Connecting Language and Geography with Region-Topic Models

Michael Speriosu, Travis Brown, Taesun Moon, Jason Baldridge, and Katrin Erk

The University of Texas at Austin
Austin TX 78712, USA
{speriosu,travis.brown,tsmoon,jbaldrid,katrin.erk}@mail.utexas.edu

Abstract. We describe an approach for connecting language and geography that anchors natural language expressions to specific regions of the Earth, implemented in our *TextGrounder* system. The core of the system is a region-topic model, which we use to learn word distributions for each region discussed in a given corpus. This model performs toponym resolution as a by-product, and additionally enables us to characterize a geographic distribution for corpora, individual texts, or even individual words. We discuss geobrowsing applications made possible by TextGrounder, future directions for using geographical characterizations of words in vector-space models of word meaning, and extending our model to analyzing compositional spatial expressions.

Keywords: Geobrowsing, Toponym Resolution, Topic Models

1 Introduction

Incredible amounts of text are now readily available in digitized form in various collections spanning many languages, domains, topics, and time periods. These collections are rich sources of information, much of which remains hidden in the sheer quantity of words and the connections between different texts. Techniques that reveal this latent information can transform the way users interact with these archives by allowing them to more easily find points of interest or previously unnoticed patterns. In this paper, we describe our preliminary progress in developing our *TextGrounder* system, which we use to create geospatial characterizations and visualizations of text collections. We also discuss the potential for using the representations produced by our system to inform or learn models of how language encodes spatial relationships.

The spatial meaning of an utterance depends on many factors. The expression *a barbecue restaurant 60 miles east of Austin* has a compositional analysis in which one must: (1) identify whether *Austin* refers to a person or place and which person or place it is, including determining the correct latitude and longitude associated with it; (2) identify the location that is 60 miles to the east of that location; and (3) possibly identify a restaurant that serves barbecue in that vicinity. We do not tackle such compositional analysis yet; instead we begin with a standard bag-of-words model of texts that allows us to use the geographic focus

of words like *barbecue* and *restaurant* and other terms in the document to disambiguate (potential) toponyms like *Austin* and landmarks like *the Eiffel Tower*.[1] Our model *learns* that locations are highly associated with certain vocabulary items without using labeled training material; it relies only on a gazetteer. To do this, we use a simple topic model [2] that construes regions of the Earth's surface as topics. We refer to this as the *region-topic model*.

There are at least two linguistically interesting outcomes that could arise from this modeling strategy. The first is that it directly provides a light-weight form of grounding natural language expressions by anchoring them to (distributions over) locations on the Earth. This presents an opportunity to add spatially relevant features into recent vector space models of word meaning (e.g. [4]). Typically, the dimensions of vector space models are not interpretable, and the only way that a vector representation of a word can be interpreted is through its distance to the vectors of other words. In contrast, dimensions relating to locations on Earth will be informative and interpretable in themselves. This will allow us to explore the question of whether such vector space models support additional inferences informed by world knowledge. Second, our approach is language independent, and the fact that expressions are grounded geographically presents the opportunity—without using labeled data, e.g. as with SpatialML [9]—to eventually learn the meaning of expressions like *X 60 miles east of Y*, based on texts that express many different referential noun phrases *X* and *Y*, some of which will be locations which we can resolve accurately.

We aim to use TextGrounder to improve information access for digitized text collections. We are working with a collection of ninety-four British and American travel texts from the nineteenth and early twentieth centuries that were digitized by the University of Texas libraries.[2] These texts are replete with references to locations all around the Earth, so they are an ideal target for geobrowsing applications (e.g. in Google Earth) that display the relative importance of different locations and the text passages that describe them. This kind of analysis could be used to provide "distant reading" interfaces for literary scholarship [12], to support digital archeology [1], or to automatically produce geographic visualizations of important historical events, such as mapping survivor testimonies of the Rwandan genocide. It could also enable users to create mashups of temporally and generically diverse collections, such as Wikipedia articles about the Civil War with contemporary accounts by soldiers and narratives of former slaves.

2 System

TextGrounder performs *geolocation* in a very general sense: it connects natural language texts, expressions, and individual words to geographical coordinates and distributions over geographical coordinates. The most basic and concrete application of geolocation is *toponym resolution*, the identification and disambiguation of place names [7]. For instance, there are at least forty places around

[1] Which could be in Paris (France), Paris (Texas), Las Vegas (Nevada), etc.

[2] `http://www.lib.utexas.edu/books/travel/index.html`

the world called *London*; a toponym resolver must identify that a particular mention of London refers to a place (and not a person, like *Jack London*) and identify which *London* was intended as the referent (e.g., London in Ontario or England). Most systems focus solely on recognizing the places associated with texts based on matching known names to known locations. Typically, simple pattern matching or heuristics are used to identify and disambiguate places.

TextGrounder performs toponym resolution as a by-product; it automatically interprets references to places, landmarks, and geographic features in free text, and uses that information to provide location information on digital maps. Because it learns from raw text, the system uses information and representations that support a much more general connection between language and geography than toponym resolution alone. The system thus performs a light-weight form of grounding computational representations of words in the real world.

The underlying model, depicted in Figure 1, is an adaptation of probabilistic topic models [2]. Topics are simple distributions over the vocabulary for which some particular words have higher probability than others—for example, a topic related to sports would have high probability for words like *team*, *game*, and *ball*. To adapt this approach for geolocation, we represent the Earth as a set of non-overlapping 3-by-3 degree regions, where each region corresponds to a topic. Each document is thus a mixture of region-topics, so different locations discussed in the same document can be modeled. Ultimately, this means that we associate word distributions with specific locations such that words that are more relevant to that location have higher probability. We do not retain all region-topics; instead, given a gazetteer, such as World Gazetteer[3], we consider only region-topics that spatially contain at least one entry in the gazetteer.

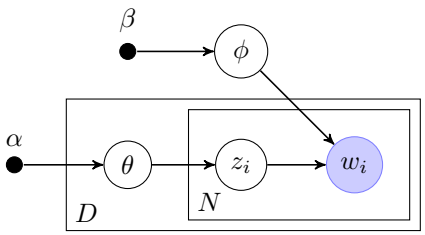

Fig. 1: Graphical representation of the region-topic model with plate notation. The N word observations w_i over D documents is conditioned on the word-level region assignments z_i and a word-by-region prior $\phi|z, \beta \sim \text{Dirichlet}(\beta)$. The topics are drawn from a multinomial on the region-by-document prior $\theta|d, \alpha \sim \text{Dirichlet}(\alpha)$ where $d \in D$. Structurally, the model is identical to a standard topic model—however, the initialization and interpretation of the topics is anchored by actual regions on Earth rather than arbitrarily assigned latent semantic concepts.

To analyze a corpus, we first run the Stanford named entity recognizer[4] (NER) and extract all expressions identified as locations. We then learn the region-topics for each word and toponym. Unlike standard topic models, where topics are not explicitly linked to an external representation, region-topics are

[3] `http://world-gazetteer.com/`
[4] `http://nlp.stanford.edu/ner`

anchored to specific areas of the Earth's surface. This allows us to initialize the inference procedure for our model by seeding the possible topics to only those for which we have some evidence; this evidence comes via toponyms identified by the NER system and the regions which contain a location indexed by those toponyms. The word distributions for non-toponyms in a text conditioned over regions are then inferred along with distributions for the region-constrained toponyms through a collapsed Gibbs sampler. Note that we do not consider the topology of the regions themselves (i.e. our model has no knowledge of the systems of neighborhoods which are inherent in the definition of regions over the globe); the present model is an intermediate step towards that goal.

Toponym resolution is performed implicitly by this model because the identified toponyms in a text are constrained to have positive joint probability only with the regions that enclose the corresponding, possibly ambiguous, coordinates in the gazetteer for those toponyms. If each toponym in a document is associated with multiple regions, the topic model will learn a topic and word distribution that assigns high probabilities to regions that coincide among the possible regions. For example, *London*, *Piccadilly* and *Hyde Park* might occur in the same document; each of these toponyms are ambiguously mapped to more than one region. There are different mixtures of regions that contain all these toponyms; the topic model will assign higher probability to an analysis that accounts for all of them in a single region (namely, the one containing London, UK). After a burn-in period for the Gibbs sampler, we take a single sample (or average over multiple samples) and geolocate the toponyms by placing the toponym on the coordinates which are resolved by the gazetteer and the region assignment.

The region-topic distributions include both toponyms and standard vocabulary items (non-toponyms). Because non-toponyms are unconstrained over regions, they provide additional evidence for determining the set of region-topics required to explain each document. Thus, they aid in toponym resolution *and* the model discovers the words that are most associated with each region. For example, the region-topic containing Austin, Texas would have high probability for words like *music*, *barbecue*, and *computers*, whereas for San Francisco, we'd expect *bay*, *finance*, and *tourism* to be prominent words. Based on these distributions, we can determine additional relationships, such as the distribution of a word over the Earth's surface (by considering its probability in each of the region-topics) or the similarity of different regions based on their corresponding region-topics (e.g. through information divergence measures).

3 Datasets and output

We seek to make the British and American travel collection more useful for scholars of the period through TextGrounder-generated KML (Keyhole Markup Language) files that may be loaded into a geobrowser like Google Earth, including (1) plotting the prominence of different locations on Earth in the collection, (2) embedding text passages at their identified locations for discovery, and (3) plotting the region-topic word distributions (see Figure 2). These preliminary

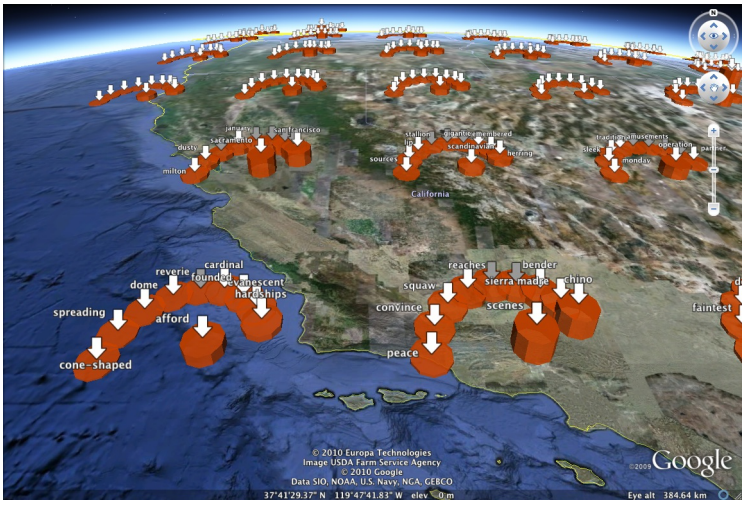

Fig. 2: TextGrounder visualization in Google Earth for John Beadle's *Western Wilds, and the Men who Redeem Them*. The top ten words associated with each region are shown, with red 3D bars indicating their relative frequency for that region.

visualizations provide subjective characterizations of the quality of the system output, which has been useful as we develop and refine our approaches. To obtain an objective measure of performance on the specific task of toponym resolution, we are now interfacing TextGrounder with the TR-CoNLL corpus [7], which contains 946 English-language newspaper articles that contain human-annotated ground truth for 5,963 toponym mentions.

To test the cross-lingual applicability of TextGrounder, we will create a multilingual geotagged subset of Wikipedia (see [13] for an extensive discussion of Wikipedia's geotagged articles and modeling based on them) that we can use as a test corpus. TextGrounder associates multiple regions with each document, but some regions will tend to dominate each document; we can thus choose a location that is most central to each document and check the geospatial distance from that location to the one annotated in Wikipedia. We will create the corpus by extracting pages in English, German, and Portuguese that have similar geographic coverage in each language (this is necessary because the English Wikipedia is much larger than the others and has more complete geotagging). We will identify a subset of pages in all three languages that discuss the same locations, using their geotagged information. This will be a reasonably large set: there are currently over 170,000 articles in English (and 1.2 million across all languages) that are annotated with a geotag for the main subject of the article.

The approach and methodology we advocate are general and flexible—the same methods can be applied relatively independently of the particular corpus being analyzed and the task at hand. The resulting robustness gives us confidence

that our approach will scale well, allowing us to provide geographical searching and browsing for a much wider range of documents than has been possible in traditionally curated literary or historical collections. The unsupervised methods we use allow a more useful mapping of texts because they do not base grounding entirely on toponyms; this means we can characterize the relative importance of different locations using a much wider array of evidence than those that simply resolve toponyms. Furthermore, incorporation of more diverse evidence is of retroactive benefit to toponym resolution, and we believe it will be mutually beneficial to jointly learn a textual hidden space and a geospatial model.

4 Spatial features and word meaning

Vector space models are a popular framework for the representation of word meaning, encoding the meaning of lemmas as high-dimensional vectors [6, 8]. In the default case, the components of these vectors measure the co-occurrence of the lemma with context features over a large corpus. Vector spaces are attractive because they can be constructed automatically from large corpora; however, the interpretation of the representation for a word is based solely on its distance in space to other words. The region-topic model provides an opportunity to represent the meaning of words through *grounded* features: words can be represented as a vector whose dimensions are region topics, and the coordinates are the word probabilities under the topics. This model overcomes the dichotomy of corpus-derived but uninterpretable versus human-generated and interpretable features: it is automatically derived, but offers directly interpretable geographical features.

We will use the region-topic models as a vector space model to study three sets of issues. (1) Traditional vector space models characterize the meaning of a word intra-textually, solely through other words. How do grounded representations compare on traditional tasks like word similarity estimation? Are they perhaps less noisy simply by virtue of pointing to extra-textual entities? (2) Similarity measures typically used in vector space models, such as Cosine and Jaccard, treat dimensions as opaque. In a model where dimensions are regions, we can exploit world knowledge in measuring similarity, for example by taking the distance between regions into account. Can this fact be used to derive better estimates of word similarity? (3) While most vector space models derive one vector per word, conflating senses of polysemous words, it is also possible to derive vectors for a word in a particular context [11, 3]. In a context of *eat apple*, the vector of *apple* would focus on the fruit sense of apple, suppressing features that speak to the company sense. This raises the question of whether it is possible to determine contextually appropriate interpretable features. In the example above, features like *Michigan*, *California* or *New Zealand* should be strengthened, while *Cupertino* (associated with Apple Inc.) should be suppressed. On the technical side, the main challenge will lie in the difference in strength between dimensions, due to different corpus frequencies of different senses of a polysemous word.

5 Related work

There has been quite a bit of research addressing the specific problem of toponym resolution (see [7] for an overview). Of particular relevance is the Perseus Project, which uses a heuristic system for resolving toponyms and creating automatically generated maps of texts written around the time of the Civil War [14].

The two current approaches that are most similar to ours are the location-aware topic model [10] and the location topic model [5], but the form of our model is different from both of these. The location-aware topic model assumes that every document is associated with a small set of locations, so its representation of geography is discrete and quite restricted. The location topic model is more similar to ours: they also seek to learn connections between words and geography using a topic model, and the visualizations they produce (for travel blogs) have a similar flavor. Interestingly, they model documents as mixtures of location-based topics and more general topics: this of course allows them to characterize words that do not have compelling specific geographical meaning. They preprocess their data, and perform toponym disambiguation using a heuristic system (the details of which are not given). Our model uses a different representation that actually grounds topics explicitly, because each topic is directly tied to a specific region on Earth. As a result, our model connects language to geography and performs toponym disambiguation as a by-product. We are interested in combining these two models to see how the learned word distributions differ and the effects they have on toponym disambiguation and our visualizations.

6 Conclusion

The Internet has become a repository of information in many of the world's languages, but the sheer quantity of written material—especially when considering multilingual contexts—also makes it harder to find or digest information of interest. We seek to create meaningful abstractions of language that allow large text collections to be browsed with respect to the places they discuss. These abstractions are learnable from unannotated texts, which greatly facilitates their use for any language with digitized material.

The historically and politically relevant collections that we are examining provide diverse materials that are replete with references to real people and places. This makes them an ideal target for geospatial resolution. Our model performs this resolution, but more importantly, it uses representations that enable many alternative ways of relating language to geography. This in turn supports many different ways to visualize texts geospatially, including seeing the geographic centrality of an entire collection or for a single word or expression, as well as exploring the text passages most relevant for a given location in context. These kinds of visualization will enable scholars to interact with massive text collections in novel ways, and will test the potential of maps to serve "not as all-encompassing solutions, but as generators of ideas" [12].

Additionally, these representations create the possibility to anchor natural language expressions to the real world in a light-weight fashion—this has the

potential to make them useful for inclusion in vector space models of word meaning. By starting at this level, using very simple assumptions about the dependencies between words (by treating texts as bags-of-words), we can analyze many texts and many languages. However, we ultimately are interested in deriving the geospatial meaning of *compositional* expressions—a very difficult task, but one which we hope our current models will help us eventually address.

TextGrounder is an ongoing effort. The system, example output and updated documentation are available on the project's website.[5]

Acknowledgments. We acknowledge the support of a grant from the Morris Memorial Trust Fund of the New York Community Trust.

References

1. Barker, E., Bouzarovski, S., Pelling, C., Isaksen, L.: Mapping an ancient historian in a digital age: the herodotus encoded space-text-image archive (hestia). Leeds International Classical Studies 9(1) (2010)
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Erk, K., Padó, S.: A structured vector space model for word meaning in context. In: Proceedings of EMNLP. Honolulu, HI (2008)
4. Erk, K.: Representing words as regions in vector space. In: Proceedings of CoNLL-2009. pp. 57–65. Association for Computational Linguistics, Boulder, Colorado (June 2009)
5. Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J.M., Pang, Y., Zhang, L.: Equip tourists with knowledge mined from travelogues. In: Proceedings of WWW 2010. pp. 401–410 (2010)
6. Landauer, T., Dumais, S.: A solution to Platos problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104(2), 211–240 (1997)
7. Leidner, J.: Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Dissertation.com (2008)
8. Lowe, W.: Towards a theory of semantic space. In: Proceedings of CogSci. pp. 576–581 (2001)
9. Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., Wellner, B.: Spatialml: Annotation scheme, corpora, and tools. In: Proceedings of LREC'08. Marrakech, Morocco (May 2008)
10. Mei, Q., Liu, C., Su, H., Zhai, C.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: Proceedings of WWW '06. pp. 533–542. ACM, New York, NY, USA (2006)
11. Mitchell, J., Lapata, M.: Vector-based models of semantic composition. In: Proceedings of ACL. Columbus, OH (2008)
12. Moretti, F.: Atlas of the European Novel 1800-1900. Verso (1999)
13. Overell, S.: Geographic Information Retrieval: Classification, Disambiguation and Modelling. Ph.D. thesis, Imperial College London (2009)
14. Smith, D.A., Crane, G.: Disambiguating geographic names in a historical digital library. In: Proceedings of ECDL'01. pp. 127–136 (2001)

[5] `http://code.google.com/p/textgrounder/`