Named Entity Recognition and Part of Speech Recognition in Tweets

Abstract

Named Entity Recognition (NER) is part of Natural Language Processing and is a form of information extraction that helps locate and classify named entities in unstructured text into categories such as locations, people, organizations etc. While the performance of conventional NLP tools is rigorous for formal pieces of literature such as articles and long pieces of texts, it is severely degraded in a noisy, informal corpus of 140 character messages that are tweets. Coupled with the insufficient information in a tweet, named entities being out-of-vocabulary or OOV and the lack of training data, NER becomes all the more challenging. Recently several works have been posited to implement POS tagging, Conditional Random Fields (CRFs), normalization in Named Entity Recognition and other forms of distant supervision or unsupervised learning. In this paper we propose conducting domain adaptation where we prefer the Broad twitter corpus or BTC (Derczynski et al. 2016) as a means of training, development and test data over the Ritter et al. 2011 dataset as the former is not only significantly bigger than the latter but also sampled across different regions, temporal periods, and types of Twitter users. To further delve into the consideration of named entities we use domain transfer by modifying the corpus from Ritter et al. 2011 to match the 3 named entities specified in the BTC (Person, Location, Organization) and using algorithms put forward in Ritter to evaluate the BTC data. Using this new dataset we hope to test state-of-the-art natural language processing algorithms and machine learning algorithms.

1 Introduction

For humans the task of being able to identify and name entities from unstructured text is not difficult at all- whether it be recognizing a city, a person or a company by its name or description. But the pace of technology in keeping up with the same level of computational ability humans possess is still lagging. The advent of social media conglomerates like Facebook, Twitter, and Instagram and the creation of new text types such as status messages and user posts have posed difficult challenges for language technology because of the aforementioned informal and noisy nature that is inherent to these text types. Still, because of the easily accessible platform of tweets, they can provide information which is more up-to-date and a faster mode of communication than news articles. Even more so, the rapidly increasing number of tweets in existence warrants data-mining through NER and information extraction.

Currently there is less than 100k tokens publicly available with the added constraints of high performance systems such as that posited by Liu et al. 2012 not being available for evaluation and thus

not reproducible, single-annotators and low-levels of agreements between multiple annotators (Derczynski et al., 2016).

Prior research in this field is dominated by distant supervision and semi-supervised learning algorithms. A commonly used dataset in Twitter NER is the Ritter dataset and has been used for rebuilding the NLP pipeline (Ritter et al. 2011), KNN algorithm and Conditional random fields (Liu et al. 2011). Refer to more prior research under relevant works.

Machine learning is vital in the case of conducting Twitter NER because of the limited amount of annotated data, so the computer needs to have the ability to learn from training sets without being underfitted or overfitted and still have a satisfactory F-1 score on different test datasets. The combination of a supervised and unsupervised learning system will allow for the computer to adjust to data it hasn't seen annotated before and still produce satisfactory results.

1	#BoycottFakeStars Unfollow this fake stars who come in industry with the help of his family nor his/her talent. Like sushant singh rajput , rajkumar rao nd many more come in film industry bcz of their talent nor from their back . #BoycottFakeStars
2	When someone wants to make his name have a stardom through tiktok.. why don't u let tht happen? It isn't to gain fandom through tiktok..it maybe easy to get one vedio viral but maintaining it .. gaining popularity isn't tht easy as u think #TrollingCauseDepression
3	I don't know how many people remember the importance of this day! But two years ago on this day we lost 18 soldiers including my friend Gangadhar Dalui in #UriAttack Paying homage to all d fallen soldiers & taking pledge to support their families within our limits! #UriMartyrsDay

Table 1: Example of Noisy Text from Tweets sourced from Indian Twitter

2 Related Work

Named entity recognition has been vastly researched and its solutions can be categorized into 1). Rule Based (Krupka and Hausman 1998), 2). machine learning based (Finkel and Manning, 2009; Ritter et al. 2010; Liu et al. 2011) and 3). hybrid methods (Jansche and Abney, 2002). Moreover, with the availability of annotated corpora such as CoNLL03 (Tjong Kim Sang and De Meulder, 2003), data-driven methods have increasingly become the norm.

2.1 NER on non-twitter corpi

Heuristic based approaches dominated the field of NER in the 20th century but Bikel et al. 1999 was crucial in revolutionizing conventional approaches to become flexible with each new source of text through employing learning algorithms. Handcrafting finite state patterns such as <proper-noun>+ <corporate designator> ==> <corporation> for recognizing names, locations(...) lacked the factoring in of typical naming conventions such as how organizations choose to have names representative of the industry they are in or type of service/good they are offering. That's why Bikel et al. 1999 reasoned that in preventing the use of excessive resources towards fitting rules over different data and minimizing the significant tweaking that was required with the introduction of each new text, to implement a hidden markov model. Through this they were able to construct a bigram language model which would compute "the likelihood of a sequence of words… and there would be a probability associated with every transition from the current word to the next word".

Currently research into NER has mostly focused on formal texts such as news articles (Mccallum and Li, 2003) but has also diversified into the biomedical NER systems with Yoshida and Tsujii 2007 publication utilizing shallow parsing, POS tagging and orthographic features. In another case, because of the nature of elements of supervision on Twitter based NER approaches, which requires the availability of labeled data, Finin et al. 2010 used a crowd-sourcing way utilizing Amazon Mechanical Turk Services and CloudFlower to prepare labeled data and train a CRF model for testing the effectiveness of human done labeling. Still, NER has been gaining renewed interest from the challenging task posed by tweets.

2.3 Rebuilding the pipeline

Ritter et al. 2011 presents a novel way of "rebuilding the NLP pipeline" through POS tagging, chunking and NER which would outperform the conventional Stanford NER system, which because of its unreliable capitalization and misclassification of nouns and proper nouns posed large losses in performance metrics. This would first use Conditional Random Fields for named entity segmentation as a sequence labeling task and then a distantly supervised approach applying LabeledLDA so as to add

constraints from Freebase, an open-domain database, on the dataset as a form of supervision and to classify named entities.

2.4 Semi Supervised Learning

Liu et al. 2011 proposed another multifaceted approach regarding the implementation of normalization of tweets which corrects "ill-formed words" using a global linear model, combination of KNN algorithm with a linear conditional random fields (CRFs) model and a semi supervised learning framework that makes up for the lack of training data. The K-Nearest-Neighbors algorithm is used for pre labeling the over 12k tweet corpus which is then used as the input for the CRF model in performing sequential labeling. With the added introduction of 30 gazetteers-representing general knowledge across a host of different domains-into the mix, the method Liu proposes aims at combining global information from KNN and gazetteers with contextual information from the tweets to subsidize the lack of training data. In a later publication, building up from her research, Liu et al. 2012 constructs a named entity normalization method for tweets that would allow for more efficient and accurate entity recognition that would account for the variations of NEs in tweets. This proves to be successful in increasing the F1 score by a margin of 3.4% from the baseline as it implements NER and NEN jointly using a factor graph as their model, thereby also solving for the limitation of errors propagating from the entity recognition to NEN.

Shubhanshu Mishra and Jana Diesner also took a similar approach to Liu et al. with a semi-supervised NER system and CRFs but also introduced "leverage random feature dropout for up-sampling the training data" which allows for understanding new tokens into the system via unsupervised learning. Furthermore, we can analyze the empirical analysis from Derczynski et al. 2015 for named entity recognition and disambiguation to see different systems's performances on noisy texts. They concluded that the most significant drop in the performance of NER approaches comes from poor capitalization and that slang contributes only a minor drop in the performance readings. Still, they continue that improvements did come with micro-blog trained POS tagging and normalization. Possible avenues of more improvement to the capitalization problem can be training a mirco-blog specific recaser.

3 Data

The data that will be used in this research will be based on the Broad Twitter Corpus (BTC; Derczynski et al. 2016). The commonly used Ritter dataset in comparison has a mere 45000 tokens, which is just 15% the size of the CoNLL'2003 dataset that is popular for news NER. Because of the need for a

sizeable, highly diverse and quality annotation dataset, the BTC boasts gold-standard named entity annotations from both NLP experts and crowd workers and is stratified for time over a six-year period, for different places that account for the different variations of English across the world, and is socially segmented for non-professional content and news. We use the recommended train and development test split of section H and use the entirety of section F for testing. Section H is stratified for the month, time of day and day of the week allowing for copious amount of varied data across "temporal cycle types" (Derczynski et al. 2016). Section F offers content from individuals providing twitter-based commendary-or the *twitterati* and is stratified across regions of the UK, as well as authors from backgrounds ranging from sports and journalism to music and politics.

Corpus	Tokens	Entity schema	Annotator type	Annotator qty.	Notes
Finin et al. (2010)	7K	PLO (3)	Crowd only	Multiple	Low IAA (Fromreide, 2014)
Ritter et al. (2011)	46K	Freebase (10)	Expert	Single	IAA unavailable
Liu et al. (2011)	12K	PLO (3) + Product	?	?	Private corpus
Rowe et al. (2013)	29K	PLO (3) + Misc	Expert	Multiple	No hashtags/usernames
Broad Twitter Corpus	165K	PLO (3)	Expert + Crowd	Multiple	Source JSON available

Table 2: Comparison of Different Openly available Corpi (PLO is Person, Location, Organization)

	Feature	Count
Dataset	Documents	9 551
	Tokens	165 739
Entities	Person	5 271
	Location	3 114
	Organization	3 732
	Total	12 117

Table 3: BTC statistics

4 Baseline Construction

4.1 Annotation Scheme

In creating a baseline of our own we randomly sampled 150 tweets from Twitter's India platform. In doing so we aimed at analyzing a major English speaking nation that was exempted from the Broad Twitter Corpus. In the creation of our baseline we further attempted to understand named entity recognition and POS tagging in the regional differences inherent to the English language in India as

compared to other countries analyzed in the BTC. Similarly to the annotation system implemented in the BTC, we had multiple annotators, (student and NLP expert) and our data was stratified to have content from a host of different entities such as celebrities, sports, news, and politics. Moreover, entity classifications were done with the same three aforementioned factors: person, location and company.

In the case of polysemous entities, annotators would classify an entity after understanding it in the context that it's used. For example: *"...and success of dhoni movie the perfect dhoni for the movie…".* Here "dhoni movie" references a movie about Mahendra Singh Dhoni, but is not a person, whereas further ahead in the tweet, "the perfect dhoni for the movie" is illustrative of a person.

An important thing to also note is that although there were measures taken to source tweets from different time periods such as the 2016 Demonetization issued under the Modi administration, the Triple Talak controversy, and/or the URI surgical strike on Pakistan from India, the number of tweets collected in that past month have composed a far larger proportion of the baseline data. This invariably exposes the data for entity drift (Masud et al., 2010) where the selected entities may be prevalent currently but change in the future. For example, in the case of 2014, the PM of India was Manmohan Singh, but today it is Narendra Modi. In another case, it might be veneration of a deity during Diwali in October and Santa during Christmas. Because of this, there is the possibility of overfitting the data as the model would be trained on data not stratified over different periods of time, and so on the introduction of a testing set from a different period of time, the results will not be at par.

Day of the month (June 2020)	14	15	16	17	18	19	20
Baseline	23	27	25	10	14	18	15

Table 4: Volume of tweets collected by day of month

Note: 5 Tweets on Demonetization (2016), 1 tweet on Triple Talak (2018), 3 tweets on Ram Mandir (2017), 5 tweets on Uri surgical strike (2016), 4 tweets from Pulwama attack (2019).

4.2 Adjudication

Adjudication is very important in creating reliable and accurate annotations. Currently there are automated adjudication methods such as MACE (Hovy et al., 2013) but they do not provide the satisfactory processing that a human would in understanding the impacts of different circumstances. That's why we primarily chose to implement human adjudication.

Furthermore, in order to test the efficacy of annotations there were implications for utilization of naïve inter-annotator agreement or IAA. We measured the proportion of annotators classifying each token unanimously as the same entity. This would then be averaged across the entire corpus to find the inter-annotator agreement.

5 Learning Approach

Learning approach taken in this research paper will be utilizing the algorithms posited from Ritter, and retraining them on the BTC dataset. In the Ritter approach, Support Vector Machines were used with features used including: fraction of words that had tweets capitalizalized, the fraction that appear in a dictionary as lowercase/or uppercase but are not so in tweets, the frequency of the word 'I' appearing lower case and whether or not the first word is capitalized. This is really significant as Derczynski et al 2013 concluded that the most significant drop in the performance of NER approaches comes from poor capitalization. With features based on capitalization, performance would improve at named entity segmentation (Ritter et al. 2011). Although the number of named entities was significantly more in the Ritter paper (10) we will prefer the BTC dataset which has 3 entities that are more clearly delineated.

6 Results

6.1 Reproducing Ritter Original Train and Development Sets

Statistics of Ritter Original Dataset Features

Number of data sets (groups): 1
Number of instances: 2393
Number of items: 46462
Number of attributes: 98251
Number of labels: 21

processed 16261 tokens with 661 phrases; found: 538 phrases; correct: 204.
accuracy: 93.59%; precision: 37.92%; recall: 30.86%; FB1: 34.03

Table 5

Named Entity	Precision	Recall	FB1	Number of entities found

company	42.86%	23.08%	30.00	21
facility	13.89%	13.16%;	13.51	36
geo-loc	42.45%	50.86%	46.27	139
movie	12.50%	6.67%	8.70	8
musicartist	0.00%	0.00%	0.00	6
Other	32.95%	21.97%	26.36	88
person	47.52%	56.14%	51.47	202
product	4.76%	2.70%	3.45	21
sportsteam	40.00%	5.71%	10.00	10
tvshow	0.00%	0.00%	0.00	7

6.2 Testing Ritter Modified Development Data And Ritter Modified Training Data

Statistics of Ritter Modified Dataset Features

Number of data sets (groups): 1
Number of instances: 2393
Number of items: 46462
Number of attributes: 98251
Number of labels: 7

processed 16250 tokens with 326 phrases; found: 271 phrases; correct: 141.
accuracy: 97.32%; precision: 52.03%; recall: 43.25%; FB1: 47.24

Table 6

Named Entity	Precision	Recall	FB1	Number of entities found
company	69.23%	23.08%	34.62	13
geo-loc	54.17%	44.83%	49.06	96
person	49.38%	46.78%	48.05	162

6.3 Derczynski BTC Train and Developmental Data Test with Ritter Algorithms

Statistics the data set(s)

Number of data sets (groups): 1
Number of instances: 998
Number of items: 14441
Number of attributes: 40575
Number of labels: 7

processed 15002 tokens with 1732 phrases; found: 1618 phrases; correct: 1151.
accuracy: 94.25%; **precision: 71.14%; recall: 66.45%; FB1: 68.72**

Table 7

Named Entity	Precision	Recall	FB1	Number of entities found
LOC	48.15%	33.33%	**39.39**	108
ORG	42.75%	15.53%	22.78	138
PER	75.80%	86.96%	**81.00**	1372

6.4 Derczynski BTC Train and Test Data Test with Ritter Algorithms

Number of data sets (groups): 1
Number of instances: 998
Number of items: 14441
Number of attributes: 40575
Number of labels: 7

processed 12308 tokens with 1462 phrases; found: 1345 phrases; correct: 930.
accuracy: 93.62%; **precision: 69.14%; recall: 63.61%; FB1: 66.26**

Table 8

Named Entity	Precision	Recall	FB1	Number of entities found
LOC	45.39%	34.33%	**39.09**	152
ORG	46.97%	18.45%	26.50	132

PER	75.31%	86.38%	**80.46**	1061

7 Analysis

As shown in table 6 and table 7, tests using the BTC training data on the BTC test data and development data have the model outperforming the Ritter original and modified data sets using the Ritter algorithms. We believe the main reason for this is the differing approaches each uses for data selection. Ritter et al. 2011's dataset is anachronistic to some extent having collected their data in a short period of time one day, which also means that the data is also constrained from the fact that it only includes information from those who were active at that period of time (Derczynski et al. 2016). Also unlike in the annotations of the Ritter dataset, in order to incorporate previously missing essential entity classification for user mentions and hashtags, Derczynski designed the BTC to separate preceding symbols such as # or @ into separate and individual tokens.

Although holistically testing on BTC shows improvement from the current results, there are still instances where the performance metrics of an entity fared worse than on the Ritter datasets. Specifically, in the Derczynski development data set and test set the recall and FB1 scores for the named entity of organization was less than half that of the recall and FB1 score on the Ritter dataset. The reason behind this is most probably because in the Ritter algorithms there wasn't a specific tag for organization like there was in the BTC and the closest thing to it was the company tag. Because of this, the model would run on the BTC data and try to find entities that it could classify as a company but because it was annotated for organization there would have been many instances where it missed the classification. Another possibility or factor working in conjunction could be that the lack of information on organizations in the entities dictionaries provided by Ritter makes it so that the model doesn't have sufficient capability to delineate an entity as an organization.

Year	1996	2009	2010	2011	2012	2013	2014
Our corpus	0	3	5	127	2414	275	6022
Ritter (2011)	0	0	6902	0	0	0	0
CoNLL'03	1358	0	0	0	0	0	0

Table Describing Spread of data collection by Derczynski et al. 2016

The favorable results of the BTC data are not isolated in our experiments. In tests conducted by Roth et al. 2017 where their NER and POS tagging algorithms were not based off tweets consideration, the utilization of the BTC data for testing and CoNLL 2003 data for training still shows that despite having a cross-domain experiment their model could still function better on out-of-domain data.

Evaluating on the level of mention detection, the BTC allowed their team to see an 8 point increase in their F1 score in their third experiment one and a 3.5 point F1 increase in their fifth experiment.

Using Ziph's law as well we can analyze the statistical distribution of words in a corpus in which the frequencies of words are inversely proportional to their ranks. For example: some very high frequency words account for most of the tokens in a certain piece of text can be "the, of, I" etc. or can be very low frequency words such as "Dalit" or "barbarism" in our baseline Indian tweets dataset (Piantadosi, NCBI). Following the general formula $f(r) \propto \frac{1}{r^a}$ we can see in figure 1 and figure 2 the comparison between the named entity mentioned in the newswire dataset which is classically used for non-tweet based and more formal NER procedure and the BTC dataset.

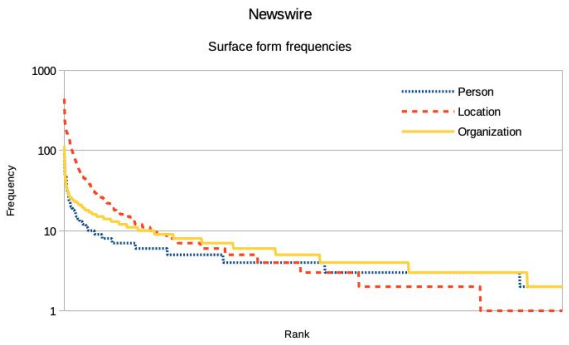

Figure 1: Frequency-Rank curve for entities in CoNLL'03 data.

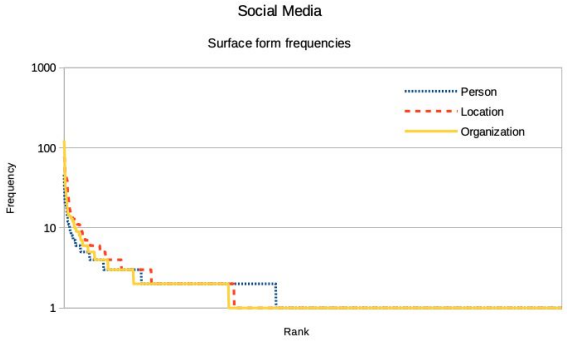

Figure 2: Frequency-Rank curve for entities in the Broad Twitter Corpus.

Retrieved from Derczynski et al. 2016

9 References

1. Still compiling all of them