneficiall in many different settings. One of the main applications of NER to the real world is information extraction. Because Twitter is such a large source of data with a huge range of topics, it is a great text to draw from in Information Extraction (IE). NER with a basis of

normalization will make information extraction much simpler to accomplish.

III. Twitter

Twitter is a micro-blogging social networking site which is greatly useful for Information Extraction and other such Natural Language Proccesing tasks because it is a huge database for information written by the average person. These tweets are also a source of information for research because they are posted very quickly after the involved events have occurred. However, this massive amount of data is very hard to analyze because of a few differences between Twitter text and the text of standard English.

Tweets are restricted to 140 characters, and because of this restriction, its users need to express their social goings-on in as few words as possible.

Thus words are often mispelled, either accidentally or to shorten length, acronyms are substitued for phrases, and non-grammatical scentence structures are used instead of those that are conventional. This often thwarts the identification or labeling of the words in these Twitter posts.

Twitter has grown from 5,000 tweets per day, in the opening year of this social micro-blogging site in 2007, to a starteling 200 million tweets per day in 2011. It is because of this massive growth that Twitter is becoming too large of source of data to be ignored by the research of the day.

IV. Problem Definition

There are many Named Entity Recognition tools already in existence which are available on the web. So why not use one of those? Normal NER tools are very ineffective when used on Twitter posts. NER tools depend greatly on sentence structure and context to determine named entities. However, tweets are short in nature and tend to be wildly grammatically incorrect. Because of the little context in tweets and the sloppy sentence structure, a normal NER tool performs poorly, and a different approach must be taken.

The purpose of this project is to test the effectiveness of using a normalizer as a preproccessor to a NER tool. And, if time permits, to consolidate both the normalizer and Named Entity Recognition tools in to one single NER super-tool. The use of a normalizer should increase the effectiveness of the NER tool because, although the context is still very low, the grammatical changes will increase the usability of the sentence structure in the recognition process.

This project also intends to increase the effectiveness of a NER tool through the use of cross-referencing classifiers. Multiple classifiers which are trained on different data, such as CRF classifiers, should allow for more accurate results. These classifiers would be able to increase the Recall and Percision because each classifier would find named entities where the other had missedand each would overlook falsely catagorized tokens where one had been mistaken. Thus more named entities would be able to be pulled from a given set of tweets, while less tokens would be falsly recognized

V. Related Research

A. Named Entity Recognition

Up untill now, the majority of the study of Named Entity Recognition has been in relation to documents on the web. Lui et al. implemented a classifier based approach to NER [12]. They used a combination of both K-Nearest Neighbors (KNN) and Conditional Random Fields (CRF) based classifiers. However, Downey et al. used a statistical model to extract these named entities from the web. This approach out-performed a semi-supervised CRF by 73 percent [4]. While still other methods of functional relations were implemented by Hasegawa et al. by tagging named entities and learning the context behind named entities which occur in a similar phrase [7].

There are three significant measurements which are used to evaluate the effectiveness of a NER program.

- Precision (P): Precision is the proportion of the number of correctly identified named entities to the total number of entities identified (the sum of the number of correctly identified and incorrectly identified named entities).

$$P = \frac{N_{correct}}{N_{correct} + N_{incorrect}} \tag{1}$$

- Recall (R): The recall is the proportion of correctly identified named entities to the total number of named entities (the Key count).

$$R = \frac{N_{correct}}{N_{key}} \tag{2}$$

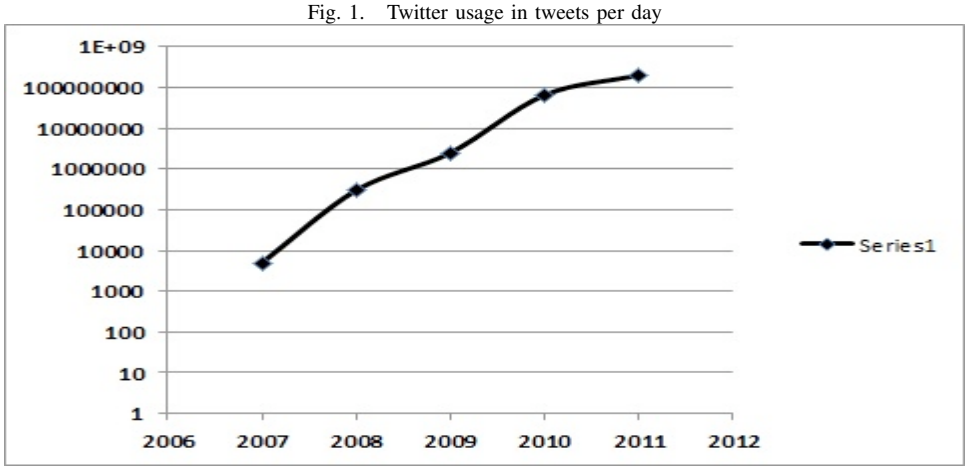
- F-Measure (F): The F-Measure is a measurement involving both R and P as combined in the following equation.

$$F = \frac{2RP}{R + P} \tag{3}$$

TABLE III
Named entity strategies

Strategy	Precision	Recall	F-Measure
CRF on standard texts	91.7	92.0	91.8
CRF on Twitter text	46.3	45.3	45.8
Combined CRF and KNN on Twitter texts	81.6	77.8	80.2

Fig. 1. Twitter usage in tweets per day

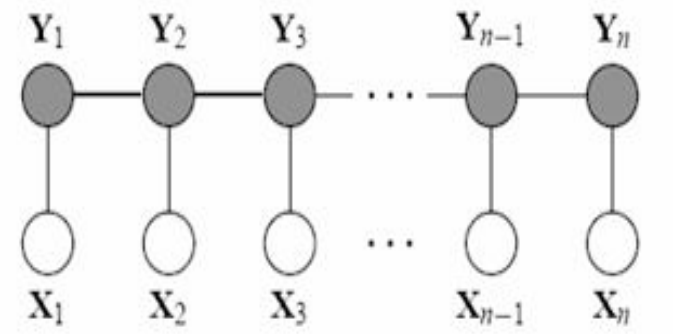


B. Conditional Random Fields

Conditional Random Fields are a type of classifier which involves a probabalistic graphical model and is often used in Natural Language Proccessing. This model is used most often for assigning labels to data. It is implemented in place of Hidden Markov Models.

This classifier uses a large amount of training data to draw from as its knowlage base, and then, based on that training data a probabalistic model is formed. Thus, when text is fed into this Condidtional Random Field model, the probabalistic model created by the training data can be used to determine how a spesific word should be tagged or how a sentence should be parsed.

Fig. 2. Y and X of Conditional Random Fieldsl



A Conditional Random Field is similar to a Markov Random Field contained on a random variable $\mathbf{X}$ which represents the observation sequences. We define $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ to be a undirected graphical model such that there exists a node $\mathbf{n}$ in $\mathbf{V}$ which in turn corresponds to another random variable $\mathbf{Y_n}$ in the set of $\mathbf{Y}$ [**?**].

C. Text Normalization

Because Tweets are often written in colloquial English with shortened and altered words, normal NER tools are very unsucessful at determining named entities. This difficulty is also due to the short nature of these Twitter posts, because they have very little context. Liu et al. states that the aim of text normalization is to substitue meaning-consistent standard English for non-standard tokens [11]. There are several ways in which non-standard tokens are used in Tweets according to Kaufmann and Kalita [10].

- The shortening of words
- Spelling errors
- Repetition for emotion

The meaning of shortened words can be found by consulting tables of text acronyms constructed during the research of Choudhury et al. on the structure of texting language [3]. Spelling errors are checked by looking at near letters for reverse order. While repetition of lettedsr is addressed by reducing to 3 repeated letters and checking against other sources [10].

The process of the normalization tool which is used in this project can be seen in figure 2 [10].

Text Normalization is not necessary for NER when the source is simply documents on the web, but when our source is moved to colloquial language sites such as Twitter, conventional machine learning doesn't perform well [6]. Because of this, text normalization is necessary for successful results.

VI. Approach

In this project, a normalizer which translated the text into standard English was used in partnership with a voting CRF classifyer system. These voting classifiers are classifiers which all come up with results on the same selection of text and then, based on the summed results, one classification is reached.

This type of classification is used in the research of Ekbal and Bandyopadhyay [9]. By using this voting system they were able to gain fantastic results of 92.03 in F-measure when

Fig. 4. The process path of a tweet once it has been introduced to the system

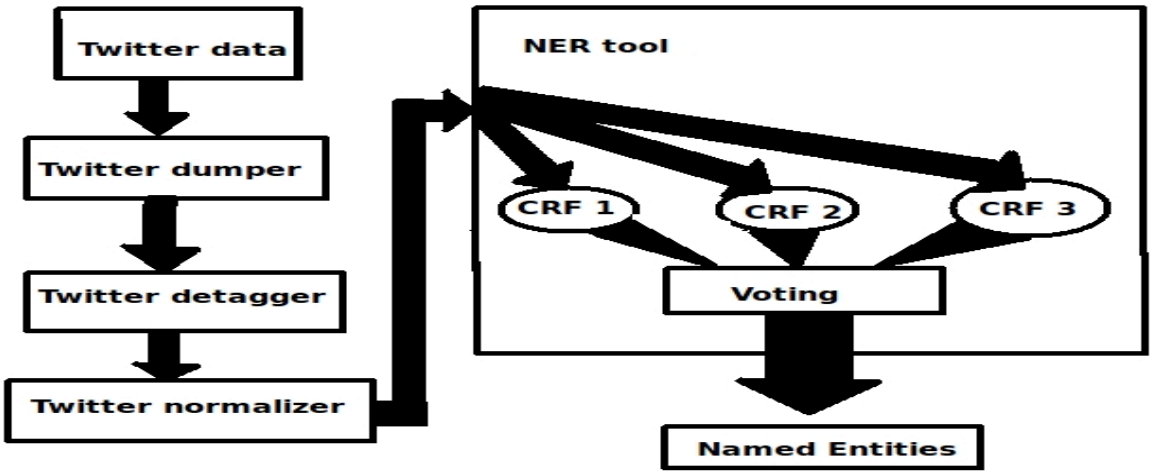
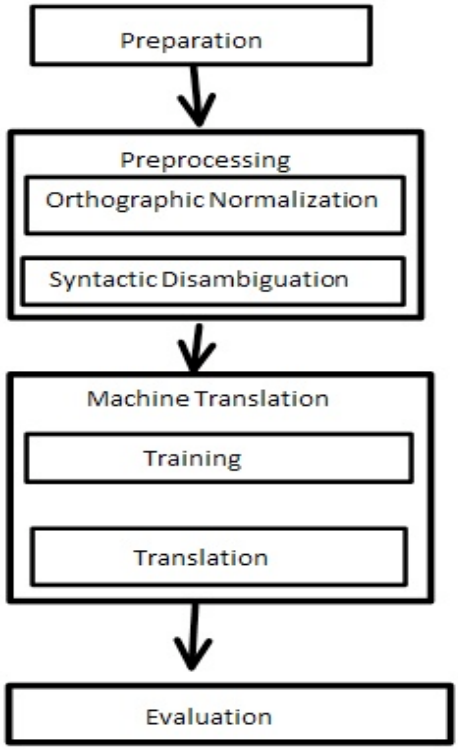


Fig. 3. The process path of the normalization tool



evaluating Named Entity Recognition for Indic laguages. Thus, due to the success of Named Entity Recognition on Twitter posts in our project, we had decided to also experiment with the use of voting classifiers in addition to normalized Twitter posts.

Cross referenced Conditional Random Fields classifiers were used; however, each classifier had been trained on a different set of training data which gave this system more information to draw from. This did increase the effectivness of the given system by .3 percent and will allow for even better results in the future.

During this process a tweet goes through many changes. Table IV shows an example of an actual tweet's changes as it runs through the system.

In previous research by Liu et al CRF and KNN classifiers are combined to create a better NER system, because the f-measure of NER tools drops from 90.8% to 45.8% when used on tweets. In their project they were able to increase the f-measure on tweets to 80.2% when combining the previously mentioned classifiers.

However, The results of this project show that the use of a normalizer with a simple CRF classified NER tool raises the f-measure on tweets to 83.6%. This shows that not only does a normalizer increase the effectiveness of a NER tool, but it can increase it so that it outpreforms a NER tool developed specifically for tweets.

This is a great improvement, and, through the voting tecniques we used with three differently trained Conditional Random Fields classifyers, we were able to increase the F-measure by .3 %. Overall, in this paper we were able to increase the the total F-measure of a NER tool used on twitter posts to 83.9%. This should greatly improve the studies of twitter in the field of Natural Language Proccessing in the future.

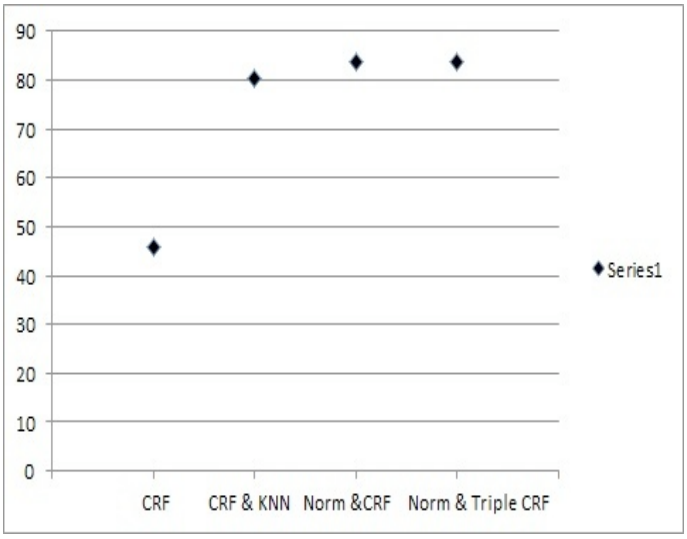
TABLE IV
TWEET NORMALIZATION AND NAMED ENTITY RECOGNITION

Wowwwwwwww!!! U guys R done with all the Harry Potter books? lol I hardly finished one!!!!
wow You guys R done with all the Harry Potter books? lol I hardly finished one!!!!
Wow you guys are done with all the Harry Potter books? i hardly finished one!!!!
Wow you guys are done with all the Harry Potter books? I hardly finished one!
Person: Harry Potter

TABLE V
NAMED ENTITY STRATEGIES INCLUDING THIS STATE OF THE ART SYSTEM

Strategy	Precision	Recall	F-Measure
CRF on standard texts	91.7	92.0	91.8
CRF on Twitter text	46.3	45.3	45.8
Combined CRF and KNN on Twitter texts	81.6	77.8	80.2
Triple Voting CRF on Normalized Twitter texts	84.4	83.4	83.9

Fig. 5. The F-measure of Compared NER Systems Having Only a CRF Classifier, Having Dual Classifiers, Having a Normalizer with a CRF Classifier, and Having a Noramlizer with a Triple Voting CRF Classifier Respectively



VII. IMPROVMENTS

A. This project

In this project we set out to increase the effectiveness of a Named Entity Recognition tool on tweets. In order to do this we first created a program which stripped the raw tweets from the tags that are included with them when they are dumped into a file.

This program considered the common characteristics of these tweets and removed the unnessisary peices, the tags. After this, these tweets, which were then just raw tweet text, were placed in the file which input to the Twitter Normalizer.

After this program was completed it was time to set up the Twitter Normalizer, provided by a previous paper written by Kaufmann and Kalita [10]. It took some work to transfer the program to this system, but soon it was in place. A method was then added which would send the normalized text into the input file for the NER tool and run that code.

After this the tweet file would run through the normalizer three times, each with a different CRF classifier which had been traind on different training data. Finally, the classifiers would vote, and any token on which at least two of the classifiers had aggreed as a named entity was considered as such. All those with one or less votes were disregarded.

B. Future projects

There are many future impovements which would be greatly benificial for Named Entity Recognition as it applies to Twitter. Future Improvements include the following:

- Implementing a CRF KNN voting NER which could be combined with the NER tool, (possibly with the addition of a third classifier which would help in the case of ties.)
- Add into the Normalization tool a checker which would take into account common slang and common shortend words which are not included in the program at present.
- Package all of the system into an exicutable .jar file to allow for easier protability and access.
- Create a GUI from which the entire system can be run allowing for easier use by those who do not have an in depth knowlage of programming.

VIII. CONCLUSION

This project has greatly improved the productivity of NER tools on non-standard text such as posts from the micro-

blogging site of Twitter. The addition of a normalization tool for tweets allowed for these posts to be translated into standard text which made analysis easier. The voting of differently trained classifiers allowed for higher recall and precition, resulting in an overal higher F-measure on this programs ability to recognise named entities in tweets.

Named Entity Recognition is a large part of information extraction and these advances in relation to NER from Twitter posts will help with many future research problems. Being able to determine if a word that is mentioned in a text is a named person, location, or organization is a huge step forward in Information Extraction and will greatly help analysis of any posts on Twitter. This will be useful in future research such as that involving Natural Language Proccessing and Information Extraction.

Because of this improvement many more research oppertunities in this field will be able to be executed in a much more accurate fashion than in previous such projects. Overall we would judge this project to be a success, and hope that it will bring help to many future projects in this field

REFERENCES

[1] M. Banko, M Cafarella, S. Soderland, M. Broadhead, O. Etzioni. 2007. *Open Information Extraction from the Web.* In Procs. of IJCAI 2007.

[2] M. Cafarella, D. Downey, S. Soderland, O. Etzioni. 2005. *KnowItNow: Fast, Scalable Information Extraction from the Web.* In Procs. of the Human Language and Technology Conference 2005.

[3] M.Choudhury, R. Saraf, V. Jain, A. Mukherjee, S. Sarkar, A. Basu. 2007. *Investigation and modeling of the structure of texting language.* Int. J. Dpc. Ama;. REcognit., 10(3):157-174, 2007.

[4] D. Downey, M. Broadhead, O. Etzioni. 2007. *Locating Complex Named Entities in Web Text.* In Procs. of IJCAI 2007.

[5] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. Weld, A. Yates. 2005. *Unsupervised Named-Enitiy Exstraction from the Web: An Expirimental Study.* Artificial Intelligence, 165(1):91134, 2005.

[6] B. Han, T. Baldwin. 2011. *Lexical Normalizations of Short Text Messages: Makn Sens a #twitter.* In Proc. of ACL-HLT 2011.

[7] T. Hasegawa, S. Sekine, R. Grishman. 2004. *Discovering Relations among Named Entities from Large Corpora.* In Proc. of the 42nd Annual Meeting on Association for Computational Linguistics 2004.

[8] H.M. Wallach. 2004. *Conditional Random Fields: An Introduction.* University of Pennsylvania. CIS Technical Report MS-CIS-04-21.

[9] A. Ekbal, S. Bandyopadhyay. 2009. *Voted NER System using Appropriate Unlabeled Data.* In Proc. of the 2009 Named Entities Workshop: Shared Task on Translaiteration (NEWS 2009), ACL-IJCNLP, pp. 210 (2009)

[10] M. Kaufmann, J. Kalita. 2010. *Synatactic Normalization of Twitter Messaqges.* International Conference on Natural Language Proccessing (ICON 2011), Kharagpur, India, December, pp. 149-158.

[11] F. Lui, F. Weng, B. Wang, Y. Lui. 2011. *Insertion, Deletion, or Substitution? Normalizing Text Messages without Pre-categorization nor Supervision.* In the Proc. of The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies

[12] X. Liu, F. Wei, S. Zhang, M. Zhou. 2011 *Recognising Named Entities in Tweets.* Harbin Institute of Technology, Harbin, China.