Crowdsourcing and annotating NER for Twitter #drift

Hege Fromreide, Dirk Hovy and Anders Søgaard
Center for Language Technology, University of Copenhagen
hege.fromreide@gmail.com, dirkh@cst.dk, soegaard@hum.ku.dk

Abstract

We present two new NER datasets for Twitter; a manually annotated set of 1,467 tweets ($\kappa = 0.942$) and a set of 2,975 expert-corrected, crowdsourced NER annotated tweets from the dataset described in Finin et al. (2010). In our experiments with these datasets, we observe two important points: (a) language drift on Twitter is significant, and while off-the-shelf systems have been reported to perform well on in-sample data, they often perform poorly on new samples of tweets, (b) state-of-the-art performance across various datasets can be obtained from crowdsourced annotations, making it more feasible to "catch up" with language drift.

Keywords: NER, Twitter, crowdsourcing

1. Introduction

Linguistic conventions are constantly challenged and renegotiated at social media platforms, and it seems the out-of-vocabulary rate of any fixed Twitter corpus, for example, just keeps increasing over time (Eisenstein, 2013). This is a challenge to named entity recognition (NER), i.e., the task of identifying and classifying names of people, companies, etc., in text. Named entities are low frequency items, and in 140 character tweets there is little linguistic context to give away whether a word is a named entity or not. State-of-the-art NER systems trained on annotated newswire data therefore perform badly on Twitter (Ritter et al., 2011).

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 (43.1%). This huge drop in performance is obviously prohibitive for down-stream IE in Twitter. The system proposed in Ritter et al. (2011), 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).

The main reason for the drop from news to Twitter is a change in topics and linguistic conventions (Ritter et al., 2011). Eisenstein (2013) shows that topics and linguistic conventions on Twitter change *very* rapidly. This explains the poor performance of the system proposed by Ritter et al. (2011) on our data. A few months old training data from Twitter is almost useless if you want to induce a model for identifying names in tomorrow's tweets. In other words, evaluation of NER for Twitter on held-out data from the same sample of tweets may be very misleading.

This paper presents two new NER datasets and shows how we can train models with state-of-the-art performance across available datasets using crowdsourced training data.

2. NER for Twitter

Twitter data is extremely challenging to NLP with widespread use of sentence fragments, creative abbreviations, misspellings, unconventional capitalization, usernames, so-called hashtags, use of 'RT' to indicate retweeting, URLs and smileys. POS tags are common features in state-of-the-art NER systems, and predicted POS in Twitter will be of lower quality than usual. Ritter et al. (2011) report a drop of the state-of-the-art Stanford tagger from 97% to 80% due to unreliable capitalization, sentence structure and out-of-vocabulary (OOV) words. Similar observiations were made by Foster et al. (2011).

A few papers have been published on NER for Twitter. Ritter et al. (2011), which we will use as a baseline system below, use domain-specific preprocessing tools and distant supervision to adapt a CRF model to the Twitter domain. They also rely on 1,800 manually annotated gold-standard tweets (doing 4-fold cross-validation over 2,400 tweets). They also rely on dictionaries and word clusters. The idea of using word clusters for cross-domain NER has been explored elsewhere (Turian et al., 2010; Rüd et al., 2011). Liu et al. (2011) and Liu et al. (2012) assume (filtered) gold-standard training data ($> 6,120$ tweets) and achieve an average F_1-score of 75.1%, but their result seems to rely on tuning parameters to test data.

Crowdsourcing The Finin et al. (2010) data set is annotated using Amazon Mechanical Turk (AMT),[1] which is a crowdsourcing marketplace for work that requires human intelligence. AMT allows the user to divide the task into several small tasks called Human Intelligence Tasks (HIT). Each worker in the crowd completes one or more HITs and receives a reward. All the combined HITs constitute the outsourced task. Finin et al. assigned 5 tweets to one HIT and paid 5 cents for every HIT completed. Every HIT was annotated at least twice. AMT also requires 10% of the price. Thus $100 gives 4400 annotated tweets. In total the dataset consists of 12,800 unique tweets annotated by 266 different annotators.

One problem with crowdsourcing is that answers sometimes get submitted by spammers, who complete tasks randomly just to collect the reward. To deal with this, AMT provides a threshold on worker agreement as a form of

[1] http://mturk.com

quality control, but this is not sufficient to exclude bad annotators in complicated tasks. Finin et al. (2010) try to overcome this by measuring inter-worker agreement, and supplements the un-annotated tweets with 400 gold standard tweets. One gold standard tweet is added to every HIT, and in this way they can control the quality of the workers. Every worker has a unique identification, so it is possible to exclude bad annotators and remove their annotations. Below we use MACE (Hovy et al., 2013) for additional quality control.

Crowdsourced NER annotations We manually examined 2,974 of the tweets from Finin et al. (2010) and found examples of both spammers with random annotations, and annotators who obviously did not understand the task sufficiently. Examples found include pronouns annotated as persons, products (e.g. 'iPhone') tagged as ORG, and lack of understanding of the context. For instance 'china' was in one occasion labeled LOC when it referred to porcelain. NY was also mistakenly classified as LOC when it referred to the baseball team New York Yankees and should be labeled ORG according to the annotation guidelines.

MACE Hovy et al. (2013) suggest to use EM to evaluate redundant annotations, detect which annotators are trustworthy and recover the most likely answer. They design a model with three variables, namely the annotated label, the true label and a binary label indicating whether the annotator is a spammer or not. If the annotator is not a spammer, the true label is assumed to determine the annotated label completely. This model is called MACE. MACE treats the correct labels as latent variables and does not need supervision. On our data MACE leads to a small, but significant improvement over majority voting, e.g., an F_1 score of 74.1% rather than 72.9% on FMKKM11-TEST. We use the default settings of MACE (50 iterations, 10 restarts, no confidence threshold). In majority voting, we break ties by prefering the majority class.

3. Annotation

For evaluation, we manually annotated a subset of 2,975 tweets of the data from Finin et al. (2010) (FMKKM11-TEST) in accordance with the guidelines used to annotate the CoNLL 2003 data. FMKKM11-TEST consists of 51,056 tokens. We followed the CoNLL 2003 annotation guidelines, but in line with Finin et al. (2010), we only used the labels LOC, ORG and PER (*not* MISC). We also evaluate our system on the test data used in Ritter et al. (2011) (46,469 tokens) (RCEE11-TEST), as well as a new data set (New-TEST) sampled June 14 2013 and manually annotated following the CoNLL 2003 guidelines for the purposes of this paper. This dataset contains 20,664 tokens with 1,581 tokens part of named entities. Following Finin et al. (2010), we again ignore the class of MISC (miscellaneous) named entities such as movie titles and festivals, though abundant in tweets.

We doubly annotated 10% of the data. Our raw agreement was 0.988 and our Kappa score (κ) 0.942.

In our annotations, PER is used to label first, last and middle names, names of fictional characters and aliases. Titles, like mr., president and officer, are not said to be part of the named entity. The organization class, labeled ORG, refers to entities with organization structure. This includes companies, political movements, musical bands and orchestras, sport clubs, government bodies and public organizations (like schools). The location class (LOC) covers geographical names, like names of countries, regions, oceans and mountains, as well as man-made locations like monumental structures, roads, bridges and buildings. Public and commercial places like schools, museums, hospitals and airports are also covered by this class. Note that many words are ambiguous out of context, e.g., *Washington* can be both PER, LOC and ORG.

If hashtags are named entities they are also annotated. The annotators of Finin et al. (2010) crowdsourced dataset have occasionally annotated usernames as named entities; we annotate them as PER or ORG. Note also that named entities are often referred to by abbreviations or spelling variatins, e.g., *fb*, *Fb*, *fbk*, *Fbook* and *facebook* – or *suju* and *UOfM* below.

(1) @Pet_PandaLOVE yeah they were my three emergency days or in case the days I wanted to keep for when suju/ORG come to the UK/LOC

(2) Omw Back From UOfM/ORG They Love Me They Want Me To Come Back:) Just Couldnt Really Show out Cause I Was Still Hurting From Last Night!

The word *suju* is short for Super Junior (a South Korean boy band). *UOfM* is an abbreviation for University of Michigan.

4. Experiments

Data The crowdsourced Twitter data from Finin et al. (2010) consists of 12,800 unique tweets annotated by 266 different annotators. The split used for training is referred to as FMKKM11-TRAIN below. Most of these tweets are annotated at least twice (95%). The CoNLL 2003 training data contains 14,989 manually annotated sentences from the Reuters corpus. In total, we train on 369,706 tokens.

The three test datasets differ in whether Twitter user accounts ("@-...") are annotated as PER. When we train our model for RCEE11-TEST and New-TEST, we therefore remove annotations of user accounts as persons from FCKKM11-TRAIN.

The OOV rates for the three test sets relative to the crowdsourced data are presented in Figure 1. As we would expect, the OOV rates for words and bigrams are slightly lower on FMKKM11, with the exception that the OOV rate for PER is similar to RCEE11. The reason is that user accounts are annotated as PER in our in-sample experiments; ignoring user accounts, the OOV rate is 69.5% rather than 71.7%.

System Our system is called MACE-CRF below. We use the default settings of MACE (50 iterations, 10 restarts, no confidence threshold) to get training labels on the FMKKM11-TRAIN. The model is learned using CRF++[2] with default parameters. In addition to word forms, we use POS tags, word lists, as well as word clusters. Gimpel et al. (2011) provides a POS-tagger designed especially for

[2] https://code.google.com/p/crfpp/

	FMKKM11-TEST				RCEE11-TEST				New-TEST			
	LOC	ORG	PER	All	LOC	ORG	PER	All	LOC	ORG	PER	All
MACE-CRF	68.8	52.8	84.0	**74.2**	62.0	32.2	73.5	59.7	65.4	39.3	78.7	**66.9**
RCEE11-CRF	54.7	39.2	39.6	43.1	68.0	46.8	75.0	**67.1**	51.1	25.6	62.4	52.4
CoNLL-CRF	32.0	56.8	52.5	67.1	59.7	16.8	60.5	44.2	64.8	24.8	60.6	48.8

Table 1: Results on three Twitter NER datasets (Finin et al., 2011; Ritter et al., 2011; new dataset)

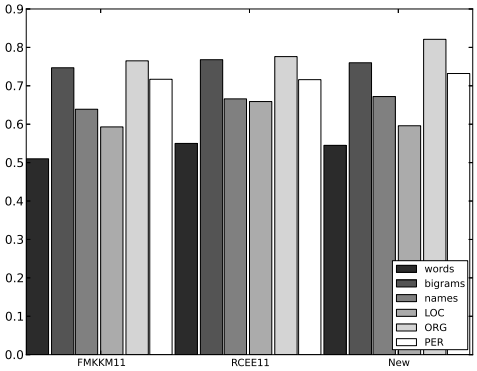

Figure 1: Out-of-vocabulary rates for the three test sets

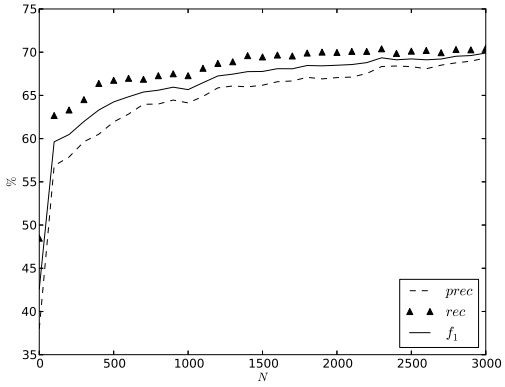

Figure 2: MACE learning curve (N is number of crowdsourced data points, F_1 along the y-axis)

Twitter data and reports an accuracy of nearly 90%. This tagger is applied to the Twitter data and provides features for the NER. The CoNLL data set is already tagged with POS, but the tagset differs from the one used by Gimpel et al. (2011). Therefore, both tagsets are converted to the universal part-of-speech tagset provided by Petrov et al. (2011). As additional features we use word clusters learned running the Brown clustering algorithm[3] on the UKWAC corpus[4], as well as publicly available Twitter word clusters.[5] Finally, we also use occurrence in gazeteers[6][7], filtered using frequency lists.

Baselines Our baseline systems include CRF++ trained only on CoNLL 2003 data (CoNLL-CRF), using the same features and the same parameters, and the full system in Ritter et al. (2011) (RCEE11-CRF). Ritter et al. (2011) report that the Stanford NER system obtains an F_1-score of 29% on RCEE11-TEST. Our baseline CRF model thus performs better than the Stanford NER system on this data, probably because of the word clusters.

Results One result of this paper is that while NER for Twitter works better than supervised systems trained on newswire, we see significant performance drops when evaluating NER for Twitter systems on tweets sampled differently from the data used to train such systems. This seems to motivate using crowdsourcing in the loop when doing NER for Twitter.

Our main result is that a simple CRF model trained on crowdsourced data seems to do at least as well at this task as a state-of-the-art NER model for Twitter that relies on gold-standard Twitter data. The reason probably is that while our training data is of worse quality, we have about 20 times more data. Our word clusters are also obtained from an Internet corpus rather than from newswire, and we use fewer features, making over-fitting to a particular sample of Twitter data less likely. However, note that it would be easy to obtain crowdsourced training data of more recent data to boost performance (to an expected 70-75% F_1). The poor results on ORG are primarily due to low recall, probably explained by the high OOV rates for organization names (see Figure 1 for comparison).

The MACE learning curve is presented in Figure 2. We see that a 1000 crowdsourced tweets, corresponding to an annotation cost of $22, is enough to close half of the performance gap between in- and out-of-domain performance.

5. Conclusion

We showed that there is considerable population drift on Twitter. Consequently, state-of-the-art NER systems suffer from significant out-of-sample performance drops. More than half of the gap between state-of-the-art NER systems' performance on news data (90-93%) and Twitter data (40-45%) can, however, be closed using crowdsourced data for bias correction with an annotation cost of approx. $200. On Twitter data sampled two years later than the crowdsourced training data, we observe a performance drop of almost 10%, but interestingtly the crowdsourced model still performs better than a state-of-the-art NER system for Twitter trained on gold-standard Twitter data.

[3] https://github.com/percyliang/brown-cluster
[4] http://wacky.sslmit.unibo.it/
[5] http://www.ark.cs.cmu.edu/TweetNLP/cluster_viewer.html
[6] http://geonames.org
[7] http://optima.jrc.it/data/entities.gzip

6. References

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *NAACL*.

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Josef Le Roux, Joakim Nivre, Deirde Hogan, and Josef van Genabith. 2011. From news to comments: Resources and benchmarks for parsing the language of Web 2.0. In *IJCNLP*.

Kevin Gimpel, Nathan Schneider, Brendan OConnor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter. In *ACL*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *NAACL*.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *ACL*.

Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, and Xiangyang Zhou. 2012. Joint inference of named entity recognition and normalization. In *ACL*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. CoRR abs/1104.2086.

Alan Ritter, Sam Clark, Mausam Etzioni, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *EMNLP*.

Stefan Rüd, Massimiliano Ciaramita, Jens Müller, and Hinrich Schütze. 2011. Piggyback: Using search engines for robust cross-domain named entity recognition. In *ACL*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *ACL*.