Broad Twitter Corpus: A Diverse Named Entity Recognition Resource

Leon Derczynski
Dept. of Computer Science
University of Sheffield
S1 4DP, UK
leon.d@shef.ac.uk

Kalina Bontcheva
Dept. of Computer Science
University of Sheffield
S1 4DP, UK
k.bontcheva@shef.ac.uk

Ian Roberts
Dept. of Computer Science
University of Sheffield
S1 4DP, UK
i.roberts@shef.ac.uk

Abstract

One of the main obstacles, hampering method development and comparative evaluation of named entity recognition in social media, is the lack of a sizeable, diverse, high quality annotated corpus, analogous to the CoNLL'2003 news dataset. For instance, the biggest Ritter tweet corpus is only 45 000 tokens – a mere 15% the size of CoNLL'2003. Another major shortcoming is the lack of temporal, geographic, and author diversity. This paper introduces the Broad Twitter Corpus (BTC), which is not only significantly bigger, but sampled across different regions, temporal periods, and types of Twitter users. The gold-standard named entity annotations are made by a combination of NLP experts and crowd workers, which enables us to harness crowd recall while maintaining high quality. We also measure the entity drift observed in our dataset (i.e. how entity representation varies over time), and compare to newswire. The corpus is released openly, including source text and intermediate annotations.

1 Introduction

Businesses, governments, and communities increasingly need real-time information from dynamic, large-volume media data streams, such as blogs, Facebook, and Twitter. In particular, the automatic detection of mentions of people, organizations, locations, and other entities (i.e. Named Entity Recognition) is a key step in numerous social media analysis applications, e.g. competitor and brand monitoring (Mostafa, 2013), debate and election analysis (Mascaro and Goggins, 2012; Tumasjan et al., 2010), disaster response (Kedzie et al., 2015; Neubig et al., 2011), and health- and well-being applications (Coppersmith et al., 2014; Choudhury et al., 2013).

NER methods (typically trained on longer texts, such as news articles), have been shown to perform poorly on shorter and noisier social media content (Ritter et al., 2011). Therefore, recent Twitter NER work (Ritter et al., 2011; Liu et al., 2011; Derczynski et al., 2015) has focused on improving the state-of-the-art, through new methods. The challenges come from named entities (NEs) typically being out-of-vocabulary (OOV) as compared to the training newswire data; the shorter context; and lack of sufficiently large NE annotated social media datasets.

In more detail, Table 1 shows that there are less than 90 thousand tokens of publicly available NE-annotated tweet datasets, and even those have shortcomings in terms of annotation methodology (e.g. singly annotated), low inter-annotator agreement, and stripping of important entity-bearing hashtags and user mentions. At the same time, Ritter et al. (2011) demonstrated that blending gold-standard newswire training data with social media training data - in order to increase the size of the dataset and better generalize - leads to worse performance. Lastly, as this paper demonstrates, existing NER datasets suffer from poor temporal diversity, making the trained NER models sensitive to the entity drift phenomenon.

We address these challenges through building a sizeable manually-annotated corpus – the Broad Twitter Corpus (BTC). In order to maximize diversity, the BTC is stratified for time, including social media posts from a six-year period, drawn from different parts of year, days of the month, and times of the day. Secondly, it is drawn from different places, accounting for different variants of English and the

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 1: Characteristics of openly-available social media NER corpora. PLO standards for Person, Location, Organization NE types.

different entities found in various regions of the English-speaking world. Finally, it is partially socially segmented, including reactions to news stories, non-professional content, and text from the "twitterati".

The corpus is made freely available in various formats; the source text is included under Twitter's revised 2015 licensing guidelines, as are the intermediate annotations.

2 Corpus Construction

The goal of the corpus is to provide a representative example of named entities in social media. Social media is said to contain more variance than some other text types, like newswire. It certainly is authored by a broader demographic than newswire (Eisenstein, 2013), and contains a variety of styles and formality registers, unlike other user-generated content such as Youtube comments or SMS (Hu et al., 2013). Changes in general discourse subject are also said to present more quickly in social media, creating topic shifts of both greater magnitude and higher frequency than other text types. We focus on Twitter, using this as the "model organism" of social media text (Tufekci, 2014), to assemble a corpus that is capable of catching these variances.

2.1 Annotation Scheme

The BTC corpus is divided into segments, where the documents within each segment share a common theme (see Table 2). The documents consist of the social media message – i.e. tweet – complete with its JSON metadata. The text of the tweet is then annotated with into sentences and tokens, using the TwitIE tokenizer (Bontcheva et al., 2013).

Tokenisation presents some issues in tweets. Classic schemas like PTB do not work well with constructs like smilies or URLs. To address this, we use the TwitIE tokeniser (Bontcheva et al., 2013) which is roughly based on TweetMotif and the twokeniser tool (O'Connor et al., 2010). Of note, we separate the preceding symbol in mentions and hashtags (the @ or # characters) as a distinct token, but still include this in entity spans.

The main question was which entity classes should be covered in the corpus. Some Twitter corpora have used ten top-level Freebase categories (Bollacker et al., 2008; Ritter et al., 2011)[1] or have included products (see Table 1). For expert-based annotation methodologies Hovy (2010) recommend at most ten, ideally seven, categories. In crowdsourcing, successful tasks tend to present even fewer choices – in most cases between two and five categories (Sabou et al., 2014).

Therefore, we chose to focus on the three most widely used entity categories: Person, Location, and Organization (Table 1). Apart from being well understood by annotators, these three categories offer straightforward mapping to the other existing Twitter datasets, so all could be used in combination, as training data, if needed. An initial pilot (Bontcheva et al., 2014a) also included a fourth class, "Product", but crowd workers struggled to annotate these correctly and with good recall, so they were dropped.

For polysemous entities, our guidelines instructed annotators to assign the entity class that corresponds to the correct entity class in the given context. For example, in "*We're driving to Manchester*", Manchester is a location, but in "*Manchester are in the final tonight*", it is a sports club – an organization.

Special attention is given to username mentions. Where other corpora have blocked these out (Rowe et al., 2013) or classified them universally as person (Ritter et al., 2011), our approach is to treat these as named entities of any potential class. For example, the account belonging to Manchester United football club would be labeled as an organization.

[1] Categories used: company, facility, geo-location, movie, music artist, person, product, sports team, TV show and other.

Section	Region	Collection period	Description	Annotators	# Tweets
A	UK	2012.01	General collection	Expert	1000
B	UK	2012.01-02	Non-directed tweets	Expert	2000
E	Global	2014.07	Related to MH17 disaster	Crowd & expert	200
F	Stratified	2009-2014	Twitterati	Crowd & expert	2000
G	Stratified	2011-2014	Mainstream news	Crowd & expert	2351
H	Non-UK	2014	General collection	Crowd & expert	2000

Table 2: Segments of the corpus. A region of "stratified" indicates that data was taken from six regions in the English-speaking world that had a sufficient crowdsourcer workforce to ensure annotator diversity.

2.2 Corpus Diversity

In order to maximize diversity of English social media content, our methodology samples tweets along three dimensions: spatial, temporal and social.

The spatial dimension aims to account for linguistic variation in social media across English-speaking countries. In particular, the discussed entities vary from one region to the next, e.g *Justin Trudeau* in Canada vs. *Theresa May* – in the UK. For these reasons, data is collected from the USA, the UK, New Zealand, Ireland, Canada and Australia.[2]

The temporal dimension is key for being able to make models more resilient towards temporal entity drift (Masud et al., 2010; Magdy and Elsayed, 2016), i.e. the change in entities mentioned at different time periods. For example, *George Bush* or *Pamela Anderson* in the 1990s vs Angelina Jolie and Tiger Woods in 2010, or Santa near Christmas and Guy Faulks in UK tweets in the early winter.

We aim to capture this drift in the BTC by collecting posts over a number of years,[3] as well as being taken from different times of years, days in the month, and times of day. In contrast, previous social media NE corpora were gathered during narrow contiguous time periods (Ritter et al., 2011).

The final, social dimension again aims to account for variation in linguistic styles across different kinds of Twitter users. For instance, verbal communication behaviors such as g-dropping are often copied into typed social media messages (Eisenstein et al., 2010). To try to capture these, the corpus collects data from different segments, explicitly taking in content from well-known public figures, news outlets, well-known social media figures, plus a large volume of randomly-selected posts.

2.3 Corpus Segmentation

The dataset is organized into multiple segments (Table 2) to reflect the diversity criteria and annotation approach (expert vs. crowd). English tweets were filtered using langid.py (Lui and Baldwin, 2012).

Segment A comprises a random sample of UK tweets, collected after New Year, annotated by multiple NLP experts. This data was used for calibration, so includes both expert input and crowd correction.[4]

Segment B is similar to segment A. In this segment we focused on non-directed tweets – i.e. those that are not private replies and so do not begin with a username mention. These were found to be more likely to contain a named entity based on sampling the Ritter et al. (2011) corpus.

Segment E is a small sample focused on a specific event, the crash of flight MH17 over Ukraine. It contains commentary from different places and social levels, contributing different kinds of language and reactions. The entities here have a wide range and are generally from outside the English-speaking world, often instead being Ukrainian, Dutch or Malaysian. This provides needed diversity in names, as well as examples of L2 English language use around NEs (i.e., from non native speakers).

Segment F comprises content from popular individuals that provide twitter-based commentary – the "twitterati". These are stratified across the six English-speaking regions listed above, as well as a general global section. Authors are from the worlds of celebrity, music, politics, sports, journalism; they are a

[2]While there are other countries with English as a first language, such as Botswana and Singapore, we were constrained by the number of local crowd workers we could access (see Section 3.1).

[3]Taken from an archive of Twitter "garden hose" data, a fair 10% sample (Kergl et al., 2014).

[4]The crowd correction is via feedback from its use as gold test data; see Section 3.1.

Annotator group	Recall over final annotations	F1 Agreement
Expert	0.309	0.835
Crowd	0.837	0.350

Table 3: Comparison of expert and crowd annotators over segment A, for calibration. Note higher recall but lower agreement in crowd. Agreement is F1 lenient with micro-averaging.

mix of both principals in the fields and also commentators. This content is reasonably unique to social media, often being too low-impact, speculative or controversial to reach mainstream news.

Segment G contains, in contrast, tweets from mainstream news in the six English-speaking regions, e.g. CNN in the US, SMH in Australia, RTE in Ireland, CBC in Canada and so on. Many local outlets that do not have an international edition are also included.

Segment H is the most varied. To balance the UK bias of segment B, this segment excludes tweets of UK origin (according to the Twitter metadata). The segment is stratified for month of year, time of day, and day of week, giving an even spread over many temporal cycle types in the collection period.

3 Annotation Process

To make annotation scalable and of high quality, while ensuring sufficient annotator variety, corpus annotation was carried out using a mix of NLP experts and paid-for crowdsourcing. The annotation process follows general best practices in crowdsourced corpus annotation (Callison-Burch and Dredze, 2010; Alonso and Lease, 2011; Sabou et al., 2014). For example, task design is kept clean (a critical factor, more important than e.g. interface language – (Khanna et al., 2010)), and the process developed over pilot and refinement iterations.

Tasks were built in GATE and jobs automatically managed through through Crowdflower (Bontcheva et al., 2014b). First segments were entirely annotated and adjudicated by experts as calibration. To maximize annotator focus, images attached to tweets or featuring in content (e.g. news stories) linked to from tweets are shown alongside the task, for worker priming (Morris et al., 2012).

A majority of crowd workers use crowd work as their primary income. We calculate task mean completion times and reward work so that it pays *above* the minimum wage in our country.

3.1 Annotator recall

To annotate text with diverse content, annotators with diverse knowledge are needed. Typical named entity cues, such as capital first letters, are often missing in social media; extra knowledge is one way of compensating for these. We introduce the concept of "annotator recall", which describes the ability of annotators to identify NE mentions, partly based on their own knowledge. The breadth of social media content makes annotation harder for experts. For example, none of the experts we asked during annotation (or the attendees of talks we gave on the topic) could initially explain the entity *KKTNY* in the text "*KKTNY in 45mins!!!!!*", without referring to external resources. However, the crowd, being more diverse, was sometimes able to identify and ground this expression.[5]

To assess annotator recall, we compared expert and crowd annotation over segment A. First, this segment was annotated by a mixture of expert annotators, with each document doubly-annotated and results expert adjudicated. Crowd annotators were then also asked to annotate this data using the same guidelines, and the results compared, shown in Table 3.

In general, the crowd found more entities than experts (Table 3). That is to say, crowd recall over the oracle annotation of the data was higher. However, agreement was lower in the crowd. This was handled by including human expert adjudication over all segments. Further, to improve corpus-level knowledge diversity, each worker was limited to a maximum of 50 tasks.[6] Details of the adjudication process are given in Section 3.2.

[5]The entity colloquially refers to a television program, "Kim and Kourtney Take New York"; from Ritter et al. (2011) data.

[6]Each task involves annotating five tweets for one entity type; each tweet was annotated by five to seven different workers.

Agreement level	IAA (max-recall vs. expert)	IAA (naïve)
Whole document	0.839	n/a
Person entity	0.920	0.799
Location entity	0.963	0.861
Organization entity	0.936	0.954
All entities	**0.940**	**0.877**

Table 4: Inter-annotator agreement breakdown.

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

Table 5: Corpus size statistics

After merging in these crowd annotations, a random set of 100 documents were taken from this segment and used as gold test data in later tasks. Gold test data is used to assess performance of individual workers online, by being seeded among real annotation tasks. Performance on gold test data estimates the quality of an annotator's work. Annotators are made aware of results of this process and can highlight incorrect gold test data. The gold test data also provides qualification training for workers; they must perform well on gold test data before accessing real tasks. This gold test set was used throughout the remainder of the corpus annotation, providing quality control and crowd annotator training.

Spatial variation also played a role in crowd annotator recall. Based on the crowd vs. expert comparison data, there were multiple cases where the annotation was deemed "easy" by experts, but the crowd would still miss it. Investigation suggested that annotators based in the same country as the origin tweet had better recall on these entities. Conversely, annotators working on documents from other countries had lower recall. As matched geographic contexts tend to produce better results, during annotation, the geographically stratified parts of corpora were issued only to crowd workers in the same region, in order to maximize recall and local knowledge.

3.2 Adjudication

Adjudication is an important step in refining annotator data. In the case of crowdsourced annotations, further economies of scale can be afforded by automating adjudication; tools already exist for this, such as MACE (Hovy et al., 2013). However, auto-adjudication is poorly equipped to handle exceptional circumstances; further, it is hard to judge its impact without human intervention. The construction of the BTC involved a combination of automatic and human adjudication.

Primarily, we found there were problems with recall. Often, only a single crowd worker or expert would annotate a given (correct) entity. Under traditional agreement-based measures, this singleton annotation would be in the minority, and so likely removed in a typical adjudication step. Given our and others' experiences with annotator recall and diversity (Balog et al., 2012; Difallah et al., 2013), we find this method inappropriate. Further, this annotator behavior has implications for inter-annotator agreement (IAA); perfectly adequate aggregate annotations will nevertheless have reduced agreement. Therefore, naïve IAA measures (such as Fleiss' κ) risk under-reporting multiple annotator performance in high-diversity text. Also, while IAA is sometimes used as a correlate of task understanding, here it also reflects annotator world knowledge.

Workers are continually monitored using 100 documents from segment A, which removes workers having too low performance on reference tasks. This ensures annotator quality while simultaneously permitting variance in annotator knowledge.

To manage annotator sparsity (see Section 3.1), we experimented with a simple automatic adjudication rule. Any time a span is annotated by a worker, that span is placed in the final set. Adjacent annotations of the same type are concatenated. This is the "max-recall" method. We apply max-recall to all original expert and crowd annotations; then, an expert adjudicator evaluates annotator max-recall output to provide a final gold version. Hapax legomena are not unusual, given the diverse subject matter, though only retained after best-effort verification by the expert. Agreement figures are given in Table 4, as well as standard annotator performance.

In our experiments, we measured naïve IAA as the proportion of annotators agreeing on each token of an entity. This gives a per-entity agreement figure. The per-entity is then micro-averaged across the entire corpus, to show the level of annotator agreement, with some ability to account for variance in

entity bounds. This method does not give generous score, but fits our scenario of having many annotators, which makes pairwise comparison awkward. Indeed, even under this measurement regime, IAA levels were respectable.

3.3 Results

In total, the final corpus contained 9 551 documents. In these, there were 165K tokens and 12K entity mentions. Details are given in Table 5; the comparison of this dataset to other social media named entity corpora is in Table 1. Over the annotation process, we collected 125K annotations from 755 workers. These comprised eight experts and a total of 747 crowd workers. Inter-annotator agreement was 0.94, with max-recall autoadjudication followed by a final expert confirmation adjudication step.

4 Drift Analysis

We have described three dimensions for variation: temporal, spatial, and social. This section examines the difference in entities found across the corpus.

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 6: Volume of tweets in corpus by year; darker background means higher volume.

4.1 Data Selection Comparison

The selection of texts affects the kinds of entities and entity contexts that will be found. Getting a good range is important, as discussed in Section 2. The corpus aims to establish a good spatial, temporal and social range of NEs. This section describes the variation present in the corpus along each of these dimensions, and compares it to two other major NER datasets: the CoNLL 2003 shared task data (Tjong Kim Sang and Meulder, 2003), and Ritter et al. (2011)'s NER data.[7] The other social media corpora mentioned earlier are distributed as plain text, without timestamps or messages IDs, which precludes their analysis.

The temporal distribution is described and compared to the CoNLL2003 and Ritter datasets by year (Table 6), month (Table 7), day of month (Table 8), day of week (Table 9), and time of day (Table 10).[8] Temporal data is based on the local time. Note that both the Ritter and CoNLL data are confined to narrow ranges, the former having been collected in a short period on one day, and the latter having training data from a few days in the summer of 1996 and test data from later the same year (6/7 December). These two corpora are therefore constrained to anachronistic terminology, with the Ritter data also having socially constrained data (i.e., it is only from people active during one Friday evening – September 17, 2010).

4.2 Spatial Entity Diversity

Different regions are likely to mention different entities. The corpus contains two segments that are subdivided into various country-specific strata. To measure the spatial diversity of named entities, we compare the density of entity mentions that are specific to each region, against the rest of the segment.

[7] Many thanks to Alan Ritter for providing references to the original tweet IDs, which allowed this metadata to be captured.
[8] No time of day data was present in the CoNLL dataset.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Our corpus	2308	68	502	862	1074	1056	1321	850	342	419	23	21
Ritter (2011)	0	0	0	0	0	0	0	0	6902	0	0	0
CoNLL'03	0	0	0	0	1	0	0	1131	1	0	1	224

Table 7: Volume of tweets in corpus by month; darker background means higher volume.

Day of month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Our corpus	559	162	187	163	186	191	541	174	187	197	200	545	201	165	274	265	905	276	254	253	203	170	667	304	217	273	303	250	207	230	137
Ritter (2011)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6902	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CoNLL'03	1	0	0	0	8	165	51	0	0	0	0	0	0	0	0	0	0	0	0	0	7	125	103	66	85	111	121	168	130	107	110

Table 8: Volume of tweets in corpus by day of month; darker background means higher volume.

Weekday	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Our corpus	999	1115	2062	2016	1019	1027	608
Ritter (2011)	0	0	0	0	6902	0	0
CoNLL'03	111	122	174	262	376	227	86

Table 9: Volume of tweets in corpus by day of week.

Time of day	morning	afternoon	evening	night
Our corpus	1873	3299	2156	1518
Ritter (2011)	0	0	6902	0

Table 10: Volume of tweets by time of day.

Novel entities are calculated as the proportion of all entities that occur uniquely in one region. This is based on the assumption that if entities are distributed evenly across the world, i.e. region has no effect, then entity mentions sampled from any given region will roughly match the total. Conversely, if entity mentions do vary by region, then the proportion of novel entities in any coherent region will be higher. For calibration, we also include the global portion of segment F and all of the global segment H. Singleton mentions are removed to smooth the data. Results are given in Table 11, showing high regional variance. Most regions' surface forms are novel, i.e. unique to that region; the U.S. may have the lowest proportion of novel NEs because U.S. users make up a large proportion of global users. The "global" category has the largest size. This is required to give a broad enough picture of globally common entities to make inter-regional comparisons meaningful. The side effect a disproportionate novel part.

5 Entities in Social Media

Having examined diversity in the underlying text, we next analyze characteristics of entities. We qualitatively examine surface forms, and compare entity distribution in social media to that in newswire.

5.1 Common Surface Forms

Table 12 presents the most frequent surface forms in our corpus and also in the CoNLL'03 NER annotated data. The latter comes from news, based on the RCV1 corpus, which is largely US-based newswire from the 1990s (Rose et al., 2002) written by white working-age men (Eisenstein, 2013).

Temporal concept drift (Masud et al., 2010) is evident here. For example, the most frequently-mentioned person entities have different surface forms, while referring to the same concept. The lexical representation of "the President of the US" has changed from *Clinton* to *Obama*. Similarly, the leader of Russia is present but with a different word; *Yeltsin* in the older newswire, *Putin* in modern social media.

The top locations mentioned remain largely the same level and granularity, being countries that are major actors on the global scale or in the Anglosphere. Extra presence in the social media data of items like *Dublin*, *Ontario* and *Melbourne* may be attributable to our sampling, which more evenly distributed across English speaking nations than a population-weighted or social media presence-weighted approach.

We also see that celebrity figures, such as *@justinbieber* and *Kate Middleton*, are more prevalent in the social media top ranking – as are journalists, such as *@timhudak* and *David Speers*. However, the CoNLL data does contain a large number of sportsmen, as it is rich in cricket reportage; e.g. *Wasim*

[9]Segment E has Russian and Ukrainian media coverage, promoting mentions of this entity. This excludes RT as "retweet".

Region	Distinct surface forms	Forms only in this category	% Novel
Australia	116	95	81.90
Canada	109	94	86.24
Ireland	105	83	79.05
New Zealand	58	52	89.66
United Kingdom	203	144	70.94
United States	135	84	62.22
Global	628	582	92.68

Table 11: Proportions of entities specific to individual regions on social media.

Person		Location		Organization	
Our corpus	**CoNLL'03**	**Our corpus**	**CoNLL'03**	**Our corpus**	**CoNLL'03**
Obama	Clinton	UK	U.S.	Independent	Reuters
President Obama	Yeltsin	Ukraine	Germany	Irish News	U.N.
Kate Middleton	Arafat	US	Australia	RT *(Russia Today)*[9]	CHICAGO
@justinbieber	Lebed	London	France	Twitter	NEW YORK
Putin	Dole	Canada	England	Facebook	OSCE
JudithCollins	Wasim Akram	Iraq	Russia	Malaysia Airlines	NATO
@SimoLove	Waqar Younis	Russia	Britain	BBC	Interfax
Prince William	Mushtaq Ahmed	Australia	Italy	YouTube	EU
James Foley	Dutroux	Irish	China	Reuters	Barcelona
Harper	Croft	Gaza	LONDON	twitter	KDP
David Cameron	Netanyahu	U.S.	Spain	Liverpool	DETROIT
Cameron	Bill Clinton	Dublin	Japan	Labour	USDA
Princess Beatrice	Aamir Sohail	NZ	Sweden	CNN	PUK
Tony Abbott	Mullally	Ireland	Israel	Arsenal	MINNESOTA
Kate	Mother Teresa	Ontario	Pakistan	WorldCup	BOSTON
@timhudak	Wang	Melbourne	NEW YORK	Apple	ST LOUIS
@RossMarowits	Saeed Anwar	USA	Iraq	UN	PHILADELPHIA
NICOLA	Moin Khan	China	Belgium	Guardian	TORONTO
@David_Speers	Salim Malik	Syria	London	Google	Surrey
Zara Phillips	Rubin	Scotland	United States	EU	European Union

Table 12: Top 20 most frequent surface forms of each entity type, in the CoNLL'03 data (newswire from RCV1) and from our Twitter data.

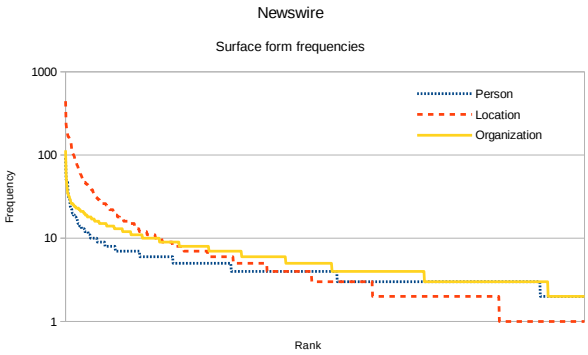

Figure 1: Frequency-Rank curve for entities in CoNLL'03 data.

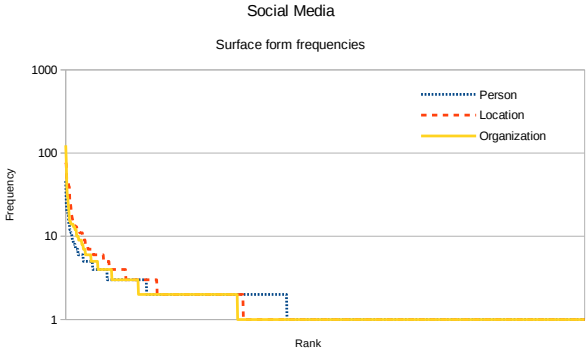

Figure 2: Frequency-Rank curve for entities in the Broad Twitter Corpus.

Akram and *Moin Khan*. This may be attributable to the specific social stratification of the newswire subsample found in the CoNLL data; i.e., it covers little outside of global and business affairs, and only one sport. Alternatively, a cricket championship may have been underway during the narrow date range from which this data was sampled. In any event, one would not expect these names to occur in modern datasets. Further, context around cricketers probably makes poor training examples for general Person entities. In contrast, social media data focuses on a broad range of popular figures and even personal names not found in the public sphere (like *NICOLA*).

Regarding organizations, both datasets are rich in sports clubs, with the typical location/organization metonymy found in that regard (e.g. *DETROIT* can refer to a place or sports club). The CoNLL data is rich in major league baseball data, but poor in its editorial reportage, explaining the high incidence of U.S. sport club mentions but with no frequent baseball personalities. BTC data includes large social media and other internet-based enterprises among its most frequently mentioned organizations. Both datasets also frequently discuss news outlets, possibly due to self-promotion in their own coverage.

5.2 Surface Form Distribution

The frequencies of entity mentions can be used to describe the corpus. We measure frequency-rank curves over the Broad Twitter Corpus and compare them to newswire data (the CoNLL'03 dataset). Frequency rank curves for all three entity types are shown in Figures 1 and 2.

	Social media			Newswire		
	Person	**Location**	**Organization**	**Person**	**Location**	**Organization**
Total weight	5 271	3 114	3 732	10 053	10 572	9 268
h-index	11	15	14	18	44	23
Tail weight	5 048	2 547	3 263	9 399	6 098	8 425
Head proportion	4.23%	18.21%	12.57%	6.51%	42.32%	9.10%

Table 13: Measuring the contribution of the head of entity surface form frequency/rank curve.

In a corpus covering a very narrow topic range, one might expect a small range of entities to be mentioned frequently, and comprise the majority of entity mentions in that corpus. This could be called head-heavy, as the entity mention frequency mass is concentrated in the head of the curve, not the tail. To measure the head-heaviness of these distributions, we first determine the h-index for each entity class (i.e. the lowest entity frequency that is larger than the number of times it is seen) and use this to bisect the curve into head and tail. The frequency mass of the head is then measured as a proportion of the whole. Results are given in Table 13, without disambiguating metonymies. Note that, while in general the head makes up for fewer of the entities in social media when compared to newswire, this is not the case for organization entities. These are slightly more diverse in newswire, with the mass of the head making up only 9.1% of all mentions, compared to 12.57% in tweets. This may be due to the focus on major league baseball teams in the news data.

6 Conclusion

Social media is an important resource and presents many challenges, though the paucity and bias of existing datasets hampers NER research in this text type. This paper presents a large, high-quality corpus for social media that addresses these problems, and demonstrated the breadth and quality of the annotated data. The dataset can be freely downloaded from `http://www.gate.ac.uk/wiki/broad-twitter-corpus.html`, including both all original text (under Twitter license) and also tools for reconstructing the annotated corpus.

Acknowledgements

We are grateful to all our annotators, including among others Dominic Rout, Diana Maynard, Konstantin Yershov, Marta Sabou, Roland Roller, Michał Łukasik, Johann Petrak, Patrick Paroubek, Nattapong Sanchan, Gerhard Wohlgennant, Adam Funk, Amel Fraisse, Arno Scharl, and our many crowd workers across the globe. Alan Ritter of Ohio State University gave us excellent help with references to the original JSON for the 2011 corpus, pivotal to our analyses. Graham Dove of Aarhus University gave tips for our visualisations, which benefit both us and hopefully the reader. This research was supported by the EU under FP7 grant No. 611223, PHEME, and by the UK EPSRC under the CHIST-ERA scheme from grant No. EP/K017896/1, uComp.

References

Omar Alonso and Matthew Lease. 2011. Crowdsourcing for information retrieval: principles, methods, and applications. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1299–1300. ACM.

Krisztian Balog, Yi Fang, Maarten de Rijke, Pavel Serdyukov, and Luo Si. 2012. Expertise retrieval. *Foundations and Trends in Information Retrieval*, 6(2–3):127–256.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani. 2013. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics.

Kalina Bontcheva, Leon Derczynski, and Ian Roberts. 2014a. Crowdsourcing named entity recognition and entity linking corpora. *The Handbook of Linguistic Annotation (to appear)*.

Kalina Bontcheva, Ian Roberts, Leon Derczynski, and Dominic Rout. 2014b. The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12.

Munmun De Choudhury, Scott Counts, Eric Horvitz, and Michael Gamon. 2013. Predicting depression via social media. In *Proceedings of ICWSM*. AAAI, July.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Associational for Computational Linguistics Workshop of Computational Linguistics and Clinical Psychology*.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing and Management*, 51:32–49.

Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. Pick-a-crowd: tell me what you like, and I'll tell you what to do. In *Proceedings of the 22nd international conference on World Wide Web*, pages 367–374. ACM.

Jacob Eisenstein, Brendan O'Connor, Noah Smith, and Eric P. Xing. 2010. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369.

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88.

Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating NER for Twitter #drift. In *Proceedings of LREC*, pages 2544–2547. European Language Resources Association.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proc. of NAACL-HLT*, pages 1120–1130.

Eduard Hovy. 2010. Annotation. In *Tutorial Abstracts of ACL*.

Yuheng Hu, Kartik Talamadupula, Subbarao Kambhampati, et al. 2013. Dude, srsly?: The surprisingly formal nature of Twitter's language. *Proceedings of ICWSM*.

Chris Kedzie, Kathleen McKeown, and Fernando Diaz. 2015. Predicting salient updates for disaster summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1608–1617. Association for Computational Linguistics.

Dennis Kergl, Robert Roedler, and Sebastian Seeber. 2014. On the endogenesis of Twitter's Spritzer and Gardenhose sample streams. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 357–364. IEEE.

Shashank Khanna, Aishwarya Ratan, James Davis, and William Thies. 2010. Evaluating and improving the usability of Mechanical Turk for low-income workers in India. In *Proceedings of the first ACM symposium on computing for development*. ACM.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 359–367.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *The 50th Annual Meeting of the Association for Computational Linguistics*, volume 4, pages 25–30. ACL.

Walid Magdy and Tamer Elsayed. 2016. Unsupervised adaptive microblog filtering for broad dynamic topics. *Information Processing & Management*, 52(4):513–528.

Christopher Mascaro and Sean Patrick Goggins. 2012. Twitter as virtual town square: Citizen engagement during a nationally televised republican primary debate. In *APSA 2012 Annual Meeting Paper*.

Mohammad M Masud, Qing Chen, Latifur Khan, Charu Aggarwal, Jing Gao, Jiawei Han, and Bhavani Thuraisingham. 2010. Addressing concept-evolution in concept-drifting data streams. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 929–934. IEEE.

Robert R Morris, Mira Dontcheva, and Elizabeth M Gerber. 2012. Priming for better performance in microtask crowdsourcing environments. *Internet Computing, IEEE*, 16(5):13–19.

Mohamed M. Mostafa. 2013. More than words: Social networks text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241 – 4251.

Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. 2011. Safety information mining – what can NLP do in a disaster. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 965–973. Asian Federation of Natural Language Processing.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *Proceedings of the Fourth AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 384–385.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*, Edinburgh, UK.

Tony Rose, Mark Stevenson, and Miles Whitehead. 2002. The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources. In *LREC*, volume 2, pages 827–832.

Matthew Rowe, Milan Stankovic, Aba Sah Dadzie, B.P. Nunes, and Amparo Elizabeth Cano. 2013. Making sense of microposts (#msm2013): Big things come in small packages. In *Proceedings of the WWW Conference - Workshops*.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the 9th international conference on language resources and evaluation (LREC14)*, pages 859–866.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.

Zeynep Tufekci. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185.