

Location Extraction from Twitter Messages using Bidirectional Long Short-Term Memory Model

Zi Chen^a, Badal Pokharel^b, Bingnan Li^c and Samsung Lim^d

School of Civil & Environmental Engineering, University of New South Wales, Kensington, Sydney, Australia

Keywords: Named Entity Recognition, Location Extraction, Social Media, Deep Learning Model.

Abstract: Texts are a common form to encode location information which can be used crucially in disaster scenarios. While Named Entity Recognition (NER) has been applied to location extraction from formal texts, its performance on informal and colloquial texts such as social media messages is unsatisfactory. The geo-entities in social media are often neglected or categorized into unknown or ‘other’ entity types such as person or organisation. In this paper we proposed a Bidirectional Long Short-Term Memory (LSTM) Neural Network to identify location information especially aiming to recognize rarely known local places in social media messages. The contribution of both syntactic and semantic features to the classification results was explored as well. The proposed method was validated on a Twitter dataset collected from typhoon-affected areas, showing promising performance in detecting location information.


1 INTRODUCTION


Social media can be beneficial in disaster risk reduction, response and recovery process (Ahmed, 2011), and Twitter stands out as an effective social media platform because of its global extension and the speed in which information gets disseminated (Chatfield and Brajawidagda, 2013). While location information can be crucial in disaster management, Twitter offers three types of information for extracting the location where the incident happens: (i) geo-tagged texts or geo-coordinates (ii) users’ locations in their profiles (iii) location mentions in the tweets. The previous studies show that the explicit and accurate information about the place where an event has happened can be gained with the tweets having geo-coordinates (Nakaji and Yanai, 2012). However, the tweets containing geo-coordinates are rare among the whole Twitter stream. The location information from the user’s profile can be an alternative to detection of the place of event, but its accuracy is not guaranteed. Therefore, the location mentions in the tweet text accounts the most for the


location recognition, which requires further text processing.


The text mining technique that is commonly used to extract place names from texts is Named Entity Recognition (NER). This method analyzes the text based on Part-Of-Speech (POS) tagging and labels a certain group of words as an entity category including person, organisation or location. The technique performs well for well-structured sentences and well-known places, but not so good for the extraction of local geo-entities from social media messages. It is because the texts are usually written in informal or random format due to geographic or non-geographic ambiguities.

In order to improve the extraction of unknown place names from the social media, we proposed a model based on Bidirectional Long Short-Term Memory (LSTM) Neural Network to recognize the local geo-entities mentioned in social media messages. Stanford Named Entity Recognizer was used to label the training data. The experimentation was carried out to incorporate both syntactic and semantic features in the model. In addition to word

^a  <https://orcid.org/0000-0001-5100-8393>

^b  <https://orcid.org/0000-0001-5010-9205>

^c  <https://orcid.org/0000-0003-3417-3295>

^d  <https://orcid.org/0000-0001-9838-8960>

embedding, POS tags, letter cases and prepositions are considered as well to augment the model.

2 RELATED WORK

The extraction of location mentions from texts is a long-studied problem since texts are one of the most common forms to encode geographic information, but in this process the geo-referents of the location mentions become ambiguous and even the boundaries of the location mentions in text are difficult to recognize without enough background information, especially when the place name is abbreviated for brevity. Naturally the place name extraction falls into sub-problems: entity delimitation and toponym disambiguation. The former delimitates the boundaries of a place name, which is the focus of our study, and the latter decides the most possible geo-referent to the place name.

While entity delimitation can be solved by matching with gazetteer (Sultanik and Fink, 2012) or hard-coded rules (Cunningham et al., 2001), the current NER tools or systems are generally more powerful in extracting location mentions from formal texts in terms of recall. One of the most renowned NER systems is Stanford NER (Finkel, Grenager and Manning, 2005) where a Conditional Random Field (CRF) model incorporates long distance features to identify named entities including location. Some researchers have also attempted further augmentation to the NER capability of extracting locations by constructing a gazetteer from word clusters (Kazama and Torisawa, 2008). Web NER aims at separating complex place names from web pages by utilizing capitalization cues and lexical statistics (Downey, Broadhead and Etzioni, 2007).

However, the NER systems have shown unsatisfactory performances on social media messages (Bontcheva et al., 2013) mainly owing to the informal and irregular expressions as well as the short texts. An outstanding performance has been achieved by Stanford NER when it is retrained by annotated tweets (Lingad, Karimi and Yin, 2013), but the annotation of social media messages is time-consuming and elusive on large scale text processing. Meanwhile LSTM is becoming a popular choice of NER owing to its suitability for sequence classification. Bi-LSTM-CRF networks using contextualized embeddings were employed for chemical NER (Awan et al., 2019). An exploratory study on Indonesian Twitter Posts (Rachman et al., 2018) used LSTM for NER and experimented on a

small dataset, on which it achieved a F1-score of 0.81 for location recognition.

Recent years have witnessed a lot of efforts attempting to address the problem of extracting location mentions from social media streams. A model to predict the occurrence of location mentions in a tweet was introduced and was found that this preliminary process can enhance the accuracy of entity delimitation (Hoang and Mothe, 2018). A statistical language model was built in (Al-Olimat et al., 2017) based on augmented and filtered region-specific gazetteers from online resources such as OpenStreetMap (OSM) to extract place names from tweets, a F1-score of 0.85 was achieved while no training data is required.

Unlike the studies above, we mainly focus on the location mentions referring to geo-entities of small scale, which rarely appear in most gazetteers. And a deep learning model is constructed and trained on Stanford NER annotated tweets, without manual annotation.

3 METHODOLOGY

3.1 Model

We used a Bidirectional LSTM (Bi-LSTM) model, which belongs to the category of Recurrent Neural Networks (RNNs), to label whether a word is an element of a location mention or not. RNN is an appropriate approach for sequence classification due to its capability of passing the output of one node to its successor, which can be interpreted as the influence of a word to the successive word. LSTM is a specialized RNN which performs better on long sequences while the impact of previous layers decay along with vanishing gradient in RNN. LSTM employs a mechanism called forget gates to control the information flow in the neural network.

On the basis of LSTM, Bi-LSTM model is trained on two directions of the input sequence, forward and backward, providing comprehensive context information of the target word. Bi-LSTM model has been implemented on NER tasks and shown competitive performances on benchmark datasets (Chiu and Nichols, 2016).

As seen in Figure 1, features of the input sequence are passed to the forward and backward LSTM layers respectively in our model, the two outputs are subsequently concatenated and passed to a fully connected layer. Finally, a softmax layer outputs the probability distribution over all the class labels.

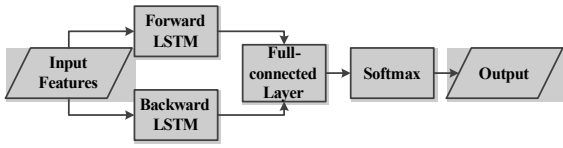


Figure 1: Bi-LSTM model architecture.

3.2 Features

Apart from the Bi-LSTM model, we tested the impacts of four categories of features on the classification results: word embedding, POS tags, capitalization and prepositions. Word embedding is a word representation method that maps words to continuous vectors in higher dimensions. It is grounded on Distributional Hypothesis that words with similar contexts are inclined to have similar semantic meanings (Harris, 1954), and word embedding encodes the context information in the vector. Therefore, the similarity of two words can be evaluated by the cosine value of their word vectors.

One of the most popular approaches of word embedding is the Skip-Gram model. It is a fully connected neural network that takes the one hot encoding vector of the target word w_i as input and produces the conditional probability that an arbitrary word w_j from the vocabulary occurs in the context window of w_i . Parameters θ of the model are optimized by maximizing the log-likelihood sum of $P(w_j|w_i; \theta)$.

The other three features, POS tags, capitalization and prepositions are regarded as categorical data and encoded into integers. POS tags are the labels that denote the part of speech of a word such as noun, verb or adjective. Since the location mentions are mostly comprised of nouns, POS tags can signify the possible occurrence of location mentions.

Another factor that could influence the identification of location mentions is the capitalization or case of a word. We consider the capitalization in three categories: upper case, lower case and title case.

Prepositions that can describe places or directions such as ‘at’ or ‘in’ are also important indicators of location mentions. We collect all the prepositions regarding locations in a list and encode the occurrence of prepositions in tweets as the list index. The three indexed categorical features are mapped and trained into corresponding embeddings in the Bidirectional LSTM model and concatenated with the pre-trained word embedding for each token or word as can be seen in Figure 2.

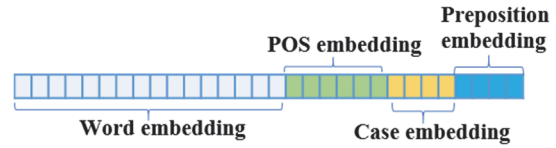


Figure 2: Concatenated feature embeddings.

As the features employed to describe the data are heuristic, we also explore the contribution of each feature to the correct location mention labelling by adjust the dimension of the feature in the following experiments.

4 EXPERIMENTS

4.1 Dataset

The dataset we used in the experiments was collected from Twitter via its official application programming interfaces (APIs). From to 21st to 30th of August 2017 two consecutive typhoons *Hato* and *Pakhar* affected Southern China area while two populated coastal cities Hong Kong and Macau were severely impacted. We collected and deduplicated tweets from the two areas during the typhoon-impacted period and extracted typhoon-related tweets by an augmented Convolutional Neural Network (CNN), resulting in 10,996 tweets ready for location extraction.

4.2 Pre-processing

The named entities including location mentions in the tweets were annotated via Stanford NER tools, and POS tagging was implemented likewise using Stanford POS tagger. Among 10,996 tweets location mentions were found in 4,215 tweets. Considering that Stanford NER tools can ignore the unknown places only familiar to locals, we only used the 4,215 positive tweets as training data.

In order to test the proposed model on its capability of detecting local place names, we manually selected and labelled 100 tweets in which Stanford NER misclassified the place names. The word embeddings with dimension of 200 were pre-trained by the Skip-Gram model on the Twitter stream we collected. The embeddings of POS tags, capitalization and prepositions were trained along with the model.

4.3 Model Training

In the training process we leave one tenth of the training data, which is 422 tweets, for validation, and

3,793 for training. If the validation loss is not improved in 5 epochs the training process will be terminated.

Figure 3 shows the variation of accuracy and loss for training set and validation set when all the four features are employed. The validation accuracy is improved to 96% in comparison to 92% when only word embedding is used, which proves that the utilisation of syntactic features such as POS tags can boost the model performance. In the following subsection we will examine how the features influence the model on the test set.

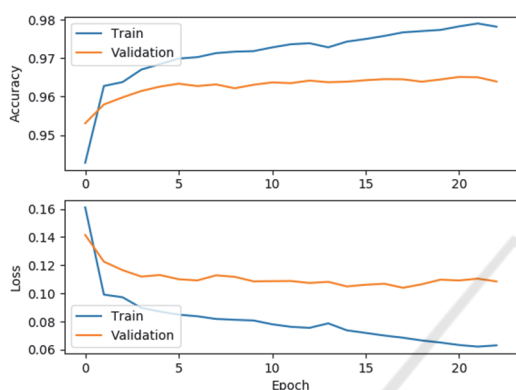


Figure 3: Model training.

4.4 Feature Evaluation

4.4.1 Individual Application of Feature

We first verify every feature separately on the test set. In the training process all the tweets are padded to the same length for alignment, the padded element is labelled 0 as its location mention category. A word that is part of a location mention is labelled 1, and other words are 2. The positive label 1 accounts for an extreme small proportion in the true labels of test set at 1.18%.

In Table 1 we present the precision, recall and F1-score of each label when only word embedding is applied. The proposed model performs well on label 0 and 2 which are the majority of labels but fails to predict any location mentions, resulting in all the metrics of label 1 ending up being 0%.

Table 1: Classification results on word embedding.

Label	Precision	Recall	F1-score
0	0.96	0.98	0.97
1	0.00	0.00	0.00
2	0.92	0.92	0.92

We further investigate in the classification performance of other features on label 1 as shown in Table 2. Best performance is achieved by POS tags on every metric and the employment of prepositions produces greatly inferior outcomes compared to POS tags and capitalization.

Table 2: Classification results on other features.

Feature	Precision	Recall	F1-score
POS tags	0.31	0.51	0.38
Capitalization	0.27	0.44	0.33
Prepositions	0.23	0.10	0.14

In general, individual application of feature produces unsatisfactory results, which is mostly caused by the imbalanced nature of data where location mentions are exceedingly infrequent.

4.4.2 Combination of Features

After examining the individual application of the semantic and syntactic features, the effects of hyper-parameters of combined features are validated on the test set to reach an optimal outcome for label 1. The four hyper-parameters to be altered and their candidate values are listed in Table 3.

Table 3: Hyper-parameters to be tested.

Padded length	POS embedding	Case embedding	Preposition embedding
50	30	20	10
60	40	30	20
70	50	40	30

Padded length is the alignment length that every tweet will be padded or truncated to. The other parameters are the dimensionalities of POS embedding, case embedding and preposition embedding.

In Table 4 with dimensionalities of case embedding and preposition embedding fixed, padded length and POS embedding dimension vary in respective scopes. Overall improvements up to 12% on F1-score are observed in comparison with the best performance obtained by individual application of POS tags.

Although no general rules on the variation of padded length and POS embedding are discovered, it can be obviously concluded that the incorporation of semantic and syntactic features enhances the model performance.

Table 4: Classification results under different padded lengths and POS embedding dimensions.

Padded length	POS embedding	Precision	Recall	F1-score
50	30	0.42	0.46	0.44
50	40	0.42	0.61	0.50
50	50	0.35	0.61	0.44
60	30	0.38	0.58	0.46
60	40	0.35	0.67	0.46
60	50	0.32	0.53	0.40
70	30	0.31	0.60	0.41
70	40	0.36	0.58	0.44
70	50	0.37	0.55	0.44

Given fixed padded length and POS embedding dimension, F1-score does not vary considerably along with case or preposition embedding in Table 5. But case embedding and preposition embedding of 20 or 30 are preferred.

Table 5: Classification results under different case embedding and preposition embedding dimensions.

Case embedding	Preposition embedding	Precision	Recall	F1-score
20	10	0.31	0.58	0.41
20	20	0.36	0.58	0.44
20	30	0.33	0.61	0.43
30	10	0.31	0.53	0.39
30	20	0.35	0.61	0.44
30	30	0.41	0.47	0.44
40	10	0.32	0.58	0.41
40	20	0.32	0.44	0.37
40	30	0.27	0.47	0.34

We also compare the proposed model to one of off-the-shelf NER tools TwitterNLP (Ritter et al., 2011) which is specialised for Twitter data. The experiments are conducted on the same test set. In Table 6 it can be seen that precision, recall and F1-score have been largely increased, which proves the capability of the proposed model on identifying location mentions from social media messages.

Table 6: Comparison to TwitterNLP.

Model	Precision	Recall	F1-score
Bi-LSTM	0.42	0.61	0.50
TwitterNLP	0.30	0.14	0.19

5 CONCLUSION

In this paper we proposed a model based on Bi-LSTM Neural Network to detect the local geo-entities mentioned in social media messages that are ignored by Stanford NER. We have attained competitive results on classification performance of the proposed model with noisy training data. However, there is still room for enhancement in the Bi-LSTM model especially on precision. To cope with the issue of imbalanced datasets, we can further under-sample the majority negative labels, or integrate our model with methods such as decision trees which are suitable for imbalanced datasets. On the other hand, the training data labelled by Stanford NER possibly contains many false negative labels, where the concept of multiple instance learning can be considered to solve this problem.

ACKNOWLEDGEMENTS

This research is sponsored by China Scholarship Council (CSC).

REFERENCES

- Ahmed, A., 2011. Use of social media in disaster management. In: *International Conference on Information Systems 2011, ICIS 2011*. pp.4149–4159.
- Al-Olimat, H.S., Thirunarayan, K., Shalin, V. and Sheth, A., 2017. Location Name Extraction from Targeted Text Streams using Gazetteer-based Statistical Language Models. [online] Available at: <<http://arxiv.org/abs/1708.03105>> [Accessed 9 Dec. 2019].
- Awan, Z., Kahlke, T., Ralph, P.J. and Kennedy, P.J., 2019. Chemical named entity recognition with deep contextualized neural embeddings. In: *IC3K 2019 - Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. pp.135–144.
- Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M.A., Maynard, D. and Aswani, N., 2013. *TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text*. [online] Available at: <<https://gate.ac.uk/wiki/twitie.html>> [Accessed 9 Dec. 2019].
- Chatfield, A.T. and Brajawidagda, U., 2013. Twitter early tsunami warning system: A case study in Indonesia's natural disaster management. In: *Proceedings of the Annual Hawaii International Conference on System Sciences*. pp.2050–2060.
- Chiu, J.P.C. and Nichols, E., 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, [online] 4, pp.357–370. Available at:

- <<http://nlp.stanford.edu/projects/glove/>> [Accessed 9 Dec. 2019].
- Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V., 2001. GATE: an Architecture for Development of Robust HLT Applications. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, [online] (July), p.168. Available at: <<http://gate.ac.uk/>> [Accessed 9 Dec. 2019].
- Downey, D., Broadhead, M. and Etzioni, O., 2007. Locating complex named entities in web text. In: *IJCAI International Joint Conference on Artificial Intelligence*. pp.2733–2739.
- Finkel, J.R., Grenager, T. and Manning, C., 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In: *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. pp.363–370.
- Harris, Z.S., 1954. Distributional Structure. *Distributional Structure, WORD*, [online] 10(3), pp.146–162. Available at: <<https://www.tandfonline.com/action/journalInformation?journalCode=rwr20>> [Accessed 2 Jul. 2019].
- Hoang, T.B.N. and Mothe, J., 2018. Location extraction from tweets. *Information Processing and Management*, 54(2), pp.129–144.
- Kazama, J. and Torisawa, K., 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In: *ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*. pp.407–415.
- Lingad, J., Karimi, S. and Yin, J., 2013. Location extraction from disaster-related microblogs. In: *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*. [online] New York, New York, USA: ACM Press, pp.1017–1020. Available at: <<http://dl.acm.org/citation.cfm?doid=2487788.2488108>> [Accessed 8 Jul. 2019].
- Nakaji, Y. and Yanai, K., 2012. Visualization of real-world events with geotagged tweet photos. In: *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2012*. pp.272–277.
- Rachman, V., Savitri, S., Augustianti, F. and Mahendra, R., 2018. Named entity recognition on Indonesian Twitter posts using long short-term memory networks. In: *2017 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017*. pp.228–232.
- Ritter, A., Sam, C., Mausam and Etzioni, O., 2011. Named entity recognition in tweets: An experimental study. In: *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, pp.1524–1534.
- Sultanik, E.A. and Fink, C., 2012. Rapid geotagging and disambiguation of social media text via an indexed gazetteer. In: *ISCRAM 2012 Conference Proceedings - 9th International Conference on Information Systems for Crisis Response and Management*.