Location Mention Detection in Tweets and Microblogs

Shervin Malmasi
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
shervin.malmasi@mq.edu.au

Mark Dras
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
mark.dras@mq.edu.au

***Abstract*—The automatic identification of location expressions in social media text is an actively researched task. We present a novel approach to detection mentions of locations in the texts of microblogs and social media. We propose an approach based on Noun Phrase extraction and n-gram based matching instead of the traditional methods using Named Entity Recognition (NER) or Conditional Random Fields (CRF), arguing that our method is better suited to noisy microblog text. Our proposed system is comprised of several individual modules to detect addresses, Points of Interest (e.g. hospitals or universities), distance and direction markers; and location names (e.g. suburbs or countries). Our system won the ALTA 2014 Twitter Location Detection shared task with an F-score of 0.792 for detecting location expressions in a test set of $1{,}000$ tweets, demonstrating its efficacy for this task. A number of directions for future work are discussed.**

***Keywords*-location; detection; identification; Twitter; tweet; microblog; social media;**

I. Introduction

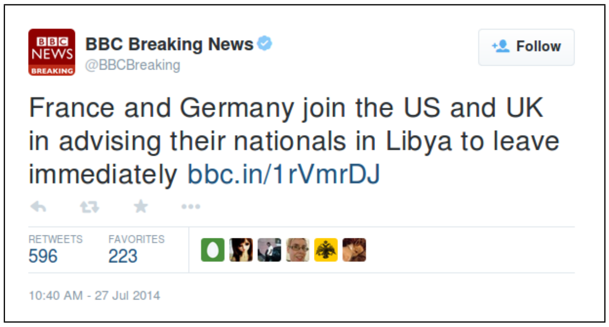
Locations are a key piece of information in social media discourse, often linked to specific events or news that are being discussed. In this context, the identification of location expressions in social media has attracted the attention of researchers and the extraction of this data from Twitter messages, called tweets, is actively researched [1] [2].

The specific goal of this task is to identify all mentions of locations in the text of tweets. A *location* can be defined as any specific mention of a country, region, city, suburb, street address, or other POI (Point of Interest). A POI can be a library, such as “Central Library” or the name of an airport such as “Manchester Airport”. These location expressions can appear in the text itself, or in hashtags (e.g. #china) and mentions (e.g. @Visit_Japan). Some example tweets and their identified locations are shown in Figure 1. Some tweets can contain multiple locations, as shown in Figure 2. Applications of such systems include the early detection of emergencies, crises and natural disasters in real time [3]–[5]. They could also be employed for targeted advertising purposes [6].

The overarching aim of the present work is to propose and evaluate a methodology for the detection of such location mentions in microblogs and social media.

Tweet	Location
`France and Germany join the US and UK in advising their nationals in Libya to leave immediately http://bbc.in/1rVmrDJ`	`France, Germany, US, UK, Libya`
`Dutch investigators not going to MH17 crash site in eastern Ukraine due to security concerns, OSCE monitors say`	`MH17 crash site, eastern Ukraine`
`Seeing early signs of potential flash flooding with stationary storms near St. Marys, Tavistock, Cambridge #onstorm pic.twitter.com/BtogIxgQ5G`	`St. Marys, Tavistock, Cambridge`

Figure 1. Several example tweets and the location expressions that they contain. Reproduced from [18].



BBC Breaking News
@BBCBreaking
Follow

France and Germany join the US and UK in advising their nationals in Libya to leave immediately bbc.in/1rVmrDJ

RETWEETS 596
FAVORITES 223

10:40 AM - 27 Jul 2014

Figure 2. An example tweet which contains more than one distinct location expression. Reproduced from [18].

II. Related Work

Researcher have been actively working on detecting such location mentions in both social media data as well as formal texts. In this section we briefly look at some previous approaches to this task.

One approach to this task has been based on Named Entity Recognition (NER), which is the process of identifying names, locations and organizations within texts. When applied to our target problem, this can be viewed as a subtask of NER where we are only interested in locations.

A set of tools for performing tasks such as NER specifically on Twitter was developed by [7]. The system, known as *T-NER*, was designed to also perform geo-location detection in tweet data. This system augmented the Stanford NER system with information from Freebase [8] to improve performance and achieved an F-score of 0.77 in detecting locations.

Another approach proposed by [9] has a 2-stage architecture and makes use of Conditional Random Field (CRF) modelling. Furthermore, they also used gazetted resources from Wikipedia to augment their system.

The authors of [10] also applied NER to the task and compared various tools with the standard models as well as NER models trained only on Twitter data. They conclude that existing NER tools should be re-trained on microblog data before being applied to Twitter data.

We should also note that such content-based approaches are not limited to English data or Twitter; other researchers have also tested them on other languages and microblogs such as Weibo [11].

III. Data

Data for the task included a training set of 2,000 tweets with manually annotated location information and 1,000 test tweets to be processed. To ensure a blind evaluation, the location annotations for the test tweets were not made available until after testing. This data was collected as part of the research presented by [10] and more details can be found in their work.[1]

IV. Methodology

In contrast with the work described in section II, we take a different approach to this problem. Instead, we opt to use syntactic parse trees to identify potential location information. Parse trees have been used in other NLP tasks such as Native Language Identification [12] [13] and other tree representations such as parent-annotated trees [14] have also been tested.

It is well known that microblog data is noisy and contains large proportions of non-standard words which pose challenges for most NLP systems trained on well-formed text. These include misspellings, hashtags, abbreviations, malformed sentences and other slang and colloquial terms. Although NER methods are highly effective in detecting locations in formal texts, they do not perform as well for Twitter data [10]. It is most likely this noisy nature of tweets and microblog data that makes it more challenging to distinguish the syntactic environments that predict locative arguments.

Yet another disadvantage of supervised NER systems is the requirement for sufficient amounts of annotated training data, preferably sourced from microtext sources if they are to be trained specifically for such target texts.

Given the above reasoning, we opt to develop an *unsupervised* approach based on a combination of syntactic parse trees and gazetteer information.

During the last decade, there has been growing interest and work in the development of geo-information databases and resources which could be utilised for such tasks. These *gazetteers* are usually made available in machine-readable format or web services. GeoNames[2] is one such data source and we use it in the present work.

The GeoNames geographical database[3] contains over 10 million geographical names and consists of over 9 million unique features of 2.8 million populated places and 5.5 million alternate names. The database is updated regularly and the information is sourced from dozens of unique sources.[4]

The remainder of this section focuses on describing how we achieve this through the various components of our system.

A. Preprocessing

As a first step, non-English tweets are detected using a dictionary-based language identification approach and discarded.

The tweets are then processed to normalize mentions and hashtags within the text by removing the @ and # symbols. These tokens are also stored separately in the original form for further processing in later stages. URLs are also stripped from the text and we do no process them.

B. Syntactic Parsing

Next, the Stanford CoreNLP[5] suite of NLP tools and the provided pre-trained English models are used to tokenize, POS tag and parse each tweet. This information is stored on separate annotation layers from the original tweet text. This is so that we can recover the untokenized strings in the original tweet.

C. Noun Phrase Extraction

The extraction of noun phrases is a critical component of our system. This is due to the fact that locative information is generally expressed through nouns and we can exploit this by discarding tokens that have been identified as other phrase types, such as verbs. After parsing, we use the generated constituency parses to extract the noun phrases (NPs) from within each tweet.

[1] Requests for the data should also be directed to the authors of [10].

[2] http://www.geonames.org/

[3] Available for download free of charge under a creative commons attribution license.

[4] http://www.geonames.org/data-sources.html

[5] http://nlp.stanford.edu/software/corenlp.shtml

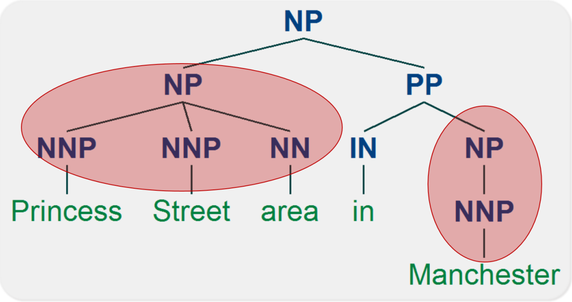


Figure 3. An example of a complex noun phrase (NP) which contains additional NPs embedded within it. In such cases it may not be possible to match the entire NP as a single location since it contains constituent NPs with different locations.

Many of the NPs found in the data can be considered complex NPs,[6] and in these cases we only extract the constituent NPs they contains. This is achieved through a rule-based tree splitting method that breaks the tree at certain branches. Our method works by recursively breaking down the NP at non-NP branches, such as prepositions, in order to extract only the simpler constituent NPs. Figure 3 shows an example of a complex NP and its constituent NPs.

One important advantage of this approach is that the parser will tag any words that it does not recognize (such slang), or tokens that are not part of a sentence (e.g. a trailing list of tags after a post) or incomplete text fragments as NPs. These tokens may contain locations that would likely not be identified by NER or CRF systems due to the lack of appropriate syntactic context.

D. *N-gram based location matching*

The extracted noun phrases may still contain more than one location or other non-location tokens – *e.g.* “Christchurch New Zealand earthquake” or “Bangkok residents” – making it difficult to precisely match the locations. We resolve this by using an n-gram based matching approach. Here, we first attempt to match the whole NP as a single location, and if no exact match is found, we consider all of its n-gram subsets. For an NP of N tokens, this include all n-grams of order $N-1$ through to unigrams. It is important to process the subsets in this descending in order to match maximal subsets of the NPs. If an n-gram is matched as a location, its subsets will not be considered.

Let us illustrate this with an example noun phrase “Buckleys Rd Dunmore”, as shown in Figure 4. This NP contains two location mentions (Buckleys Road in the suburb of Dunmore) within a single phrase. As a first step we attempt to match the entire NP as a location, but no precise match can be found. Consequently, we then consider the subset 2-grams and 1-grams, as shown in the second and third rows of Figure 4. These two location mentions are then matched by two separate components of our system: the address matching and geographic lookup modules, respectively. These components are described later this section.

3-grams	“Buckleys Rd Dunmore”
2-grams	“Buckleys Rd”, “Rd Dunmore”
1-grams	“~~Buckleys~~”, “~~Rd~~”, “Dunmore”

Figure 4. A demonstration of how our n-gram based matching works. Successful location matches are shown in red. The first (row 1) attempts to match the whole string as a unique location. When this fails, its constituent n-grams are considered (row 2) and an address is matched. Any n-grams that were not matched are further broken down and checked (row 3).

Greater Manchester Fire Service
Dandenong train crash
Queensland flood report
Guatemala Earthquake

Figure 5. Some examples of how our partial matching method can detect the location tokens (shown in red) within noun phrases. Without using an n-gram based matching approach, some of the non-location tokens may be erroneously detected as locations, increasing the detector’s false positive rate.

In our experiments, not processing the phrases via this n-gram matching procedure leads to a higher false positive rate as some extra parts of NPs may be matched as locations.

This procedure is also helpful when processing noun phrases with partial location information, e.g. “China earthquake report”. Only one noun in the NP is a location expression here. Some examples of how this method can capture the location-relevant subsets of NPs is shown in Figure 5.

We now describe several subcomponents of our system that are used to determine if these n-gram candidates are location expressions.

1) Address Matching: Addresses are a crucial piece of location information. The generally structured format of addresses makes them suitable for matching via regular expressions. To this end we developed a set of regular expressions to capture NPs containing address expressions using a wide array of street types along with their abbreviations. Examples of such road types include Arcade, Avenue, Boulevard, Road, Street, Highway, Overpass, *etc.* The regular expressions are also designed to capture street numbers. Some sample addresses extracted by our system are listed in Table I.

2) Point Of Interest Matching: Another type of location we are interested in are Points of Interest (POIs). A POI can be, *inter alia*, a hospital, airport, river, university, park or shopping center. We compiled a list of such locations and created a set of regular expressions to match NPs that

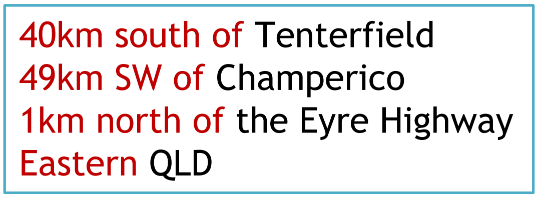
[6] A noun phrase that contains other NPs, for example, within prepositions.

Table I
SOME EXAMPLE ADDRESSES MATCHED BY OUR REGULAR EXPRESSIONS.

Orchard Rd	Yanilla Ave
Warrego Hwy	Hawkesbury River Bridge
Forge Creek Rd	North West Coastal Highway
Wedderburn-brenanah Rd	Kaban Rd.
School Rd overpass	Princess Street
333 Manly Road	Batemans Bay

Table II
SOME EXAMPLE POINT OF INTEREST (POI) LOCATIONS MATCHED BY OUR REGULAR EXPRESSIONS.

Kumbarilla State Forest	Nudgee Golf Club
Princess Alexandra Hospital	Melbourne Airport
Wong Wong bakery	Thomas Jefferson University Hospital
Navigator College	Khartoum arms factory



40km south of Tenterfield
49km SW of Champerico
1km north of the Eyre Highway
Eastern QLD

Figure 6. Some example of various distance and direction markers, highlighted in red, as matched by our matching module. These segments provide important info for pinpointing a specific location within a broad geographical area.

contain them. Some example results are shown in Table II.

3) Location name matching: We employ the above-described GeoNames database to match non-address locations, such as suburbs, countries and other landmarks. To do this, we utilize the advanced search features offered by the web service, including fuzzy matching to help address misspellings.[7] The location candidates are sent via the API and they are marked as locations if a match is reported.

Some example locations matched by GeoNames include Guatemala, Wahroonga, Syria, Ultimo, Greece, Brisbane, New Jersey, and Manchester.

4) Distance and Direction Marker Matching: The final component of our system matches distance and direction markers, which were also annotated in our training data. This type of information, *e.g.* "*25 km North of* Beijing", is often found within complex locative NPs.

We compiled a list of such directional and distance markers and created a rule-based module to match them, again using regular expressions. Some example of markers found in our data are shown in Figure 6.

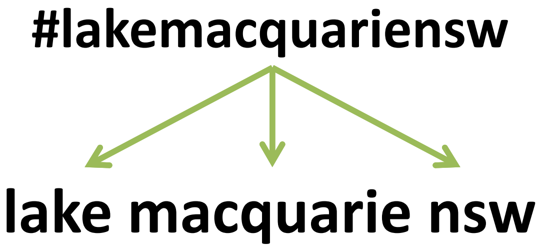
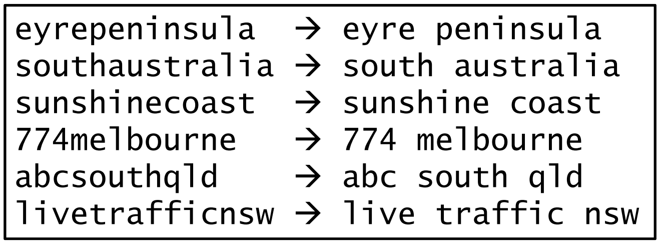


Figure 7. An example of applying word segmentation to a Twitter hashtag. The aim here is to find the perfect segmentation boundaries to recover the intended words in a concatenated hashtag. In this example the hashtag has been correctly segmented into the three words and it refers to a location.



eyrepeninsula	→	eyre peninsula
southaustralia	→	south australia
sunshinecoast	→	sunshine coast
774melbourne	→	774 melbourne
abcsouthqld	→	abc south qld
livetrafficnsw	→	live traffic nsw

Figure 8. Some examples of hashtags/mentions and their segmentations on the right. This is an important step in finding tokens that are location expressions, particularly those that are contained in tags with multiple words, such as "nsw".

E. Hashtag and Mention matching

In developing the above-described components we discovered that these methods could not match locations that were embedded within hashtags and mentions that included multiple concatenated words, *e.g.* "#ChinaFlooding". The key issue here is the concatenation of the words which prevent our modules from detecting the location words [15]. To address this, these compound word tokens need to be segmented to decompose them into the constituent words. An example of this segmentation is shown in Figure 7.

We attempt to address this issue by applying a word segmentation method. More specifically, we employ an approach based on language models, as described by [16]. Using this method a segmenter is built using unigram and bigram models of word frequency and attempts to find the word boundaries using a naive Bayes approach. We augment our language models with additional location information from GeoNames and other tokens that have been detected by our system.

We apply this method in our system to process hashtags and mentions before passing them to our detection modules. Some example segmentation results extracted from our data are shown in Figure 8.

[7] The web service offers a number of advanced features that can help increase search specificity.

F. Caching

Optionally, the the matched locations can be cached for faster lookups in processing future entries. There are many common location mentions that appear with great frequency and storing a cached mapping of NPs/hashtags/mentions to their particular location mentions can provide a significant improvement in processing large amounts of data.

V. Evaluation Method

Evaluation for this task is usually performed using the the F1 score. This is a metric based on precision – the ratio of true positives (tp) to predicted positives (tp + fp) – and recall – the ratio of true positives to actual positives (tp + fn). The F1 metric is calculated as:

$$F1 = 2\frac{pr}{p+r} \quad \text{where} \quad p = \frac{tp}{tp+fp}, \quad r = \frac{tp}{tp+fn}$$

Here p refers to precision and r is a measure of recall.[8] Results that maximize both will receive a higher score since this measure weights both recall and precision equally. It is also the case that average results on both precision and recall will score higher than exceedingly high performance on on measure but not the other.

Furthermore, the evaluation here is conducted on a per-token basis and partial location mentions are also included. This means for a text with a location mention “Northern Canada”, annotating just “Canada” would receive a precision of $\frac{1}{1}$ and recall of $\frac{1}{2}$.

VI. Experiment and Results

Our system was used to enter the Twitter Location Detection competition at the 2014 Australasian Language Technology Association (ALTA) Workshop [18]. We run our system on the test set of the data which contains $1,000$ tweets. The location annotations were not made available to us. Our described system achieved an F-score of 0.792 on the test set, ranking first among the shared task entries and winning the competition.

We believe that this is a good result which proves the efficacy of our proposed system in a demonstrable manner.

An analysis of our system results shown that all components contribute to the system. The GeoNames components is one of the most important modules and responsible for much of the performance.

We also want to emphasize the important of hashtag segmentation for this task; our results improved by around 0.05 through the addition of the compound word decomposition functionality, making it an important component.

VII. Discussion and Conclusion

We presented a novel unsupervised approach for detecting location mentions in microblogs and social media texts.

A key contribution here is the definition of various location expression types and methods to detect them independently. The inclusion of hashtag segmentation was also found to be a key factor in maximizing performance.

There are a number of directions for future work. The application of lexical tweet normalization techniques could help improve the parsing results which could in turn improve the accuracy of our NP extraction.

Information from other services such as Yahoo BOSS Geo Services[9] could also be incorporated into the system. Data sourced from more granular gazetteers that include street-level information, such as OpenStreetMap[10] could help improve the accuracy of the location expression matching. This can help overcome some limitations of our address matching modules. The following tweet is a particular example which highlights a weakness of this module:

> “The road to **Easy Street** goes through the sewer. It is a **rough road** that leads to the heights of greatness.”

Here the tokens in bold have been erroneously marked as location expressions, even though they are only figurative expressions. Having street level data could help reduce these false positives.

We also note that conducting a comprehensive error analysis could also provide to be a fruitful line of future inquiry. This analysis could provide valuable insights about the most common errors being committed by the current system – similar to the above example – thus helping guide future efforts in this area.

Displaying the locations on a map, in conjunction with an interactive system, is an interesting idea for future work which can help users find tweets pertaining to a specific geographic space. Such methods are also useful for visualization and can help find trends within the data.

Acknowledgments

We would like to thank our three anonymous reviewers for their valuable comments.

The data and the task’s original idea is from John Lingad’s Honours project (The University of Sydney) co-supervised with Jie Yin (CSIRO). The shared task prize was sponsored by IBM Research.

[8] See [17] for more details about these metrics.

[9] https://developer.yahoo.com/boss/geo/

[10] http://www.openstreetmap.org/

References

[1] J. Mahmud, J. Nichols, and C. Drews, "Where Is This Tweet From? Inferring Home Locations of Twitter Users," in *ICWSM*, 2012.

[2] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 1079–1088.

[3] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power, "Using social media to enhance emergency situation awareness," *IEEE Intelligent Systems*, vol. 27, no. 6, pp. 52–59, 2012.

[4] M. Núñez-Redó, L. Díaz, J. Gil, D. González, and J. Huerta, *Discovery and integration of Web 2.0 content into geospatial information infrastructures: a use case in wild fire monitoring*. Springer, 2011.

[5] S. Middleton, L. Middleton, and S. Modafferi, "Real-time Crisis Mapping of Natural Disasters using Social Media," 2014.

[6] T. L. Tuten, *Advertising 2.0: social media marketing in a web 2.0 world*. Greenwood Publishing Group, 2008.

[7] A. Ritter, S. Clark, O. Etzioni *et al.*, "Named entity recognition in tweets: an experimental study," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1524–1534.

[8] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 1247–1250.

[9] 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*. Association for Computational Linguistics, 2011, pp. 359–367.

[10] 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*. International World Wide Web Conferences Steering Committee, 2013, pp. 1017–1020.

[11] J. Ao, P. Zhang, and Y. Cao, "Estimating the Locations of Emergency Events from Twitter Streams," *Procedia Computer Science*, vol. 31, pp. 731–739, 2014.

[12] S. Malmasi, S.-M. J. Wong, and M. Dras, "NLI Shared Task 2013: MQ Submission," in *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 124–133. [Online]. Available: http://www.aclweb.org/anthology/W13-1716

[13] S. Malmasi and M. Dras, "Chinese Native Language Identification," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL-14)*. Gothenburg, Sweden: Association for Computational Linguistics, April 2014, pp. 95–99. [Online]. Available: http://aclweb.org/anthology/E14-4019

[14] S. Malmasi and A. Cahill, "Measuring Feature Diversity in Native Language Identification," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver, Colorado: Association for Computational Linguistics, June 2015.

[15] G. Berardi, A. Esuli, D. Marcheggiani, and F. Sebastiani, "ISTI@ TREC Microblog Track 2011: Exploring the Use of Hashtag Segmentation and Text Quality Ranking," in *TREC*, 2011.

[16] P. Norvig, "Natural language corpus data," *Beautiful Data*, pp. 219–242, 2009.

[17] D. A. Grossman, *Information retrieval: Algorithms and heuristics*. Springer, 2004, vol. 15.

[18] D. Molla and S. Karimi, "Overview of the 2014 ALTA Shared Task: Identifying Expressions of Locations in Tweets," in *Proceedings of the Australasian Language Technology Workshop (ALTA)*, Melbourne, Australia, 2014, p. 151.