# A Language Modeling Approach for Georeferencing Textual Documents

Duarte Dias, Ivo Anastácio, and Bruno Martins
{duarte.dias, ivo.anastacio, bruno.g.martins}@ist.utl.pt

Instituto Superior Técnico, INESC-ID
Av. Professor Cavaco Silva
2744-016 Porto Salvo, Portugal

**Abstract.** Most text documents can be said to be related to some form of geographic context, although traditional text mining methods simply model documents as bags of tokens. Recently, geographic information retrieval has captured the attention of many different researchers that work in fields related to text mining and data retrieval, envisioning the support for tasks such as map-based document indexing, retrieval and visualization. In this paper, we empirically compare automated techniques, based on language models, for assigning geospatial coordinates of latitude and longitude to previously unseen textual documents, using only the raw text of the documents as input evidence. We measure the results obtained by character or token-based language models using a collection of georeferenced Wikipedia articles, with the best performing method obtaining an average prediction error of 278 Kilometers, and a median prediction error of just 26 Kilometers.

## 1 Introduction

Most text documents can be said to be related to some form of geographic context and, recently, Geographical Information Retrieval (GIR) has captured the attention of many different researchers that work in fields related to retrieving and mining contents from large document collections. We have for instance that 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 cartographic visualization of text documents [11]. However, place reference resolution presents several non-trivial challenges [10,12,1], due to the inherent ambiguity of natural language discourse (e.g., place names often have other non geographic meanings, different places are often referred to by the same name, and the same places are often referred to by different names). Moreover, we have that there are many vocabulary terms, besides place names, that can frequently appear in documents related to specific

This work was partially supported by the Fundação para a Ciência e a Tecnologia (FCT), through project grant PTDC/EIA-EIA/109840/2009 (SInteliGIS)

geographic areas, and GIR applications can also benefit from this information. Instead of trying to correctly resolve the individual references to places that are made in textual documents, it may be interesting to instead study methods for assigning entire documents to geospatial locations [16].

In this paper, we compare automated techniques 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 discrete binned representation of the Earth's surface. The bins from this representation, corresponding to equally-distributed areas of the Earth's surface, are initially associated to textual documents (i.e., we use all the documents from a training set that are known to refer to each particular bin). We build compact representations (i.e., character-based or token-based language models) from these georeferenced sets of documents, capturing their main statistical properties. New documents are then assigned to the most similar bin. We finally assign documents to their respective coordinates of latitude and longitude, with basis on the centroid coordinates associated to the bins. Experiments with a collection of Wikipedia articles showed good results for a hierarchical method based on this general approach, with the best performing configuration obtaining an average prediction error of just 278 Kilometers, and a median error of 26 Kilometers.

The rest of this paper is organized as follows: Section 2 presents related work, while Section 3 details the proposed approach. Section 4 presents the experimental validation of the proposed method, describing the considered Wikipedia dataset, the evaluation protocol, and the obtained results for the different variations of the proposed method. Finally, Section 5 presents our conclusions and points directions for future work.

## 2 Related Work

The relationship between language and geography has long been a topic of interest to linguists [8]. Many studies have, for instance, shown that geography has an impact on the relationship between vocabulary terms and semantic classes. For instance the term *football*, in the United States, refers to the particular sport of American football. However, in regions such as Europe, the term *football* is usually associated to different sports (e.g., soccer or, less frequently, rugby football). Terms such as *beach* or *snow* are also more likely to be associated to particular locations. In this study, we are interested in seeing if vocabulary terms and textual contents in general can be used to predict geographical locations.

Eisenstein et al. investigated the dialectal differences and the variations in regional interests over Twitter users, using a collection of georeferenced *tweets* and probabilistic models [6]. These authors tried to georeference USA-based Twitter users with basis on their *tweet* content, concatenating all the *tweets* for each single user and using Gaussian distributions to model the locations of the Twitter users. The approaches proposed in this paper are instead based on a discrete representation for the Earth's surface, and on relatively simple probabilistic models built over these discrete representation.

Overell investigated the use of Wikipedia as a source of data for georeferencing textual articles, in addition to article classification by category, and individual place reference resolution [13]. Overell's main goal was to resolve place references in documents, for which global document georeoreferencing could serve as an input feature. For document georeferencing, Overell proposed a simple model that uses only the metadata available (e.g., article title, incoming and outgoing links, etc.) and not the actual text.

Anastácio et al. surveyed heuristic approaches to assign documents to geographic scopes, based on recognizing place references in the documents and afterwards combining the recognized references [2]. The authors specifically compared approaches based on (i) the occurrence frequency for the references, (ii) the spatial overlap between bounding boxes associated to the references, (iii) hierarchical containment between the references, using a taxonomy of administrative divisions, and (iv) graph-propagation methods using again a taxonomy of administrative divisions. Experiments with a collection of Web pages from the Open Directory Project showed that hierarchical containment achieved very good results. In this paper, we are also studying approaches for georeferencing the entire contents of textual documents, but using only the raw text as input, instead of relying on place references recognized in the texts.

Wing and Baldridge, in a very similar study to the one that is reported in this paper, compared different approaches for automatically georeferencing documents, also based on statistical models derived from a large corpus of already georeferenced documents, such as Wikipedia [16]. The Kullback-Leibler divergence between the language model for a test document, and language models for each cell in a discrete gridded representations for the Earth's surface, was used to predict the most likely grid cell for a document. A similar approach was proposed for the temporal resolution of documents, capable of determining the date of publication of a story, based on its text [9]. Again, the authors built histograms encoding the probability of different temporal periods for a document, later using the Kullback-Leibler divergence to make the predictions. The work reported in this paper is very similar to that of Wing and Baldridge, but we propose to use (i) a different scheme for partitioning the set of documents into bins of equal area, according to their geospatial locations, (ii) a different language modeling approach for classifying documents according to the most similar bins, and (iii) a hierarchical decomposition approach for improving the computational performance of the classification method.

## 3 The Proposed Method for Document Geocoding

The proposed approach is based on discretizing the surface of the Earth into a set of bins, allowing us to predict locations with standard approaches for discrete outcomes. However, unlike previous authors such as Serdyukov et al. [14] or Wing and Baldridge [16], which used a grid of square cells of equal degree, we use the hierarchical triangular mesh[1] approach to discretize the Earth's sur-

[1] `http://www.skyserver.org/htm/Old_default.aspx`

face [5,15]. This strategy results in a grid that roughly preserves an equal area for each bin, instead of variable-size regions that shrink latitudinally, becoming progressively smaller and more elongated as they get closer towards the poles. Notice that our binned representation ignores all higher level semantic regions, such as states, countries or continents. Nonetheless, this is appropriate for our purposes, since documents can be related to geographical regions that do not fit into an administrative division of the Earth's surface.

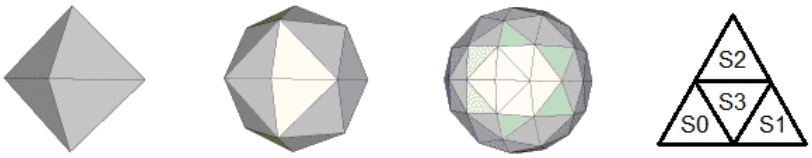


Fig. 1: Decompositions of the Earth for triangular meshes with resolutions of zero, one and two, obtained by recursively dividing circular triangles (adapted from figures available from the URL `http://www.skyserver.org/htm/Old_default.aspx`).

The hierarchical triangular mesh offers a multi level recursive decomposition of a spherical approximation to the Earth's surface. It starts at level zero with an octahedron and, as one projects the edges of the octahedron onto the sphere, it creates 8 spherical triangles, 4 on the Northern and 4 on the Southern hemispheres. Four of these triangles share a vertex at the pole and the sides opposite to the pole form the equator. Each of the 8 spherical triangles can be split into four smaller triangles by introducing new vertices at the midpoints of each side, and adding a great circle arc segment to connect the new vertices with the existing one. This sub-division process can repeat recursively, until we reach the desired level of resolution. The triangles in this mesh are the bins used in our representation of the Earth, and every triangle, at any resolution, is represented by a single numeric ID. For each location given by a pair of coordinates on the surface of the sphere, there is an ID representing the triangle, at a particular resolution, that contains the corresponding point.

Notice that the proposed representation scheme contains a parameter $k$ that controls the resolution, i.e. the area of the bins. Having course grained bins can lead to very rough estimates, but classification accuracy with a thin-grained resolution can also decrease substantially, due to insufficient data to build the language models associated to each bin. In our experiments, we ranged the $k$ parameter from 4 to 8, with 0 corresponding to a first-level division. Table 1

| Resolution | 4 | 8 | 9 |
|---|---|---|---|
| Total number of categories | 2,048 | 524,288 | 2,097,152 |
| Average area of each bin ($km^2$) | 28774.215 | 1041.710 | 261.675 |

Table 1: Number of bins and corresponding area.

presents the number of bins that would be generated at each of the considered levels of resolution. The number of bins $n$ is given by $n = 8 * 4^k$. Table 1 also shows the area in squared kilometers corresponding to each bin.

With the above discrete representation of the Earth's surface, we used the LingPipe[2] package to build character-based or token-based language models, afterwards using these models for associating to each bin the probability of being the best class for a given novel document.

In brief, the LingPipe classifiers perform joint probability-based classification of textual documents into categories, based on either character-based or token-based language models (i.e., in our experiments, we tested these two different classification approaches). The general idea is to build a language model $p(text|cat)$ for each category $cat$, afterwards building a multinomial distribution $p(cat)$ over the categories, and finally computing joint log probabilities for the classes according to Bayes's rule, yielding:

$$\log_2 P(cat, text) \propto \log_2 P(text|cat) + \log_2 P(cat) \quad (1)$$

In the formula, $P(text|cat)$ is the probability of seeing a document $d$ in the language model for category $cat$, and $P(cat)$ is the marginal probability assigned by the multinomial distribution over the categories.

The book by Carpenter and Baldwin [3] has full details on the language models used for estimating $p(text|cat)$, and on the multinomial distribution $p(cat)$ over the categories (i.e., the bins from our representation of the Earth), which is estimated using a maximum a posteriori probability (MAP) estimate with additive (i.e., Dirichlet) priors.

In terms of character-based language models, they are essentially generative language models based on the chain rule, which smooth estimates through linear interpolation with the next lower-order context models, and where there is a probability of 1.0 to the sum of the probability of all sequences of a specified length. Our character-based language models used 6 character sequences. As for the token-based language models, we captured sequences of tokens with a 2-gram model, and modeled white-spaces and unknown tokens separately. The reader can refer to the book by Carpenter and Baldwin for more detailed information about the classification method [3].

After having probabilities assigned to each of the bins in our representation, we compute the latitude and longitude coordinates with basis on the centroid coordinates for the most probable bin.

[2] `http://alias-i.com/lingpipe`

Although the classification method introduced above could be used directly to assign documents to the most probable bins, it would be very inefficient in practice when considering a thin-grained resolution, due to the very large number of classes – see Table 1 – and due to the need for estimating, for each document, its probability of having been generated by the language model corresponding to each class. In this paper, we propose to use a hierarchical classification approach, where instead of a single classifier considering all bins from a detailed triangular mesh encoding the Earth's surface, we use a hierarchy of classifiers with two levels. The first level corresponds to a single classification model using bins from a coarse-grained division of the Earth, and the other level corresponds to different classifiers, one for each class from the first level, encoding different parts of the Earth with a thinner granularity. With this hierarchical scheme, classification can be made much more efficiently, as documents need to be evaluated with less language models, and each language model can be trained with a smaller number of examples. We also take advantage of the properties from the hierarchical triangular mesh in order to reduce the number of classes in each of the models from the second level of our classification hierarchy. If a given bin from the decomposition of the Earth does not contain any training documents assigned to it, and if only one of its neighbouring bins in the mesh contains documents, then we use a single class from the hierarchical triangular mesh of the immediately smaller resolution, in order to represent this region in the classification model.

In a related previous work, Wing and Baldridge [16] reported on very accurate results (i.e., a median prediction error of just 11.8 Kilometers, and a mean of 271 Kilometers) with a similar but non-hierarchical classification approach, based on the Kullback-Leibler divergence between language models. However, these authors also claim that a full run with all their experiments (i.e., six different strategies) required about 4 months of computing time and about 10-16 GB of RAM when run on a 64-bit Intel Xeon E5540 CPU. Our hierarchical classification approach can substantially reduce the required computational effort.

## 4 Experimental Evaluation

We now describe the experimental methodology used for comparing the proposed methods, as well as the obtained results. For the experiments reported in this paper, we used a sample of the articles from a Wikipedia dump from 2011. Included in this dump are a total of 4,080,270 articles, of which 393,294 were associated to latitude and longitude coordinates. Previous studies have already shown that Wikipedia articles are a well-suited source of textual contents for the purpose of georeferencing documents [13,16].

We processed the Wikipedia dump to extract the raw text from the articles and for extracting the geospatial coordinates, using manually-defined patterns to capture some of the multiple templates and multiple formats for expressing latitude and longitude in Wikipedia. Considering a random order for the articles, about 90% of the georeferenced articles that we could process were used for model training (i.e., a total of 353,294 articles) and the other 10% were used

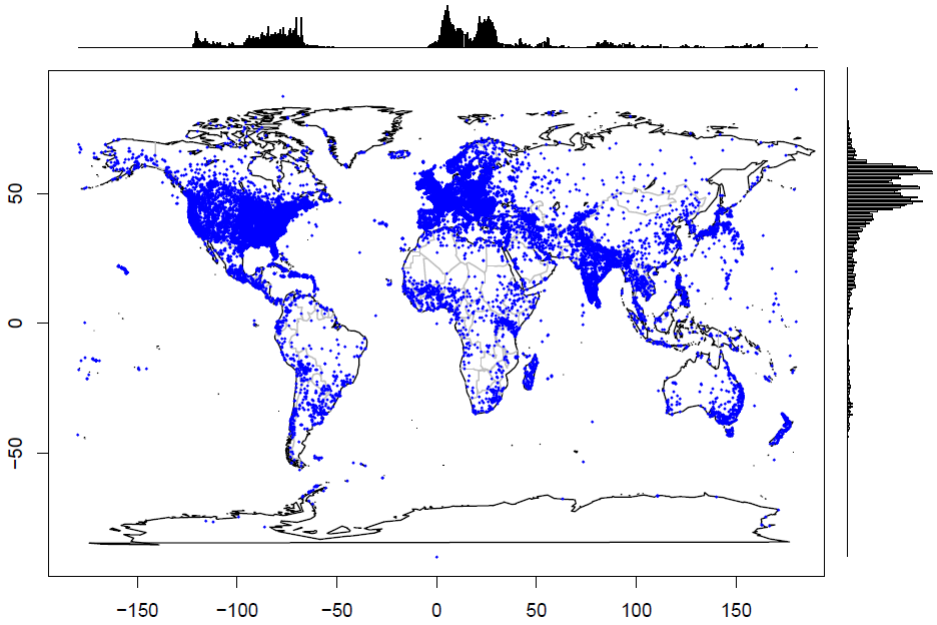


Fig. 2: Geographic distribution of the Wikipedia documents.

| Statistic | Train Data | Test Data |
|---|---|---|
| Number of documents | 353,294 | 40,000 |
| Number of words | 150,531,698 | 16,884,437 |
| Average words per document | 426 | 422 |
| Standard deviation of words per document | 77.964 | 231.704 |

Table 2: Statistical characterization of the considered Wikipedia dataset.

for model validation (i.e., a total of 40,000 articles). Table 2 presents a statistical characterization for the considered dataset, while Figure 2 illustrates the geospatial distribution of the locations associated to the Wikipedia documents. Notice that some geographic regions (e.g, North America or Europe) are considerable more dense in terms of document associations than others (e.g, Africa). Moreover, 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 in Table 1. In our dataset, there are a total number of 1,123 bins containing associations to documents at resolution level 4, and a total of 42,331 and 82,253 bins, respectively at resolutions 8 and 9.

| Method | Resolution | | Classifier Accuracy | | Geospatial Distance | |
|---|---|---|---|---|---|---|
| | k1 | k2 | First Level | Second Level | Average | Median |
| Character models | 0 | 4 | 0.9316 | 0.8225 | 414.014 | 238.149 |
| Character models | 1 | 8 | 0.9215 | 0.3996 | 278.568 | 26.245 |
| Character models | 2 | 9 | 0.9101 | 0.2625 | 283.163 | 26.556 |
| Token models | 0 | 4 | 0.9411 | 0.5235 | 707.358 | 300.751 |
| Token models | 1 | 8 | 0.9070 | 0.2812 | 402.187 | 44.089 |
| Token models | 2 | 9 | 0.8875 | 0.2013 | 475.124 | 39.944 |

Table 3: The obtained results for document geocoding.

Using the Wikipedia dataset, we experimented with classification models relying on bin sizes of varying granularity. Table 3 presents the obtained results for the different methods under study, with the error values for each bin size. The prediction errors shown in Table 3 correspond to the distance in Kilometers, computed through Vincenty's formulae[3], from the predicted locations to the locations given at the gold standard. The accuracy values correspond to the relative number of times that we could assign documents to the correct bin (i.e., the bin where the document's true geospatial coordinates of latitude and longitude are contained). The $k1$ and $k2$ values correspond to the resolution for the Earth's representation used at each level of the hierarchical classifier.

The results from Table 3 show that the method corresponding to the usage of character-based language models obtained the best results, with a prediction accuracy of approximately 0.4 in the task of finding the correct bin, while assigning documents to the correct geospatial coordinates had an error of 278 Kilometers on average. The documents that could be assigned to the correct bin had an average distance towards the correct coordinates of 14.150 Kilometers.

A visualization of our results can be seen in Figure 3, where the map represents the geospatial distribution for the predicted locations. The figure shows that errors are evenly distributed, and also that Europe and North America remain the regions of highest density. Figures 4 and 5 illustrate the distribution for the errors produced by the character-based (Figure 4) and token-based (Figure 5) classifiers, in terms of the distance between the estimated coordinates and the true geospatial coordinates. These figures plot the number of documents whose error (i.e., distance) is greater or equal than a given value, using doubly logarithmic axes. Figure 4 shows that our language models based on characters classify the majority of documents with a small error in terms of distance, with about 100 documents having an error greater than 10,000 Kilometers. Figure 5 shows worse results for the token-based models, with about 200 documents having a error greater than 10,000 Kilometers.

[3] `http://en.wikipedia.org/wiki/Vincenty's_formulae`

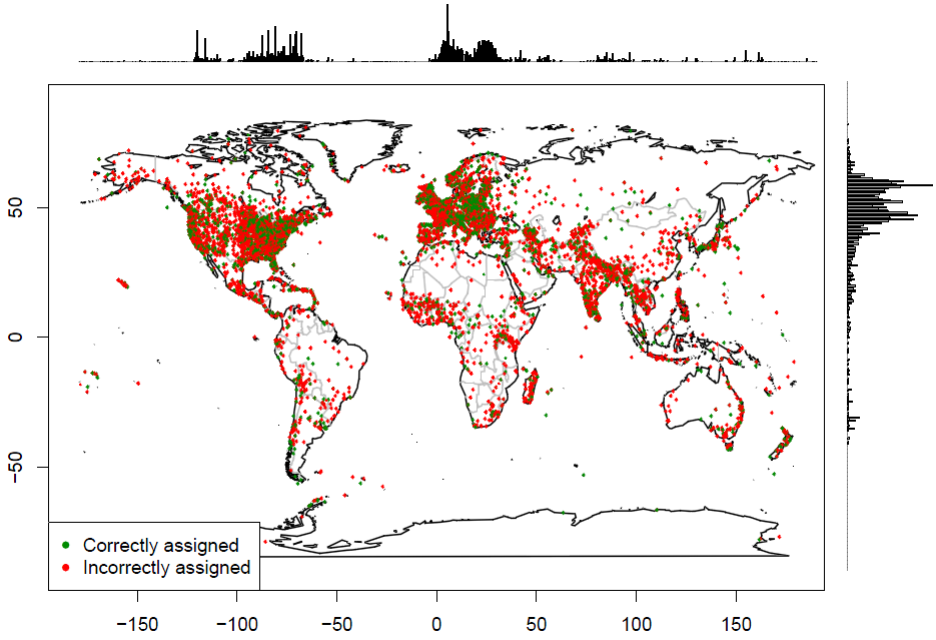


Fig. 3: Estimated positions for the Wikipedia documents.

## 5 Conclusions and Future Work

This paper evaluated methods for georeferencing textual documents. We have shown that automatic identification of the geospatial location of a document, based only on its text, can be performed with high accuracy using simple supervised methods based on language modeling, and using a discrete binned representation of the Earth's surface based on a hierarchical triangular mesh. The proposed method is simple to implement, and both training and testing can be easily parallelized. Our most effective georeferencing strategy uses language models based on character n-grams, and assigns coordinates of latitude and longitude through the centroid coordinates for the most probable bin in the triangular mesh used to represent the Earth.

There are many possible applications for the method described in this paper. A particular application that we are currently pursuing relates to the usage of the probability distributions over the bins from our representation of the Earth, in order to build thematic maps showing geographic incidence of particular constructs extracted from texts (e.g., maps showing the geographic distribution of opinions towards certain themes). However, it should be noticed that the proposed classification approach, based on language models, does not provide accurate probability estimates for the different classes involved in the problem, instead focusing only on the simpler task of predicting which class is the most

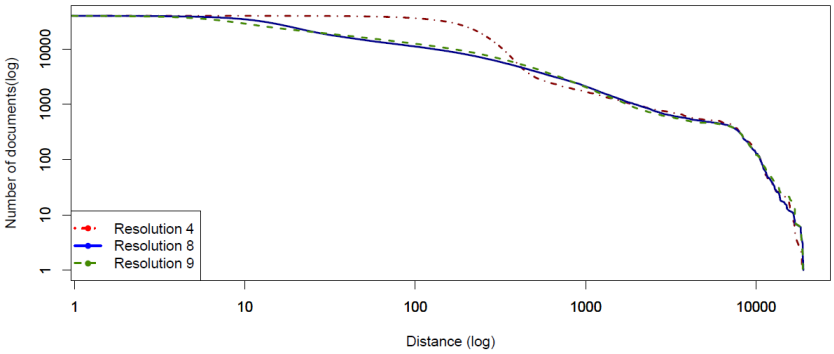


Fig. 4: Errors in coordinates estimated with character-based models.

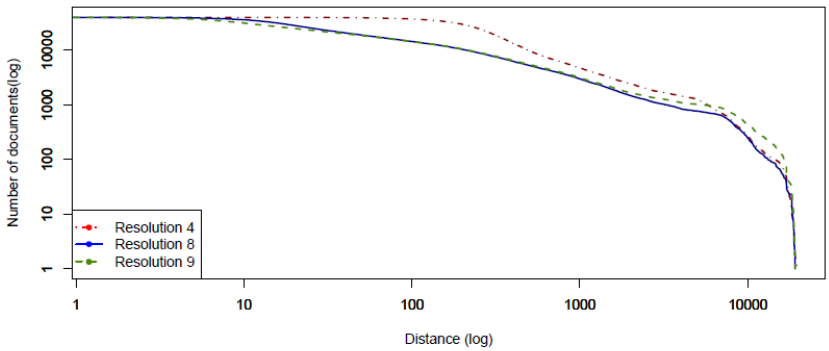


Fig. 5: Errors in coordinates estimated with token-based models.

likely. For future work, we would like to experiment with other classification approaches for assigning documents to the most likely bin(s), for instance max-margin methods such as Support Vector Machines, or Maximum Entropy models. Having well calibrated probability estimates, we could also explore different approaches for choosing the geospatial coordinates that should be assigned to documents, for instance considering (i) a weighted average of the centroid coordinates for all the bins in the model, where the weights come from the probabilities assigned to each of the bins, or considering (ii) a weighted average of the centroid coordinates for the most probable bin and for its adjacent neighbors in the decomposition of the Earth given by the triangular mesh.

There are also many other ideas for future work. We would like, for instance, to experiment with maximum entropy models using either expectation constraints specifying affinities between words and labels [4], or with posterior

regularization [7], leveraging the fact that the presence of the words corresponding to place names is a strong indicator for the document belonging to a particular class. We would also like to experiment with document expansion techniques, particularly for small documents, that could build pseudo-documents by constructing a neighborhood around the original ones (e.g., using linkage information between hypermedia documents, although we should note that we are primarily interested in georeferencing documents using only the text because there are a great many situations in which linkage information is unavailable, a particular example being historical documents in digital libraries).

Although identifying a single location for an entire document can provide a convenient way for connecting texts with locations, useful for many different applications, many other applications could benefit from the complete resolution of place references in textual documents [10]. The probability distributions over bins, provided by our method, can for instance be used to define a document-level prior for the resolution of individual place names.

## References


1. E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.
2. I. Anastácio, B Martins, and P. Calado. A comparison of different approaches for assigning geographic scopes to documents. In *Proceedings of the 1st Simpósio de Informática*, 2010.
3. Bob Carpenter and Breck Baldwin. *Natural Language Processing with LingPipe 4*. LingPipe Publishing, draft edition, 2011.
4. Gregory Druck, Gideon Mann, and Andrew McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in Information Retrieval*, 2008.
5. G. Dutton. Encoding and handling geospatial data with hierarchical triangular meshes. In M. J. Kraak and M. Molenaar, editors, *Advances in GIS Research II*. Taylor and Francis, 1996.
6. Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2010.
7. Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11, 2010.
8. B. Johnstone. Language and place. In R. Mesthrie and W. Wolfram, editors, *Cambridge Handbook of Sociolinguistics*. Cambridge University Press, 2010.
9. Abhimanu Kumar, Matthew Lease, and Jason Baldridge. Supervised language modeling for temporal resolution of texts. In *Proceeding of the 20th ACM conference on Information and knowledge management*, 2011.
10. J. Leidner. *Toponym Resolution in Text*. PhD thesis, University of Edinburgh, 2007.
11. Michael D. Lieberman and Hanan Samet. Multifaceted toponym recognition for streaming news. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011.

12. B. Martins, I. Anastácio, and P. Calado. A machine learning approach for resolving place references in text. In *Proceedings of the 13th AGILE International Conference on Geographic Information Science*, 2010.
13. Simon Overell. *Geographic Information Retrieval: Classification, Disambiguation and Modelling*. PhD thesis, Imperial College London, 2009.
14. Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in Information Retrieval*, 2009.
15. Alexander S. Szalay, Jim Gray, George Fekete, Peter Z. Kunszt, Peter Kukol, and Ani Thakar. Indexing the sphere with the hierarchical triangular mesh. Technical Report MSR-TR-2005-123, Microsoft, 2005.
16. Benjamin Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.