How Noisy Social Media Text, How Diffrnt Social Media Sources?

Timothy Baldwin,[♠♡] **Paul Cook,**[♡] **Marco Lui,**[♠♡] **Andrew MacKinlay**[♠♡] **and Li Wang**[♠♡]
♠ NICTA Victoria Research Laboratory
♡ Department of Computing and Information Systems, The University of Melbourne
tb@ldwin.net, paulcook@unimelb.edu.au, mhlui@unimelb.edu.au,
Andrew.MacKinlay@nicta.com.au, li.wang.d@gmail.com

Abstract

While various claims have been made about text in social media text being noisy, there has never been a systematic study to investigate just how linguistically noisy or otherwise it is over a range of social media sources. We explore this question empirically over popular social media text types, in the form of YouTube comments, Twitter posts, web user forum posts, blog posts and Wikipedia, which we compare to a reference corpus of edited English text. We first extract out various descriptive statistics from each data type (including the distribution of languages, average sentence length and proportion of out-of-vocabulary words), and then investigate the proportion of grammatical sentences in each, based on a linguistically-motivated parser. We also investigate the relative similarity between different data types.

1 Introduction

Various claims have been made about social media text being "noisy" (Java, 2007; Becker et al., 2009; Yin et al., 2012; Preotiuc-Pietro et al., 2012; Eisenstein, 2013, inter alia). However, there has been little effort to quantify the extent to which social media text is more noisy than conventional, edited text types. Moreover, social media comes in many flavours — including microblogs, blogs, and user-generated comments — and research has tended to focus on a specific data source, such as Twitter or blogs. A natural question to ask is how different the textual content of the myriad of social media types are from one another. This is an important first step towards building a general-purpose suite of social media text processing tools.

Most research to date on social media text has used very shallow text processing (such as keyword-based time-series analysis), with natural language processing (NLP) tools such as part-of-speech taggers and parsers tending to be disfavoured because of the perceived intractability of applying them to social media text. However, there has been little analysis quantifying just how hard it is to apply NLP to social media text, or how intractable the data is for NLP tools.

This paper addresses the two issues above. We build corpora from a variety of popular social media sources, including microblogs, user-generated comments, user forums, blogs, and collaboratively-authored content. We then compare these corpora to more conventional texts through a variety of statistical and linguistic analyses to quantitatively assess the relative extent to which they are "noisy", and quantify similarities between them. Our findings indicate that there are certainly differences between social media sites, but that if we focus our attention on English text, there are striking similarities, and that even sources such as Twitter may be more "NLP-tractable" than they are often portrayed.

2 Background

Natural language processing (NLP) has been applied to a wide range of applications on social media, especially Twitter. Numerous studies have attempted to go beyond simple keyword and burstiness models to identify real-world events from Twitter (Benson et al., 2011; Ritter et al., 2012; Petrovic et al., 2012). Recent efforts have considered identifying user location based on the textual content of tweets (Wing and Baldridge, 2011; Roller et al., 2012; Han et al., 2012b) and user metadata (Han et al., 2013). Related work has examined models of the relationships between words and locations for the purpose of identifying and studying regional linguistic variation (Eisenstein et al., 2010; Eisenstein et al., 2012).

Given the abundance of non-standard language

International Joint Conference on Natural Language Processing, pages 356–364,
Nagoya, Japan, 14-18 October 2013.

on social media, including lexical variants (e.g. *supa* for *super*) and acronyms (e.g. *smh* for *shaking my head*), as well as genre-specific phenomena such as the usage of hashtags and mentions on Twitter, standard NLP tools cannot be immediately applied. Efforts to address this problem have taken two main approaches: modifying social media data to more closely resemble standard text, and building social media-specific tools.

Lexical normalisation is the task of converting non-standard forms such as *tlkin* and *touchdooown* to their standard forms (*talking* and *touchdown*, respectively), in the hopes of making text more tractable to NLP (Eisenstein, 2013). Approaches to normalisation have exploited various sources of information including the context in which a given instance of a lexical variant occurs (Gouws et al., 2011; Han and Baldwin, 2011), although the best results to date have been achieved by automatically discovering lexical variant–standard form pairs from a large Twitter corpus (Han et al., 2012a). This latter approach is particularly appealing because it allows for very fast normalisation, suitable for processing large volumes of text.

Conversely, Owoputi et al. (2013) and Ritter et al. (2011) developed part-of-speech (POS) taggers for Twitter that are better able to handle properties of this text type such as the higher out-of-vocabulary rate compared to conventional text. Ritter et al. further developed a Twitter shallow parser and named-entity recogniser. Foster et al. (2011) evaluated standard parsers on social media data, and found them to perform particularly poorly on Twitter, but showed that their performance can be improved through a retraining strategy.

Another natural question to ask is how similar the characteristics of social media text are to those of other domains. More specifically, we may be interested in a numerical measurement of how closely the language used in one corpus matches that of another. Kilgarriff (2001) proposed a method for calculating both inter-corpus similarity and intra-corpus homogeneity, and language modelling has also been used as the basis for calculating how well one corpus models another. We discuss both of these options below.

3 Datasets

In order to evaluate the characteristics of text in different social media sources, we assembled the following datasets from across the spectrum of popular social media sites, varying in terms of document length, the number of authors/editors per document, and the level of text editing:

TWITTER-1/2: micro-blog posts from Twitter, crawled using the Streaming API over two discrete time periods (TWITTER-1 = 22 September 2011 and TWITTER-2 = 22 February 2012) to investigate the temporal-specificity of the data — documents up to 140 characters in length, single author per document, and no facility for post-editing

COMMENTS: comments from YouTube, based on the dataset of O'Callaghan et al. (2012), but expanded to include all comments on videos in the original dataset[1] — documents up to 500 characters in length, single author per document, and no facility for post-editing

FORUMS: a random selection of posts from the top-1000 valid vBulletin-based forums in the Big Boards forum ranking[2] — documents of variable length (with a site-configurable restriction on maximum post length), single author per document, and optional facility for post-editing (depending on the site configuration)

BLOGS: blog posts from tier one of the ICWSM-2011 Spinn3r dataset (Burton et al., 2011) — generally no restriction on length, single author per document, and facility for post-editing

WIKIPEDIA: text from the body of documents in a dump of English Wikipedia — no restriction on document length, usually multiple authors/editors per document, and facility for post-editing

As a reference corpus of English from a non-social media source, we also include documents from the British National Corpus (Burnard, 2000):

BNC: all documents from the written portion of the British National Corpus (BNC) — documents of up to 45K words from a variety of sources, mostly by a single author, with editing.

We present the number of documents and average document size for each dataset in Table 1.

[1] We post-processed the retrieved comments to remove all occurrences of the unicode U+FEFF codepoint (which is used either as a byte order marker at the start of messages or a zero-width no-break space when used elsewhere in a document), as it skewed the results of the language identification.

[2] `http://rankings.big-boards.com`

Corpus	Documents	Average words per document
TWITTER-1	1 000 000	11.8 ± 8.3
TWITTER-2	1 000 000	11.6 ± 8.1
COMMENTS	874 772	15.8 ± 18.6
FORUMS	1 000 000	23.2 ± 29.3
BLOGS	1 000 000	147.7 ± 339.3
WIKIPEDIA	200 000	281.2 ± 363.8
BNC	3141	$31\,609.0 \pm 30\,424.3$

Table 1: Number of documents and average document size (mean±standard deviation, in words) for each dataset

TWITTER-1/2 and COMMENTS, predictably, contain the shortest documents, with 12–16 words per document on average. Forum posts are around twice the length on average (but the spread of document lengths is considerably greater). Blog posts, on average, contain around ten times the number of words of a forum post, with a greater spread again of document lengths and longer sentences. Amongst our social media sources, Wikipedia documents are by far the longest, but considerably shorter than BNC documents.

4 Corpus Pre-processing

We first pre-process each dataset using the following standardised methodology.[3] In the case that the corpus comes with tokenisation and POS information, we strip this and perform automatic preprocessing to ensure consistency in the quality and composition of the tokens/tags.

We first apply `langid.py` (Lui and Baldwin, 2012) — an off-the-shelf language identifier — to each document to detect its majority language. We then extract all documents identified as English for further processing.

We next perform sentence tokenisation. In line with the findings of Read et al. (2012a) based on experimentation with a selection of sentence tokenisers over user-generated content, we sentence-tokenise with `tokenizer`.[4]

Finally, we tokenise and POS tag the datasets using `TweetNLP` 0.3 (Owoputi et al., 2013).

One particularly important property of `TweetNLP` is that it identifies content such as mentions, URLs, and emoticons that aren't typically syntactic elements of a sentence. Moreover, it is able to distinguish between usages of hashtags which are elements of a sentence, and those which are not, as in the case of Examples (1) and (2) below, respectively.

(1) love this #awesome view out of my window

(2) Swinging with the besties! #awesome

We POS tag each sentence in each corpus using `TweetNLP`, and remove all tokens identified as non-linguistic.[5] In our examples above, e.g., we remove the token *#awesome* from (2) but not (1).

To normalise for corpus size, we extract a random sample of sentences totalling 5M tokens from each dataset, and further partition this sample into 5 equal-sized sub-corpora.

5 Analysis

In this section, we analyse the characteristics of the language used in the respective data sources.

5.1 Language Mix

First, we analyse the breakdown of languages found in each data source based on the predictions of `langid.py`, as detailed in Table 2. Note that these results are based on the full datasets without language filtering. Also note that WIKIPEDIA and the BNC are intended to be monolingual English collections, and that FORUMS has a strong bias towards English due to the crawling methodology. For the remainder of the datasets, we expect the results to be representative of the language bias of the respective data sources.

All data sources are dominated by English documents, although in the case of TWITTER-1/2, less than half of the documents are in English (`en`), with Japanese being the second most popular language, and strong representation from languages such as Portuguese (`pt`), Spanish (`es`), Indonesian (`id`), Dutch (`nl`) and Malay (`ms`). These results are largely consistent with earlier studies on the language distribution in Twitter (Semiocast, 2010; Hong et al., 2011).

That the BNC is predicted to be 100% English is a validation of the accuracy of `langid.py`. WIKIPEDIA is more interesting, with tiny numbers (around 0.2% in total) of documents which are predicted to have a majority language of Latin (`la`), German (`de`), etc. Manual analysis of these

[3] Acknowledging that superior domain-specific approaches exist, e.g. for Wikipedia sentence tokenisation using markup (Flickinger et al., 2010).

[4] `http://www.cis.uni-muenchen.de/~wastl/misc/`

[5] Specifically, we remove any token tagged as #, @, ~, U, or E.

TWITTER-1		TWITTER-2		COMMENTS		FORUMS		BLOGS		WIKIPEDIA		BNC	
en	.406	en	.439	en	.757	en	.914	en	.784	en	.998	en	1.000
ja	.144	ja	.124	de	.034	de	.016	ru	.050	la	.000		
pt	.098	es	.091	es	.028	es	.011	fr	.025	de	.000		
es	.093	pt	.072	fr	.023	ro	.009	zh	.022	fr	.000		
id	.031	id	.029	ru	.023	it	.007	de	.019	es	.000		
nl	.025	nl	.022	pt	.020	nl	.007	es	.017	no	.000		
ms	.016	ar	.019	pl	.012	fr	.006	ja	.010	he	.000		
ko	.015	ko	.018	ar	.011	pl	.003	it	.010	zh	.000		
de	.015	ms	.015	it	.011	da	.002	pt	.009	ja	.000		
it	.013	fr	.015	nl	.006	sv	.002	sv	.008	pt	.000		

Table 2: Top-10 languages (by ISO-639-1 identifier) in each dataset

documents reveals that most are made up of lists of different types: names of people from a variety of ethnic backgrounds, foreign place names, or titles of artworks/military honours in various languages. As such, the language tags are actually overwhelmingly correct,[6] in the sense that the predominant language is indeed that indicated.

The implications of these results for text processing of social media are profound. While English clearly dominates the data, there are significant amounts of non-English text in all our social media sources, with Twitter being the most extreme case: the majority of documents are *not* English. Additionally for TWITTER-1/2 and COMMENTS, instances of all 97 languages modelled by `langid.py` were found in the dataset. At the very least, this underlines the importance of language identification as a means of determining the source language in cases where language-specific NLP tools are to be used.

5.2 Lexical Analysis

Next, we analyse the lexical composition of the English documents. Hereafter, we focus exclusively on the 5M token subsample of each dataset.

In Table 3 we present simple statistics on the average word length (in characters) and average sentence length (in words) for each dataset. We also analyse the relative occurrence of out-of-vocabulary (OOV) words, based on the GNU aspell dictionary v0.60.6.1 with case folding. We strip all "online-specific" markup (hashtags, user mentions and URLs), on the basis of the output of the POS tagger (i.e. any hashtags etc. that are *not* part of the syntactic structure of the text are removed).[7] To filter out common misspellings/social media usages such as *ur* for *your*, we optionally include a pre-step of "lexical normalisation" based on the dictionary of Han et al. (2012a) which gives the standard form for a given OOV, based on combined information from slang dictionaries and automatically-learnt correspondences ("+norm").

Corpus	Word length	Sentence length	%OOV	
			−norm	+norm
TWITTER-1	3.8±2.4	9.2±6.4	.246	.225
TWITTER-2	3.8±2.4	9.0±6.3	.240	.222
COMMENTS	3.9±3.2	10.5±10.1	.198	.184
FORUMS	3.8±2.3	14.2±12.7	.181	.171
BLOGS	4.1±2.8	18.5±24.8	.206	.203
WIKIPEDIA	4.5±2.8	21.9±16.2	.190	.188
BNC	4.3±2.8	19.8±14.5	.169	.168

Table 3: Average word and sentence length, and proportion of OOV words (optionally with lexical normalisation) in each dataset

There is remarkably little difference in word length between datasets, but sentence length in TWITTER-1/2 and COMMENTS is around half that of the more formal WIKIPEDIA/BNC and also BLOGS, with FORUMS splitting the difference. The average word length for all of TWITTER-1/2, COMMENTS and FORUMS is remarkably similar. In terms of OOV words, FORUMS and COMMENTS are comparable to WIKIPEDIA and the BNC (where OOV words are dominated by proper nouns), and actually lower than BLOGS. TWITTER-1/2 has the highest OOV rate of all our datasets, although when we include lexical normalisation, it is only 2–4 percentage points higher than the other social media sources. The impact of lexical normalisation is most noticeable for TWITTER-1/2 and COMMENTS, indicating that informal text and "ad hoc" spellings are more prevalent in them than the other data sources.

[6] With the notable exception of Latin, where many of the documents contain lists of names from a variety of European language backgrounds, but little that is identifiable as Latin.

[7] This step reduced the OOV rate in TWITTER-1/2 by about one third; it also reduced the OOV rate in COMMENTS by around 10%.

These results are broadly in agreement with the findings of Rello and Baeza-Yates (2012), who used the relative frequency of a set of common misspellings to estimate the lexical quality of social media, and arrived at the conclusion that social media text is on average "cleaner" than many other web sites, and becoming progressively cleaner over time.

5.3 Grammaticality

A natural next question to ask is how grammatical the text in each of our datasets is. We measure this using the English Resource Grammar (ERG: Flickinger et al. (2000)), a broad-coverage HPSG-based grammar. One aspect of the ERG which makes it highly suited to testing grammaticality is that, unlike most NLP parsers, it is "generative", i.e. it explicitly models grammaticality, and is developed relative to both positive and negative test items to ensure it does not "overgenerate". We can therefore use it as a proxy for grammaticality judgements. Further to this, the ERG makes active use of 'root conditions' to indicate how much the grammar had to relax particular assumptions to produce a derivation for the sentence. These conditions vary on the dimensions of: (1) strict versus informal (corresponding to whether the sentence uses standard punctuation and capitalisation, or not); and (2) full sentences vs. fragments (e.g. isolated noun phrases). All of our experiments are based on the '1111' version of the grammar, and the CHEAP parsing engine (Callmeier, 2002).

In order to maximise the lexical coverage of the ERG, we used POS-conditioned generic lexical types (Adolphs et al., 2008), whereby a generic lexical entry is created for each OOV word on the basis of the output of a POS tagger. To accommodate the `TweetNLP` POS tags, we manually created a new set of mappings to generic lexical entries.[8] We additionally re-tokenised the output of `TweetNLP` to split apart contractions (e.g. *won't* and possessive clitics (e.g. *Kim's*), in line with the Penn Treebank tokenisation strategy.

In Table 4 we show the results of parsing 4000 randomly selected English sentences from each corpus using the ERG with the parsing setup we have described.[9]

The highest parse coverage was observed for the BNC (with only 23.2% not able to be parsed), closely followed by WIKIPEDIA. At the other end of the scale are the TWITTER-1 and TWITTER-2 variants, which are most likely to contain ungrammatical sentences, with up to 15% more sentences unable to be parsed, although this is only marginally higher than FORUMS and BLOGS, all of which contain more ungrammatical text than COMMENTS.

Corpus	Parseable				Unparseable
	strict		informal		
	full	frag	full	frag	
TWITTER-1	13.8	23.9	22.2	2.5	37.4
TWITTER-2	13.9	23.8	22.8	1.7	37.6
COMMENTS	18.0	22.2	26.4	1.4	31.9
FORUMS	23.9	14.1	24.7	1.5	35.6
BLOGS	25.6	17.5	18.8	2.7	35.3
WIKIPEDIA	48.7	4.5	18.9	1.5	26.2
BNC	38.4	12.0	24.0	2.2	23.2

Table 4: Percentage of sentences (from a random sample of 4000) which can be parsed using the ERG, broken down by the root condition of the top-ranked parse for the parseable sentences

Between these extremes are some mild surprises — BLOGS and FORUMS, which contain data produced in a more enduring and editable format than TWITTER-1/2, are, according to our metric, only marginally more grammatical. In addition, the non-editable and relatively transient COMMENTS sentences are substantially more likely to be grammatical than either FORUMS or BLOGS. A large part of this effect however is probably due to the sentence length differences between the corpora. As shown in Table 3, the average length for COMMENTS is only 10.5 words, on par with TWITTER-1/2 (but according to this evidence, more carefully constructed). However, in the longer sentences of FORUMS and BLOGS, there is more scope for the authors to introduce anomalies into the text, increasing the chances of the sentence being unparseable.

Examining the root conditions related to formality and fragment analyses also gives us im-

[8] The original POS mappings are based on the Penn POS tagset and have been tested and fine-tuned extensively; our POS mapping for the `TweetNLP` POS tags is much more immature, and has potentially contributed to a slight loss in parser accuracy relative to the "canonical" ERG.

[9] Note that the reported results differ significantly from the coverage numbers reported by Read et al. (2012b) for WIKIPEDIA in particular, through a combination of a generic sentence and word tokenisation strategy, a potentially lower-accuracy/coarser-grained POS tagger, and a less mature POS mapping. The impact of these factors should be constant across datasets, however, meaning that the relative numbers should be truly indicative of the relative grammaticality of their text content.

Corpus	Fragment	Preprocessor error	Resource limitations	Ungrammatical inputs	Extra-grammatical	Grammar gaps
TWITTER-1	0.16	0.24	0.00	0.32	0.09	0.18
TWITTER-2	0.19	0.22	0.00	0.31	0.10	0.17
COMMENTS	0.13	0.32	0.00	0.31	0.04	0.20
FORUMS	0.05	0.31	0.01	0.36	0.03	0.24
BLOGS	0.09	0.22	0.11	0.11	0.22	0.25
WIKIPEDIA	0.08	0.11	0.10	0.06	0.06	0.59
BNC	0.15	0.05	0.15	0.04	0.05	0.56

Table 5: A breakdown of the causes of parser error in the unparseable sentences for each dataset

portant insights into the corpora. WIKIPEDIA has by far the highest percentage of sentences with a strict, non-fragment analysis, much higher (10.3%) than the BNC even. In the less-edited corpora, of those sentences which are able to be parsed, a much smaller percentage are strict or full analyses, with the strict fragment analyses being most prevalent in TWITTER-1/2 and informal full analyses dominating in COMMENTS and FORUMS.

The spread of grammaticality numbers is perhaps not as large as we might have expected. There are a few reasons for this. One important point is that the POS-tagging using a very coarse-grained tag set has inevitably led to very general lexical entries for handling unknown words (so we are not even sure of the person, number and tense associated with a verb). This means that it is possible that some of the sentences have been spuriously identified as grammatical, since the very general types for unknown words give the grammar great flexibility in fitting a parse tree to the sentence, even where it may not be appropriate. Secondly it is possible that this POS-tagging has led to an explosion in the number of candidate parse trees, which can paradoxically lead to a small decrease in coverage over longer sentences of WIKIPEDIA and the BNC due to the risk of exceeding the parser timeout or memory limit.

In line with Baldwin et al. (2005), it is possible to shed further light on the quality of the grammaticality judgements, and also stylistic differences between the different corpora by manually analysing the unparseable sentences according to the cause for parse failure, as being due to: (1) a syntactic fragment (not explicitly handled by the ERG; e.g. noun and verb phrase fragments such as *coming home ...*, or standalone expletives such as *wow!*); (2) a preprocessor error (e.g. in sentence tokenisation or POS tagging); (3) parser resource limitations (usually caused by the grammar running out of edges in the chart, or timing out); (4) ungrammatical strings; (5) extragrammatical strings (where non-linguistic phenomena associated with the written presentation, such as bullets or HTML markup, interface unpredictably with the grammar); and (6) lexical and constructional gaps in the grammar. A breakdown of parse failure over a randomly-selected subset of 100 unparseable sentences from each of the datasets, carried out by the first author, is presented in Table 5.

It is clear that the proportion of ungrammatical sentences is an underestimate, especially in the case of WIKIPEDIA and the BNC, where more than half of the "failures" are attributable to lexical or constructional gaps in the grammar.[10] For TWITTER-1/2, COMMENTS and FORUMS, however, the proportion of grammar gaps and genuinely ungrammatical inputs, respectively, is roughly equivalent, suggesting that our original findings for these datasets are an underestimate of the actual proportion of ungrammaticality, but that the relative proportions are accurate.

An additional observation that can be made from Table 5 is that preprocessing is a common cause of parser failure, primarily in sentence tokenisation (with multiple sentences tokenised into one), and to a lesser extent in POS tagging, and also occasional errors in language identification (only observed in the TWITTER-1/2 data).

Reflecting back over the combined results for grammaticality, we can conclude that there is less syntactic "noise" in social media text than we may have thought, and that while there is no doubt that WIKIPEDIA and the BNC contain less ungrammatical text than the other datasets, the relative occurrence of syntactically "noisy" text in TWITTER-1/2, COMMENTS, FORUMS and

[10] Or, indeed, shortcomings in our POS mapping for unknown words, although again, the relative impact of this should be constant across datasets.

Corpus	Homogeneity
TWITTER-1	549
TWITTER-2	553
COMMENTS	613
FORUMS	570
BLOGS	716
WIKIPEDIA	575
BNC	542

Table 7: Corpus homogeneity using χ^2 (smaller values indicate greater self-similarity)

BLOGS is relatively constant.

There is partial concordance between these findings and those of Hu et al. (2013), who examined textual properties of Twitter messages relative to blog, email, chat and SMS data, and also a newspaper. They found that Twitter messages were more formal than chat and SMS messages, and more similar to email and blog text in composition, in making prevalent use of standard constructions and lexical items.

5.4 Corpus Similarity

So far we have examined the datasets individually. Next, we investigate how intrinsically similar in style and content the different datasets are. One possible approach to this is via calculation of "corpus similarity" between datasets and homogeneity within a given dataset. In one of the very few studies of measuring corpus similarity and homogeneity, Kilgarriff (2001) introduced a method based on χ^2, whereby we measure the similarity of two corpora as the χ^2 statistic over the 500 most frequent words in the union of the corpora. One limitation of Kilgarriff's method is that it is only applicable to corpora of equal size. We therefore use the five 1M token sub-corpora of each corpus in these experiments. We measure the similarity of two corpora as the average pairwise χ^2 similarity between their sub-corpora. We measure the homogeneity (or self-similarity) of a corpus as the average pairwise similarity between sub-corpora of that corpus.

The homogeneity scores in Table 7 indicate that social media text exhibits greater lexical variation (as captured by the χ^2 measure), and hence is less homogenous, than conventional text types (i.e. the BNC). TWITTER-1 and TWITTER-2 are the most homogenous of the social media corpora, and only fractionally less homogeneous than the BNC. BLOGS are much more diverse than the other corpora.

Turning to corpus similarity (Table 6), there appears to be a roughly linear partial ordering in the relative similarity between the corpora: TWITTER-1/2 $\equiv$ COMMENTS $<$ FORUMS $<$ BLOGS $<$ BNC $<$ WIKIPEDIA (as in, TWITTER-1/2 is more similar to FORUMS than it is to BLOGS, but more similar to BLOGS than the BNC, etc.). This can be observed most clearly based on the similarities of each other corpus with TWITTER-1/2 and WIKIPEDIA, but the similarities for all corpus pairs are consistent with this ordering. TWITTER-1 and TWITTER-2 are unsurprisingly the most similar corpora, with very little difference between the two crawls, suggesting that despite the real-time nature of Twitter, it is reasonably homogenous across time. We further see relatively high similarity between TWITTER-1/2 and COMMENTS, COMMENTS and FORUMS, and FORUMS and BLOGS.

5.5 Language Modelling

Language modelling provides an alternative to estimating corpus similarity, based on the perplexity of a dataset relative to language models (LMs) trained over other partitions from the same dataset, and also partitions from other datasets. We construct open-vocabulary trigram LMs with Good-Turing smoothing using SRILM (Stolcke, 2002).

For each corpus, we build 5 LMs, each trained on 4 of the available 1M word sub-corpora. We then use each model to compute the perplexity of the held-out sub-corpus from the same dataset, as well as all sub-corpora for each other dataset. The results are presented in Figure 1 in the form of a box plot over the 5 LMs for a given training corpus (although the variance between LMs is usually so slight that the "box" appears as a single point).

For each corpus, the lowest perplexity is obtained on the held-out data from the same corpus. Overall, these results agree with those for χ^2 similarity, namely that there is a continuous spectrum, with TWITTER-1/2 and WIKIPEDIA as the two extremes and COMMENTS, FORUMS, BLOGS and the BNC between them, in that order. Along this spectrum, COMMENTS, FORUMS and BLOGS form a cluster, as do the BNC and WIKIPEDIA.

Combining these results with those for χ^2 similarity, it would appear that FORUMS is the "median" dataset, which is most similar to each of the other datasets. The implication of this finding is that if a statistical model (e.g. for POS dis-

	TWITTER-1	TWITTER-2	COMMENTS	FORUMS	BLOGS	WIKIPEDIA
TWITTER-2	4.0	—	—	—	—	—
COMMENTS	63.7	62.4	—	—	—	—
FORUMS	91.8	90.6	62.3	—	—	—
BLOGS	115.8	119.1	128.4	61.7	—	—
WIKIPEDIA	347.8	360.0	351.4	280.2	157.7	—
BNC	251.8	258.8	245.2	164.1	78.7	92.5

Table 6: Pairwise corpus similarity ($\times 10^3$) using χ^2

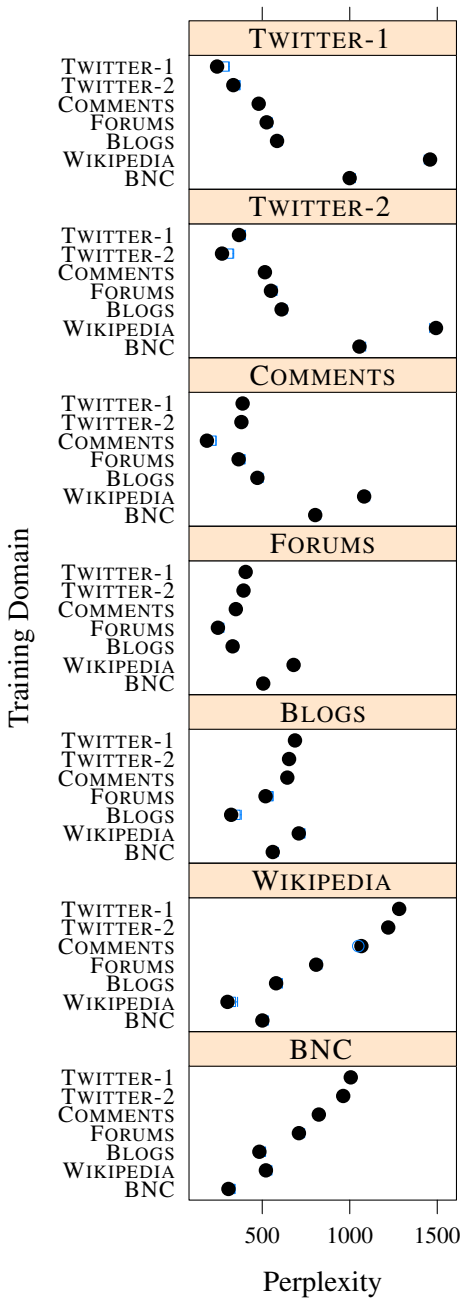

Figure 1: Trigram language model perplexity of test data conditioned on a given training corpus

ambiguation or parse selection) were to be trained on a single data type and applied to the other data types, FORUMS should be the data of choice, as with the possible exception of WIKIPEDIA, it models the other corpora remarkably well. It also provides evidence for why methods based on edited text collections such as the BNC or newswire text perform badly on Twitter data.

6 Conclusions

In this paper we built corpora from a range of social media sources — microblogs, user-generated comments, user forums, blogs, and collaboratively-authored content — and compared them to each other and a reference corpus of more-conventional, edited documents. We applied a variety of linguistic and statistical analyses, specifically: language distribution, lexical analysis, grammaticality, and two measures of corpus similarity. This is the first such systematic analysis and cross-comparison of social media text.

We analysed the widely-acknowledged "noisiness" of social media texts from a number of perspectives, and showed that NLP techniques — including language identification, lexical normalisation, and part-of-speech tagging — can be applied to reduce this noise. Crucially, this suggests that although social media is indeed noisy, it appears to be possible to use NLP to "cleanse" it. Moreover, once rendered less noisy, (further) NLP on social media text might be more tractable than it is conventionally believed to be.

In terms of grammaticality, our results confirmed that social media text is less grammatical than edited text, but also suggested that the disparity is relatively small.

Both of our more-general corpus similarity analyses revealed that the social media text types analysed appear to lie on a continuum of similarity ranging from microblogs to collaboratively-authored content. This finding has potential implications on the selection of training data for statistical NLP systems.

Acknowledgements

NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT centre of Excellence programme.

References

Peter Adolphs, Stephan Oepen, Ulrich Callmeier, Berthold Crysmann, Dan Flickinger, and Bernd Kiefer. 2008. Some fine points of hybrid natural language parsing. In European Language Resources Association (ELRA), editor, *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 1380–1387, Marrakech, Morocco.

Timothy Baldwin, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2005. Beauty and the beast: What running a broad-coverage precision grammar over the BNC taught us about the grammar — and the corpus. In Stephan Kepser and Marga Reis, editors, *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, pages 49–69. Mouton de Gruyter, Berlin, Germany.

Hila Becker, Mor Naaman, and Luis Gravano. 2009. Event identification in social media. In *Proceedings of the 12th International Workshop on the Web and Databases (WebDB 2009)*, Providence, USA.

Edward Benson, Aria Haghighi, and Regina Barzilay. 2011. Event discovery in social media feeds. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 389–398, Portland, USA.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Kevin Burton, Niels Kasch, and Ian Soboroff. 2011. The ICWSM 2011 Spinn3r dataset. In *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM 2011)*, Barcelona, Spain.

Ulrich Callmeier. 2002. PET – a platform for experimentation with efficient HPSG processing techniques. In Stephan Oepen, Dan Flickinger, Jun'ichi Tsujii, and Hans Uszkoreit, editors, *Collaborative Language Engineering*. CSLI Publications, Stanford, USA.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA.

Jacod Eisenstein, Brendan O'Connor, Noad A. Smith, and Eric P. Xing. 2012. Mapping the geographical diffusion of new words. *Arxiv preprint arXiv*, 1210.5268.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 359–369, Atlanta, USA.

Dan Flickinger, Stephan Oepen, Hans Uszkoreit, and Jun'ichi Tsujii. 2000. On building a more efficient grammar by exploiting types. *Journal of Natural Language Engineering* (Special Issue on Efficient Processing with HPSG), 6(1):15–28.

Dan Flickinger, Stephan Oepen, and Gisle YtrestØol. 2010. WikiWoods: Syntacto-semantic annotation for English wikipedia. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. From news to comment: Resources and benchmarks for parsing the language of web 2.0. In *Proc. of the 5th International Joint Conference on Natural Language Processing*, pages 893–901, Chiang Mai, Thailand.

Stephan Gouws, Dirk Hovy, and Donald Metzler. 2011. Unsupervised mining of lexical variants from noisy text. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 82–90, Edinburgh, UK.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 368–378, Portland, USA.

Bo Han, Paul Cook, and Timothy Baldwin. 2012a. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012 (EMNLP-CoNLL 2012)*, pages 421–432, Jeju, Korea.

Bo Han, Paul Cook, and Timothy Baldwin. 2012b. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1045–1062, Mumbai, India.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. A stacking-based approach to twitter user geolocation prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013): System Demonstrations*, pages 7–12, Sofia, Bulgaria.

Lichan Hong, Gregorio Convertino, and Ed H. Chi. 2011. Language matters in Twitter: A large scale study. In *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM 2011)*, Barcelona, Spain.

Yuheng Hu, Kartik Talamadupula, and Subbarao Kambhampati. 2013. Dude, srsly?: The surprisingly formal nature of Twitters language. In *Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM 2013)*, Boston, USA.

Akshay Java. 2007. A framework for modeling influence, opinions and structure in social media. In *Proceedings of the 22nd Annual Conference on Artificial Intelligence (AAAI-07)*, pages 1933–1934.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea.

Derek O'Callaghan, Martin Harrigan, Joe Carthy, and Pádraig Cunningham. 2012. Network analysis of recurring YouTube spam campaigns. In *Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM 2012)*, pages 531–534, Dublin, Ireland.

Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, Atlanta, USA.

Sasa Petrovic, Miles Osborne, and Victor Lavrenko. 2012. Using paraphrases for improving first story detection in news and Twitter. In *Proc. of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 338–346, Montréal, Canada.

Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. 2012. Trendminer: An architecture for real time analysis of social media text. In *Proceedings of the ICWSM 2013 Workshop on Real-Time Analysis and Mining of Social Streams*, Dublin, Ireland.

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars JØorgen Solberg. 2012a. Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India.

Jonathon Read, Dan Flickinger, Rebecca Dridan, Stephan Oepen, and Lilja Øvrelid. 2012b. The WeSearch corpus, treebank, and treecache – a comprehensive sample of user-generated content. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1829–1835, Istanbul, Turkey.

Luz Rello and Ricardo Baeza-Yates. 2012. Social media is NOT that bad! the lexical quality of social media. In *Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM 2012)*, Dublin, Ireland.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 1524–1534, Edinburgh, UK.

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1104–1112, Beijing, China.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012 (EMNLP-CoNLL 2012)*, pages 1500–1510, Jeju Island, Korea.

Semiocast. 2010. Half of messages on twitter are not in English — Japanese is the second most used language. Technical report, Semiocast.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. of the International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver, USA.

Benjamin Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 955–964, Portland, USA.

Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. 2012. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 27(6):52–59.