Geocoding Textual Documents Through a Hierarchy of Linear Classifiers

Fernando Melo and Bruno Martins(✉)

Instituto Superior Técnico and INESC-ID, Universidade de Lisboa, Lisbon, Portugal
{fernando.melo,bruno.g.martins}@ist.utl.pt

Abstract. In this paper, we empirically evaluate an automated technique, based on a hierarchical representation for the Earth's surface and leveraging linear classifiers, for assigning geospatial coordinates to previously unseen documents, using only the raw text as input evidence. We measured the results obtained with models based on Support Vector Machines, over collections of geo-referenced Wikipedia articles in four different languages, namely English, German, Spanish and Portuguese. The best performing models obtained state-of-the-art results, corresponding to an average prediction error of 83 Kilometers, and a median error of just 9 Kilometers, in the case of the English Wikipedia collection.

Keywords: Text mining · Document geocoding · Hierarchical text classification

1 Introduction

Geographical Information Retrieval (GIR) has recently captured the attention of many different researchers that work in fields related to language processing and to the retrieval and mining of relevant information from large document collections. For instance, the task of resolving individual place references in textual documents has been addressed in several previous works, with the aim of supporting subsequent GIR processing tasks, such as document retrieval or the production of cartographic visualizations from textual documents [5,6]. However, place reference resolution presents several non-trivial challenges [8,9], due to the inherent ambiguity of natural language discourse. Moreover, there are many vocabulary terms, besides place names, that can frequently appear in the context of documents related to specific geographic areas [1]. Instead of resolving individual references to places, it may be interesting to instead study methods for assigning entire documents to geospatial locations [1,11].

In this paper, we describe a technique for assigning geospatial coordinates of latitude and longitude to previously unseen textual documents, using only the raw text of the documents as evidence, and relying on a hierarchy of linear models built with basis on a discrete hierarchical representation for the Earths surface, known in the literature as the HEALPix approach [4]. The regions at each level of this hierarchical representation, corresponding to equally-distributed curvilinear

F. Pereira et al. (Eds.) EPIA 2015, LNAI 9273, pp. 590–596, 2015.
DOI: 10.1007/978-3-319-23485-4_59

and quadrilateral areas of the Earths surface, are initially associated to textual contents (i.e., we use all the documents from a training set that are known to refer to particular geospatial coordinates, associating each text to the corresponding region). For each level in the hierarchy, we build classification models using the textual data, relying on a vector space model representation, and using the quadrilateral areas as the target classes. New documents are assigned to the most likely quadrilateral area, through the usage of the classifiers inferred from training data. We finally assign documents to their respective coordinates of latitude and longitude, taking the centroid coordinates from the quadrilateral areas.

The proposed document geocoding technique was evaluated with samples of geo-referenced Wikipedia documents in four different languages. We achieved an average prediction error of 83 Kilometers, and a median error of just 9 Kilometers, in the case of documents from the English Wikipedia. These results are slightly better than those reported in previous state-of-the-art studies [11,12].

2 Previous and Related Work

While most work on geographic information retrieval relies on specific keywords such as place names, Adams and Janowicz proposed an approach for geocoding documents that uses only non-geographic expressions, concluding that even ordinary textual terms may be good predictors of geographic locations [1]. The proposed technique used Latent Dirichlet Allocation (LDA) to discover latent topics from general vocabulary terms occurring in a training collection of geo-referenced documents, together with Kernel Density Estimation (KDE) to interpolate a density surface, over each LDA topic. New documents are assigned to the geospatial areas having the highest aggregate density, computed from the per-document topic distributions and from the KDE surfaces.

Wing and Baldridge evaluated approaches for automatically geocoding documents based on their textual contents, specifically leveraging generative language models learned from Wikipedia [11]. The authors applied a regular geodesic grid to divide the Earth's surface into discrete rectangular cells. Each cell can be seen as a virtual document that concatenates all the training documents located within the cell's region. Three different methods were compared in the task of finding the most similar cell, for a new document, namely (i) the Kullback-Leibler divergence, (ii) naïve Bayes, and (iii) a baseline method corresponding to the average cell probability. Method (i) obtained the best results, i.e. a median prediction error of just 11.8 Kilometers, and a mean error of 271 Kilometers, on tests with documents taken from the English Wikipedia. More recently, Dias et al. [2] reported on experiments with an adapted version of the method described by Wing and Baldridge, which used language models based on character n-grams together with a discrete representation for the surface of the Earth based on an equal-area hierarchical triangular mesh approach [3]. Another improvement over the language modeling method was latter reported by Roller et al. [7], where the authors collapsed nearby training documents through the

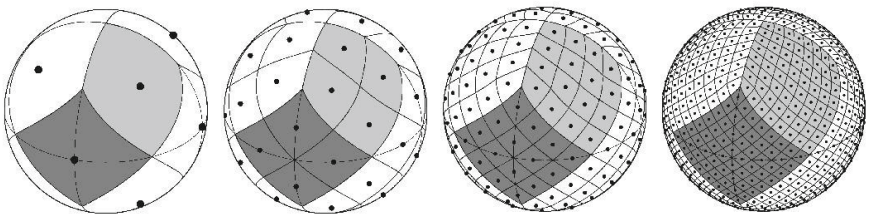

Fig. 1. Orthographic views associated to the first four levels of the HEALPix sphere tessellation.

usage of a k-d tree data structure. Moreover, Roller et al. proposed to assign the centroid coordinates of the training documents contained in the most probable cell, instead of just using the center point for the cell. These authors report on a mean error of 181 Kilometers and a median error of 11 Kilometers, when geocoding documents from the English Wikipedia.

More recently, Wing and Baldridge also reported on tests with discriminative classifiers [12]. To overcome the computational limitations of discriminative classifiers, in terms of the maximum number of classes they can handle, the authors proposed to leverage a hierarchical classification procedure that used feature hashing and an efficient implementation of logistic regression. In brief, the authors used an hierarchical approach in which the Earth's surface is divided according to a rectangular grid (i.e., using either a regular grid or a k-d tree), and where an independent classifier is learned for every non-leaf node of the hierarchy. The probability of any node in the hierarchy is the product of the probabilities of that node and all of its ancestors, up to the root. The most probable leaf node is used to infer the final geospatial coordinates. Rather than greedily using the most probable node from each level, or rather than computing the probability of every leaf node, the authors used a stratified beam search. This procedure starts at the root, keeping the b highest-probability nodes at each level, until reaching the leafs. Wing and Baldridge report on results over English Wikipedia data corresponding to a mean error of 168.7 Kilometers and a median error of 15.3 Kilometers.

3 The Proposed Document Geocoding Method

The proposed document geocoding approach is based on discretizing the surface of the Earth into hierarchically organized sets of regions, as given by the HEALPix procedure and where each set corresponds to a different partitioning resolution. Having documents associated to these discrete regions allows us to predict locations with standard discriminative classification approaches (e.g., with linear Support Vector Machines classifiers).

HEALPix is an acronym for Hierarchical Equal Area isoLatitude Pixelization of a sphere, and the procedure results on a multi level recursive subdivision

Table 1. Number of regions and approximate area for HEALPix grids of different resolutions.

Resolution	4	64	256	1024
Total number of regions	192	49,152	786,432	12,582,912
Approximate area of each region (Km^2)	2,656,625	10,377	649	41

for a spherical approximation to the Earth's surface, according to curvilinear quadrilateral regions, in which each resulting subdivision covers an equal surface area. Figure 3, adapted from an original illustration provided in the HEALPix website[1], shows from left to right the resolution increase by three steps from the base level with 12 different regions [4].

The HEALPix representation scheme contains a parameter N_{side} that controls the resolution, i.e. the number of divisions along the side of a base-resolution region that is needed to reach a desired high-resolution partition, and which naturally will also define the area of the curvilinear quadrilaterals. In our experiments, we used a hierarchy of 4 different representations with different resolutions, equaling the N_{side} parameter to the values of 4, 64, 256 and 1024. Table 1 presents the total number of regions in each of the considered resolution levels (i.e., $n = 12 \times N_{side}^2$), together with the approximate area, in squared Kilometers, corresponding to each region.

Another important question relates to the choice of how to represent the textual documents. We used a vector space model representation, where each document is seen as a vector of features. The feature weights in the vectors that represent each document are given according to the term frequency times inverse document frequency (TF-IDF) scheme, where the weight for a term i on a document j can be computed as:

$$\text{TF–IDF}_{i,j} = \log_2(1 + \text{TF}_{i,j}) \times \log_2\left(\frac{N}{n_i}\right) \tag{1}$$

In the formula, $\text{TF}_{i,j}$ is the term frequency for term i on document j, N is the total number of documents in the collection, and n_i is the number of documents containing the term i. The TF-IDF weight is 0 if $\text{TF}_{i,j} = 0$.

With the hierarchy of discrete representations given by the HEALPix method, together with the document representations based on TF-IDF, we then used linear classification algorithms to address the document geocoding task. We trained a separate classification model for each node in the hierarchy of discrete representations, taking all documents whose coordinates lay within the region corresponding to each node, as the training data for each classifier. When geocoding a test document, we first apply the root-level classifier to decide the most likely region, and then proceed greedily by applying the classifier for each of the most likely nodes, up to the leafs. After reaching a leaf region from the

[1] http://healpix.jpl.nasa.gov

hierarchical representation, the geospatial coordinates of latitude and longitude are assigned by taking the centroid coordinates of the leaf region.

Support Vector Machines (SVMs) are one of the most popular approaches for learning classifiers from training data. In our experiments, we used the multi-class linear SVM implementation from scikit-learn[2] with default parameters (e.g., with the default regularization constant), which in turn is a wrapper over the LIBLINEAR[3] package.

4 Experimental Validation

In our experiments, we used samples with geocoded articles from the English (i.e., 847,783 articles), German (i.e., 307,859 articles), Spanish (i.e., 180,720 articles) and Portuguese (i.e., 131,085 articles) Wikipedias, taken from database dumps produced in 2014. Separate experiments evaluated the quality of the document geocoders built for each of the four languages, in terms of the distances from the predictions towards the correct geospatial coordinates. We processed the Wikipedia dumps to extract the raw text from the articles, and for extracting the geospatial coordinates of latitude and longitude from the corresponding infoboxes. We used 90% of the geocoded articles of each Wikipedia for model training, and the other 10% for model validation.

In what regards the geospatial distributions of documents, we have that some regions (e.g, North America or Europe) are considerable more dense in terms of document associations than others (e.g, Africa), and that oceans and other large masses of water are scarce in associations to Wikipedia documents. This implies that the number of classes that has to be considered by our model is much smaller than the theoretical number of classes given by the HEALPix procedure. In our English dataset, there are a total of 286,966 regions containing associations to documents at a resolution level of $N_{side} = 1024$, and a total of 82,574, 15,065, and 190 regions, respectively at resolutions 256, 64, and 4. These numbers are even smaller in the collections for the other languages.

Table 2 presents the obtained results for the different Wikipedia collections. The prediction errors shown in Table 2 correspond to the distance in Kilometers, computed through Vincenty's geodetic formulae [10], from the predicted locations to the true locations given in Wikipedia. The accuracy values correspond to the relative number of times that we could assign documents to the correct region (i.e., the HEALPix region where the document's true geospatial coordinates of latitude and longitude are contained), for each level of hierarchical classification. Table 2 also presents upper and lower bounds for the average and median errors, according to a 95% confidence interval and as measured through a sampling procedure.

The results attest for the effectiveness of the proposed method, as we measured slightly inferior errors than those reported in previous studies [2,7,11,12], which besides different classifiers also used simpler procedures for representing

[2] http://scikit-learn.org/
[3] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Table 2. The results obtained for each different language.

	Classifier accuracy				Errors in terms of distance	
	1st	2nd	3rd	4th	Average	Median
English	0.966	0.785	0.540	0.262	82.501 (±4.340)	8.874 [5.303 - 15.142]
German	0.972	0.832	0.648	0.396	62.995 (±5.753)	4.974 [3.615 - 8.199]
Spanish	0.950	0.720	0.436	0.157	165.887 (±16.675)	13.410 [8.392 - 22.691]
Portuguese	0.951	0.667	0.336	0.104	105.238 (±10.059)	21.872 [13.611 - 33.264]

textual contents and for representing the geographical space. It should nonetheless be noted that the datasets used in our tests may be slightly different from those used in previous studies (e.g., they were taken from different Wikipedia dumps), despite their similar origin.

5 Conclusions and Future Work

Through this work, we empirically evaluated a simple method for geo-referencing textual documents, relying on a hierarchy of linear classifiers for assigning documents to their corresponding geospatial coordinates. We have shown that the automatic identification of the geospatial location of a document, based only on its text, can be performed with high accuracy by using out-of-the-box implementations of well-known supervised classification methods, and leveraging a hierarchical procedure based on HEALPix [4].

Despite the interesting results, there are also many ideas for future work. The geospatial coordinates estimated from our document geocoding procedure can for instance be used as prior evidence (i.e., as document-level priors) to support the resolution of individual place references in text [8]. In terms of future work, we would also like to experiment with other types of classification approaches and with different text representation and feature weighting schemes.

Acknowledgments. This work was supported by Fundação para a Ciência e a Tecnologia (FCT), through project grants with references EXCL/EEI-ESS/0257/2012 (DataStorm research line of excellency), EXPL/EEI-ESS/0427/2013 (KD-LBSN), and also UID/CEC/50021/2013 (INESC-ID's associate laboratory multi-annual funding).

References

1. Adams, B., Janowicz, K.: On the geo-indicativeness of non-georeferenced text. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (2012)
2. Dias, D., Anastácio, I., Martins, B.: A language modeling approach for georeferencing textual documents. Actas del Congreso Español de Recuperación de Información (2012)
3. Dutton, G.: Encoding and handling geospatial data with hierarchical triangular meshes. In: Kraak, M.J., Molenaar, M., (eds.) Advances in GIS Research II. CRC Press (1996)

4. Górski, K.M., Hivon, E., Banday, A.J., Wandelt, B.D., Hansen, F.K., Reinecke, M., Bartelmann, M.: HEALPIX - a framework for high resolution discretization, and fast analysis of data distributed on the sphere. The Astrophysical Journal **622**(2) (2005)
5. Lieberman, M.D., Samet, H.: Multifaceted toponym recognition for streaming news. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (2011)
6. Mehler, A., Bao, Y., Li, X., Wang, Y., Skiena, S.: Spatial analysis of news sources. IEEE Transactions on Visualization and Computer Graphics **12**(5) (2006)
7. Roller, S., Speriosu, M., Rallapalli, S., Wing, B., Baldridge, J.: Supervised text-based geolocation using language models on an adaptive grid. In: Proceedings of the Conference on Empirical Methods on Natural Language Processing (2012)
8. Santos, J., Anastácio, I., Martins, B.: Using machine learning methods for disambiguating place references in textual documents. GeoJournal **80**(3) (2015)
9. Speriosu, M., Baldridge, J.: Text-driven toponym resolution using indirect supervision. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (2013)
10. Vincenty, T.: Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. Survey Review **XXIII**(176) (1975)
11. Wing, B., Baldridge, J.: Simple supervised document geolocation with geodesic grids. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (2011)
12. Wing, B., Baldridge, J.: Hierarchical discriminative classification for text-based geolocation. In: Proceedings of the Conference on Empirical Methods on Natural Language Processing (2014)