Location Extraction From Disaster-Related Microblogs

John Lingad
School of EIE
University of Sydney, Australia
Jlin1546@uni.sydney.edu.au

Sarvnaz Karimi
IE Laboratory
CSIRO ICT Centre, Australia
Sarvnaz.Karimi@csiro.au

Jie Yin
IE Laboratory
CSIRO ICT Centre, Australia
Jie.Yin@csiro.au

ABSTRACT

Location information is critical to understanding the impact of a disaster, including where the damage is, where people need assistance and where help is available. We investigate the feasibility of applying Named Entity Recognizers to extract locations from microblogs, at the level of both geo-location and point-of-interest. Our experimental results show that such tools once retrained on microblog data have great potential to detect the *where* information, even at the granularity of point-of-interest.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text Analysis

Keywords

Location Extraction, Named Entity Recognition, Social Media Mining

1. INTRODUCTION

One of the first pieces of information that is broadcast when a disaster happens is *where* it hit. Following the broad announcement come the details. Where exactly is affected? Which suburb, street, building or area? People in close proximity to the incident but not directly affected ask a similar set of questions. For example, during a bushfire or flood they want to know if it is moving towards them, how far it is currently, and if their properties are likely to be affected. Extracting location information during a disaster is therefore crucial for keeping people on the scene and authorities informed. In return, more informed people often keep themselves safe, and more informed authorities act more effectively and allocate resources and services more efficiently.

We investigate how to extract location information from microblogs —in particular Twitter messages also known as tweets— posted during disasters. This is vitally important because currently less than one percent of tweets are geo-tagged [9], and even if they are geo-tagged, users' whereabouts could be different from the locations of the events they refer to. We define a *location* as both *geographic location*, such as country, city, river, or suburb, and *point-of-interest* (POI) such as hotels, shopping centers, and restaurants. This task has two stages: first, identifying references to the locations in the text, also known as toponym recognition; and second, geoparsing to assign geographic coordinates as latitude-longitude to the identified locations. This work concentrates on the first step, leaving the second as future work.

The main challenges of dealing with tweets are their brevity (maximum 140 characters), predominance of colloquial language, extensive use of abbreviation, and frequent deviation from grammatical rules. Natural Language Processing (NLP) tools therefore need to be adapted to work effectively on such text. We make two main contributions: we annotate a large set of tweets for all mentions of locations and then we benchmark major NLP tools in different settings for toponym recognition in Twitter.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

2. RELATED WORK

Named Entity Recognition (NER) has been well studied in the area of Natural Language Processing in the last decade. Its aim is to recognize and classify different types of entities, such as people or organizations, in a text. Conventional NER tools have proved to be very successful for identifying named entities in formal text. In recent years, the prevalence of microblogs has brought new challenges for the NER task. Because existing NER tools are typically trained on formal text such as news articles, their performance drops remarkably when applied on informal, and very short microblog data such as tweets. Therefore, recent studies, such as TwiNER [7] and TwitterNLP [11], have attempted to tackle tweets' unique characteristics and develop tweet-specific models for named entity recognition.

Identifying location information in microblog posts has attracted much attention in recent years. Location is a very important aspect that helps better understand the *where* information about events. TwitterStand [12] was one of the first to investigate how to geotag tweets' content at a geographical location on a map. They argued that state-of-the-art NER methods can not be directly applied to geotag tweets, because they are trained on text documents that are very different from tweets. Due to a lack of annotated corpus of tweets, they used TF-IDF information to extract key phases from tweets. Kinsella et al. [5] proposed to learn a language model of locations using coordinates extracted from geotagged tweets, and such a model was in turn used to predict the location of a tweet. Gelernter and Mushegian [2] explored the feasibility of applying Stanford NER out of the box [1] to automatically geoparse tweets for identifying locations in disaster-related tweets. Built upon this study, our work retrains existing NER tools using our annotated

Tool	Entity Type	Model	Training Data	Retrainable
Stanford NER	PERSON, ORGANIZATION, LOCATION	CRF	News data	✓
OpenNLP	PERSON, ORGANIZATION, LOCATION	Maximum Entropy	News Data	✓
TwitterNLP	PERSON, GEO-LOCATION, COMPANY, PRODUCT, FACILITY, TV-SHOW, MOVIE, SPORTSTEAM, BAND, and OTHER.	LabeledLDA	Twitter Data	✗
Yahoo! PlaceMaker	N/A	N/A	N/A	N/A

Table 1: Main characteristics of the tools applied in our study.

tweets and performs a systematic comparison to evaluate the effectiveness of several different tools.

Other research studies have attempted to estimate users' locations based on the content of tweets that they have posted. Ritter et al. [11] presented a probabilistic approach to estimate a user's city-level location based on place information registered in user profiles. Li et al. [8] considered the problem of identifying point-of-interest from tweets. They built a unigram language model for each POI and then applied a ranking technique to predict the POI of a tweet's origin. Ikawa et al. [3] proposed to infer the user location based on messages sent from third-party location services such as Foursquare. Unlike these studies, we focus on extracting locations strictly from the text of tweets which directly reflect where an incident has happened.

3. METHODOLOGY

Our objective is to evaluate the effectiveness of existing NER tools on extracting locations from disaster-related tweets. We choose four available off-the-shelf tools that are largely known to be effective in past studies:

Stanford NER Stanford NER[1] is a Java implementation of a Named Entity Recognizer to identify three major classes of named entities: PERSON, ORGANIZATION, and LOCATION [1]. It implements a linear chain Conditional Random Field (CRF) [6] model to label sequences of words in text into entity types.

OpenNLP OpenNLP[2] is a Java based library for various natural language processing tasks, such as tokenization, part-of-speech (POS) tagging, and named entity recognition. For named entity recognition, it trains a Maximum Entropy model using the information from the whole document to recognize entities in documents.

Yahoo! PlaceMaker Yahoo! PlaceMaker[3] is a geoparsing service that identifies place names in a given free-form text.

TwitterNLP TwitterNLP is a specific toolkit developed for performing natural language processing on Twitter data [11]. It applies a supervised topic model, called LabeledLDA [10], together with Freebase as a source of distant supervision, to classify entity mentions in tweets. TwitterNLP provides three options of using classification, POS tagging, and chunking.[4]

The main characteristics of these tools are summarized in Table 1. Stanford NER and OpenNLP, two state-of-the-art NER tools, are originally trained on news data (formal text). They both provide a retraining option for other types of text. TwitterNLP, based on topic modeling, can only be customized by augmenting the dictionary in Freebase.

As we deal with both geo-location and point-of-interest in our location extraction task, LOCATION and ORGANIZATION, tagged by Stanford NER and OpenNLP, and GEO-LOCATION, COMPANY and FACILITY, by TwitterNLP, are considered as *locations*. For Yahoo! PlaceMaker we only use the locations found in the tweets we submitted to this service. Other information specific to geoparsing is ignored.

4. EXPERIMENTS

This section describes the dataset annotated for the location extraction task, and presents experiments that compare the effectiveness of different tools we tested.

4.1 Dataset and Annotation

To evaluate the effectiveness of different tools, we created a gold standard dataset for a large set of tweets from late 2010 until late 2012. A random subset of these tweets were manually annotated as being disaster-related or not [4]. The disaster-related tweets comprised a dataset of 3,203 tweets covering a variety of disasters including, but not limited to 2012 flooding in Queensland, Australia, 2011 earthquake in Christchurch, New Zealand, 2011 England riots, 2012 flooding in York, England, and 2012 Hurricane Sandy, US. This dataset did not contain any retweets.

We annotated these tweets via CrowdFlower,[5] a crowdsourcing service over Amazon Mechanical Turk. The annotation task involved selecting the words, including hashtags and URLs which contained location information, to form a location word set for each tweet. For example, the hashtag #eqnz is selected because it contains the country New Zealand abbreviated as "nz". Agreement between annotators is defined as whether all the locations they identified for a particular tweet are exactly the same.

[1] `http://nlp.stanford.edu/software/CRF-NER.shtml`

[2] `http://opennlp.apache.org`

[3] `http://developer.yahoo.com/geo/placemaker/` visited in January 2013. Now it is migrated to Yahoo! Boss at `http://developer.yahoo.com/boss/geo/`.

[4] A shallow parsing that incorporates the structure of sentences.

[5] `http://crowdflower.com`

	Precision	Recall	F-Measure
Hashtags removed			
Stanford NER 4-class	0.699	0.682	0.691
Stanford NER retrained	0.906	0.841	**0.872**
OpenNLP out of the box	0.928	0.220	0.356
OpenNLP retrained	0.888	0.760	0.819
TwitterNLP	0.900	0.429	0.581
Yahoo! PlaceMaker	0.936	0.473	0.628
Hashtags without #			
Stanford NER 4-class	0.706	0.487	0.576
Stanford NER retrained	0.935	0.873	**0.902**
OpenNLP out of the box	0.930	0.156	0.268
OpenNLP retrained	0.912	0.767	0.833
TwitterNLP	0.903	0.301	0.451
Yahoo! PlaceMaker	0.941	0.378	0.540

Table 2: Comparison of the four tools.

Annotations were done in two stages: controlled and crowdsourced. Controlled annotations were performed by the authors each annotating the same random set of 450 tweets. If a tweet had two out of three annotators agreeing exactly on the labeled set of location words, this was taken to as the ground truth for the tweet. The rest of annotations were performed by the workers through CrowdFlower, which associated each worker with a trust level to provide quality control. The annotation task presented five tweets at a time to the workers, where one of the tweets, referred to as the gold tweet, was taken from the set of controlled annotations. The answer provided by a worker for the golden tweet was checked against its ground truth. If it matched the ground truth, the answers for the entire batch would be accepted. Otherwise the worker's trust level would be decreased. If the trust level fell below a threshold, all the annotations would be disregarded. Out of 3,203 tweets selected for annotations, we obtained a set of 2,878 tweets having a majority agreement between annotators. In this set 89% of the tweets contained at least one location.

4.2 Experimental Setup

We evaluate the effectiveness of the tools using two settings:

- Out of the box: Stanford NER involved using three included language models. OpenNLP utilized an English language model trained on locations that was downloaded from the website. TwitterNLP was tested in three modes: classification only, classification and POS, and classification, POS and chunking.
- Retraining: since Stanford NER and OpenNLP were trained on formal text, we retrained both tools using our annotated Twitter dataset. The trained tools were evaluated using 10-fold cross validation. TwitterNLP was also customized by augmenting the Freebase with an Australian gazetteer.

Since hashtags that indicate locations were included in our annotation scheme, we investigated the effect of hashtags on the performance of the tools: first, we removed all the hashtags from the entire dataset; second, we removed the # symbol and treated hashtags as normal words.

The effectiveness of the four tools was evaluated using precision, recall and F-Measure. These metrics were measured on the word level rather than entity level.

4.3 Results

Table 2 shows the comparison of applying the four tools on our dataset for the task of location extraction. The best performing out of the box language model for Stanford NER was the 4-class model that achieved an F-Measure of 0.691 removing hashtags. It was followed by Yahoo! PlaceMaker for both hashtags removed and hashtags without hash symbol settings. Surprisingly, TwitterNLP which is built for Twitter data performed worse than Stanford NER trained on news data. It was even behind Yahoo! PlaceMaker for both of the settings. We only show the results for TwitterNLP with all its three options (classification, POS tagging, and chunking) activated. The results for the other two modes followed the same pattern and thus are not shown for brevity. Adding gazetteer data harmed TwitterNLP (not shown). OpenNLP out of the box had the weakest result overall with an F-Measure of 0.356 with hashtags removed.

However, both retrained Stanford NER and OpenNLP outperformed the out of the box configurations. Stanford NER was the winner scoring 0.872 and 0.902 respectively for when hashtags were removed and only hash symbols were removed. This result highlights the importance of using appropriate training data for these tools. It also shows that the underlying models of these tools once retrained are effectively able to handle the noisy and short text from tweets.

Our data is a combination of tweets from different disasters that happened from 2010 until late 2012. It is expected that tweets of the same incident share location names and hashtags. Therefore, in a more realistic experimental setting, we should eliminate the effect of adding information from the future incidents. To do this, the tweets were sorted chronologically and partitioned into 10 sets in a manner similar to 10-fold cross validation. Stanford NER was then trained incrementally using these partitions. For example, in the first instance only the first partition was used for training. In the next iteration, the first two were used for training. This continued until nine parts were used for training. In all cases we used the last partition that contained the latest tweets for testing.

Figure 1 shows the effect of the size of the training data in the realistic setting without knowing about the incident included in the testing. When the training size is less than 50% of the data the F-Measure is between 0.5 to 0.6. When 70% of the data is seen in the training (approximately 2000 tweets), the effectiveness becomes stable. We emphasize on

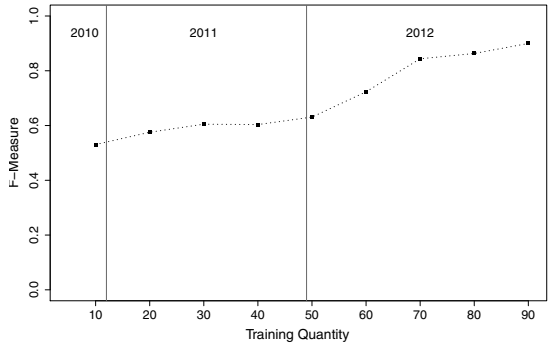

Figure 1: Effect of size of the training data when tweets are sorted in chronological order on the effectiveness of Stanford NER. Lines are added to show the trend and do not represent any datapoint.

	Stanford NER		TwitterNLP		Yahoo! PM	
Granularity	TP	FN	TP	FN	TP	FN
Country	86	14	22	78	41	59
State	94	6	18	82	22	78
City	81	19	62	38	80	20
Area	100	0	40	60	0	100
Suburb	100	0	100	0	100	0
POI	69	31	42	58	42	58
Hashtag	88	12	6	94	9	91

Table 3: Percentage of correct and incorrect locations found by three tools by granularity of locations and by hashtags. TP represents True Positive and FN is False Negative.

two points: first, to evaluate a machine learning tool on disaster related Twitter data, the factor of tweet publication time should not be overlooked; and second, the quantity of the training data also affects the effectiveness with a minimum of annotated tweets should be available for the NER tools to perform effectively.

4.4 Error Analysis

To gain an insight into what mistakes these tools made and how geographic granularity affects their performance, we carried out an error analysis. We randomly picked 121 tweets from our dataset for manual inspection. We divided the granularity of the locations into five categories: country, state, city, area, suburb, and POI. We also considered an extra category of location references inside hashtags.

Table 3 shows both the percentage of correctly identified locations or True Positives (TP), and the percentage of locations that were missed by the tool or False Negatives (FP) for three of the tools. We chose the output of the best run from Stanford NER, TwitterNLP, and Yahoo! PlaceMaker.

Stanford NER identified POIs best, detecting 69% of them, and was also the best at detecting location in hashtags. TwitterNLP was worst in handling locations in hashtags and it was even weak with a high-level location such as country (78% incorrect). It was also poor at finding POIs (58% incorrect). Yahoo! PlaceMaker was not much better in handling country names either with 59% missed country names. Hashtags were handled badly as well with 91% false negatives. In this small sample every tool found all suburbs correctly.

5. CONCLUSION AND FUTURE WORK

Extracting locations from disaster-related microblogs is important for increasing situation awareness. We presented an experimental study to quantify the potential of Named Entity Recognizers in location extraction in tweets. Our results show that if an NER tool is retrained using a set of annotated tweets, it is able to recognize locations effectively. Stanford NER in particular had an F-Measure of over 0.9 in a dataset of 2,878 disaster-related tweets. We also conducted an error analysis on the output of the tools we applied for this task. Finding POI was hardest, even for our best performing system, retrained Stanford NER.

In the future, we will investigate how to extract location information that is hidden in hashtags. We will then continue with geoparsing to infer a geographical focus for each tweet from these recognized locations.

References

[1] J. Finkel, T. Grenager, and C. Mannin. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proc. of ACL*, pages 363–370, Stroudsburg, PA, 2005.

[2] J. Gelernter and N. Mushegian. Geo-parsing messages from microtext. *Transactions in GIS*, 15(6):753–773, 2011.

[3] Y. Ikawa, M. Enoki, and M. Tatsubori. Location inference using microblog messages. In *Proc. of WWW Companion*, pages 687–690, Lyon, France, 2012.

[4] S. Karimi and J. Yin. Microtext annotation. Technical Report EP13703, CSIRO, 2012.

[5] S. Kinsella, V. Murdock, and N. O'Hare. "I'm eating a sandwich in Glasgow": Modeling locations with tweets. In *Proc. of SMUC Workshop*, pages 61–68, Glasgow, Scotland, UK, 2011.

[6] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, pages 282–289, Williamstown, MA, June–July 2001.

[7] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. TwiNER: Named entity recognition in targeted Twitter stream. In *Proc. of SIGIR*, pages 721–730, Portland, Oregon, 2012.

[8] W. Li, P. Serdyukov, A. de Vries, C. Eickhoff, and M. Larson. The where in the tweet. In *Proc. of CIKM*, pages 2473–2476, Glasgow, Scotland, UK, 2011.

[9] J. Mahmud, J. Nichols, and C. Drews. Where is this tweet from? Inferring home location of Twitter users. In *Proc. of ICWSM*, pages 511–514, Dublin, Ireland, 2012.

[10] D. Ramage, D. Hall, R. Nallapati, and C. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proc. of EMNLP*, pages 248–256, Stroudsburg, PA, 2009.

[11] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Proc. of EMNLP*, pages 1524–1534, Edinburgh, UK, 2011.

[12] J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, and J. Sperling. TwitterStand: news in tweets. In *Proc. of SIGSPATIAL GIS*, pages 42–51, Seattle, WA, 2009.