Automatic Identification of Locative Expressions from Social Media Text: A Comparative Analysis

Fei Liu♠, Maria Vasardani♡ and Timothy Baldwin♠
♠ Department of Computing and Information Systems
♡ Department of Infrastructure Engineering
The University of Melbourne

fliu3@student.unimelb.edu.au maria.vasardani@unimelb.edu.au tb@ldwin.net

ABSTRACT

With the proliferation of smartphones and the increasing popularity of social media, people have developed habits of posting not only their thoughts and opinions, but also content concerning their whereabouts. On such highly-interactive yet informal social media platforms, people make heavy use of informal language, including when it comes to locative expressions. Such usage inhibits the ability of traditional Natural Language Processing approaches to retrieve geospatial information from social media text. In this research, we: (1) develop a medium-scale corpus of "locative expressions" derived from a variety of social media sources; (2) benchmark the performance of a range of geoparsers over the corpus, with the finding that even the best-performing systems are substantially lacking; and (3) carry out extensive error analysis to suggest ways of improving the accuracy and robustness of geoparsers.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Algorithms, Experimentation

Keywords

locative expression, geoparsing, social media

1. INTRODUCTION

The coming together of social media, mobile devices and ubiquitous connectivity has led to users posting not only their thoughts and opinions, but also content relating to their whereabouts. For example, according to Gelernter et al. [14], nearly 27% of Twitter messages relating to the February 2011 Christchurch, New Zealand earthquake contained a reference to a street or building, or a toponym mention. However, due to the informal nature of social media text and widespread use of acronyms, word shortenings and irregular spellings [12, 15], various claims have been made about the ability of NLP to reliably extract information such as location mentions from social media text [17, 4, 47, 37, 3, 14]. To add to the complexity, in informal communication, people make heavy use of vague and informal place references (e.g. *my cozy room*, *my place*). While there is little hope of fully geolocating such mentions without detailed knowledge of the author, they are crucial to the task of automated extraction of spatial information and the ultimate goal of understanding place descriptions [19].

Our focus in this paper is the automatic identification of locative expressions in social media text. We build on the work of Herskovits and others [16, 32, 49] in defining a locative expression (LE) to be an expression which physically geolocates an implicit or explicit entity in the text. That is, it provides information on WHERE a given entity is located or action takes place, relating a "relatum" (the location) to a "locatum" (the entity that is being located or the agent of the action), generally via a relational word such as a preposition. For example, in *I live in the East*,[1] the relatum is *the East*, the locatum is *I* (the first person), and the relational word is *in*; in the traditional analysis of LEs, this would be represented as a triple such as (_I,*in,the East*), where _Irefers to the first-person pronoun.

While we are ultimately interested in the extraction of fully-specified locative triples, for the purposes of this work, we focus on the simpler task of identifying "degenerate locative expressions" [19], namely the relatum and relational word only, leaving the locatum underspecified; for simplicity, we will refer to degenerate locative expressions as LEs for the remainder of this paper. LEs must refer to a geophysical location (whether it is identifiable or not). As such, *my place* in the context of *We could all meet at my place ...* is an LE, but *US* in the context of *US officials ...* is not, as it refers to the government rather than the region and the officials may not be physically located in the US. In the case that locative words are part of a larger non-locative named entity (NE) (e.g. *Organisation of American States*), they are not considered to be LEs.

The relatum generally takes the form of a noun phrase and the relational word a preposition, as in *Near Petersham Gate,*

[1] All examples used in this paper are taken from the dataset used for evaluation.

Liu, Fei, Maria Vasardani and Timothy Baldwin (to appear) Automatic Identification of Locative Expressions from Social Media Text: A Comparative Analysis, In Proceedings of The 4th International Workshop on Location and the Web (LocWeb 2014), Shanghai, China.

we saw … or … *i am nowhere in my cozy room where my mom would come …* (where the LE is underlined in each case). It is also possible for the relatum to take the form of an adjective or noun modifier (e.g. … *as slaves in European markets*), or a complex noun phrase (e.g. … *near Downtown, Chinatown and Comiskey Park*). In line with work on chunk parsing, we assume that LEs cannot be nested. As such, the identification of LEs relates to each of natural language parsing, named entity recognition, semantic role labelling (SRL) and geoparsing. It differs in that many LEs are not NEs, not all LEs are governed by verbs (as is generally assumed in SRL [35]), and LEs are often informal.

Vasardani et al. [44] make the case for an LE recogniser, capable of processing unrestricted natural language (NL) text, in the task of automatically translating verbal descriptions into two-dimensional sketch maps. However, to the best of our knowledge, there has been little analysis so far on either: (1) the distribution of LEs in different social media text types; and (2) the ability of geoparsers to identify LEs in social media text. In this research, we examine the output of six existing geoparsers over social media text, based on the corpora assembled by Baldwin et al. [3] from five popular social media sources and one balanced corpus of English. The geoparsers range in intent from LE recognisers trained over informal text, to off-the-shelf named entity recognisers, to geoparsers designed to identify more formal spatial descriptions (such as addresses). We evaluate each according to the same criterion, in terms of its ability to identify LEs of all types in the different text sources. First, we manually annotated 500 randomly-selected sentences from each of the six corpora. We then used this data to empirically evaluate the geoparsers. Next, we identified typical patterns of LEs in the manually-anntotated data that each tool is able to (correctly) recognise and also analysed the most frequent LEs in the output of each geoparser across the full corpora. Our findings indicate that there is substantial room for improvement for all geoparsers, and that each has its quite distinct strengths and weaknesses.

2. BACKGROUND

We are interested in the automatic identification of LEs from unrestricted NL text, especially from social media posts. Earlier work on the identification of LEs focused on the identification of specific and application-dependent geospatial information from restricted language place descriptions [42, 18, 24]. The identification of LEs relates closely to geoparsing, namely the process of automatically identifying spatial references within unstructured text. Recent research has been focused on geoparsers which are able to process unrestricted NL descriptions. Amitay et al. [2] and Li et al. [23] demonstrated their approaches to identifying geographic terms in either web pages or news articles. More recently, with the rise of social media, the focus has been shifted to processing text in user-generated social media posts. However, as pointed out by Baldwin et al. [3], traditional NLP tools tend to struggle when applied directly due to user-generated social media text, with Twitter being a particularly hard target. Simplistically, geoparsing can be considered to be a sub-task of named entity recognition (NER), that is, the task of identifying (mostly proper) names of people, organisations, locations, etc. State-of-the-art NER models used structured classification approaches such as linear-chain conditional random fields [22] or structured perceptron [9]. Of particular relevance over social media are Liu et al. [28] and Ritter et al. [40], who report F-scores of 77–78% at identifying locative named entities in tweets.

Another widely-used approach is to match place references in a gazetteer. An attempt was made by Paradesi [36] to combine NER and external gazetteers. The system, TwitterTagger, first assigns part-of-speech (POS) tags to words in tweets to locate proper nouns, and then matches noun phrases to the USGS database[2] to identify nouns that are likely to be places (e.g. prepositions preceding place names).

Other researchers have explored approaches based on language models. Kinsella et al. [20] created a model by estimating the distribution of words associated with a location and then the probability that the tweet is related to the location. Their model performs well at the city level but suffers at the neighbourhood level. Gelernter et al. [14] built a geoparser combining the results of four parsers: a lexico-semantic named location parser, a rule-based street name parser, a rule-based building name parser and a trained NER.

3. DATASETS

In this section, we detail the datasets involved in this research, namely: (1) the Tell Us Where corpus [46]; and (2) the social media corpora used in this research.

3.1 The *Tell Us Where* Dataset

Tell Us Where (henceforth TellUsWhere) is a location-based mobile game where participants were asked to provide a NL description of their location, in answer to the question *Tell us where you are* [46]. The descriptions submitted by the participants are therefore rich in LEs. The game resulted in the submission of a total of 1,858 place descriptions, focused primarily around Victoria, Australia. These place descriptions were manually annotated for LEs [43], and this data is used to both train some of the LE identification systems (see Section 4.1), as well as to evaluate the different tools.

3.2 Social Media Corpora

The social media corpora used in this research were originally constructed by Baldwin et al. [3] to measure (among other things) the degree of lexical and syntactic noise in text from different social media sources, as compared to text from a balanced corpus of English. A dataset of around 1M documents was assembled from each source, which was then restricted down to English documents based on automatic language identification [29]. The (putatively) English documents were then sentence-tokenised using `tokenizer`.[3] Our analysis in this paper is based on 100K randomly-selected sentences from each social media source, and the balanced corpus of English. In each case, we additionally hand-annotated 500 sentences for LEs based on Penn Treebank-style word tokenisation, to evaluate the accuracy of the geoparsers.

Below, we briefly describe the five social media sources and balanced corpus of English.

[2] `http://geonames.usgs.gov/`

[3] `http://www.cis.uni-muenchen.de/~wastl/misc/`; this was found to be the most reliable sentence tokeniser for user-generated content by Read et al. [39], although Baldwin et al. found the output to be noisy and the notion of sentence to be somewhat ill-defined for social media content.

TWITTER-1/2. two sets of micro-blog posts from Twitter, crawled using the Twitter Streaming API over disjoint time periods (TWITTER-1 = 22 September 2011 and TWITTER-2 = 22 February 2012) to investigate the impact of time on the composition of the data.

COMMENTS. comments from YouTube, based on the dataset of O'Callaghan et al. [31], but expanded to include all comments on videos in the original dataset.[4]

FORUMS. posts from the top-1000 valid vBulletin-based forums in the Big Boards forum ranking.[5]

BLOGS. blog posts from tier one of the ICWSM-2011 Spinn3r dataset [7].

WIKIPEDIA. wiki markup-stripped text from the body of documents in a dump of English Wikipedia.

BNC. as our balanced corpus of the English language, all documents from the written portion of the British National Corpus [6]; note that most documents were authored in the 1980s and are edited text.

3.3 Manual Annotation

LEs were hand-annotated over the Penn Treebank-style word tokenisation, as contiguous token extents. While the focus of this research is on identification rather than grounding of the LEs, we made use of two interactive web-based map services in the annotation process in cases of uncertainty over whether an expression was locative or not: OpenStreetMap[6] and Google Maps.[7] All sentences were annotated by three annotators, with pairwise inter-annotator agreement measured at Cohen's $\kappa = 0.69$. The annotated data is available in CoNLL format at:

`http://people.eng.unimelb.edu.au/tbaldwin/etc/locexp-locweb2014.tgz`

including Penn-style POS tags based on `ARK Tweet NLP POS Tagger v0.3` [33] and full-text chunk tags based on `OpenNLP`.

4. TOOLS

A total of six geoparsers were used to automatically identify LEs, which we separate into two types: (1) end-to-end locative expression recognisers, and (2) geospatial named entity recognisers. In the first case, the tool identifies LEs (e.g. *They did the good folks [in Albany, GA], proud.*) as a first order output, whereas in the second case, the tool identifies locative *entities* (e.g. *They did the good folks in [Albany, GA], proud.*) and requires postprocessing to map these into LEs (*in Albany, GA* in this case). We describe the geoparsers below, and outline the postprocessing used to generate LEs for the geospatial named entity recognisers.

4.1 End-to-end LE Recognisers

Locative Expression Recogniser

The Locative Expression Recogniser (`LER`) is a geoparser developed by the first author to automatically identify full LEs from informal text [27]. It is trained on the manually-annotated TELLUSWHERE dataset (see Section 3.1), and has been used in research on extracting "spatial triplets" from place descriptions [19]. In addition to the sentence and word tokenisation, `LER` requires POS tagging and full-text chunk parsing information. Acknowledging that the accuracy of standard NLP tools tends to drop appreciably on social media text, we use POS tag with the `ARK Tweet NLP POS Tagger v0.3` [33] using the Penn Treebank tagset model. We use `OpenNLP`[8] as our chunk parser.

Retrained StanfordNER

The Stanford named entity recogniser [13] has been found to be both robust out of the box, and highly effective when retrained over data containing LEs [25]. In line with these findings, we retrain the Stanford named entity recogniser over the manually-annotated TELLUSWHERE dataset (the same dataset as was used to train `LER` above). We will refer to this system as `Re-StanfordNER`. Note that, unlike the pre-trained model, `Re-StanfordNER` natively recognises fully-formed LEs, and no other entity type.

4.2 Geospatial Named Entity Recognisers

In order to use geospatial named entity recognisers to identify LEs, we need to have some way of determining the syntactic context of each locative NE. In the case that it is embedded in a noun phrase headed by a relational noun (e.g. *On the north east corner of Mira Mesa Blvd. and Flanders Dr.*, where the whole expression is a single [c] ontaining the locative NEs are *Mira Mesa Blvd.* and *Flanders Dr.*), the relatum should include both the locative NEs and the relational noun. To combine locative NEs into [[] s], we apply the following heuristics (which were also used to construct TELLUSWHERE from the annotation of [43]):

1. Recursively combine locative NEs which are linearly connected with commas (e.g. [*Albany*] , [*GA*], apostrophes or the preposition *of* (e.g. [*the South Side*] *of* [*Chicago*]) into a single complex locative NE (e.g. [*Albany, GA*] or [*the South Side of Chicago*], resp.)

2. If a (possibly complex) LE is immediately preceded by a prepositional chunk (as identified via the POS tags `IN` and `TO`), combine the two into a single LE

Full-text chunk and POS information is based on the output of `OpenNLP`.[9]

StanfordNER

The Stanford Named Entity Recogniser [13] (`StanfordNER`) is based on a linear-chain conditional random field with a heavily-engineered feature set. In this research, we use `StanfordNER` with a trained 3-class (Location, Person and Organisation) model with distributional similarity features.[10]

[4]Baldwin et al. [3] removed all occurrences of the unicode U+FEFF codepoint from the documents prior to language identification, as they found that it biased the results.

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

[6]`http://www.openstreetmap.org/`

[7]`http://maps.google.com/`

[8]`http://opennlp.apache.org/index.html`

[9]Note that we use the `OpenNLP` POS tagger only for LE aggregation, and we expect any differences over `ARK Tweet NLP POS Tagger v0.3` to be very minor in this context.

[10]`http://www-nlp.stanford.edu/software/CRF-NER.shtml#Download`

We ignore all other than Location entities in the output of the system.

GeoLocator

GeoLocator is a geoparser designed for the purpose of geoparsing short, informal messages in social media posts [14]. In order to boost robustness and better handle abbreviations, non-standard spelling and highly localised LEs, it makes use of the results of four parsers: a lexico-semantic named location parser, a rule-based street name parser, a rule-based building name parser, and a trained named entity recogniser. The training data consists of Twitter messages posted following the February 2011 earthquake in Christchurch, New Zealand. Note that in addition to identifying LEs, GeoLocator predicts the location of each expression, which we ignore in our evaluation.

Unlock Text

UnlockText,[11] developed by the Language Technology group at the School of Informatics of the University of Edinburgh, is a geoparser based on gazeteers such as GeoNames[12] and Ordnance Survey Open Data.[13] The geoparser identifies place references in NL using external gazetteers. As with GeoLocator, we ignore the predicted locations of each expression, and use it simply as a NE recogniser.

TwitterNLP

TwitterNLP is a multi-purpose NLP tool tuned specifically for processing Twitter messages [40], and based on labelled LDA [38]. In addition to the NLP tasks of POS tagging and chunk parsing, it is also capable of performing classification of ten categories of named entities. In this research, we use TwitterNLP with the option of POS and chunk tags to achieve higher quality. To evaluate the tool, we focus on named entities classified as GEO-LOC.

5. ANALYSIS

In this section, we first compare the relative prevalence of LEs in different social media sources, then carry out an empirical evaluation of the geoparsers. Finally, we perform error analysis of the different geoparsers to better understand the limitations of current methods and provide pointers for future work in this space.

5.1 Occurrence of LEs in the Manually-annotated Data

First, we analyse the relative occurrence of spatial expressions in the annotated datasets, by calculating the total number of tokens in each dataset, the raw count of LEs, and the proportion of tokens that are part of an LE. The results are shown in Table 1.

Wikipedia has the most LEs, with 6.2% of tokens being contained in LEs. BNC has the next highest prevalence of LEs, at around 69% that of Wikipedia, followed closely by Blogs. Twitter-1/2 have around half the number of LEs again, followed by Comments and Forums.

Comments and Forums are the sparsest in terms of LE density per document, just below Twitter-1/2. The cause for this is that Comments documents tend to refer to the individuals and actions taking place in the associated videos (or other commenters!), rather than the spatial context of the video. A possible explanation for Wikipedia having more LEs than BNC is its encyclopaedic nature, describing places and events, as compared to BNC, which includes a balance of sources such as fiction, pamphlets and reviews, all of which tend to contain fewer LEs.

Dataset	Sentences	Tokens	LEs	LE token %
Twitter-1	500	4646	40	1.9
Twitter-2	500	4382	31	2.1
Comments	500	5219	29	1.7
Forums	500	7548	43	1.7
Blogs	500	9030	97	3.7
Wikipedia	500	10632	183	6.2
BNC	500	9782	126	4.3

Table 1: Composition of the datasets ("LE token %" = the percentage of tokens that are contained in LEs)

5.2 Geoparser Accuracy over the Social Media Datasets

The performance of each geoparser is evaluated by comparing its output against the manual annotations for each dataset, including the step of mapping locative NEs to LEs via our heuristics in the case of the geospatial named entity recognisers. We evaluate the geoparsers based on exact match with the full token extent of the manually-annotated LEs, based on chunk-level precision ($\mathcal{P}$), recall ($\mathcal{R}$) and F-score ($\mathcal{F}$). The results over each of our datasets are presented in Table 2.

The first thing to notice is that the best of the systems (StanfordNER) achieves a macro-averaged F-score of around only .31 overall, much lower than the numbers reported in by Finkel et al. [13] of around .90, over newswire and seminar announcement datasets. One possible explanation for this result is the nature of the text — the text content of social media sources such as Twitter and YouTube comments tends to be "noisy" [3]. However, the results over datasets such as Twitter-1/2 and Blogs are actually slightly higher than the edited Wikipedia and BNC, putting this claim into doubt. In fact, a large part of the disparity is that many of the LEs in our respective datasets are not NEs, but rather informal "relational" LEs such as *at home* or *around the city*, which the NE recognisers, rightly, fail to identify.

The second thing to notice is the imbalance between precision and recall between our geoparsers, ranging from LER with very high recall (macro-averaged $\mathcal{R}$ = 0.73) but very low precision (macro-averaged $\mathcal{P}$ = 0.06), to StanfordNER with moderate precision (macro-averaged $\mathcal{P}$ = 0.40) and low recall (macro-averaged $\mathcal{R}$ = 0.26). Overall, the most balanced (and hence overall best performer) is StanfordNER, followed by TwitterNLP and UnlockText. The reason for the high recall and low precision of LER and Re-StanfordNER is the training data (the TellUsWhere dataset), which contains a much higher proportion of LEs than the other datasets (the LE token proportion is 63.9%, meaning that for a randomly-selected token, it is more likely to be con-

[11] http://unlock.edina.ac.uk/texts/introduction
[12] http://www.geonames.org/
[13] http://www.ordnancesurvey.co.uk/

	Twitter-1			Twitter-2			Comments			Forums			Blogs			Wikipedia			BNC		
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
`LER`	.05	**.65**	.09	.04	**.71**	.08	.04	**.79**	.07	.04	**.81**	.08	.07	**.76**	.12	.09	**.73**	.16	.07	**.67**	.13
`Re-StanfordNER`	.05	.33	.09	.05	.39	.09	.07	.52	.13	.07	.42	.12	.11	.51	.18	.11	.49	.18	.08	.40	.14
`GeoLocator`	.08	.45	.14	.04	.29	.07	.03	.21	.05	.06	.40	.11	.12	.38	.18	.12	.35	.18	.13	.27	.17
`StanfordNER`	**.54**	.33	**.41**	**.42**	.26	**.32**	.33	.14	.20	.35	.26	.30	.40	.27	.32	**.33**	.31	**.32**	**.45**	.25	.32
`UnlockText`	.29	.25	.27	.19	.19	.19	.20	.14	.16	.17	.19	.18	.41	.30	**.35**	.25	.27	.26	.43	.28	**.34**
`TwitterNLP`	.48	.28	.35	.33	.19	.24	**.50**	.14	**.22**	**.39**	.26	**.31**	**.47**	.28	**.35**	.30	.26	.28	.39	.20	.26

Table 2: Chunk-level precision ($\mathcal{P}$), recall ($\mathcal{R}$) and F-score ($\mathcal{F}$) of the geoparsers over the manually-annotated subset of the different datasets (the best-performing system in each column is boldfaced)

Geoparser	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
`LER`	**.77**	**.76**	**.77**
`Re-StanfordNER`	.72	.68	.70
`GeoLocator`	.52	.41	.46
`StanfordNER`	.34	.02	.04
`UnlockText`	.33	.01	.03
`TwitterNLP`	.33	.03	.06

Table 3: Performance on TellUsWhere

tained in an LE than not!). This leads to overfitting, and the models over-aggressively identifying LEs.

A third, and perhaps even more surprising, thing to notice is that there aren't great differences between the datasets, in terms of the raw performance of the best-performing system, or even the breakdown of results for individual systems. This is despite the varying levels of lexical and grammatical noise in the different datasets observed by Baldwin et al. [3], and also differences in the relative density of LEs observed in Table 1. Perhaps the biggest difference is actually for `StanfordNER` and `GeoLocator` between Twitter-1 and Twitter-2 (the two Twitter datasets constructed at a 6-month interval), where precision jumps in both cases. It is possible that the popular topics in the respective time periods differ significantly in their locative composition, although the size of the annotated datasets precludes detailed error analysis in this regard.

A more specific finding is that, in contrast with the findings of [25], over our datasets and under our definition of "locative expression", `Re-StanfordNER` is nowhere near the best-performing geoparser. Detailed comparison of the respective task definitions and datasets is an intriguing area for future research.

5.3 Geoparser Accuracy over the Tell Us Where Data

We further evaluate the performance of the six geoparsers on TellUsWhere. `LER` and `Re-StanfordNER` are both trained over the full dataset, so we present results based on 10-fold cross-validation over the dataset, retraining for each test fold; the remainder of the geoparsers are evaluated as is. The performance of the six systems is presented in Table 3.

As can be observed, `LER` and `Re-StanfordNER` vastly outperform the other geoparsers and achieve much more impressive and balanced results, due to their in-domain advantage; this supports our overfitting hypothesis from Section 5.2. `GeoLocator` is the only pre-trained geolocator that achieves competitive results over this dataset: `StanfordNER`, `UnlockText` and `TwitterNLP` achieve very low recall, and lower precision even than the other methods, culminating in a very low F-score. The reason for `GeoLocator`'s robustness over this data appears to be its heavy use of gazetteers, allowing it to deal with the highly localised, largely Melbourne-specific place references in TellUsWhere. In fact, the analysis of Tytyk and Baldwin [43] would suggest that only 62.4% of the LEs in the data are "formal" (the remainder being informal LEs such as *at home* or *at uni*), making the result of $\mathcal{F} = .46$ even more impressive.

5.4 Error Analysis

We next turn to error analysis of the respective geoparsers, to better understand the causes of error and identify possible areas for future research on robust LE identification. Below, we identify common causes of errors, and discuss their impact on the different geoparsers as well as possible solutions to each issue. In general, there is of course a lot of scope for system combination, particularly given the large spread of precision and recall between the different systems.

5.4.1 Improperly Capitalised Formal LEs

English NE recognisers trained on edited text make use of capitalisation information to detect NEs, but social media data is notoriously unreliable when it comes to capitalisation [40]. An example of an uncapitalised formal place reference is shown in Example (1) (from Twitter-2):

(1) are you on your way [to leeds] right now?

Here, only `LER` and `GeoLocator` are able to recognise the improperly capitalised LE, even though it is correctly spelt and a clear-cut case of an LE.

One possible solution to this issue is to add message-level features to capture the "informativeness" of capitalisation in the message (i.e. if the message is all lower-case, it suggests that the user may not be making use of capitalisation). `TwitterNLP` actually incorporates such a feature, but fails to recognise the NE in this case. Another possible solution would be to retrain all of the systems over case-folded training data, and remove all capitalisation from the input; this approach has been shown to be effective for POS tagging over Twitter data [11]. An alternative approach would be to attempt to normalise the casing in each message prior to geoparsing [26, 45].

5.4.2 Acronyms

Acronyms are widely used in social media messages, particularly in Twitter, due to the 140-character limit [1]. An example of such usage is presented in Example (2) (from Forums):

(2) Most people can only afford 1 hour a week indoor since the cost is high [in NYC] for indoor time.

Here, *NYC* stands for *New York City*, and the LE *in NYC* is only identified by `LER`, `GeoLocator` and `TwitterNLP`, despite it being a relatively common and indeed "official" acronym for the city.

Cases such as *NYC* can be handled through the use of gazetteers that include place name abbreviations, as can be seen by the success of `GeoLocator` to identify the mention. Deabbreviation [41, 34] may also be effective for dealing with acronyms, in particular when dealing with vernacular acronyms.

5.4.3 Informal LEs

While much work has been focused on the task of recognising formal place references [48, 30, 40], people tend to use informal place references (e.g. *my bedroom*, *home*, *the wall*), particularly in social media posts. Example (3) from BLOGS contains two informal LEs:

(3) I'm eyeing a new one on ebay which is much narrower and will fit [in the corner] [between the bed and wall] inshaa Allah.

Only `LER` is able to correctly recognise the two LEs in this case. The other five geoparsers either incorrectly mark irrelevant words as LEs or are unable to identify any at all: `UnlockText`, for example, identifies *Allah* as an LE.

The approach taken by `LER` (training over data rich with informal LEs, the incorporation of semantic class features, etc.) was directly targeted at better capturing informal LEs, and appears to be successful in terms of its recall, but has obvious limitations in terms of precision. To better balance precision and recall, possible approaches are: (a) training over data that is more representative of the relative frequency of LEs in general text; and (b) using domain adaptation techniques [5, 10] to adapt models trained over TELLUSWHERE to other sources.

5.4.4 Complex LEs

A similar, yet more challenging, example is shown in Example (4) (from WIKIPEDIA), where we have complex LEs (*in the English county of Suffolk*) and also expressions which can potentially be locative but are used attributively (*a small village*) and complex prepositions (*close to*):[14]

(4) Snape is a small village [in the English county of Suffolk], [on the River Alde] [close to Aldeburgh].

Here, `LER` and `Re-StanfordNER` correctly recognise the first two LEs, but aren't able to identify the complex preposition, and incorrectly identify *Snape* and *a small village* as LEs, as does `GeoLocator`. `StanfordNER`, `UnlockText` and `TwitterNLP`, on the other hand, can detect the formal LEs *Aldeburgh* and *River Alde* but are unable to detect the complex LE *in the English county of Suffolk*, as *the English county* is not a NE.

Also relevant are coordinated LEs, possibly involving a mix of informal and formal place names, such as *near Downtown, Chinatown and Comisky Park* in Example (5) (from BLOGS):

[14]Note that we consider *Snape* to not be an LE in this instance, as the sentence introduces the entity *Snape* rather than geolocates any relatum relative to the location of Snape.

(5) I am located [in the South Side of Chicago], [near Downtown, Chinatown and Comisky Park]

Only `LER` and `Re-StanfordNER` are able to identify this LE; `GeoLocator` and `StanfordNER` are only capable of recognising *Comiskey Park* and *Chinatown* respectively while `UnlockText` and `TwitterNLP` fail to spot any NEs.

The solution here would appear to be full syntactic parsing, which has been found to be very difficult for social media text [3], although very recent work on dependency parsing over social media text suggests that the task may be within reach of NLP [21].

5.4.5 Temporal Expressions

As pointed out by Khan [19], temporal expressions can be the cause of false positives for geoparsers. For example, *in the moment* in Example (6) (from BLOGS) is incorrectly analysed as an LE by both `LER` and `Re-StanfordNER`, and, in general for these two systems, informal temporal expressions are a common cause of false positives.

(6) Knowing what it means to live in the moment.

`GeoLocator` is less susceptible to false positives, but there are cases where it systematically mistakes temporal expressions for LEs, e.g. when the message starts with *on*, followed by a full-formed date (e.g. *on 13 June 1986* or *on June 16 2007*).

One possible solution to this problem, at least for formal temporal expressions, would be to add temporal analysis to the processing pipeline [8]. Indeed, part of the reason the NE recognisers don't suffer from this issue is that they tend to have an explicit representation of temporal expressions, which suppresses false positives.

6. CONCLUSIONS

In this research, we set out to investigate the distribution of LEs in various social media text types and evaluate the performance of six geoparsers at LE identification over such text. To this end, we manually annotated 3500 randomly-selected sentences from corpora collected from popular social media sites and a balanced corpus of English. Based on this data, we found that WIKIPEDIA is much richer in LEs than the other data sources, with around one token in 16 forming part of an LE. FORUMS had the smallest proportion of LEs, at around one quarter the frequency of WIKIPEDIA. We then empirically evaluated the geoparsers over this annotated data, and found a wide spread in terms of the precision and recall achieved by the different systems, with `StanfordNER` being the best system overall, at a modest F-score of around .31. As such, the identification of LEs is still very much an open research task. We further carried out error analysis to better understand the causes of errors, based on which, we suggested directions for future research in this area.

7. ACKNOWLEDGMENTS

We thank Yi Lin and Li Wang for their assistance with this research. This research was supported in part by funding from the Australian Research Council.

8. REFERENCES

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of Twitter data. In

Proceedings of the Workshop on Languages in Social Media (LSM 2011), pages 30–38, Portland, USA, 2011.

[2] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: Geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 273–280, Sheffield, UK, 2004.

[3] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang. How noisy social media text, how diffrnt social media sources. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 356–364, Nagoya, Japan, 2013.

[4] H. Becker, M. Naaman, and L. Gravano. Event identification in social media. In *Proceedings of the 12th International Workshop on the Web and Databases (WebDB 2009)*, Providence, USA, 2009.

[5] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia, 2006.

[6] L. Burnard. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services, 2000.

[7] K. Burton, N. Kasch, and I. Soboroff. The ICWSM 2011 Spinn3r dataset. In *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM 2011)*, Barcelona, Spain, 2011.

[8] A. X. Chang and C. D. Manning. SUTIME: A library for recognizing and normalizing time expressions. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3735–3740, Istanbul, Turkey, 2012.

[9] M. Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 1–8, 2002.

[10] H. Daume III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, 2007.

[11] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206, Hissar, Bulgaria, 2013.

[12] J. Eisenstein. 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, 2013.

[13] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, Ann Arbor, USA, 2005.

[14] J. Gelernter and S. Balaji. An algorithm for local geoparsing of microtext. *Geoinformatica*, 17(4):635–667, 2013.

[15] B. Han, P. Cook, and T. Baldwin. Lexical normalisation of short text messages. *ACM Transactions on Intelligent Systems and Technology*, 4(1):5:1–5:27, 2013.

[16] A. Herskovits. Semantics and pragmatics of locative expressions. *Cognitive Science*, 9(3):341–378, 1985.

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

[18] J. D. Kelleher. *A perceptually based computational framework for the interpretation of spatial language*. PhD thesis, Dublin City University, Dublin, Ireland, 2003.

[19] A. Khan, M. Vasardani, and S. Winter. Extracting spatial information from place descriptions. In *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place (COMP 2013)*, pages 62–69, New York, USA, 2013.

[20] S. Kinsella, V. Murdock, and N. O'Hare. I'm eating a sandwich in Glasgow: modeling locations with tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents (SMUC 2011)*, pages 61–68, Glasgow, UK, 2011.

[21] L. Kong, N. Schneider, S. Swayamdipta, A. Bhatia, C. Dyer, and N. A. Smith. A dependency parser for tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, to appear.

[22] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, Williamstown, USA, 2001.

[23] H. Li, R. K. Srihari, C. Niu, and W. Li. Location normalization for information extraction. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1 (COLING 2002)*, pages 1–7, Taipei, Taiwan, 2002.

[24] H. Li, T. Zhao, S. Li, and J. Zhao. The extraction of trajectories from real texts based on linear classification. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pages 121–127, Tartu, Estonia, 2007.

[25] J. Lingad, S. Karimi, and J. Yin. Location extraction from disaster-related microblogs. In *Proceedings of the 22nd International Conference on World Wide Web companion*, pages 1017–1020, Rio de Janeiro, Brazil, 2013.

[26] L. V. Lita, A. Ittycheriah, S. Roukos, and N. Kambhatla. tRuEcasIng. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 152–159, Sapporo, Japan, 2003.

[27] F. Liu. Automatic identification of locative expressions from informal text. Master's thesis, The University of Melbourne, Melbourne, Australia, 2013.

[28] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing

named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (ACL 2011)*, pages 359–367, Portland, USA, 2011.

[29] M. Lui and T. Baldwin. 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, 2012.

[30] A. Mikheev, M. Moens, and C. Grover. Named entity recognition without gazetteers. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics (EACL 1999)*, pages 1–8, Bergen, Norway, 1999.

[31] D. O'Callaghan, M. Harrigan, J. Carthy, and P. Cunningham. 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, 2012.

[32] P. Olivier and J. Tsujii. A computational view of the cognitive semantics of spatial prepositions. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL 1994)*, pages 303–309, Las Cruces, USA, 1994.

[33] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. 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)*, pages 380–390, Atlanta, USA, 2013.

[34] S. Pakhomov. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Philadelphia, USA, 2002.

[35] M. Palmer, D. Gildea, and N. Xue. *Semantic Role Labeling*. Morgan & Claypool, San Rafael, 2010.

[36] S. M. Paradesi. Geotagging tweets using their content. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2011)*, pages 355–356, Palm Beach, USA, 2011.

[37] D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, and M. Niranjan. Trendminer: An architecture for real time analysis of social media text. In *Proceedings of 1st International Workshop on Real-Time Analysis and Mining of Social Streams (RAMSS 2012)*, Dublin, Ireland, 2012.

[38] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 248–256, Singapore, 2009.

[39] J. Read, R. Dridan, S. Oepen, and L. J. Solberg. Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India, 2012.

[40] A. Ritter, S. Clark, Mausam, and O. Etzioni. 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, 2011.

[41] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333, 2001.

[42] D. A. Tappan. *Knowledge-based Spatial Reasoning for Automated Scene Generation from Text Descriptions*. PhD thesis, New Mexico State University, Las Cruces, USA, 2004.

[43] I. Tytyk and T. Baldwin. Component-wise annotation and analysis of informal placename descriptions. In *Proceedings of the International Workshop on Place-Related Knowledge Acquisition Research (P-KAR 2012)*, Kloster Seeon, Germany, 2012.

[44] M. Vasardani, S. Timpf, S. Winter, and M. Tomko. From descriptions to depictions: A conceptual framework. In *Proceedings of the 11th International Conference on Spatial Information Theory (COSIT 2013)*, pages 299–319, Scarborough, UK, 2013.

[45] W. Wang, K. Knight, and D. Marcu. Capitalizing machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 1–8, New York City, USA, 2006.

[46] S. Winter, K.-F. Richter, T. Baldwin, L. Cavedon, L. Stirling, A. Kealy, M. Duckham, and A. Rajabifard. Location-based mobile games for spatial knowledge acquisition. In *Location-Based Mobile Games for Spatial Knowledge Acquisition*, Belfast, USA, 2011.

[47] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using social media to enhance emergency situation awareness. *Intelligent Systems, IEEE*, 27(6):52–59, 2012.

[48] G. Zhou and J. Su. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 473–480, Philadelphia, USA, 2002.

[49] J. Zlatev. Spatial semantics. *The Oxford Handbook of Cognitive Linguistics*, pages 318–350, 2007.