NER in Tweets Using Bagging and a Small Crowdsourced Dataset

Hege Fromreide and Anders Søgaard

Center for Language Technology, University of Copenhagen

Abstract. Named entity recognition (NER) systems for Twitter are very sensitive to cross-sample variation, and the performance of off-the-shelf systems vary from reasonable (F_1: 60–70%) to completely useless (F_1: 40–50%) across available Twitter datasets. This paper introduces a semi-supervised wrapper method for robust learning of sequential problems with many negative examples, such as NER, and shows that using a simple conditional random fields (CRF) model and a small crowdsourced dataset [4], leads to good NER performance across datasets.

Keywords: Twitter, semi-supervised learning, bagging, crowdsourcing, named entity recognition, unlabeled data.

1 Introduction

Supervised named entity recognition (NER) is the task of learning to identify and classify names of people, companies, locations, products, etc., in text from manually annotated data. Supervised NER systems are useful in information extraction (IE), but performance is very domain-dependent [11]. Standard datasets like CoNLL 2003,[1] MUC-7[2] and ACE 2004[3] are annotated news corpora, and models induced from such corpora have not proven successful for NER in social media like Twitter [12]. To illustrate the drop in performance from news to Twitter, we train a CRF model on the CoNLL 2003 training data and evaluate it on the (in-domain) CoNLL 2003 test data, as well as (out-of-domain) manually annotated Twitter data. Named entities are detected and labeled as either location (LOC), organization (ORG) or person (PER). While the model has close to state-of-the-art performance on in-domain data (average F_1 across LOC, ORG and PER: 90.1%), it performs much worse when evaluated on an out-of-domain Twitter dataset annotated for the purpose of this paper (53.7%). This huge drop in performance is obviously prohibitive for down-stream IE in Twitter. The system proposed in Ritter et al. [12], which is an attempt to adapt NER to Twitter using manually annotated tweets, does not improve over our supervised baseline. On the same data, their system obtains a similar result (see Table 1 below).

[1] http://www.cnts.ua.ac.be/conll2003/ner/

[2] LDC2001T02.

[3] LDC2005T09.

A. Przepiórkowski and M. Ogrodniczuk (Eds.): PolTAL 2014, LNAI 8686, pp. 45–51, 2014.

The main reason for the drop from news to Twitter is a change in topics and linguistic conventions [12]. Eisenstein [3] shows that topics and linguistic conventions on Twitter change *very* rapidly. This may explain the relatively poor performance of the system proposed by Ritter et al. [12] on our data, which is sampled differently than their training data. Language drift reduce the utility of a few months old training data from Twitter when applied to tweets sampled differently. In other words, evaluation of NER for Twitter on held-out data from the same sample of tweets may be very misleading.

Our contributions in this paper are as follows:

– We are, to the best of our knowledge, the first to consider using only crowdsourced data, available at larger volumes, and labeled data from the newswire domain to learn named entity taggers for Twitter, but we are, nevertheless, still able to outperform state-of-the-art supervised taggers,
– we evaluate a wide range of combinations of semi-supervised wrapper methods across several datasets,
– and finally, we introduce two new sizeable evaluation datasets for Twitter NER.

2 Our Approach

The standard baseline model in NER is a linear CRF [7, 14]. This model is similar to structured perceptron [2], but linear CRF minimizes a logistic loss function and provides probability estimates, making re-ranking and semi-supervised learning with confidence thresholds possible. The linear CRF is induced from sequences of words (sentences) labeled manually with symbols indicating whether words are named entities or not. Since this manually labeled data is costly to produce, as it typically requires trained linguists (however, see [4] and Rodrigues et al. [13]), several authors have proposed using semi-supervised learning algorithms for NER model induction [15, 16, 8]. The algorithm presented here is a combination of well-known techniques, but we are the first to show that a semi-supervised approach to NER can make expert annotations of in-domain data superfluous — or at least that systems induced from crowdsourced in-domain data (and a little bit of out-of-domain labeled data) in some cases can outperform state-of-the-art supervised systems.

Our approach is sketched in Fig. 1. We begin by creating five bootstrap samples from the concatenation of the crowdsourced Twitter data and our labeled newswire data. The Twitter data (T) is resampled with replacement, and each sample has the same size (N) as the original dataset. To each sample, we add a copy of the high-quality newswire data (T') that is never altered through the semi-supervised procedure (Fig. 1, line 8). From each bootstrap sample, we learn a linear CRF model (line 9). These five models now form a product-of-experts model. In each iteration of semi-supervised learning, we use this ensemble model to label the unlabeled data (line 12). In each iteration, we add unlabeled data points with predicted labels to our labeled data. The parameter N' denotes the number of unlabeled tweets to be added in each iteration. To prevent our model from becoming too conservative, we balance the unlabeled data by removing low-confidence negative predictions. The parameter M decides how many low-confidence negative predictions to be removed from the new labeled data. After the semi-supervised procedure, we return the final product-of-experts model (line 16).

```
 1: $T, T'$ labeled training data, $T_i = \emptyset$
 2: $C$ crowdsourced data
 3: $U$ unlabeled data
 4: $S$ evaluation data
 5: $\sigma(N, \cdot)$ bootstrap sample $N$ datapoints with replacement
 6: for $iter \in I$ do
 7:     for $i \in [1...5]$ do
 8:         $T_i \leftarrow \sigma(|C|, T \oplus C) \oplus T'$
 9:         $\mathbf{w}_i^* = \Sigma_{i=1}^{n} \log p(y_i \mid \mathbf{x}_i; \mathbf{w}) - \frac{\lambda}{2}||\mathbf{w}||^2$
10:     end for
11:     $U_{iter} \leftarrow \sigma(N', U)$
12:     $LU_{iter} = \{\langle \arg\max_y \prod_i^5 \Sigma_j^m \mathbf{w}_i \cdot \Phi(\mathbf{x}, i, \mathbf{x}_{i-1}, \mathbf{x}_i), \mathbf{x}\rangle \mid \mathbf{x} \in LU_{iter}\}$
13:     $T \leftarrow T \oplus$ remove_lowconf_negs$(M, LU_{iter})$ with $M < N'$
14: end for
15: for $(y, \mathbf{x}) \in S$ do
16:     $y_s = \arg\max_y \prod_i^5 \Sigma_j^m \mathbf{w}_i \cdot \Phi(\mathbf{x}, i, \mathbf{x}_{i-1}, \mathbf{x}_i)$
17: end for
```

Fig. 1. CRF bagging and bootstrapping (parameter setting: $I = 30$, $\lambda = 1$)

3 Other Related Work

Several authors have proposed rule-based NER systems for Twitter, e.g. [10]. Off-the-shelf rule-based approaches may actually be less sensitive to drift than current state-of-the-art data-driven approaches, but we see this as motivating further research in robust data-driven approaches to NER for Twitter. [17] use distant supervision to improve NER for Twitter, but results are much worse than the ones presented here, e.g. 48% F_1 on Ritter. We do think this is an interesting direction for further research, however. The combination of distant supervision and semi-supervised learning seems like a powerful way of leveraging the information available in unlabeled data without running the risk of being led astray by this data, but here we confine ourselves to semi-supervised learning methods.

4 Data Description

The crowdsourced Twitter data provided by Finin et al. [4] were collected during 2008 and consists of 12,800 unique tweets annotated by 266 different annotators from Amazon Mechanical Turk.[4] For development, we held out and manually correct 2,900 tweets from this dataset (Dev-Finin). We used 9,715 tweets for training, containing 165,704 tokens and 8,607 named entities. Most of the tweets were annotated at least twice (95%). We call the training dataset Finin. To select the most likely labels from the redundant annotations, we used MACE [6]. MACE applies EM to detect which annotators are trustworthy, and recover the most likely answer. On our held-out data, MACE led to a small, but significant, improvement over majority voting. We used the default parameters for MACE (50 iterations, 10 restarts, no confidence threshold) to adjudicate between the turkers. The training data from CoNLL 2003 contains 12,690 sentences with 197,517 tokens and 28,039 named entities. Names are more frequent in newswire data than in Twitter, and the inclusion of the out-of-domain data more than triples the number of named entities in training. For evaluation, we use three different datasets collected at different points in time. We use the entire dataset from Ritter et al. [12] (Ritter) collected during 2010. The data were originally annotated with more fine-grained categories, but were easily mapped to our tagset. We also use the dataset from the MSM13 shared task[5] consisting of 1,450 tweets. These data were sampled in 2010 and 2011. And finally, we introduce a more recent in-house dataset, sampled in June 2013 and containing 1,545 tweets (In-House).[6]

5 Experiments

Baselines. We compare our system to two off-the-shelf baselines, namely the Stanford NER tagger [5][7] as well as Alan Ritters system [12].[8] Moreover, we use a supervised CRF model trained on a combination of crowdsourced data (finin) and newswire data (conll) as a baseline (in-house baseline). We use a fairly standard feature model, very similar to [14], but with Twitter-specific Brown clusters [9]. The concatenation of crowdsourced data and newswire data is the same we trained our system on, but in the baseline model we do not add semi-supervised learning using pools of unlabeled data. Training the baseline model only on newswire data led to much worse results, consistently lower than any of the other baseline models. Using only the in-domain Twitter data gave similar precision score as our baseline model, but the system recognized fewer entities. Thus, including gold standard out-of-domain data increased recall and F_1.

System. Our approach is a combination of bagging [1] and self-training; cf. Fig. 1. We optimized N' for F_1 and recall (to optimize robustness) leading to slightly different models, resp. Bagging-1 ($N' = 1000, M = 0$) and Bagging-2 ($N' = 5000, M = 1000$). Finally, we also compare our bagging models with a co-training

[4] https://www.mturk.com

[5] http://www2013.wwwconference.org/

[6] This dataset will be made public after the reviewing.

[7] http://nlp.stanford.edu/software/CRF-NER.shtml

[8] https://github.com/aritter/twitter_nlp

procedure with two taggers, one trained on CONLL and one trained on FININ. In each iteration, each tagger labels 1,000 unlabeled tweets and adds them to the training data of the other tagger. We also experimented with bigger pools and confidence thresholds (increasing N' and M), but did not see improvements in performance on our development data. The system was generally less confident when predicting organization names, and increasing the confidence threshold further reduced the number of new samples for this category. This resulted in lower recall without notable increase in precision.

Results. Our results are presented in Table 1 and shows the F_1 for the baselines and the semi-supervised systems evaluated on the different datasets. The last column is the macro average of the different datasets, but leaving out Dev-Finin when calcultaing the average for the semi-supervised systems and the in-house baseline. The evaluation scores are computed by the perl script `conlleval.pl` from the CoNLL 2000 shared task. Our three systems all perform significantly better than all baselines ($p < 0.01$), but we note that co-training is best on MSM13 (except for the Stanford NER system), whereas the bagging-based approaches perform best on the in-house data, as well as the Ritter dataset (except for the system from Ritter et al. 2011). BAGGING-2 gives slightly better results than BAGGING-1, mainly because removing low-confident negative predictions from the unlabeled data resulted in better recall for all categories in all datasets.

Table 1. NER results. *Ritter et al. (2011) is a supervised system, evaluated by 4-CV.

	Dev	Test			
	Finin	In-House	Ritter	MSM13	Av
Baselines					
Stanford NER	63.6	61.1	50.8	**80.4**	64.0
Ritter et al. (2011)	43.1	52.4	***67.1**	74.0	59.2
In-house baseline	69.7	66.6	60.4	70.8	65.9
Semi-supervised systems					
Co-training	70.9	65.9	61.3	79.5	68.9
Bagging-1	**72.0**	68.1	61.6	75.6	69.1
Bagging-2	71.1	**70.5**	63.5	76.7	**70.2**

6 Conclusion

We showed that it is possible to learn a named entity tagger for Twitter that outperform state-of-the-art named entity taggers without adding any new gold standard data. Adding new in-domain Twitter data to training boost the performance, but due to significant language drift in Twitter, the effect of such annotations seems to diminish over time. Thus, investing in expert annotations for Twitter seems to be a poor long-term investment if the objective is to induce a robust model for identifying named entities in Twitter. Outsourcing the task to a large crowd is a cheaper and more efficient alternative, but the annotations are of worse quality.

The drop for Ritter et al.'s system when evaluated on our training data could possibly be explained by conceptual differences in the annotation scheme, but our error analysis did not reveal any evidences for such misconceptions. The performance of our in-house baseline is also reduced when applied to later datasets. This emphasize the importance of evaluating NER systems on data sampled differently than the data used in training.

Our results shows that low quality crowdsourced data from the Twitter domain together with an existing out-of-domain dataset can be used to obtain at least as good results as state-of-the-art models that relies on gold standard annotations. Further, we showed that a more robust NER system can be induced using semi-supervised wrapper methods, exploiting the vast amount of unlabeled Twitter data freely available online. All of our three methods outperformed the baselines, and bagging gave the best overall result. Removing low-confident negative predictions from training resulted in a more robust system with better recall and F_1 for all datasets, with exception of the development data.

References

1. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
2. Collins, M.: Discriminative training methods for Hidden Markov Models. In: EMNLP (2002)
3. Eisenstein, J.: What to do about bad language on the internet. In: NAACL (2013)
4. Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., Dredze, M.: Annotating named entities in Twitter data with crowdsourcing. In: NAACL Workshop on Creating Speech and Language Data with Amazons Mechanical Turk (2010)
5. Finkel, J., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: ACL (2005)
6. Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., Hovy, E.: Learning whom to trust with MACE. In: NAACL (2013)
7. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: ICML (2001)
8. Liu, X., Zhang, S., Wei, F., Zhou, M.: Recognizing named entities in tweets. In: ACL (2011)
9. Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., Smith, N.A.: Improved part-of-speech tagging for online conversational text with word clusters. In: NAACL (2013)
10. Piskorski, J., Ehrmann, M.: Named entity recognition in targeted Twitter streams in Polish. In: ACL Workshop on Balto-Slavic NLP (2013)
11. Poibeau, T., Kosseim, L.: Proper name extraction from non-journalistic texts. In: CLIN (2000)
12. Ritter, A., Clark, S., Etzioni, M., Etzioni, O.: Named entity recognition in tweets: an experimental study. In: EMNLP (2011)

13. Rodrigues, F., Pereira, F., Ribeiro, B.: Sequence labeling with multiple annotators. Machine Learning, 1–17 (2013)
14. Sha, F., Pereira, F.: Shallow parsing with conditional random fields. In: HTL-NAACL (2003)
15. Suzuki, J., Isozaki, H.: Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In: ACL, Columbus, Ohio, pp. 665–673 (2008)
16. Turian, J., Ratinov, L., Bengio, Y.: Word representations: a simple and general method for semi-supervised learning. In: ACL (2010)
17. Wang, C.-K., Hsu, B.-J., Chang, M.-W., Kiciman, E.: Simple and knowledge-intensive generative model for named entity recognition. Technical report, Microsoft Research (2013)