Learning to Geolocalise Tweets at a Fine-Grained Level

University of Glasgow | School of Computing Science

Terrier Team

Jorge David Gonzalez Paule[1], Yashar Moshfeghi[2], Craig Macdonald[3], Iadh Ounis[3]

[1]j.gonzalez-paule.1@research.gla.ac.uk, [2]yashar.moshfeghi@strath.ac.uk, [3]{firstname.secondname}@glasgow.ac.uk

Overview

The fine-grained geolocation of tweets has become an important feature for reliably performing a wide range of tasks. Recent work adopted a basic ranking approach to return a predicted location for tweets based on their content-based similarity to already available geotagged tweets. However, this 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.

1. Motivation

Recent IR tasks require geotagged tweets at a fine-grained level (local event detection). However, **only 1% of the tweets are fine-grained geotagged.** Previous work [1] used a simple ranking approach (IDF weighting) to obtain the Top-N geotagged tweets, combined with a majority voting algorithm.

- However, considering only IDF weighting to perform the ranking can reduce the quality of the Top-N tweets [1].
- We aim to improve geolocalisation by improving the quality of the Top-N ranked tweets.

We propose a **learning to rank** [2] **approach** for **fine-grained geolocalisation.**

2. Learning to Geolocalise

- We aim to learn a ranking function (LambdaMART [3]) to re-rank geotagged tweets **(doc-tweets)** based on their geographical proximity to a given non-geotagged tweet **(query-tweet)**.
- We label pairs of geotagged tweets as positive if they are located in the same fine-grained area (i.e. <=1km distance).
- 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, and apply a majority voting algorithm to select a location within the Top-N doc-tweets.

3. Features

- We propose a set of features to model fine-grained tweet geolocalisation.
- We exploit 28 features (see table below) grouped into: **content quality features, geographical features** and **similarity features.**

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
Check-in	Whether the tweet is a Foursquare check-in.	2
Hour	The hour of the day (0 to 24 hours) the tweets was posted	2
Weekday	Number of hashtags in the text.	2
User Ratio	Number of hashtags in the text.	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 between the texts.	1
Total Features		**28**

4. Experimental Setup

We use 2 datasets of tweets collected from two US cities (Chicago and New York) in March 2016.

- **Training:** 20,982 query-tweets from New York, and 16,262 query-tweets from Chicago.
- **Testing:** 20,870 query-tweets from New York, and 16,313 query-tweets from Chicago.
- **Baseline Model** [1]**:** Uses an IDF-based ranking approach, and applies a weighted majority voting to select a location within the Top-N content-based similar geotagged tweets.

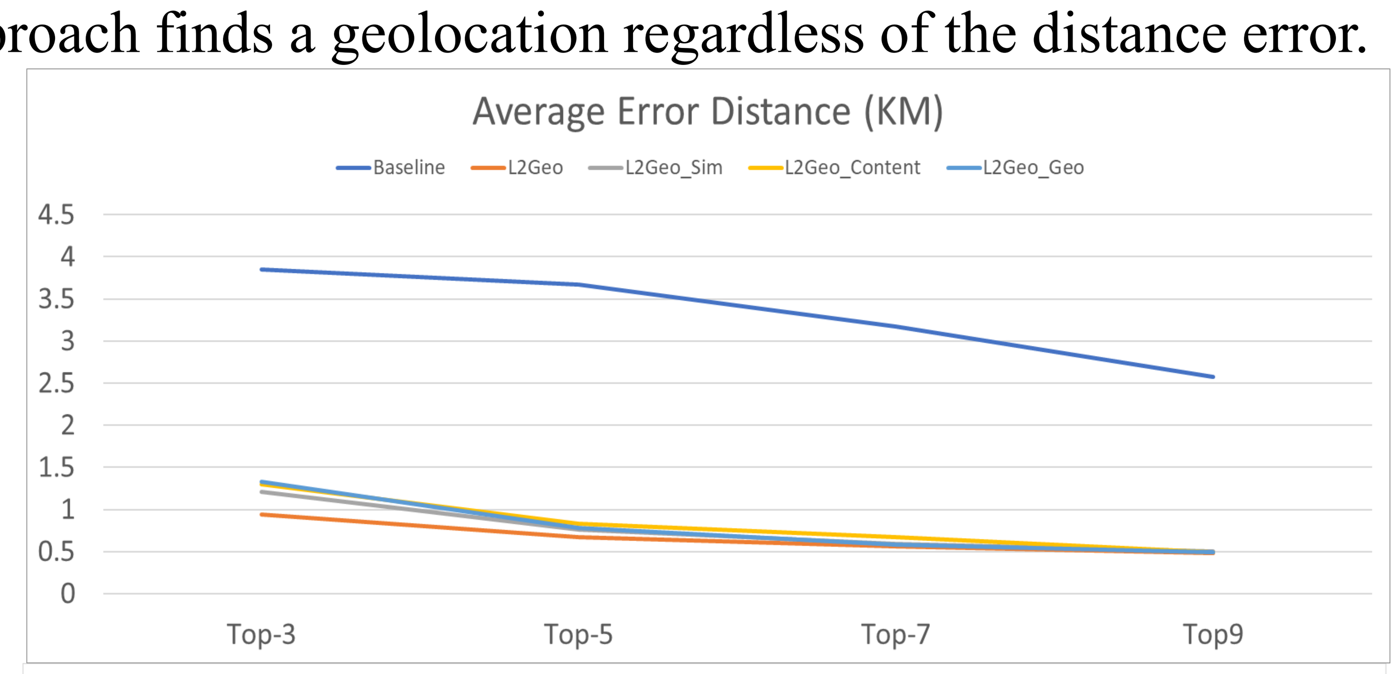
We test 4 versions of our approach with different subsets of features:

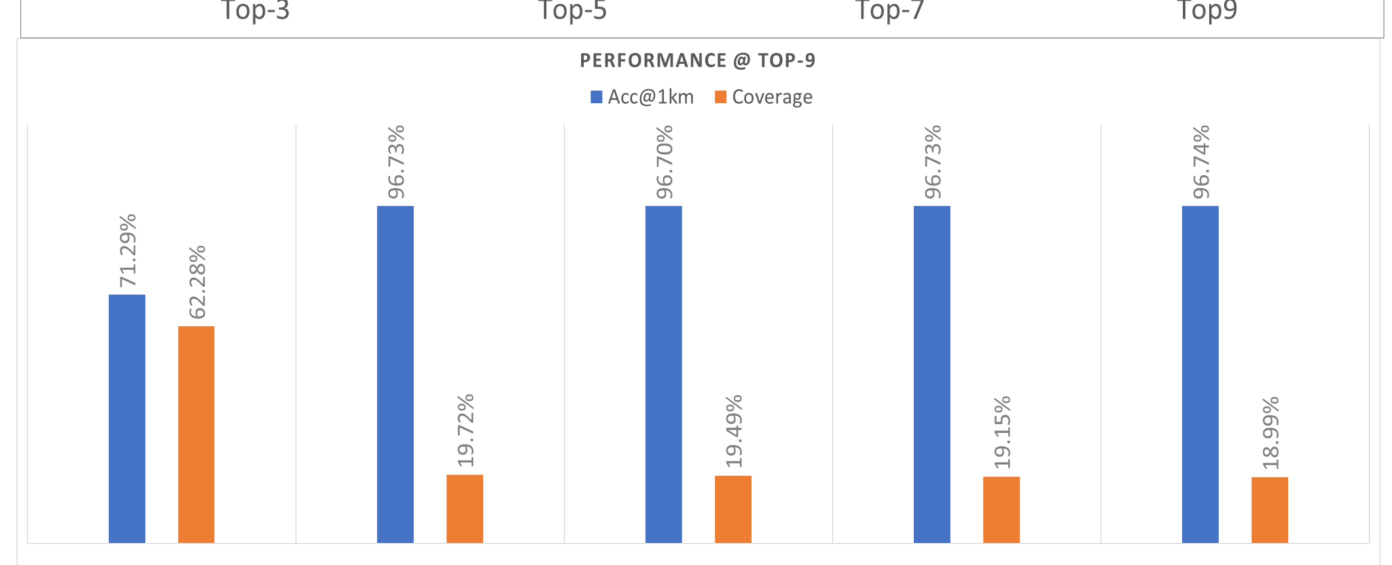
- **L2Geo:** incorporates all the features,
- **L2Geo_Sim:** uses only *similarity features*,
- **L2Geo_Content:** uses only *content quality features*
- **L2Geo_Geo:** uses only *geographical features*

5. Results

The figures below present the performance of our proposed approach (L2Geo) against the Baseline [1] on the **Chicago dataset** (similar results in New York). We report the following metrics:

- **AED**: Distance on Earth in kilometres between the predicted location and the real coordinates of the tweet in our ground truth.
- **Acc@1km**: Calculates whether the centroid of the predicted area lies within a radius of 1 km from the real location of a tweet.
- **Coverage**: The fraction of tweets in the test set from which our approach finds a geolocation regardless of the distance error.





- As the number of voting candidates (i.e. Top-N) increases, our approach achieves lower **AED**, higher **Acc@1km** but lower **Coverage** (see paper).
- **L2Geo** exhibits improvements over the rest of the learning to rank models.
- **L2Geo_Sim** shows that the *Similarity* features are the most informative subset of features.

7. Conclusions

By improving the ranking of geotagged tweets, we observed a better performance in the fine-grained geolocalisation of tweets compared to the existing prior approach [1]. Our learning to rank approach improved accuracy, but at the cost of a decrease in coverage.

References

[1]. Gonzalez Paule et al. 2017. On Fine-Grained Geolocalisation of Tweets. In Proc. ACM ICTIR.
[2] Fuxing Cheng et al. 2012. A survey of learning to rank for real-time twitter search. In Joint International. In Proc. ICPCA/SWS.
[3]. Qiang Wu et al. 2010. Adapting boosting for information retrieval measures. Information Retrieval 13, 3, 254–270.