Learning to Geolocalise Tweets at a Fine-Grained Level

Jorge David Gonzalez Paule[1], Yashar Moshfeghi[2], Craig Macdonald[3] and Iadh Ounis[4]
[1,3,4] University of Glasgow. [2] University of Strathclyde. Glasgow, Scotland, UK
j.gonzalez-paule.1@research.gla.ac.uk,yashar.moshfeghi@strath.ac.uk,{Craig.Macdonald,Iadh.Ounis}@glasgow.ac.uk

ABSTRACT

Fine-grained geolocation of tweets has become an important feature for reliably performing a wide range of tasks such as real-time event detection, topic detection or disaster and emergency analysis. Recent work adopted a ranking approach to return a predicted location based on content-based similarity to already available individual geotagged tweets. However, this work made use of the IDF weighting model to compute the ranking, which can diminish the quality of the Top-N retrieved tweets. In this work, we adopt a learning to rank approach towards improving the effectiveness of the ranking and increasing the accuracy of fine-grained geolocalisation. To this end, we propose a set of features extracted from pairs of geotagged tweets generated within the same fine-grained geographical area (squared areas of size 1 km). Using geotagged tweets from two cities (Chicago and New York, USA), our experimental results show that our learning to rank approach significantly outperforms previous work based on IDF ranking, and improves the accuracy of tweet geolocalisation at a fine-grained level.

CCS CONCEPTS

• **Information systems → Learning to rank**;

KEYWORDS

Information Retrieval; Learning to Rank; Tweet Geolocalisation

ACM Reference Format:
Jorge David Gonzalez Paule[1], Yashar Moshfeghi[2], Craig Macdonald[3] and Iadh Ounis[4]. 2018. Learning to Geolocalise Tweets at a Fine-Grained Level. In *2018 ACM Conference on Information and Knowledge Management (CIKM'18), October 22–26, 2018, Torino, Italy*. ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3269206.3269291

1 INTRODUCTION

The ever-increasing activity on Twitter has generated abundant information about the location of users in the real world at a fine-grained level (i.e. at a street, building or neighbourhood level) [19]. Given the richness of such data, new opportunities have emerged for a broad range of Information Retrieval (IR) applications such as real-time event detection [1] or disaster and emergency analysis [14]. However, only a very small sample of tweets in the Twitter stream (1% to 2%) contain geographical information [9]. Thus, inferring the geolocalisation of non-geotagged tweets has become an important yet challenging task. In this paper, we propose a novel approach for fine-grained geolocalisation of non-geotagged tweets, where a fine-grained level is defined as squared areas of size 1 km.

CIKM '18, October 22–26, 2018, Torino, Italy

Recently, Gonzalez Paule et al. [8] proposed to rank individual geotagged tweets using an IDF weighting model and combine evidence from the Top-N tweets using a weighted majority voting algorithm. However, considering only IDF weighting to perform the ranking can reduce the quality of the Top-N tweets. In this paper, we aim to improve the accuracy of geolocalisation by improving the quality of the Top-N ranked tweets. To this end, we adopt a learning to rank approach [4] to tackle fine-grained geolocalisation of tweets and propose a set of features to rank individual geotagged tweets based on their geographical proximity to a given non-geotagged tweet.

In particular, the contributions of this paper are two-fold. First, we propose a learning to rank approach for the fine-grained tweet geolocalisation problem. Also, we propose multiple types of features extracted from pairs of geotagged tweets located in the same fine-grained geographical area. Second, we evaluate our proposed approach using a ground truth of geotagged tweets gathered from two different cities and investigate the best combination of features.

2 BACKGROUND

The problem of fine-grained geolocalisation has been explored recently in the literature. Previous work followed the approach of dividing the geographical space into a set of predefined areas of a given size [11, 15]. The authors aggregated the texts of the geotagged tweets within that areas and computed the probability of a non-geotagged tweet to be generated in an area, thus returning the most likely location. Kinsella et al. [11] reduced the granularity of the areas from country level to the postal code level. However, their results showed a significant decrease in accuracy. To improve the performance, Paraskevopoulos et al. [15] reduced the size of the areas to squares of side length 1 km. Nonetheless, the drawback of these works is that they aggregated the texts of the geotagged tweets to represent an area, which is noisy and affects accuracy.

To tackle this drawback, more recently Gonzalez Paule et al. [8] proposed to treat each geotagged tweet individually and rank them by their content-similarity to a given non-geotagged tweet. Then, a weighted majority voting algorithm is used to select the most common area – squares of size 1 km associated to each tweet – within the Top-N most similar geotagged tweets. The majority voting is weighted using information about the credibility of the users that posted the tweets in the rank. This approach increased significantly the accuracy of geolocalisation. However, only two sources of evidence were used to perform geolocalisation: content-based similarity and user information. In contrast to Gonzalez Paule et al. [8], we aim to increase the accuracy of geolocalisation by adopting a learning to rank approach that learns from multiple tweet features to obtain a higher quality Top-N ranking of most similar geotagged tweets that are then fed into the majority voting algorithm.

The idea of using machine learning to learn more effective ranking functions (learning to rank) has been widely used in IR [12].

Table 1: Features computed in this paper for fine-grained geolocalisation of tweets for query-tweet and doc-tweet.

Features	Description	Total
Query Features and Document Features		
Hashtags	Number of hashtags in the text.	2
Mentions	Number of mentions in the text.	2
Urls	Number of urls in the text.	2
Entities	Number of entities in the text.	2
Verbs	Number of verbs in the text.	2
Adverbs	Number of adverbs in the text.	2
Adjectives	Number of adjectives in the text.	2
Checkin	Whether the tweet is a Foursquare checkin.	2
Hour	The hour of the day (0 to 24h) the tweet was posted.	2
Weekday	The day of the week (Monday to Sunday) the tweet was posted.	2
User Ratio	User credibility ratio [8]	2
Query-dependent (relation between query-tweet and doc-tweet)		
Hashtags	Shared number of Hashtags.	1
Mentions	Shared number of Mentions	1
User	Both tweets belong to the same user.	1
Hour	Both tweets posted the same hour of the day (0h to 24h).	1
Weekday	Both tweets posted the same day of the week (Monday to Sunday.	1
Cosine Similarity	Cosine similarity [13] between the texts of q_i and d_i.	1
Total Features		28

Learning to rank has been applied to solve retrieval tasks using web documents or large text documents. However, the ranking of tweets is challenging due to the inherent characteristics of Twitter posts – short text and informal language. This issue has been widely studied for social media search tasks, where learning to rank has been demonstrated to benefit effectiveness [4]. In this paper, we adopt learning to rank techniques to tackle the tweet fine-grained geolocalisation problem. We re-rank geotagged tweets based on their geographical proximity to a given non-geotagged tweet and propose a set of features for the fine-grained geolocalisation task.

3 LEARNING TO GEOLOCALISE

The aim of this paper is to improve the accuracy of fine-grained geolocalisation by improving the ranking of the Top-N most content-based similar geotagged tweets (denoted as a **doc-tweet**) to a given non-geotagged tweet (denoted as a **query-tweet**). To this end, we aim to learn a ranking function to re-rank doc-tweets based on their geographical proximity to the query-tweet. As our ranking function, we empirically select LambdaMART [17] as the best performing one in preliminary experiments, detailed in Section 4.2.

To train the ranking function, we use the training set described in Section 4. In order to label pairs of geotagged tweets in the training set, we first divide the geographical space of interest into a grid of fine-grained squared areas of size 1 km and associate each geotagged query-tweet and doc-tweet to their corresponding area based on their location. Then, pairs of tweets posted in the same area (i.e. distance 1 km or less) are labelled as positive. On the other hand, pairs of tweets posted in different areas (i.e. distance more than 1 km) are labelled as negative.

After the training process, we use our learned model to re-rank doc-tweets based on their probability of being posted in the same area as the query-tweet. Finally, inspired by previous work [8], we apply a majority voting algorithm to select the predicted location - a squared area of size 1km - within the Top-N doc-tweets.

We propose a set of features to model fine-grained tweet geolocalisation. In total, we exploit 28 features (see Table 1) grouped into three categories: content quality features, geographical features and similarity features. We compute document features extracted from the doc-tweet and query features extracted from the query-tweet (content and geographical features), as well as query-dependent features (similarity features) to model the relationship between query-tweets and doc-tweets.

Content Quality Features. The higher the quality of a tweet is, the more valuable information it provides. Previous research has shown the usefulness of content quality features of a tweet for learning to rank [5]. Inspired by these works, we modelled the quality of a tweet by extracting indicators of the richness of its text. First, we exploit characteristics of the Twitter social network by counting the number of hashtags, number of mentions and number of URLs of the tweet. Second, we utilise natural language techniques to count the number of entities, verbs, adjectives, nouns and adverbs in the text.

Geospecific Features. In addition to previous state-of-the-art features, we added new features as signals for geolocalisation by extracting geospecific information contained within the query-tweet and the doc-tweet. First, we check if the tweet corresponds to a Foursquare check-in. Foursquare[1] is a social media network in which users can do check-ins at venues when they visit them. Users have the option of generating a tweet sharing this information with their followers along with the geolocation of the venue. Second, following Gonzalez et al. [8] approach, we compute a credibility score for the doc-tweet which represents the posting activity of the user that generated the tweet. A doc-tweet posted by a user with a high score is more likely to be representative of a geolocalisation. The credibility score is based on the ratio of tweets posted by a user at a fine-grained distance (1 km) to other similar tweets (Top-N). We utilise the training and validation sets described in Section 4 to compute the score, using the Top-N tweets with values of N of 3, 5, 7 and 9.

Finally, different types of events tend to occur at different hours of the day or days of the week. For instance, people usually visit clubs at nights and weekends. Thus, if two tweets were posted in the same time frame, their content is likely to be related to the same type of events that are recurrent in the same location.

Similarity Features. Query-dependent features aims to model the relationship between the query-tweet and the doc-tweet. These set of features are presented in Table 1. The intuition behind these features is that when people visit a certain location, they make use of social media to describe their surroundings or events occurring

[1] http://www.foursquare.com

in the location. This means that many of the generated tweets will share the same characteristics. Therefore, the similarities between the two tweets are a strong indicator of their geolocalisation. To model the similarity between the query-tweet and the doc-tweet, we first compute their cosine similarity [13]. Second, we count the number of common entities, mentions and hashtags, and check if both tweets were posted by the same user. Finally, we calculate if the query-tweet and the doc-tweet were generated in the same hour of the day or on the same day of the week.

4 EXPERIMENTS

In this section, we describe our experiments for evaluating our learning to rank approach for fine-grained geolocalisation of tweets. Our datasets consist of a ground truth sample of English geotagged tweets[2] collected during March 2016 and located in Chicago (132,751 geotagged tweets) and New York (153,144 geotagged tweets), USA. To evaluate our approach, we divide each dataset into three subsets. First, we consider the geotagged tweets posted during the first three weeks of March as our document set, resulting of 100,176 for Chicago and 111,292 for New York. Second, we randomly divide the last week of March into background-queries set and testing-queries set to ensure the same characteristics. The background-queries set consists of 16,262 geotagged tweets for Chicago, and 20,982 geotagged tweets for New York. Finally, the testing-queries set contains 16,313 geotagged tweets for Chicago and 20,870 geotagged tweets for New York

Next, we create our training set for learning to rank by performing a retrieval task (using IDF weighting model) with the geotagged tweets in the background-queries set as query-tweets, and the geotagged tweets in the documents set as doc-tweets. We use the generated pairs of query-tweet and doc-tweet as a training set to learn our learning to rank algorithm. We then perform the same task but using the query-tweets in the testing-queries set to create the testing set for evaluating our learning to rank approach.

Lastly, we index every geotagged tweet in the documents set using the Lucene platform[3], and pre-process them by removing stopwords and applying Porter stemming. Moreover, we preserve retweets, usernames and hashtags as tokens in the dataset. The reason behind preserving retweets is that when a user retweets a content, the geolocation of the original tweets is not necessarily preserved. Moreover, the similarity between a tweet and its retweet is high, therefore we can assign the location of the original tweet to the retweet.

4.1 Metrics.

We report the following metrics for evaluating the effectiveness of our approach:

Average Error distance (*km*): We compute the distance on Earth (Haversine formula [16]) between the predicted location and the real coordinates of the tweet in our ground truth. As the predicted location is an area (see Section 4.2), the distance between the ground truth coordinate and the centroid of the area is calculated. Lower values indicates better performance.

Accuracy@1km: We calculate whether the centroid of the predicted area lies within a radius of 1 km from the real location of a tweet. Higher values indicated better performance.

Coverage: We consider Coverage as the fraction of tweets in the test set from which our approach finds a geolocation regardless of the distance error. Higher values indicated better performance.

4.2 Models

In total, we implement five approaches (explained in detail below), including a state-of-the-art model as the baseline. Following previous works definition of fine-grained level [8, 15] we create a grid structure of squared areas with a side length of 1 km (denoted by "fine-grained grid"), which is utilised in all the models.

Baseline: The baseline model is an implementation of the work by Gonzalez Paule et al. [8]. This approach uses an IR weighting model to retrieve the Top-N most content-based similar tweets to a given non-geotagged tweet. Finally, the approach applies a weighting majority voting algorithm to obtain the most voted fine-grained location. The votes are given by the tweets within the Top-N rank, which are associated with their corresponding squared area of the fine-grained grid described above. Moreover, each vote is weighted by a user credibility score that is associated to the user that generated the tweet. This score is calculated based on the ratio of tweets posted by the user that are highly similar to other tweets posted at 1 km distance.

Then, we index and preprocess each of the geotagged tweets (see Section 4) as a single document. After indexing the tweets, we generate the Top-N rank of geotagged tweets using IDF weighting model and apply the weighted majority voting algorithm to obtain the final predicted location. In our experiments, we consider the Top-3, -5, -7 and -9 for evaluation.

L2Geo: Our proposed learning to geolocalise approach, described in Section 3. We empirically select the best performing configuration for our approach. We experiment using MART [7], RankNet [3], RankBoost [6], AdaRank [18], LambdaMART [17] and Random Forests [2] as ranking functions. Also, we configure the ranking functions to re-rank the Top-10 and Top-50 geotagged tweets, and optimise NDCG@N with N with values of 3, 5, 7, 10 during the training process. Finally, LambdaMART [17] configured to optimise NDCG@3, and re-ranking the Top-10 retrieved tweets showed to be the best performing configuration. Due to lack of space, we do not report detailed results of these experiments in this paper and will be considered in future work.

Additionally, in order to assess the best set of features for fine-grained geolocalisation, we built four different versions of our approach that use different combinations of the features described in Section 3: **L2Geo** which incorporates all the features, **L2Geo_Sim** which uses only the set of *similarity features*, **L2Geo_Content** which utilises only the set of *content quality features* and **L2Geo_Geo** which uses only the set of *geographical features*.

5 EXPERIMENTAL RESULTS

Table 2 presents average error distance (*A_Err_km*), accuracy at 1km (*Acc@1km*) and *Coverage* for different configurations (*Config*) of our learning to rank approaches (*L2Geo*, *L2Geo_Sim*, *L2Geo_Content* and *L2Geo_Geo*) against the baseline model described in Section 4.

[2]Geotagged and non-geotagged tweets share the same characteristics [10].
[3]https://lucene.apache.org/

Table 2: Results for Chicago dataset (left) and New York dataset (right). The tables present the metrics described in Section 4.1 for our proposed approach (L2Geo) against our Baselines, using the Top-N (Top-N) elements in the rank. Significant differences w.r.t our best Baseline (Baseline_Top-9) are denoted by $*$ (p<0.01).

		Chicago			New York		
Model	**Config**	**A_Err_km↓**	**Acc@1km↑**	**Coverage↑**	**A_Err_km↓**	**Acc@1km↑**	**Coverage↑**
Baseline	Top-3	3.849	61.17%	**83.28%**	4.234	52.33%	**75.84%**
Baseline	Top-5	3.669	62.78%	79.08%	4.362	51.98%	75.09%
Baseline	Top-7	3.170	66.82%	70.41%	4.008	54.81%	67.83%
Baseline	Top-9	2.576	71.29%	62.28%	3.476	59.23%	59.94%
L2Geo	Top-3	0.939$*$	92.92%$*$	38.12%$*$	1.373$*$	87.18%$*$	28.27%$*$
L2Geo	Top-5	0.671$*$	95.38%$*$	28.18%$*$	0.862$*$	93.43%$*$	21.65%$*$
L2Geo	Top-7	0.557$*$	96.33%$*$	23.22%$*$	0.679$*$	95.61%$*$	18.68%$*$
L2Geo	Top-9	**0.483**$*$	96.73%$*$	19.72%$*$	**0.622**$*$	**96.64%**$*$	16.53%$*$
L2Geo_Sim	Top-3	1.207$*$	89.73%$*$	32.29%$*$	1.198$*$	88.7%$*$	26.99%$*$
L2Geo_Sim	Top-5	0.759$*$	94.18%$*$	25.27%$*$	0.824$*$	93.58%$*$	20.69%$*$
L2Geo_Sim	Top-7	0.593$*$	95.84%$*$	21.94%$*$	0.703$*$	95.28%$*$	18.09%$*$
L2Geo_Sim	Top-9	0.503$*$	96.70%$*$	19.49%$*$	0.634$*$	96.34%$*$	16.22%$*$
L2Geo_Content	Top-3	1.297$*$	88.71%$*$	32.73%$*$	1.342$*$	86.85%$*$	26.35%$*$
L2Geo_Content	Top-5	0.828$*$	93.52%$*$	24.97%$*$	0.911$*$	92.99%$*$	20.56%$*$
L2Geo_Content	Top-7	0.670$*$	95.42%$*$	21.41%$*$	0.762$*$	95.04%$*$	18.26%$*$
L2Geo_Content	Top-9	0.491$*$	96.73%$*$	19.15%$*$	0.669$*$	96.35%$*$	16.39%$*$
L2Geo_Geo	Top-3	1.333$*$	88.86%$*$	31.52%$*$	1.538$*$	86.32%$*$	27.74%$*$
L2Geo_Geo	Top-5	0.779$*$	94.00%$*$	24.31%$*$	0.926$*$	93.06%$*$	21.34%$*$
L2Geo_Geo	Top-7	0.579$*$	95.87%$*$	20.95%$*$	0.778$*$	95.19%$*$	18.43%$*$
L2Geo_Geo	Top-9	0.489$*$	**96.74%**$*$	18.99%$*$	0.668$*$	96.28%$*$	16.51%$*$

We observe that our learning to rank approach, in both datasets, outperforms the baseline in terms of accuracy and error distance, but with the cost of a decrease in coverage. In particular, compared to the best performing baseline (Baseline at Top-9) our approach (*L2Geo* at Top-3) increases accuracy from 71.29% to 92.92% in the Chicago dataset, and from 59.23% to 87.18% in the New York dataset. Additionally, the average error distance is reduced from 2.576 km to 0.939 km and 3.476 km to 1.373 km for Chicago and New York respectively. However, coverage is reduced from 62.28% to 38.12% in Chicago, and 59.94% to 28.27% in New York.

Upon analysing the above-mentioned table, we observe that *L2Geo*, which combines all the proposed features, exhibits improvements over the rest of the learning to rank models that use subsets of features. Additionally, comparing our learning to rank models that incorporate different subsets of features, the effectiveness of *L2Geo_Sim* shows that *Similarity features* are the most informative type of features compared to *L2Geo_Content* and *L2Geo_Geo*

6 CONCLUSIONS

In this work, we tackled the fine-grained tweet geolocalisation task. We proposed a set of features for modelling the task and investigated their effectiveness when integrated into a learning to rank technique combined with a majority voting algorithm. To demonstrate the effectiveness of our approach, we conducted an experiment on two datasets of English geotagged tweets. Our results showed improvements in terms of accuracy of geolocalisation using our learning to rank approach (*L2Geo*) with respect to the baseline, utilising all the proposed features. Additionally, we observed that compared to other types of features, *Similarity features* (*L2Geo_Sim*) are the most informative. Future work will examine features individually and investigate the best combination of them, as well as explore other possible features for fine-grained geolocalisation.

REFERENCES

[1] Farzindar Atefeh and Wael Khreich. 2015. A survey of techniques for event detection in twitter. *Computational Intelligence* 31, 1 (2015), 132–164.

[2] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[3] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proc. 22nd ACM ICML, pages 89–96.*

[4] Fuxing Cheng, Xin Zhang, Ben He, Tiejian Luo, and Wenjie Wang. 2012. A survey of learning to rank for real-time twitter search. In *Joint International Conference ICPCA/SWS, pages 150–164.*

[5] Fuxing Cheng, Xin Zhang, Ben He, Tiejian Luo, and Wenjie Wang. 2012. A survey of learning to rank for real-time twitter search. In *ICPCA-SWS*. Springer, 150–164.

[6] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of machine learning research* 4, Nov (2003), 933–969.

[7] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[8] Jorge David Gonzalez Paule, Yashar Moshfeghi, Joemon M Jose, and Piyushimita Vonu Thakuriah. 2017. On Fine-Grained Geolocalisation of Tweets. In *Proc. 3rd ACM ICTIR, pages 313–316.*

[9] Mark Graham, Scott A Hale, and Devin Gaffney. 2014. Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer* 66, 4 (2014), 568–578.

[10] Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research* 49 (2014), 451–500.

[11] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. I'm eating a sandwich in Glasgow: modeling locations with tweets. In *Proc. 3rd ACM SMUC, pages, 61–68.*

[12] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.

[13] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval.* Vol. 1. Cambridge university press Cambridge.

[14] Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2016. EAIMS: Emergency Analysis Identification and Management System. In *Proc. 39th ACM SIGIR, pages 1101–1104.*

[15] Pavlos Paraskevopoulos and Themis Palpanas. 2015. Fine-Grained Geolocalisation of Non-Geotagged Tweets. In *Proc. IEEE/ACM ASONAM, pages 105–11.*

[16] C Carl Robusto. 1957. The cosine-haversine formula. *The American Mathematical Monthly* 64, 1 (1957), 38–40.

[17] Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. *Information Retrieval* 13, 3 (2010), 254–270.

[18] Jun Xu and Hang Li. 2007. Adarank: a boosting algorithm for information retrieval. In *Proc. 30th ACM SIGIR, pages 391–398.* 391–398.

[19] Yu Zheng and Xing Xie. 2011. Location-based social networks: Locations. *Computing with spatial trajectories* (2011), 277–308.