On Fine-Grained Geolocalisation of Tweets

Jorge David Gonzalez Paule, Yashar Moshfeghi, Joemon M. Jose and Piyushimita (Vonu) Thakuriah
j.gonzalez-paule.1@research.gla.ac.uk

School. of Computing Science, University of Glasgow

Recently, the geolocalisation of tweets has become an important feature for a wide range of tasks in Information Retrieval and other domains, such as real-time event detection, topic detection or disaster and emergency analysis. However, the number of relevant geo-tagged tweets available remains insufficient to reliably perform such tasks. Thus, predicting the location of non-geotagged tweets is an important yet challenging task, which can increase the sample of geo-tagged data and help to a wide range of tasks. In this paper, we propose a location inference method that utilises a ranking approach combined with a majority voting of tweets weighted based on the credibility of its source (Twitter user). Using geo-tagged tweets from two cities, Chicago and New York (USA), our experimental results demonstrate that our method (statistically) significantly outperforms our baselines in terms of accuracy, and error distance, in both cities, with the cost of decrease in recall.

(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**
- Current works geolocalise tweets at a coarse-grained level (country or city level).
- Thus we **aim to geolocalise tweets at a fine-grained level** (neighbor or street level)

(2) Weighted Majority Voting for Tweet Geolocalisation

Our proposed approach consists of three steps.

1) Divide the geographical area into **fine-grained squares of size 1km.**

2) Obtain the **Top-N most content-based similar geo-tagged tweets** to a non-geo-tagged.

3) Combine evidence gathered from the Top-N tweets by adopting a **Weighted Majority Voting** algorithm (WMV).

Given the Top-N most content-similar tweets, we then **select the most frequent location within the Top-N set** and associate that as the geo-location of a given tweet.

$$Location(t_{ng}) = \underset{l_j \in L}{\operatorname{argmax}} \left(\sum_{i=1}^{N} W_{t_i} * Vote(t_i^{l_i}, l_j) \right) \qquad Vote(t_i^{l_i}, l_j) = \begin{cases} 1 & t_i^{l_i} = l_j \\ 0 & t_i^{l_i} \neq l_j \end{cases}$$

t_i^{li} – ith tweet in the Top-N rank;
L - Set of locations (l_j) in the Top-N tweets;
Vote(t_i^{li}) - The vote of each tweet is weighted (W_{ti}) by the credibility of tweet's source s_i;

(3) WEIGHTING

In our Majority Voting Algorithm, each tweet vote is weighted by the credibility of its source. The credibility is computed as follows:

1) Obtain the Top-N content-based most similar tweets for every tweet in a validation set.

2) For each source s_i, we define a set $\mathbf{TN_i}$ that contains all the tweets appearing in any of the Top-N rankings (t_{si}) produced for each tweet in the validation set (t_{vi}).

3) Finally the credibility is given by the ratio of all tweets in $\mathbf{TN_i}$ placed within less than **1km distance** from the tweets in the validation set ($\mathbf{t_v}$)

$$W_{t_i} = \frac{|\{t_{si} \in TN_i \mid distance(t_{s_i}, t_{vi}) \leq 1km\}|}{|TN_i|}$$

(4) EVALUATION

Datasets. We utilised geotagged tweets from New York (131,273) and Chicago (155,114).

Baseline Models. Paraskevopoulos et al. [1] approach using 5 different retrieval models: Divergence From Randomness (dfr), Language Model with Dirichlet Smoothing (lmd), IDF (idf), TF-IDF (tf idf) and BM25 (bm25).

WMV. We apply our weighted majority voting algorithm on top of the retrieval task. We considered the Top-3, -5, -7 and -9 content-based most similar tweets obtained from the retrieval task. Also used IDF as our retrieval model (best performance) [3].

(5) RESULTS

Table 1 and 2 shows Average Error Distance in kilometres **(A Err km)**, Accuracy at 1 kilometre **(Acc@1km)** and **Recall** for our proposed approach **("WMV")** against our Baseline **("Baseline").**

Table 1

Chicago			
Model	**A_Err_km**	**Acc@1km**	**Recall**
Baseline_tf_idf	8.100	42.40%	99.97%
Baseline_idf	14.056	13.18%	99.97%
Baseline_dfr	8.586	37.40%	99.97%
Baseline_lmd	6.185	47.79%	99.97%
Baseline_bm25	7.637	41.76%	99.97%
WMV@Top3	3.849**	61.17%**	83.28%**
WMV@Top5	3.669**	62.78%**	79.08%**
WMV@Top7	3.170**	66.82%**	70.41%**
WMV@Top9	2.576**	71.29%**	62.28%**

Table 2

New York			
Model	**A_Err_km**	**Acc@1km**	**Recall**
Baseline_tf_idf	7.505	38.39%	99.98%
Baseline_idf	12.755	12.78%	99.98%
Baseline_dfr	7.609	36.28%	99.98%
Baseline_lmd	7.169	37.29%	99.98%
Baseline_bm25	7.460	38.25%	99.98%
WMV@Top3	4.234**	52.33%**	75.84%**
WMV@Top5	4.362**	51.98%**	75.09%**
WMV@Top7	4.008**	54.81%**	67.83%**
WMV@Top9	3.476**	59.23%**	59.94%**

(6) DISCUSSION AND CONCLUSIONS

- We proposed an approach for fine-grained geolocalisation of tweets by adopting a weighted majority voting algorithm. The weight of each tweet vote is obtained by calculating the credibility of its source (i.e. Twitter user).
- Our approach ("WMV") (statistically) significantly outperforms the best performed baseline (i.e. "Baseline_lmd") in terms of accuracy and error distance, in both cities.
- As the number of voting candidates (i.e. Top-N) increases, our approach achieves lower error distance, higher accuracy but lower recall.
- Our approach is fast, effective and does not require training.

Acknowledgments This research is partially supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ ERC.

(7) REFERENCES

[1] Pavlos Paraskevopoulos and Themis Palpanas. 2015. Fine-Grained Geolocalisation of Non-Geotagged Tweets. (ASONAM'15).

[2] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. "I'M Eating a Sandwich in Glasgow": Modeling Locations with Tweets. (SMUC'11).

[3] Jesus Alberto Rodriguez Perez and Joemon M. Jose. 2015. On Microblog Dimensionality and Informativeness: Exploiting Microblogs' Structure and Dimensions for Ad-Hoc Retrieval. (ICTIR '15).

[4] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised Text-based Geolocation Using Language Models on an Adaptive Grid. (EMNLP-CoNLL '12).

[5] Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. 2013. A Multi-Indicator Approach for Geolocalization of Tweets. (ICWSM'13)