Named Entity Recognition in 140 Characters or Less*

Kelly Geyer, Kara Greenfield, Alyssa Mensch, Olga Simek

MIT Lincoln Laboratory, 244 Wood St, Lexington MA, United States
{kelly.geyer, kara.greenfield, alyssa.mensch, osimek}@ll.mit.edu

ABSTRACT
In this paper, we explore the problem of recognizing named entities in microposts, a genre with notoriously little context surrounding each named entity and inconsistent use of grammar, punctuation, capitalization, and spelling conventions by authors. In spite of the challenges associated with information extraction from microposts, it remains an increasingly important genre. This paper presents the MIT Information Extraction Toolkit (MITIE) and explores its adaptability to the micropost genre.

CCS Concepts
•**Information systems → Information extraction**

Keywords
Named entity recognition, re-training, social media, Twitter

1. INTRODUCTION
Named entity recognition (NER) is a subtask of information retrieval concerned with the automatic extraction of named mentions of entities, where the set of possible entity types originally consisted of people, organizations, and locations. Even the original work on NER from MUC-6 recognized the need for systems to be able to extract varying sets of named entity types from varying genres [1]. Since then, NER has been used to extract diverse entity types (such as diseases and products) from diverse genres (such as speech transcripts and microposts). Practitioners of NER in such diverse domains have been forced to accept that the systems must be re-trained on in-domain data in order to obtain optimal results. The ability to retrain systems has enabled the success of NER, but knowing the quantity on in-domain training data that is required is often more of an art than a science. In this paper, we examine the requirements for successfully retraining an NER system to extract an expanded set of entities from the micropost genre, a notoriously hard genre for NER [2].

*This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2. MITIE
The MIT Information Extraction Toolkit (MITIE) [3] is a free, open-source software library of state-of-the-art NLP tools developed at MIT Lincoln Laboratory. MITIE enables the automated extraction of named entities and of binary relations (for example, a person's place of birth) from unstructured text in English and Spanish. MITIE utilizes distributional word embeddings [4] to reduce dimensionality and improve performance, Conditional Random Fields and structured support vector machines for learning syntactic relationships [5], and automated hyperparameter optimization to facilitate user customization. MITIE is built on the high-performance Dlib machine learning library [6] [7], includes interfaces to C, C++, Java, R, and Python, and is easy to train on new data sets [8], such as microposts.

One of the goals in developing MITIE was to enable fast named entity recognition. To this end, MITIE is capable of processing 53,600 words per second when run single-threaded on a 2.4GHz Intel Xeon processor. Even with this speed, accuracy was not compromised and MITIE is able to achieve an F1 score of 88.1 on the CoNLL 203 NER task [3] [9]

3. RETRAINING MITIE FOR MICROPOSTS
We utilized the training data from the NEEL 2016 Challenge Data Set [10] for our experiments. This corpus consists of 5991 tweets which have been annotated for named entity mentions of types: person, organization, location, event, product, character, and thing.

Our experiments consisted of varying the number of training documents and testing on the remainder of the documents, utilizing 5-fold cross validation. No out of domain training data was used to supplement the in-domain data. Each document corresponded to a single tweet. The documents were not guaranteed to contain any mentions of named entities. For each experiment, we trained a single MITIE model to simultaneously classify all seven of the entity types under consideration.

4. Results
Across all of the entity types other than character and thing, training with in-domain data began to show diminishing (but still positive) returns with 500 training documents. This was seen in measuring F1 and precision and recall independently, as shown in Figures 1, 2, and 3. Also of note was that increasing the number of in-domain training documents benefited performance in precision significantly more than recall for all entity types.

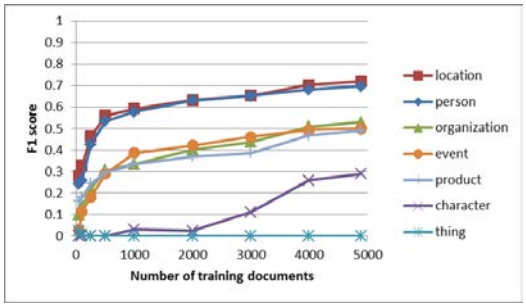

Figure 1: F1 Score

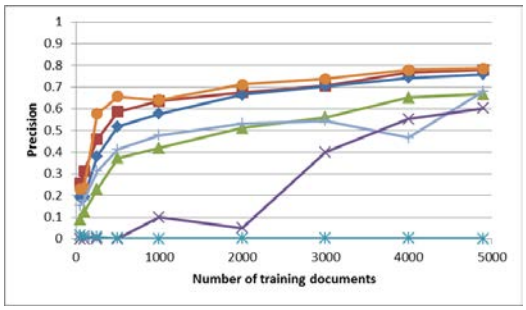

Figure 2: Precision

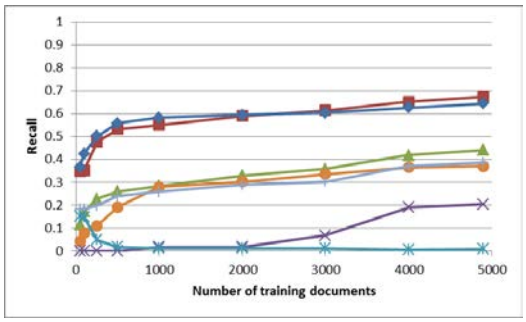

Figure 3: Recall

4.1 Comparison Between Entity Types

We considered the hypothesis that the difference in performance in correctly recognizing different types of entity mentions was due to the number of times that that entity type appeared in the training data. This hypothesis however, proved to be false. Of particular interest is the performance in recognizing event mentions. Despite the fact that this was a particularly rare entity in this corpus, MITIE excelled at recognizing event mentions, particularly with regard to precision.

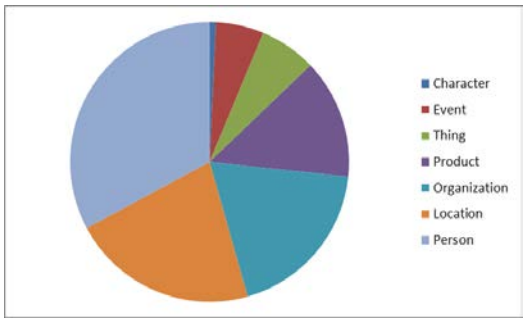

Figure 4: Entity Type Distribution

While not a sufficient condition, there is a threshold quantity of mentions of a given entity type which is necessary for NER accuracy to be significantly above chance performance. As seen in Figure 4, the character entity type is extremely rare and correspondingly begins showing large performance gains beginning with 3000 training documents.

Also of note was the consistently poor performance in recognizing mentions of thing entities. We hypothesize that this is due to thing being a poorly defined entity type, but have not yet tested that hypothesis.

5. CONCLUSIONS

In this paper, we presented exploratory analysis comparing the number of in-domain training documents used with named entity recognition performance in the micropost genre. In this analysis, we also compared performance on recognizing different entity types. Additionally, we presented the MIT Information Extraction Toolkit, an open-source structural SVM approach to named entity recognition and binary relation extraction.

6. FUTURE WORK

In future work we would like to explore other dimensions of retraining NER systems. Some particular questions of interest are examining whether the patterns seen in the number of in-domain training micropost documents required are mirrored in other domains and identifying a causal factor that explains the varying performance in recognizing entities of different types.

7. ACKNOWLEDGMENTS

We would like to thank Davis King, Arjun Majumdar, and Michael Yee for their work on developing MITIE.

8. REFERENCES

[1] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A Brief History," *COLING*, vol. 96, pp. 466-471, 1996.

[2] A. Ritter, S. Clark and O. Etzioni, "Named Entity Recognition in Tweets: An Experimental Study," in *EMNLP '11*, 2011.

[3] D. King, "MITLL/MITIE," [Online]. Available: https://github.com/mit-nlp/MITIE.

[4] P. Dhillon, D. Foster and L. Ungar, "Eigenwords: Spectral Word Embeddings," *Journal of Machine Learning Research (JMLR)*, vol. 16, 2015.

[5] T. Joachims, T. Finley and C.-N. Yu, "Cutting-Plane Training of Structural SVMs," *Machine Learning*, vol. 77, no. 1, pp. 27-59, 2009.

[6] D. King, "davisking/dlib," [Online]. Available: https://github.com/davisking/dlib.

[7] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755-1758, 2009.

[8] S. Haleen and A. Halterman, "mitie-trainer," [Online]. Available: https://github.com/Sotera/mitie-trainer.

[9] T. K. Sang, E. F. and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *The seventh conference on Natural language learning at HLT-NAACL*, 2003.

[10] R. e. al., "NEEL Challenge Data Set," 2016.