A Feature Based Simple Machine Learning Approach with Word Embeddings to Named Entity Recognition on Tweets

Mete Taşpınar[1], Murat Can Ganiz[2], and Tankut Acarman[1]

[1] Department of Computer Engineering, Galatasaray University, Istanbul, Turkey
mtaspinar@ogr.gsu.edu.tr, tacarman@gsu.edu.tr
[2] Department of Computer Engineering, Marmara University, Istanbul, Turkey
murat.ganiz@marmara.edu.tr

Abstract. Named Entity Recognition (NER) is a well-studied domain in Natural Language Processing. Traditional NER systems, such as Stanford NER system, achieve high performance with formal and grammatically well-structured texts. However, when these systems are applied to informal and noisy texts, which have mixed language with emoticons or abbreviations, there is a significant degradation in results. We attempt to fill this gap by developing a NER system with using novel term features including Word2vec based features and machine learning based classifier. We describe the features and Word2Vec implementation used in our solution and report the results obtained by our system. The system is quite efficient and scalable in terms of classification time complexity and shows promising results which can be potentially improved with larger training sets or with the use of semi-supervised classifiers.

Keywords: Named Entity Recognition, Word2Vec, Word Embeddings, Classification, Machine Learning

1 Introduction

Extracting meaningful information from social media platforms becomes more important with ever increasing amount of available data. There are several challenges in mining microblog texts such as tweets. The enormous amount of noisy data involving abbreviations, typing errors, and special characters to indicate special terms such as hashtags or mentions makes extracting operations difficult. Due to these challenges, existing NER systems [2, 3, 4, 5, 6, 8, 11, 13, 14] usually do not perform well on these domains. In this study, we present a simple yet effective machine learning based classifier using some novel features including word embeddings based features for identifying different classes of named entities in tweets. In order to evaluate our approach, we use the NEEL dataset [1] for our experimental study and we evaluate our results with respect to the studies [2, 3, 4] published in NEEL 2016 Challenge. We conduct several experiments with different subsets of the features. We illustrate that the addition of word embedding features considerably increases the accuracy, and a

simple machine learning classifier with word embedding features can compete with more complicated methods such as Conditional Random Fields (CRF). The impact of this study and usability of the results is crucial for several reasons. Firstly, more advanced classifiers; possibly ensemble learning and semi-supervised approaches can be applied to improve the achieved performance. Secondly, use of machine learning based classification algorithms such as Support Vector Machines (SVM) and word embeddings algorithms such as Word2Vec [18, 19] will allow to develop highly scalable and distributed models versus the traditional models used in this domain such as CRFs. They scale well especially with the increasing amount of training set.

2 Related Work

A special Twitter implementation of GATE Natural Language Processing (NLP) framework abbreviated as TwitIE is presented in [2]. GATE NLP is based on Stanford NER classifier and uses CRF model. In [3], a feature based approach performing Stanford NER is described and ARK is used for part-of-speech (POS) tagging. Several features such as length of the mention and when the mention is capitalized are trained with the supervised classifiers such as Random Forest, Naive Bayes, k-nearest neighbour and Support Vector Machines (SVM). The standard NER system implementations such as Stanford NER, MITIE, twitter_nlp and TwitIE are studied in [4]. The dataset is trained for MITIE.

For Turkish tweets the NER software of European Media Monitor (EMM) is used in [5] and [6]. 3 different Turkish datasets including tweets, a speech-to-text interface and the data taken from a Turkish hardware forum are applied to a machine learning algorithm named CRFs [6, 7, 8]. For Spanish formal documents, a rule based approach is applied in [9] and [10] an unsupervised feature generation is shown to improve the stand-alone performance of the process NER. In [11], the Stanford NER system performance metric F1 is improved by 25% by re-building the part-of-speech, and chunking jobs.

3 Implementation and Feature Set

We are particularly focused on NER to present a simple, feature based machine learning approach with additional word embedding features for identifying different classes of named entities in tweets. The features used in this study are as follows; StartCapital (whether the term is capitalized or not), AllCapital (If the term is all uppercase), Hashtag (If the term starts with the letter '#'), Mention (If the term starts with the letter '@'), POS (Part-of-Speech Tag of the term), Length (number of characters in the word), VowelRatio (The ratio of number consonant over the number of vowels in the word) and SimClassCentroid[i] (Cosine Similarity between term's Word2Vec vector to the centroid Word2Vec vector of class i).

We employ a 3-term wide sliding window approach in extracting features. We use Stanford POS Tagger with 36-tag tagset. Additionally, we use Word2Vec algorithm

to create vector space representations of each term in the training set. Word2vec is trained by a fairly large corpus, consist of 400 million tweets [15]. We compute class centroid vectors by averaging the term vectors belonging to a particular named entity class. For each term in the dataset, cosine similarity to each class centroid is calculated. These are used as additional features.

4 Experimental Study

For training and testing purposes, we use the NEEL 2016 twitter dataset provided by [1]. We use Python programming language with gensim library [16] for executing word2vec as an underlying word embeddings algorithm and scikit-learn library [17] for training and testing classifiers. We use the following off-the-shelf supervised classification algorithms from scikit-learn toolkit in our experimental study: Logistic Regression, Support Vector Machine, KNeighborsClassifier (k-NN), MultinomialNB, BernoulliNB, ExtraTreeClassifier and DecisionTreeClassifier.

We conduct several experiments using different subsets of the feature set and the entity types. The dataset is annotated with seven entity types: Person, Thing, Organization, Location, Product, Event and Character. The 5 features that we extracted are StartCapital, AllCapital, Hashtag, Mention, POS (Part-of-Speech) Tag, the other features that we try did not generate good scores. We have also 7 additional word2vec based features, one for each named entity class. Evaluation results with 7 NER types and different subsets of features are given through Table 1 to Table 3. In Table 1, 5 features and 7 NER types are used. A precision of 0.55 is reached and F1 is reached at level 0.49 when we use ExtraTreeClassifier algorithm. In table 2, 7 word2vec features and 7 NER types are used and KNeighboursClassifier reaches at 0.70 precision and 0.57 F1. Combination of 5 features and 7 word2vec features slightly increases precision at 0.71 and F1 at 0.58 using Logistic Regression algorithm.

Table 1. Experiment Results with 5 features and 7 NER Types.

	Precision	**Recall**	**F1**	**F1 (Micro Avg)**
Logistic Regression	0.27	0.25	0.24	0.25
SVM	0.52	0.44	0.48	0.44
Extra Tree Classifier	**0.55**	**0.46**	**0.49**	**0.46**

We also evaluate both of 5 features and 7 word2vec features with an additional class of 'NoType', which means that a term is not a named entity. Due to the nature of natural language, an overwhelming majority of the terms in tweets are not named entities. This leads to a highly-skewed class distribution where NoType class dominates with 86%. On this dataset, two models, ExtreTreeClassifier with 5 features and Logistic Regression with 5 features + 7 word2vec features can reach only 0.88 F1.

Table 2. Experiment Results with 7 word2vec features and 7 NER Types.

	Precision	**Recall**	**F1**	**F1 (Micro Avg)**
Logistic Regression	0.65	0.55	0.52	0.55
SVM	0.66	0.58	0.55	0.58
k-NN	**0.70**	**0.58**	**0.57**	**0.58**
ExtraTreeClassifier	0.58	0.52	0.49	0.52

Table 3. Experiment Results with 5 features + 7 word2vec features and 7 NER Types.

	Precision	**Recall**	**F1**	**F1 (Micro Avg)**
Logistic Regression	**0.71**	**0.56**	**0.58**	**0.56**
SVC (Support Vector Classifier)	0.63	0.56	0.58	0.56
KNeighborsClassifier	0.56	0.50	0.50	0.50
ExtraTreeClassifier()	0.54	0.51	0.49	0.51

In order to get a detailed look of the results of our best performing model (Logistic Regression, 5 features + 7 word2vec features, 7 NER classes), we provide confusion matrix and class based evaluation metrics. As given in Table 4, majority of the instances belong to Person and Product class (238 entities for each. We can see that "Organization" is the most confused named entity class by our models. We see from the first column that majority of the "Organization" entities (57 out of 122) are misclassified as "Person". Overall, the majority of the misclassifications are accumulated at the first column.

We compare our results with the three studies in the NEEL 2016 workshop [1] in Table 5. Our results are higher in comparison with respect to the algorithms using TwitIE, Stanford NER, MITIE and twitter_nlp in [2] and [4]. And precision is reached at the same level in [3] using feature-based approach.

Table 4. Confusion Matrix of Logistic Regression, 5 features + 7 word2vec features, 7 NER classes.

NER Type	Person	Thing	Organization	Location	Product	Event	Character
Person	**209**	17	4	0	2	6	0
Thing	0	**21**	1	0	1	6	0
Org.	57	12	**37**	1	4	11	1
Location	7	2	0	**15**	0	1	0
Product	4	3	8	0	**94**	129	0
Event	0	3	2	0	0	**11**	0
Character	15	1	7	0	3	1	**0**

Table 5. Comparison of the performance with respect to the studies presented in NEEL 2016 workshop [1].

Study	Precision	Recall	F1
A feature based approach performing Stanford NER, [3]	0.729	0.626	0.674
Stanford NER, MITIE, twitter_nlp and TwitIE, [4]	0.587	0.287	0.386
TwitIE (CRF Model), [2]	0.435	0.459	0.447
Our approach (Logistic Regression, 5 features + 7 word2vec features, 7 NER classes)	0.71	0.56	0.58

5 Conclusions

Our approach is based on extracting Tweet specific syntactic features along with word embeddings, in particular Word2Vec [18, 19] based semantic features and using them in machine learning based classifiers. Experimental results show that our system can outperform two of the three studies in the NEEL 2016 workshop [1] (please see Table 5) in terms of F1 metric and present close precision performance level (0.71 versus 0.729) while comparing with respect to the best performing study (see for instance, [3]), although the training set size is quite limited. One important advantage of our approach is the low training time complexity and scalability. In the future work, we are planning to use more advanced classifiers; possibly ensemble learning and semi-supervised approaches to improve classification performance.

Acknowledgements. The co-authors Mete Taşpınar and Murat Can Ganiz would like to thank Buğse Erdoğan and Fahriye Gün from Marmara University @BIGDaTA_Lab for their help. This work is supported in part by Marmara University BAP D type project.

References

1. G. Rizzo, M. van Erp, J. Plu, and R. Troncy. Making Sense of Microposts (#Microposts2016) Named Entity rEcognition and Linking (NEEL) Challenge. In 6th Workshop on Making Sense of Microposts (#Microposts2016), pages 50–59, 2016.
2. P. Torres-Tramon, H. Hromic, B. Walsh, B. Heravi, and C. Hayes. Kanopy4Tweets: Entity Extraction and Linking for Twitter. In 6 th International Workshop on Making Sense of Microposts (#Microposts), 2016.
3. S. Ghosh, P. Maitra, and D. Das. Feature Based Approach to Named Entity Recognition and Linking for Tweets. In 6 th International Workshop on Making Sense of Microposts (#Microposts), 2016.
4. K. Greenfield, R. Caceres, M. Coury, K. Geyer, Y. Gwon, J. Matterer, A. Mensch, C. Sahin, and O. Simek. A Reverse Approach to Named Entity Extraction and Linking in Microposts. In 6 th International Workshop on Making Sense of Microposts, 2016.
5. Dilek Kucuk, Guillaume Jacquet, and Ralf Steinberger. 2014. Named Entity Recognition on Turkish Tweets. In Proceedings of the Language Resources and Evaluation Conference.
6. Gokhan Celikkaya, Dilara Torunoglu, and Gulsen Eryigit. 2013. Named Entity Recognition on Real Data: A Preliminary Investigation for Turkish. In Proceedings of the 7th International Conference on Application of Information and Communication Technologies.
7. G. A. Şeker and G. Eryiğit, Initial explorations on using CRFs for Turkish Named Entity Recognition., presented at the In Proceedings of the 24th International Conference on Computational Linguistics, COLING 2012, Mumbai, India, 2012.
8. Beyza Eken and Ahmet Cüneyd Tantug. Recognizing named entities in Turkish tweets. In Proceedings of the Fourth International Conference on Software Engineering and Applications, Dubai, UAE, January 2015.
9. Moreno, I., P. Moreda, M. T. Rom´a-Ferri. 2015. MaNER: a MedicAl Named Entity Recogniser for Spanish. En Proceedings of the 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015.
10. Moreno, I., P. Moreda, M. T. Rom´a-Ferri. 2016. An Active Ingredients Entity Recogniser System Based on Profiles. Proceedings of 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016.
11. Riiter, A., Clark, S., Etzioni M., Etzioni, O. Named Entity Recognition in Tweets: An Experimental Study. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011 Edinburgh, Scotland, UK, July 27–31, 2011.
12. Siencnik., S. Adapting word2vec to Named Entity Recognition. Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015.
13. Dilek Kucuk and Ralf Steinberger. 2014. Experiments to Improve Named Entity Recognition on Turkish Tweets. Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM) @ EACL 2014, pages 71–78,Gothenburg, Sweden, April 26-30 2014.
14. Ek, T., Kirkegaard, C., Jonsson, H., Nugues, P.: Named entity recognition for short text messages. Procedia-Social and Behavioral Sciences 27, 178–187 (2011)
15. F Godin, B Vandersmissen, W De Neve, R Van de Walle. Named entity recognition for twitter microposts using distributed word representations. ACL-IJCNLP 2015, 146-153
16. https://pypi.python.org/pypi/gensim
17. http://scikit-learn.org/stable/index.html
18. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR Workshop*, 2013.
19. Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of NAACL HLT*, 2013.