Automatic Identification of Expressions of Locations in Tweet Messages using Conditional Random Fields

Fei Liu, Afshin Rahimi, Bahar Salehi, Miji Choi, Ping Tan, Long Duong
Department of Computing and Information Systems
The University of Melbourne
{fliu3,bsalehi,jooc1,pingt,lduong}@student.unimelb.edu.au
afshinrahimi@gmail.com

Abstract

In this paper, we propose an automatic identification model, capable of extracting expressions of locations (EoLs) within Twitter messages. Moreover, we participated in the competition of ALTA Shared Task 2014 and our best-performing system is ranked among the top 3 systems (2nd in the public leaderboard). In our model, we explored the validity of the use of a wide variety of lexical, structural and geospatial features as well as a machine learning model Conditional Random Fields (CRF). Further, we investigated the effectiveness of stacking and self-training.

1 Introduction

With the rise of social media, people have developed a fondness for posting not only their thoughts and opinions, but also content regarding their whereabouts (Liu et al., 2014). While the spatial information carried by tweets is crucial to a wide variety of location-based applications ranging from real-time disaster detection (Sakaki et al., 2010; Núñez-Redó et al., 2011; Yin et al., 2012) to targeted advertising (Tuten, 2008; Evans, 2012), only 26% of Twitter users specify their locations as granular as a city name (e.g. *Melbourne, Australia*) in their profile according to Cheng et al. (2010). Further, as little as 0.42% of all the tweets investigated by Cheng et al. (2010) are associated with the per-tweet geo-tagging feature (i.e. a latitude and longitude). To add to the complexity, on such highly-interactive yet informal social media platforms, people make heavy use of informal language, such as acronyms (e.g. *NYC*) and word shortenings (e.g. *St.*) due to the 140-character limit (Agarwal et al., 2011; Eisenstein, 2013; Han et al., 2013), making the identification task even more difficult. Despite the difficulties, identifying geospatial information in social media text has drawn much attention (Lingad et al., 2013).

Our focus in this paper is the automatic identification of EoLs in the text of tweets consisting of any specific reference to a geospatial location. A location, as defined by Lingad et al. (2013), consists of both *geographic location(s)*, such as country, city, river, or suburb, and *point(s)-of-interest* (POI (s)) which refer to hotels, shopping centres, and restaurants.

The task is closely related to Named Entity Recognition (NER). In this regard, Liu et al. (2011) and Ritter et al. (2011) report F-score of 77-78% at identifying spatial named entities in tweets. Matching place references in a gazetteer (Hill, 2000) is another widely-used approach. Paradesi (2011) investigated the approach of combining NER and external gazetteers. Further, Gelernter and Balaji (2013) built a geoparser incorporating the results of four parsers.

In our attempt to build an automatic EoL identification system, we employed a conditional random field (CRF) (Lafferty et al., 2001), which can be found and has proved to be successful in various Natural Language Processing (NLP) tasks (Sha and Pereira, 2003; Gimpel et al., 2011; Finkel et al., 2005; Ritter et al., 2011). In this paper, we present our approach to building such a system as well as a variety of features, such as lexical, structural and geospatial features and show major improvements on the task of EoL identification over earlier attempts. Our best-performing system is ranked among the top 3 systems (2nd in the public leaderboard).

The paper is organised as follows: the dataset and external resources used in our system is described in Section 2 and Section 3. We introduce the tools involved in this paper in Section 4. In Section 5 and Section 6, we provide the description of our system and analyse its performance with different feature sets respectively. We present

Fei Liu, Afshin Rahimi, Bahar Salehi, Miji Choi, Ping Tan and Long Duong. 2014. Automatic Identification of Expressions of Locations in Tweet Messages using Conditional Random Fields. In *Proceedings of Australasian Language Technology Association Workshop*, pages 171−176.

the conclusions in Section 7.

2 Dataset

We used the dataset introduced by (Lingad et al., 2013) to evaluate our proposed system. This dataset was also used in ALTA shared task 2014 and contains 1,942 tweets in the training set and 1,003 tweets were selected for the test set. According to (Lingad et al., 2013), around 89% of the tweets contain at least one location. The location mentions can be either in the text, in hashtags (e.g. #Australia), URLs or in mentions (e.g. @australia).

The dataset contains the list of tweet IDs and the locations mentioned in the respective tweets. At the time of extracting the tweets from twitter, 58 tweets in training set were not accessible.

3 External Resources

Apart from the training and test datasets, we introduce the additional datasets and resources involved in this project in this section.

3.1 User Meta Data

We extracted location meta information of the authors of the messages in the training data and created a list of such location mentions.

3.2 Text Retrieved from URLs

Additionally, for the purposes of self-training, we also downloaded the text of the articles whose URLs are contained in the tweets (37% contain URLs in the training set). Due to the unavailability of some URLs, we were only able to retrieve some of the articles.

3.3 GeoNames

As an external gazetteer, we adopted *GeoNames*[1] whose data can be downloaded to increase the coverage of our model since only a limited number of tweets were provided for training.

4 Tools

In this section, we introduce the tools we utilised in our system.

CRF++

`CRF++` is an open source, general-purpose implementation of CRF by Kudo (2005) and can be applied to a wide variety of NLP tasks. Since it only takes CoNLL format training and test data, we converted the training and test data.

Retrained StanfordNER

The Stanford named entity recogniser (Finkel et al., 2005) has proved to be effective when retrained over data containing EoLs (Lingad et al., 2013) even though evidence found by Liu et al. (2014) indicates otherwise. We retrained it over the training data and will refer to it as `Re-StanfordNER`.

GeoLocator

`GeoLocator` is a geoparser created by Gelernter and Balaji (2013) to geoparse informal messages in social media. The training data for this model was extracted from Twitter following the Februray 2011 earthquake in Christchurch New Zealand. It incorporates the output 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.

5 System Description

In this section, we describe our approach to creating an automatic EoL identification system.

5.1 Pre-processing

We pre-processed both the training and test dataset with lexical normalisation (using the dictionary created by Han et al. (2012)), POS tagging and full-text chunk parsing. Recognising the incompetent performance of traditional NLP tools when applied to social media text (Java, 2007; Becker et al., 2009; Yin et al., 2012; Preotiuc-Pietro et al., 2012; Baldwin et al., 2013; Gelernter and Balaji, 2013), we adopted `ARK Tweet NLP POS Tagger v0.3` (Owoputi et al., 2013) with the Penn Treebank tagset model for the task of word tokenisation and POS tagging. For chunk parsing, we used `OpenNLP`[2].

5.2 Features

We trained our model (based on `CRF++`) with various features, which can be categorised into three categories: lexical features, structural features and geospatial features. Note that we used a context window of 2 for each feature.

- Lexical features include lemmatised words (using NLTK (Bird et al., 2009), POS,

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

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

brief word class introduced by Settles (2004) where capitial and lowercase letters are replaced with 'A' and 'a', digits with '0' and all other characters with '_' and consecutive identical characters are collapsed into one (e.g. *#Adelaide* → *_Aa*), capitalisation and locative indicator (Liu, 2013).

- Structural features include position of the word in the chunk and POS of the first word in the chunk.
- Geospatial features include *GeoNames* geospatial feature class described by Liu (2013).

As pointed out by Wolpert (1992), stacking is able to generate better results than any single one of the trained model. We therefore also applyed stacking by combining the output of our `CRF++`-based model, `Re-StanfordNER` and `GeoLocator` and using them as three distinct features.

5.3 Self-training

Self-training, a semi-supervised learning algorithm, has proved to be successful, as reported by Plank et al. (2014), in Twitter POS tagging and NER tasks with an error reduction of 8–10% over the state-of-the-art system. We employed self-training using text retrieved from the URLs in the training and test dataset as the new test data. First, we train `CRF++` over the original gold-standard training data. Next, we predict on the new test data and expand the training data by including new instances from the new test data with prediction confidence higher than or equal to a threshold value. This process is repeated until there is no instance from the new test data to be added. Furthermore, we experiment with various threshold values.

5.4 Post-processing

In order to improve the recall of our model, we further include two post-processing methods: gazetteer matching and aggregation.

Gazetteer Matching

In addition to the machine learning approach, we also explored the use of external gazetteers and a matching algorithm. The algorithm, based on dynamic programming, searches the gazetteer case-sensitively for the maximum number of matched words in a sentence.

To further enable our model to detect directional words (e.g. *north*, *northern*) and common elements of toponyms (e.g. *street*, *road*), we also compiled a list of generic terms which are frequently used as part of an EoL by splitting entries in *GeoNames* into single tokens and including the top 500 most frequent words. Also, we created an algorithm capable of finding case-insensitive partial as well as whole-word matches.

Aggregation

Similar to the union operation of sets, we aggregated the prediction results of `CRF++` and `Re-StanfordNER` in the attempt to achieve higher recall, classifying a word as part of an EoL as long as it is identified in the output of at least one of two machine learning tools.

6 Evaluation

In this section, we present the performance of our system as well as analyses of the results. All the evaluation is based on the test data and the gold-standard annotations provided by the organiser. In addition to the mean F-score generated by the evaluation script provided by *Kaggle in Class*, we also include macro-averaged precision, recall and F-score to better understand the performance of our system with various feature setups.

The performance of our system is presented in Table 1. The performance attained using only word ($\mathcal{W}$) and POS ($\mathcal{P}$) with `CRF++` is better than `Re-StanfordNER` in precision but inferior in recall, resulting in a slightly lower macro-averaged F-score ($\mathcal{F}$) than that of `Re-StanfordNER`. Aggregating the two achieves a substantial gain in performance, boosting the macro-averaged F-score from 67.39 to 72.07. As we improved the performance of `CRF++` by adding more sophisticated features incrementally, the benefits of aggregation became less substantial, which is not that surprising considering the output of the `Re-StanfordNER` is already included and used as a feature in stacking. In most cases, the results with aggregation are better than those without aggregation. However, applying aggregation has negative impacts on the recall of `CRF++` with stacking, even though it enables the model to achieve a modest gain in F-score. The reasons for this remain unclear.

We also observed that stacking improved the performance on the whole test data substantially (a 3.85 increase in mean F-score without aggregation). Upon closer investigation of the impact of stacking on the performance on the test data, we discovered that stacking was less effective on the private test set (a 3.01 increase in mean F-score) than on the public one (a 4.7 increase in mean F-score), which might have been caused by the fact that `GeoLocator`, `Re-StanfordNER` and `CRF++` (with lexical, structural and geospatial features) overfit the public test data. Based on this, we suspect that the public test data is more similar to the training data than the private test data. Further, we created a Venn diagram of the output of the three systems and discovered that there is room for further improvement with stacking and that a 13.35 F1 point increase can be achieved if we had an oracle stacking algorithm.

		$\mathcal{GL}$	$\mathcal{RS}$	**CRF++**		
				+$\mathcal{W},\mathcal{P}$	+$\mathcal{L},\mathcal{S},\mathcal{G}$	+$\mathcal{ST}$
−$\mathcal{A}$	$\mathcal{P}$	61.76	62.96	65.53	68.81	72.22
	$\mathcal{R}$	65.31	72.34	69.35	72.95	**76.87**
	$\mathcal{F}$	63.48	67.32	67.39	70.82	74.47
	$\mathcal{MF}$	60.84	64.94	64.57	68.56	72.41
+$\mathcal{A}$	$\mathcal{P}$	–	–	72.33	74.14	**74.60**
	$\mathcal{R}$	–	–	71.81	74.07	76.49
	$\mathcal{F}$	–	–	72.07	74.10	**75.54**
	$\mathcal{MF}$	–	–	69.48	71.93	**73.57**

Table 1: Macro-averaged precision ($\mathcal{P}$), recall ($\mathcal{R}$), F-score ($\mathcal{F}$) and mean F-score ($\mathcal{MF}$) attained by using `GeoLocator` ($\mathcal{GL}$) and `Re-StanfordNER` ($\mathcal{RS}$) out of the box and adding each feature incrementally to `CRF++`. Features include word ($\mathcal{W}$), POS ($\mathcal{P}$), lexical features ($\mathcal{L}$), structural features ($\mathcal{S}$), geospatial features ($\mathcal{G}$) and stacking ($\mathcal{ST}$). $\mathcal{A}$ stands for aggregation. Evaluation based on the test data (the best $\mathcal{P}$, $\mathcal{R}$, $\mathcal{F}$and $\mathcal{MF}$ are in bold).

Also, we investigated the impact of the use of external gazetteers. The results are summarised in Table 2. Note that the two gazetteer matching algorithms were applied upon our best performing system so far, which is able to achieve a macro-averaged F-score of 75.54. Further, we discovered that including GeoNames was not beneficial to the overall performance as it introduces a number of false positives.

Method	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{MF}$
+$\mathcal{U},\mathcal{DT}$	**76.88**	**77.00**	**76.94**	**74.98**
+$\mathcal{U},\mathcal{G},\mathcal{DT}$	76.75	76.42	76.58	74.64

Table 2: Macro-averaged precision ($\mathcal{P}$), recall ($\mathcal{R}$), F-score ($\mathcal{F}$) and mean F-score ($\mathcal{MF}$) attained by using user meta data ($\mathcal{U}$), Geonames ($\mathcal{G}$) and the list of directional words and toponyms ($\mathcal{DT}$) (the best $\mathcal{P}$, $\mathcal{R}$, $\mathcal{F}$and $\mathcal{MF}$ are in bold).

Additionally, we also applied self-training with 4 different confidence threshold values ranging from .70 to .95 and the results are shown in Table 3. Note that self-training was applied to `CRF++` with lexical, structural and geospatial features, which results in a macro-averaged F-score of 70.82. While precision and recall fluctuate, no significant improvement can be observed in F-score despite the claim of 8–10% error reduction by Plank et al. (2014). Rather, the overall performance declined to around 68–69 in F-score.

Threshold	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{MF}$
.70	66.21	72.61	69.26	67.12
.80	65.59	**72.65**	68.94	66.76
.90	65.94	72.12	68.89	66.72
.90	**67.16**	72.23	**69.60**	**67.39**

Table 3: Macro-averaged precision ($\mathcal{P}$), recall ($\mathcal{R}$), F-score ($\mathcal{F}$) and mean F-score ($\mathcal{MF}$) attained by self-training with various threshold values (the best $\mathcal{P}$, $\mathcal{R}$, $\mathcal{F}$and $\mathcal{MF}$ are in bold).

7 Conclusions

We proposed an automatic EoL identification model which is able to work on Twitter messages. In this paper, we described our approach to building such a system based on a CRF. Moreover, we presented the performance of our system with various feature setups and discovered a variety of features which are helpful to the task, such as lexical, structural and geospatial features as well as stacking. Further, evidence indicates that the inclusion of external gazetteers and matching algorithms works well and contributes to the boost of the overall performance with the exception of *GeoNames*. Lastly, we found that self-training did not improve the performance. As future work, possible enhancement can be done on the stacking algorithm and the gazetteer matching approach.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media (LSM 2011)*, pages 30–38, Portland, USA.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. 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.

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010)*, pages 759–768, Toronto, ON, Canada.

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

Dave Evans. 2012. *Social media marketing: An hour a day*. John Wiley & Sons.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. 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.

Judith Gelernter and Shilpa Balaji. 2013. An algorithm for local geoparsing of microtext. *Geoinformatica*, 17(4):635–667.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2 (ACL 2011)*, pages 42–47, Portland, USA.

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

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalisation of short text messages. *ACM Transactions on Intelligent Systems and Technology*, 4(1):5:1–5:27.

Linda L. Hill. 2000. Core elements of digital gazetteers: Placenames, categories, and footprints. In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2000)*, pages 280–290, Lisbon, Portugal. Springer-Verlag.

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

Taku Kudo. 2005. Crf++: Yet another crf toolkit. *Software available at `http://crfpp.sourceforge.net`*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289, San Francisco, USA.

John Lingad, Sarvnaz Karimi, and Jie Yin. 2013. Location extraction from disaster-related microblogs. In *Proceedings of the 22Nd International Conference on World Wide Web Companion (WWW 2013)*, pages 1017–1020, Rio de Janeiro, Brazil.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (ACL 2011)*, pages 359–367, Portland, USA.

Fei Liu, Maria Vasardani, and Timothy Baldwin. 2014. 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)*, pages 9–16, Shanghai, China.

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

Manuela Núñez-Redó, Laura Díaz, José Gil, David González, and Joaquín Huerta. 2011. Discovery and integration of web 2.0 content into geospatial information infrastructures: a use case in wild fire monitoring. In *Proceedings of the 6th International Conference on Availability, Reliability and Security (ARES 2011)*, pages 50–68, Vienna, Austria.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. pages 380–390, Atlanta, USA.

Sharon Myrtle Paradesi. 2011. 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.

Barbara Plank, Dirk Hovy, Ryan McDonald, and Anders Søgaard. 2014. Adapting taggers to twitter with (less) distant supervision. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1783–1792, Dublin, Ireland.

Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. 2012. Trendminer: An architecture for real time analysis of social media text. In *Proceedings of 1st International Workshop on Real-Time Analysis and Mining of Social Streams (RAMSS 2012)*, Dublin, Ireland.

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

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*, pages 851–860, Raleigh, USA.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (JNLPBA 2004)*, pages 104–107, Geneva, Switzerland.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (HLT-NAACL 2003)*, pages 134–141, Edmonton, Canada.

Tracy L Tuten. 2008. *Advertising 2.0: social media marketing in a web 2.0 world*. Greenwood Publishing Group.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.

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