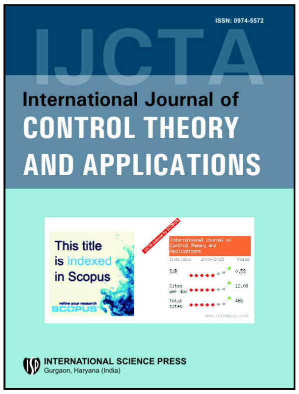

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 19 • 2017

Named Entity Recognition from Social Media Text: A Comparative Study

Ashutosh Bhoi[1] and Rakesh Chandra Balabantaray[1]

[1] Department of Computer Science and Engineering IIIT Bhubaneswar, Bhubaneswar, Odisha, India,
Email: c115001@iiit-bh.ac.in, rakesh@iiit-bh.ac.in

Abstract: Information extraction from social media text is an important application of natural language processing. Identification of named entities (NEs) from microblog contents like twitter is a challenging task due to their noisy and short nature and lack of contextual information. This paper, presents a survey on many NER systems developed by the researchers in this field considering different parameters. Here, the best performing NER systems available for microblog contents are compared along with their results mentioned below. Again a comparative analysis of all the above mentioned models is given which are tested with tweets from Allan Ritter tweet dataset and also with text data from blogs written for American presidential election available online. The precision, recall and F1-measure of all six models on tweets as well as small blogs which are near formal to newswire text are shown below. These twitter NER are also compared with the baseline NER system like Stanford NER and Natural Language Toolkit NER which are giving satisfactory results on newswire text.

Keywords: NER, NLTK, GATE POS tagger, Social media text, Information Extraction

1. INTRODUCTION

Named Entity Recognition (NER) is a part of information extraction (IE) process which takes a textual content as input and generates a named entity list as output. Named entities (NE) are the most significant words in a text document which are used for indexing the document. Proper indexing of named entity terms in a document makes the searching easier. The role of NER is very significant in information extraction tasks like relationships identification, semantics annotation, mining opinions, ontology population etc. Named entities are basically the proper nouns present in a document [8]. If we can properly identify the proper nouns present in a text document then we can easily generate the NE list. Once the NE list is generated, we can classify them into any predefined categories (3-class, 7-class or any number of class) using several machine learning classifiers used by the researchers in the literature. NER from a meticulously authored content like news article is not that much of a concern as several works has already been done by the researchers with high level of accuracy. However it has still been remained as a challenge for the natural language processing (NLP) researchers to get a satisfactory accuracy from social media text like tweets, facebook posts or any microblog content. Most of the contents of social media text is ungrammatical, misspelled and too short in length. NER systems mostly are domain specific and language specific. The performance of most of the NER systems depends fundamentally on gazetteers which make it more demanding in a social media text processing circumstances due to an implicit low coverage [13]. Due to all these kinds of restrictions, the state-of -the-art NER methods are giving unsatisfactory results in that informal and noisy content of social media text [2]. In the early days, NER researchers were using rule based system then they switched to different machine learning based systems. The machine learning algorithms may be based upon supervised, semi-supervised or reinforced learning [11]. Advanced annotation method, proper feature selection enhances the part of speech accuracy which ultimately improves the NER performance [15]. To extract named entity from tweets, two things are needed: either the existing state-of-the-art NER systems must be modified so that they can fit for tweets or these informal texts must be transformed to formal texts so that the state-of-the-art NER systems can be applied to perform the task.

The sequence of the paper is organized as follows: In section-2, the related work is discussed. Extensive analysis of survey work is done in section-3. The results and discussion about the work are briefed in section-4. And finally section-5 describes the conclusion and future work.

2. RELATED WORK

Frederik Hogenboom et al. summarized the techniques for event extraction from text data recognizing between data driven, knowledge driven and conglomerate methods and offered a qualitative assessment of them [1]. Kezban Dilek Onal et al. showed how useful word embeddings can be for NER task on informal text [2]. Sandeep Ashwini et al. developed a dedicated system to identify targetable named entities present in social media text by proper analysis [3]. A. Ritter et al. improved the performance level of NE extraction from tweets significantly by utilizing the in-bound, out-of bound and unlabeled data. They also applied Labeled-LDA using certain rules based on a public domain database for extra supervision to classify named entities in tweets [4]. K. Bontchva et al. introduced twitter specific data import, metadata handling and the requirement for domain adaptation to recognize named entities [5]. J J Jung in his paper proposed how microtext clustering using contextual associations improve the NER of a streaming texts like twitter. Contextual association among microtexts within a microtext cluster is significant information for NER task [6]. Leon Derczynski et al. developed an advanced dataset for Twitter entity disambiguation and experimented an empirical analysis for named entity recognition and disambiguation task. They found that poor capitalization is particularly the main reason for significant drops in NER (also NEL) recall for microblog content text like twitter [7]. The accuracy of NER task depends upon assisted languages, favored textual class and domain and also on preferred entity types [11]. C. Jenny et al. in their work tried to explore whether word representations can also improve (semi-) supervised NER for Spanish. To perform the above task, they used a linear Conditional Random Field (CRF) classifier in which word representation is considered as an additional feature [12]. Many researchers have used different embedding techniques which provide significant improvement especially when there is very less amount of training data [14].

3. EXTENSIVE ANALYSIS

Here, we compared different NER systems with different parameters as mentioned in the table below. The systems can be categorized based on two aspects like the differences in machine learning approaches and algorithms.

3.1. Based on the machine learning approaches used

Basically, three machine learning approaches are used in the literature. The details of each approach are given below.

3.1.1. Supervised approach

This method needs large annotated corpora to train the machine learning based model. Large annotated corpora are available for formal text due to which these methods can be successfully used in such text. Again it is very difficult and tedious task to create such corpora for informal text like tweets. So on tweets supervised methods are not giving satisfactory results due to unavailability of large and proper corpora. For formal text like newswire, biomedical text with the availability of annotated corpora, the data driven methods are the dominating methods. Support vector machine based method is used to classify whether a given word applied in a formal text is a named entity or not [23].

3.1.2. Semi-supervised approach

To overcome the constraints associated with supervised learning, semi-supervised learning methods were evolved. In these methods, we don't require large corpus like supervised methods. Here, dynamic gazetteers are used to solve the problem. A semi-supervised approach along with a combination of KNN classifier and linear conditional random fields has been tried. The KNN based classifier manages the pre-labeling to collect global coarse affirmation across tweets while the CRF model carries out sequential labeling to acquire fine grained information concealed in a tweet [17]. Taggers are adapted to twitter with not-so-distant supervision method with the help of dictionaries and linked websites to guide semi-supervised adaptation of POS and NE taggers to twitter [18]. C. Li et al. developed a two step unsupervised learning model which uses the global context acquired from Microsoft N-gram corpus and Wikipedia to segment tweets with the help of a dynamic programming approach. In the second step, they used the Random walk model to consider the clubbable property to utilize the local context obtainable from tweet stream [19].

3.1.3. Unsupervised approach

In unsupervised learning methods, no training corpora is required which is the biggest reason why this is the most suitable approach for informal text like tweets. Here, everything is taken care by the gazetteers or any dynamic online repository like Wikipedia. Tweet segmentation is used as important features along with both global and local context to improve the accuracy [16].

3.2. Based on the algorithms used

There are many classification algorithms used by researchers in the literature for NER. The most prominently used algorithms are conditional random fields (CRF) and Maximum entropy (MaxEnt) based models which are giving better results than the other models like Hidden Markov Model (HMM).

3.2.1. Conditional Random Field model

CRF is the most widely accepted model for POS tagger [3]. CRF is a prediction model which is used for structured prediction. Linear chain CRF is most widely used in the field of natural language processing. C. Li et al. used BILOU schema features for time aware POI extraction to extract fine grained location from tweets [20].

3.2.2. Maximum Entropy model (MaxEnt)

Subject to explicitly described prior data, Maximum entropy defines the probability distribution that best represents the current state of knowledge, is the one with largest entropy. Unlike Naïve Bayes classifier, MaxEnt doesn't consider the features to be conditionally independent of each other. Out of all the models, maximum entropy model selects the one that has the largest entropy. This model is appropriate in the field of natural language processing (NLP) as the most important features in text are the words and they are not independent of each other [21].

3.2.3. Hidden Markov based Model (HMM)

A statistical Markov model where the states are unobserved (hidden) is stated as Hidden Markov model. In HMM, the states are not explicitly visible rather the output, which are reliant on the states are visible. HMM based POS tagger is extended to NER system which is trained and tested on Hindi and Bengali data to show its effectiveness [22].

Table 1
Comparative Analysis of different NER systems

NER System	*Developer*	*Text*	*Algorithm used*	*Chunker/ Decoder*	*Implementation Language*	*Trained Corpus*	*Supported Languages*	*No. of classes*	*Gazetteer*	*Supervised or not*	*Remarks*
Stanford NER	Stanford NLP Group	Newswire Text	CRF	Gibbs sampling	Java	CONLL 2003 MUC-6 MUC-7 ACE	English German Spanish Chinese	3/4/7	Yes	Semi-supervised	Standard capitalization is the most contributing feature
NLTK Tagger	Team NLTK	Newswire Text	Maximum Entropy	-	Python	ACE	English Dutch Portuguese Spanish	9	Yes	Semi-supervised	POS tagger is the backbone of this NER system
Annie TwitIE	GATE	Tweets	CRF	Viterbi	Java	Ritter Dataset	English	3	Yes	Semi-supervised	Proper language filtering and normalization
Tweet NLP	CMU	Tweets	CRF	-	Java	Tweebank	English	-	Yes (Frequently used tokens)	Semi-supervised	Tweets specific features with gold annotation
Tweeter NLP	Allan Ritter & Sam Clark	Tweets	CRF	T-CHUNK	Java	Ritter Dataset	English	10	Yes	Semi-supervised (Distantly supervision)	Use Out of vocabulary (OOV) to reduce false positive
Twitter NER	Xiaohua Liuet al.	Tweets	KNNCRF	-	-	Self annotated dataset of 12245 tweets	English	4	Yes	Semi-supervised	KNN is used to collect global coarse evidence across tweets
TwiNER	Chenliang Li	Tweets stream (both region & topic based)	Random walk	T-CHUNK	-	Not required	English	No classification	Yes	Unsupervised	No linguistic features are considered here

4. RESULTS AND DISCUSSION

To evaluate the performance of all the previously mentioned NER systems, some blogs and tweets are collected as our dataset. For blogs dataset, seven hundred blogs and for tweets, the Allan Ritter tweet datasets are considered. This Allan Ritter tweet dataset contains 1827 tweets. It is also called as the Tweebank dataset. The blogs were given to some of our students of three groups to give their feedback. Finally, based on that, the precision, recall and F1 measure of all the systems for blogs are calculated.

Table 2
Comparison results of all NER systems

NER Systems	*Tweets*	*Small Blogs*				
	Precision	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
Stanford NER	0.614	0.352	0.442	0.943	0.842	0.889
NLTK Tagger	0.572	0.328	0.417	0.811	0.736	0.772
GATE Annie NER	0.763	0.624	0.686	0.914	0.818	0.863
Twitter NLP by Allan Ritter	0.734	0.608	0.672	0.793	0.682	0.733
Tweet NLP by CMU	0.634	0.463	0.535	0.774	0.672	0.719
TwitterNER by Xiaohua Liu	0.746	0.613	0.672	0.791	0.672	0.726

P: Precision R: Recall F1: F-measure

The result of the experimental study is shown in the above mentioned Table-2. When the precision of all systems are compared, Stanford NER outperforms other NER systems on blogs. However, the precision of Stanford NER and GATE tagger are close to each other on blogs, where as GATE tagger still outperforms the Stanford NER. The recall score of GATE NER is far ahead of NLTK tagger and also that of Stanford NER.

In case of F1-measure, there is a trivial improvement in results for blogs with GATE NER when compared with other approaches. However, there is a significant improvement in NER from tweets with GATE NER as compared to NLTK tagger [9] based approach and Stanford NER [10]. For tweets, GATE NER gives the best results which use Stanford POS tagger i.e. trained on Penn Treebank tweet set.

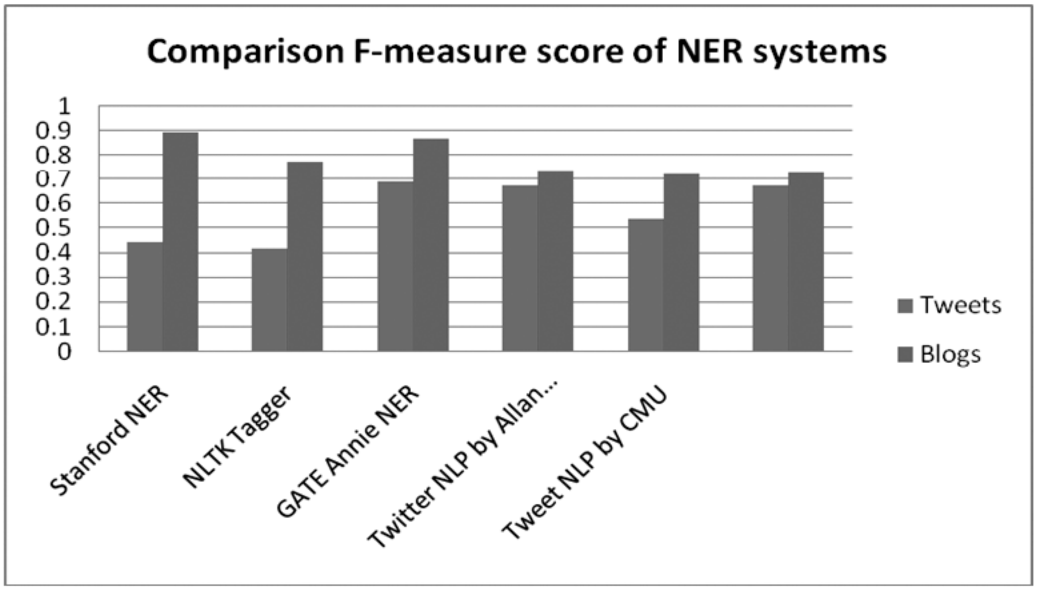

Figure 1. F-measure comparison of all NER systems

5. CONCLUSION

NER from social media text is a very tedious task due to its informal and short nature. The NLP researchers have got significant results in NER from newswire text where as a lot of work needed to get some satisfactory results in informal text like tweets or any microblog contents. Stanford NER system which gives the best results for newswire text, fails to respond effectively for tweets. The twitter based NER systems perform well for tweets, however still an ample scope of improvement is there to reduce the gap in results between formal and informal text. The best results are achieved when NER is extracted from tweets in case of GATE named entity recognizer in comparison with the other systems. There is a significant improvement in terms of recall and also in precision and F1-measure for NER from tweets with GATE NER system. The objective of checking with small blogs is that how the successful twitter NER systems perform on near formal blog text. Most interestingly, it is found that GATE twitter NER system also gives very good results on blogs because of the usage of customized version of Stanford POS tagger in the pipeline of GATE NER system.

REFERENCES

[1] F. Hogenboom , F. Frasincar, U. Kaymak , F. Jong , E. Caron "A Survey of event extraction methods from text for decision support systems," Decision Support Systems, vol. 85, pp. 12-22, May 2016.

[2] K. D. Onal, P. Karagoz "Named Entity Recognition from Scratch on Social Media," Proceedings of the 6th International Workshop on Mining Ubiquitous and Social Environments (MUSE 2015) co-located with the 26th European Conference on Machine Learning / 19th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2015), Porto, Portugal, Sep., 2015.

[3] S. Ashwini, J. D. Choi, "Targetable Named Entity Recognition in Social Media," arXiv:1408.0782[cs.CL], Aug. 2014.

[4] A. Ritter, S. Clark, Mousam, O. Etzioni, "Named Entity Recognition in Tweets: An Experimental Study," Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524-1534, Edinburg, Scotland, UK, Jul., 2011.

[5] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, N. Aswani,"TwitIE: An Open-source Information Extraction Pipeline for Microblog Text," Proceedings of Recent Advances in Natural Language Processing, pp. 83-90, Hissar, Bulgaria, September,2013.

[6] J. J. Jung, "Online named entity recognition method for microtexts in social networking services: A case study of twitter," Expert Systems with Applications, vol. 39, pp. 8066-8070, 2012.

[7] L. Derczynski, D. Maynard, G. Rizzo, M. Erp, G. Gorrel, R. Troncy, J. Petrak, K. Bontcheva, "Analysis of named entity recognition and linking for tweets," Information Processing and Management, vol. 51, pp. 32-49, 2015.

[8] M. Marrero, J. Urbano, S. S. Cuadrado, J. Morato, J. Miguel, G. Berbis, "Named Entity Recognition: Fallacies, challenges and opportunities," Computer Standards & Interfaces, vol.35, pp. 482-489, 2013.

[9] S. Bird, E. Loper , E. Klein, "Natural Language Processing with Python," O'Reilly Media Inc., 2009.

[10] J. R. Finkel, T. Grenager, C. Manning, " Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling," *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics,* pp. 363-370, 2005. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

[11] N. David , S. Satoshi, "A Survey on Named Entity Recognition and Classification," Lingvisticae Investigationes, vol.30, issue.1, pp.3-26, 2007.

[12] C. Jenny, J. Ochoa, C. Thorne, G. Glavas, "Spanish NER with Word Representations and Conditional Random Fields," Proceedings of the sixth Named Entity Workshop joint with 54th *Association for Computational Linguistics*, pp.34-40, Germany, 2016.

[13] A. Zirikly, M. Diab, "Named Entity Recognition for Arabic Social Media," Proceedings of NAACL-HLT, pp.176-185,Colorado, 2015.

[14] N. Peng, M. Dredze, "Named Entity Recognition for Chinese Social Media with Jointly Trained Embeddings," Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.548-554, Portugal, 2015.

[15] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, N. A. Smith," 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, vol. 2,*pp. 42-47,Jun.* 2011.

[16] C. Li, A. Sun, J. Weng, and Q. He,"Tweet Segmentation and its Application to Named Entity Recognition," IEEE Transactions on Knowledge and Data Engineering, vol. 27, pp. 558-570, Feb. 2015.

[17] X. Liu, F. Wei, S. Zhang, and M. Zhou, "Named Entity Recognition for Tweets,"ACM Transactions on Intelligent Systems and Technology, vol. 4, pp. 1-15, Jan. 2013.

[18] B. Plank, D. Hovy, R. McDonald, and A. Søgaard, "Adapting taggers to Twitter with not-so-distant supervision," In Proceedings of COLING 25th International Conference on Computational Linguistics: Technical Papers, pp. 1783-1982, Dublin, Ireland, Aug. 2014.

[19] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B. Lee, "TwiNER: Named Entity Recognition in Targeted Twitter Stream," Special Interest Group on Information Retrieval, pp. 721-730, Aug. 2012.

[20] C. Li, A. Sun, "Fine-Grained Location Extraction from Tweets with Temporal Awareness", Proceedings of the 37th international ACM SIGIR conference on Research & Development in Information Retrieval, pp. 43-52, Gold Coast, Queensland, Australia, July, 2014.

[21] O. Bender, F. J. Och, H. Ney, "Maximum Entropy Models for Named Entity Recognition," CONLL '03 Proceedings of the seventh conference on Natural language learning at HLT-NAACL, vol. 4, pp. 148-151, Edmonton, USA, 2003.

[22] A. Ekbal, S. Bandopadhyay, "A Hidden Markov Model Based Named Entity Recognition System: Bengali and Hindi as Case Studies," 2nd International Conference Pattern Recognition and Machine Intelligence, pp. 545-552, Dec. 2007.

[23] A. Mansouri, L. S. Affendey, A. Mamat, "Named Entity Recognition Approaches," International Journal of Computer Science and Network Security, pp. 339-344, vol. 8, Feb. 2008.