The Substantial Words Are in the Ground and Sea: Computationally Linking Text and Geography

Travis Brown, Jason Baldridge, Maria Esteva, and Weijia Xu

1. Introduction

In a recent *Digital Humanities Quarterly* essay on the future of geographical tools for scholarly research in the humanities, Tom Elliott and Sean Gillies speculate that by 2017, "all web-facing textual resources will be parsed (rightly or wrongly) for geographical content" (par. 13). Most of this geoparsing, they argue, will be done by "the search engines," which will automatically identify names and descriptions of places and match them with "coordinate data in reference datasets." Only a small body of texts produced by "academics and specialist communities" will be annotated with more sophisticated geographical information. Elliott and Gillies provide a compelling outline of the dangers and difficulties of relying on companies like Google for these kinds of research tools, and their own project, Pleiades, is a good example of a curatorial tool that members of a specialized community (in their case, classics scholars) can use to develop authoritative geographical resources. In this essay, however, we will propose a different model for the future of digital geography—and possibly digital projects more generally—in the humanities. Our more optimistic projection is that the next several years will see a narrowing, not a widening, of the divide between massively automated (and often wildly inaccurate) endeavors like Google Books, on the one hand, and scholarly projects based on professional curation, like Pleiades, on the other. We will present our own work on TextGrounder,[1] a geoparsing system that learns relationships between words and places from large bodies of unannotated text, as one step toward this convergence.

Anyone who has ever looked at the "places mentioned in this book" list for a Google Books text has seen examples of the many ways that auto-

Texas Studies in Literature and Language, Vol. 54, No. 3, Fall 2012

DOI: 10.7560/TSLL54303

mated geographical annotation can fail. Elliott and Gillies find mislabeled references to Syracuse, New York, and Tempe, Arizona, in an 1854 translation of Xenophon's *Anabasis*, for example, and a glance at the map for the *Oxford World Classics Edition of the King James Bible* similarly turns up Bethesda, Maryland; Abilene, Texas; Dothan, Alabama; and a Bathsheba in the Caribbean. The problem is that place names are highly ambiguous: even in a monolingual context, a single place name, or *toponym*, may refer to dozens or even hundreds of different locations or regions on the earth. The difficulty of selecting the correct location or region for a given toponym —a task often referred to as *toponym resolution*—is compounded by the fact that individual documents may cover a wide geographical scope, and that individual toponyms may refer to many different kinds of geographical or geopolitical features or entities. "Washington," for example, is most prominently the name of a state in the US and the nation's capital, but it also refers to hundreds of towns, cities, counties, lakes, mountains, and streets around the world. One might expect the vocabulary of a document to provide clear clues as to the distribution of its geographical reference, but making use of non-toponym context in toponym resolution has proven to be difficult for a number of theoretical and practical reasons, and many current toponym resolution systems focus primarily on the place names in a document, in some cases ignoring other expressions entirely.

The toponym resolution system used by Google for the visualizations in Google Books seems to be particularly straightforward: it possibly does nothing more than choose the most prominent or populous entry in a database that maps toponyms to coordinates on the surface of the earth. We might additionally speculate, given the kinds of errors made by the system, that Google's algorithm gives places in the US priority over the rest of the world. The value of the geographical analysis and visualization in Google Books is undermined for many users by the fact that Google does not currently provide information about the methods used to create these maps, but the data and mapping tools provided by Google have been widely useful as a platform for other projects with their own analytical tools.

One example of a more sophisticated approach that builds on Google's resources is the system developed by David A. Smith and Gregory Crane for the Perseus Project at Tufts University. The Perseus system takes the geometric center of all potential referents for toponyms in a document and then removes candidates based on simple rules and commonsense assumptions about the data (Smith and Crane 133). Such a system, for example, might encode an assumption that repeated occurrences of a toponym in a document refer to a single location, or that places referred to in a document are likely to be near each other (Leidner 83). This kind of more advanced heuristic approach represents the current state of the art for toponym resolution, and the improvement that it offers over the Google Books approach can be seen on

the "most frequently mentioned places" pages for documents in the Perseus Digital Library. The Perseus maps use the same Google Maps interface as Google Books, but the toponym resolution is generally of a markedly higher quality. It is still far from perfect, however; it finds, for example, references in a translation of Homer's *Odyssey* to both North Wind, Maryland, and West Wind, Ohio. The data-driven approach that we will describe below provides one possible strategy for reducing the frequency of these kinds of errors by extracting deeper connections between geography and natural language expressions in text. These connections, in turn, have the potential to support novel means of interacting with text collections.

Before we turn our attention to the TextGrounder system, it is worth noting that the need for (or utility of) automated geographical disambiguation for scholarly research in the humanities is not a matter of universal agreement, even when it comes to projects that are primarily quantitative or digital. In the "Maps" chapter of *Graphs, Maps, Trees*, for example, Franco Moretti describes a process of literary map-making that is entirely driven by manual selection and annotation:

> What literary maps do . . . First, they are a good way to prepare text for analysis. You choose a unit—walks, lawsuits, luxury goods, whatever—find its occurrences, place them in space . . . or in other words: you *reduce* the text to a few elements, and *abstract* them from the narrative flow, and construct a new, *artificial* object like the maps that I have been discussing. (53)

The radical nature of Moretti's project is often emphasized by both his allies and his detractors, but there is also something deeply traditional about the curatorial aspect of his methodology as it is described in passages like this. These activities—*choosing*, *finding*, *placing*, *reducing*, *abstracting*, *constructing*—are vastly more complex than telling the difference between Memphis, Tennessee, and Memphis, Egypt. Like the geographical collections that an archaeologist might build using Elliott and Gillies's Pleiades software, the maps that Moretti provides in *Graphs, Maps, Trees* and *Atlas of the European Novel, 1800–1900* are masterpieces of scholarly detail and attention.

This kind of curation—as opposed to automation—is the dominant paradigm in the humanities for the creation of digital resources in general, and for cartographic or geographical resources in particular. The Civil War Washington archive at the University of Nebraska, the Valley of the Shadow at the University of Virginia, and the Nineteenth-Century Concord Digital Archive, for example, are all the result of enormous amounts of scholarly effort, which is both a key to their value—they represent the knowledge and perspectives of experts—and a limitation, since only a small percentage of the available historical documents can be processed in this way.

Our argument here is twofold. In the first place we claim simply that developing better models of the relationships between words and places will allow tools like TextGrounder to perform more accurate toponym resolution than the heuristic-based systems currently used by projects such as Google Books and the Perseus Digital Library. We also argue, perhaps more provocatively, that the techniques we describe have the potential to transform the ways that scholars interact with text archives in much the same manner that Moretti's maps and other curated geographical resources do, but on a vastly larger scale. These techniques multiply the effectiveness of curatorial expertise: instead of laboriously annotating large datasets, scholars train and evaluate programs that perform this annotation. The advantages go beyond improvements in efficiency, however. With the right models, we claim, allowing the computer to perform the *choosing*, *finding*, and *placing* will reveal latent patterns and information in large text collections that even trained experts could not be expected to discover.

2. Application and Datasets

Our TextGrounder system is one component of TeXIT (Texas X-Lingual Interpretation of Texts), a project at the University of Texas at Austin that is supported by the Morris Memorial Trust Fund through the New York Community Trust. The project is dedicated to improving computational analyses of natural language texts and developing tools to assist in cross-cultural information exchange. The goal of TextGrounder specifically is to identify references to places and times in texts and to disambiguate them to points on the surface of the earth or on a historical timeline. The focus in this paper will be on the geographical component.

Toponym resolution as described above is the most common task in geolocation, but TextGrounder uses information and representations that support a much more general connection between language and geography than toponym resolution alone, as it is performed by Google Books or similar geoparsing applications. Unlike those systems, TextGrounder performs a lightweight form of *grounding* computational representations of words to properties of the real world—an approach that we will describe in the following section. The result is that natural language texts, expressions, and individual words are connected to geographical coordinates and distributions over geographical coordinates, and toponyms are resolved as a by-product of this more general analysis. The resulting relationships between locations, documents, toponyms, and non-toponym expressions can be visualized in a variety of ways.

The system currently accepts unannotated texts and a *gazetteer* as input, and generates as output Keyhole Markup Language (KML) files that may be loaded into a geobrowser such as Google Earth. In its simplest form,

a gazetteer is simply a mapping between place names and lists of coordinates, with some gazetteers also encoding information about geopolitical organization, population demographics, or location type (such as *populated place*, *landmark*, or *geological feature*). Gazetteers vary widely in their quality and coverage, but gazetteer development is currently a topic of much interest and effort, and there is a dizzying amount of reliable geographical information that is freely available in electronic form today. We have been working primarily with the GeoNames gazetteer, which includes over 8 million place name entries collected from sources including the CIA World Factbook and the National Geospatial-Intelligence Agency, and is distributed under a Creative Commons Attribution 3.0 License (Wick). The Google Earth geobrowsing software and services are also freely available, and they provide a preliminary form of visualization that allows us to form a subjective characterization of the quality of the output of our models as we develop and refine our approaches. In the future the system will support more advanced methods for geographical searching and browsing.

Our immediate goal for the application is to improve information access for digital text collections, with a focus on two specific corpora that have been developed at the University of Texas Libraries. The first of these, which we will refer to as the Perry-Castañeda Library (PCL) Travel corpus, is a collection of ninety-four British and American books on world travel and history from the late nineteenth and early twentieth centuries. These texts were digitized by the University of Texas Libraries and are replete with references to locations around the earth. The popular and extra-canonical nature of the texts provides a unique window into Anglo-American attitudes and prejudices in the decades preceding the First World War, at a time when technological and social developments made both travel and the consumption of travel literature available to increasingly large portions of the populations of the US and the UK. The many research questions raised by these texts in relation to the language of racial and cultural difference in this period—along with their wide-ranging geographical references—make the collection an ideal target for a geobrowsing interface that displays the relative importance of different locations and the text passages and terms that describe them.

Our other initial test corpus is *With Walt Whitman in Camden*, a uniquely wide-ranging biographical memoir by Whitman's friend and literary executor, Horace Traubel. This text spans nine volumes in over a million and a half words, and includes verbatim reproductions of hundreds of letters to and from Whitman, as well as conversations with Whitman about his travels, friends, and opinions. Traubel visited Whitman almost daily in the last years of his life (from 1888 to 1892), and during this period Whitman's health frequently confined him to his home. The letters, documents, and conversations that Traubel relates in *With Walt Whitman in Camden* trace Whitman's

retrospective reconsiderations of his travels and the world around him, and the methods that we are developing will potentially allow us to discover new patterns in Whitman's geographical imagination during these years.

3. Computational Approach

In the opening strophe of "A Song of the Rolling Earth" in *Leaves of Grass*, Walt Whitman poses (and answers) a question about the reality of words:

> Were you thinking that those were the words, those upright lines?
> those curves, angles, dots?
> No, those are not the words, the substantial words are in the
> ground and sea. (2–3)

This assertion of the insubstantiality of written text prefigures one of the central concerns of computational linguistics today: how to develop better models of word meaning. TextGrounder is an interdisciplinary project involving both computational linguists and humanities scholars, and while this paper is primarily focused on current and potential applications for literary studies, the approaches that we are developing are motivated by this long-standing question in computational linguistics. We are particularly interested in acquiring computational models of word meaning that are *grounded*, in the sense that they link natural language expressions to measurable properties of the real world. In one common approach to grounding, these properties are sensory: visual, auditory, olfactory, or tactile. This kind of model of word meaning is necessary, for example, for the building of robots that communicate in natural language and interact directly with their physical surroundings. Linking natural language to sensory properties in this manner is a difficult undertaking, but there are other ways of connecting language to the real world (or some proxy of it), such as by extracting public opinion trends from Twitter posts (O'Connor et al.), predicting movie revenues from movie reviews (Joshi et al.), or creating three-dimensional images from textual descriptions (Coyne and Sproat).

The earth's coordinate system and timeline are two particularly promising representations for grounding natural language expressions, and in our work on TextGrounder we are developing models that link texts to locations on the surface of the earth and moments in human history. This approach to grounding is especially promising given that languages and language use are intimately tied to both location and historical time: for example, the term "wireless" in a twenty-first century American document most probably refers to a cableless Internet connection, while this meaning would be impossible in a postwar British novel, where the term would mean "radio receiver." The ability to discover these

kinds of relationships among words and locations and times may prove useful beyond geolocation and toponym resolution, and we hope to apply these methods to other natural language processing tasks such as word-sense disambiguation and textual entailment.

Of course location and time are also of particular relevance for scholars in the humanities, and we are working to exploit these acquired connections between words and locations in applications for enabling humanities scholars to interact with large quantities of text. We are engaged in a cyclical development process in which the possibilities afforded by automated analysis can give rise to a range of possible literary questions, which in turn push the algorithm development as those questions become more focused and refined. In this sense, the process can be thought of as a form of algorithmic aid to curation, in which the capabilities of the system reveal possibilities for analysis, and the useful questions supported by those possibilities inform system development to produce more accurate or meaningful output.

It is possible to implement a basic form of toponym resolution in a conceptually straightforward manner, but the process involves a number of resources and technical challenges. One of the most important resources is the gazetteer. The next requirement is a natural language processing infrastructure for analyzing the text of the corpus. At a bare minimum it is necessary to have a tool that performs *named entity recognition*—in this case the task of identifying references to locations in text. A naïve implementation—such as the one currently used in Google Books—might simply search for all occurrences in the text of toponyms given in the gazetteer, but this approach will obviously have a very low precision—that is, it will incorrectly identify many non-toponyms, such as "west wind" or Jack London's last name, as toponyms.

In TextGrounder, we use a pipeline of natural language processing tools from the OpenNLP toolkit to identify sentence boundaries, tokenize the text, label each word with a part of speech, and finally identify the named entities, which may be multiword expressions. All of these language analysis steps employ models that were trained on newswire text from the 1980s and 1990s using machine-learning techniques. These natural language processing tasks have been the focus of much study in the fields of natural language processing and computational linguistics over the past several decades, and tools like OpenNLP generally produce highly accurate output on contemporary English text.[2] Their performance on more temporally distant English text—such as the nineteenth-century documents that are the focus of our current work—is often difficult to evaluate, given the scarcity of linguistically annotated historical corpora. The output of our current pipeline on a Baedeker's guide from 1909, for example, is recognizably reasonable when random samples are subjected to manual inspection, but it is difficult to establish precise measures of

quality without large amounts of annotated text in the appropriate domain, and one of our goals for the project is to develop task-based methods for evaluating these analytical steps.

Once the raw text has been processed in this way, a wide range of algorithms for resolving the toponyms can be employed. The core of the task at this point is to cycle through the identified toponyms and, for each toponym, choose the entry in the gazetteer that matches the toponym's name and is the most appropriate for that specific occurrence. As described above, most current toponym resolution systems rely on a combination of rules and encoded assumptions about the distributions of toponyms in documents. One common approach, for example, is to give priority to the more populous of the candidate locations for a toponym while simultaneously trying to minimize the physical distance between different possible locations for multiple toponyms in a single document (Andogah 12–15). The accuracy of these algorithms is still low relative to similar disambiguation tasks, even for contemporary texts, and our hypothesis is that this is—at least in part—because current approaches generally do not effectively use the words in the text surrounding the toponyms, but instead rely almost exclusively on the information in the gazetteers and the spatial relationships between candidate resolutions for the toponyms. Additionally, in our work we explicitly do not use information about population, since population figures given in gazetteers are for recent years and are likely to be irrelevant or even misleading for analyzing texts from the nineteenth century.

Though human annotators will always outperform machines for accuracy (although not speed) on tasks that involve finding and labeling, such as toponym resolution, there are machine-learning methods for extracting latent semantic information—deeper connections between text and meaning—that no human could easily match. For example, a method called Latent Dirichlet Allocation (LDA) is a probabilistic extension of the well-known Latent Semantic Analysis, and is the basis of one common approach to *topic modeling* (Steyvers and Griffiths). LDA topic models are based on two simple assumptions: that basic semantic concepts are represented as probability distributions over the entire vocabulary, and that any given document is primarily about a small number of semantic concepts. Documents are viewed, in probabilistic terms, as having been created by a generative model that first chooses proportions for the topics and then chooses words from those topics respecting those proportions.

Over the past several years topic model approaches—including particularly LDA—have been used in a variety of digital humanities projects. The fact that topic modeling works on raw text with little or no preprocessing or annotation makes it a valuable tool for scholars working in domains without large annotated corpora that could be used to train

taggers and parsers. Two prominent topic-modeling projects involve historical newspaper corpora, with Sharon Block and David J. Newman at UCLA–Irvine working on the *Pennsylvania Gazette*, and Robert K. Nelson examining the Richmond *Daily Dispatch* in the Digital Scholarship Lab at the University of Richmond. Block and Newman describe the advantages of the approach for the analysis of historical periodicals:

> Because there is no a priori designation of topics—in fact there are very few "knobs to turn" in the method—historians do not need to rely on fallible human indexing or their own preconceived identification of topics. But the most important advantage of this method is its ability to analyze orders-of-magnitude more documents than a person can reasonably view. Thus, instead of resorting to sampling to analyze a large volume of documents, the computer can analyze the entire corpus. This is especially useful in studies of print culture that strive to understand how historical actors read and understood entire publications. (766)

Nelson further emphasizes the importance of the unsupervised nature of the approach and its value as an exploratory tool:

> Topic modeling and other distant reading methods are most valuable not when they allow us to see patterns that we can easily explain but when they reveal patterns that we can't, patterns that surprise us and that prompt interesting and useful research questions. (par. 21)

Both projects track the distribution of topics over time, which is a natural line of investigation given the arrangement of a newspaper into a chronological sequence of documents. Various extensions of the LDA model have been developed to incorporate this kind of chronological information more explicitly, including Xuerui Wang and Andrew McCallum's *Topics over Time*. Our own work extends the basic model to incorporate geographical rather than chronological information, but the approaches are potentially compatible.

Before we describe our geographical topic model, however, we will first look at a concrete example of the output of a topic model on a historical corpus. Figure 1 shows a set of eight topics that were automatically extracted from the PCL Travel corpus using the Latent Dirichlet Allocation model.

Each column shows the top fifteen words for one of the one hundred topics acquired from the collection using the topic-modeling package from the MALLET machine-learning toolkit (McCallum). Each topic has been provided with a label by the authors, purely for expository purposes—these labels are not part of the model or its output. These topics are learned

WATER	LAND	WRITING	ART	EGYPT	CITY	MECCA	PLAINS
sea	mountain	cloth	church	egypt	street	mecca	city
coast	feet	book	renaissance	nile	london	city	sioux
island	rock	illustrations	madonna	cairo	station	place	chicago
steamer	valley	story	picture	egyptian	railway	day	river
bay	great	books	venice	temple	road	time	kansas
land	mountains	work	mark	ancient	hotel	desert	missouri
ship	hills	life	doge	time	city	mahomet	west
harbour	rocks	history	early	rameses	square	country	business
port	side	author	saints	water	place	caravan	miles
short	miles	crown	century	desert	house	arabia	hundred
boat	plain	illustrated	great	tombs	great	journey	line
water	river	gilt	work	years	english	damascus	dakota
board	hill	edition	wall	river	office	days	streets
islands	stone	vols	works	temples	regent	pilgrims	railroad
town	long	volume	room	luxor	club	camels	lincoln

Figure 1. Eight topics extracted from the PCL Travel corpus using the MALLET toolkit. Note that the topic labels were provided by the authors, and are not part of the output of a topic model.

without any annotation beyond the division of the text into documents; the structure of the model itself draws out the fact that certain words are correlated. The topics as presented above are quite coherent as lists of related words, but they are in fact more than just lists: each word has a corresponding probability in the topic, so they are additionally ranked with respect to their importance as representative words for the topic. Note that each topic is a distribution over the *entire* vocabulary, so the word "desert," for example, is in fact in the WATER topic (along with all other words): it just has an extremely low probability.

From the perspective of a topic model, a document is simply an unordered and unstructured collection of words that are drawn in specific proportions from various topics. For example, for a document about the Great Plains of the US and shipping routes by train and water, it could be that 50% of the words come from the PLAINS topic, 30% from the WATER topic, and 20% from the LAND topic (with 0% for all other topics). This simplification completely ignores many layers of linguistic phenomena, including syntactic relationships and discourse, but it is often a useful characterization: it is computationally extremely efficient and can provide unexpectedly compelling ways of imagining the organization of large document collections. Perhaps most importantly, the probabilistic foundation behind Latent Dirichlet Allocation makes these models extensible and able to incorporate other kinds of information, such as timestamps in a chronologically arranged corpus. It can also be extended to model the hierarchical organization of named areas (such as country, state, or city) or to use continuous probability distributions (like the von Mises-Fisher

distribution) to capture the intuition that locations that are physically near one another are frequently mentioned near each other in a document.

Note that some of these topics have a clear geographic scope. The ART topic is most obviously related to Italy, and even more particularly to Venice —to the point of including "doge," the title of the chief magistrate of Venice, as a highly probable word. The CITY topic contains many generic city terms, but also includes a number of terms specifically related to London, England. The MECCA topic includes not only the destination Mecca itself, but also words related to the journeys that pilgrims made to Mecca in the late nineteenth century. Finally, the PLAINS topic is connected to a broad area, as is clear from the range of words that come from city names in the Great Plains region. These connections between toponyms and words that describe aspects of the locations they represent are a by-product of the model and the text on which it is trained. The model in this case does not include any specific geographical components—data about population or physical distance, for example—and does not employ any form of gazetteer. The geographical information that is apparent in the lists above is learned entirely from patterns of word occurrence in the corpus.

The TextGrounder system extends standard topic models to incorporate information from a gazetteer. In this extended model, which we call the Region Topic Model (Speriosu et al.), or RTM, the topics are regions on the surface of the earth, each of which contains some number of toponyms from the gazetteer. For example, one region might include Lincoln, Nebraska, along with Omaha, Topeka, and Kansas City. A reference to "Lincoln" in a given document is ambiguous; it could refer to the dozens of cities named Lincoln around the world (not to mention the sixteenth president of the United States and various organizations), but co-occurrence of the term with "Omaha" and "Topeka" suggests that the document is in some sense related to our Nebraska-Kansas region—and therefore that it has that region as a topic. As in the LDA example above, each topic (or region) is associated with a probability distribution over the vocabulary. If, for example, the terms "corn," "plains," "prairie," and "Free-Stater" have high probabilities in our hypothetical Nebraska-Kansas topic, then the occurrence of those terms in a document will be evidence for assigning the topic to that document. The words that accompany toponyms therefore influence their resolution; crucially, this is accomplished without any human supervision about connections between specific words and specific locations.

In the simplest version of the RTM model implemented in TextGrounder, regions are simply rectangular cells on the surface of the earth. The map shown in figure 2 presents the output of this model for the PCL Travel corpus, visualized in the Google Earth geobrowser. The region displayed includes northern Italy along with parts of France, Switzerland, and other countries.

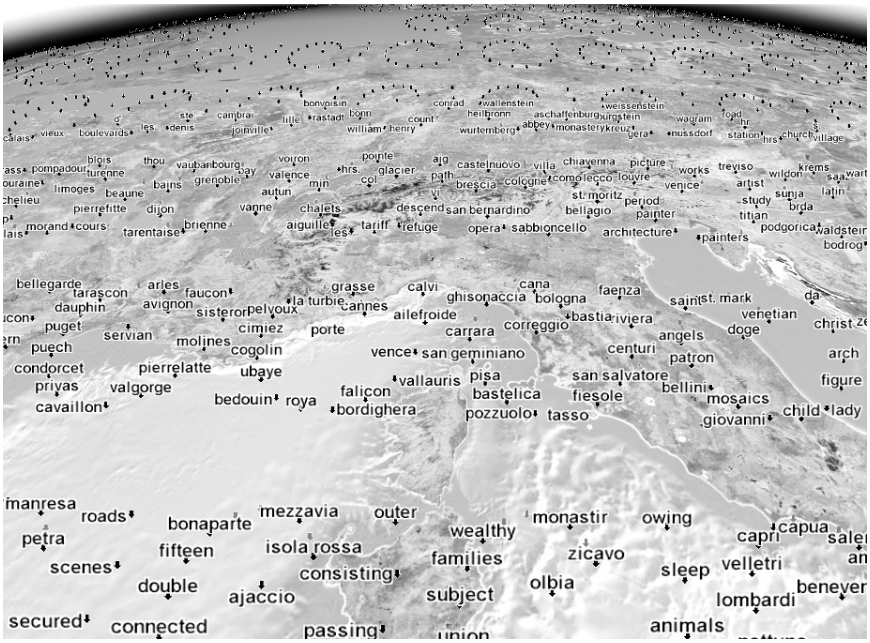

Figure 2. Word distributions from the RTM model in Google Earth.

Here we see that "bonaparte" is strongly associated with a region including Corsica and Sardinia, and that a variety of names and terms related to Renaissance art are associated with Italy. We find "glacier" and "chalet" in the Alps, and "tariff" and "refuge" in Switzerland. Many of the associations are unexpected, with some inviting further investigation, such as "sleep" in Italy and "bedouin" in southern France. Others seem simply anomalous, such as the situation of "louvre" somewhere in Austria.

For other locations this model fails dramatically, but it often does so in interesting ways that provide opportunities to validate and improve the system or discover new associations in the data. In the example shown in figure 3, the topic is clearly defined semantically, in its focus on Egypt and the Old Testament, yet it appears near the border of Texas and Louisiana. The model has fallen victim to the historical fondness of Americans for biblical city names. This failure is in part due to a gap in our training corpus—there are apparently very few documents in the PCL Travel corpus that discuss places in East Texas—but it also highlights a possible problem with our model: we treat all entries in the gazetteer equally, so that initially the term "egypt" is just as likely to refer to the town in Texas as to the nation. We are experimenting with ways of adjusting the model to address this issue. However, we would also note that such associations are driven

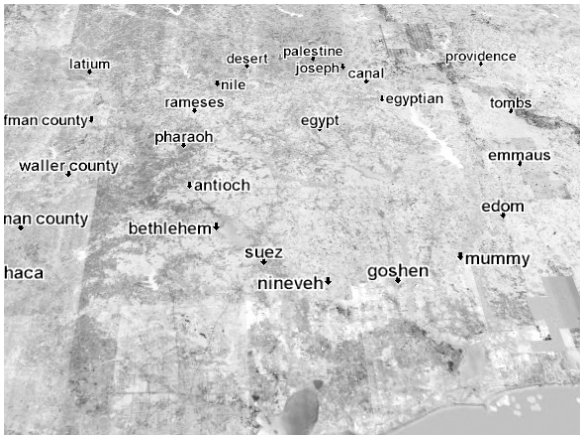

Figure 3. Egyptian toponyms in East Texas.

not by random chance, but by underlying relationships in the data that may be of interest for literary scholarship. Even with such "errors," the output of the model can be useful for the purpose of data exploration and discovery.

We have found that these kinds of visualization—which were built with the literary scholar in mind—have come to play a critical role in the process of developing, debugging, and implementing the model. In many natural language processing tasks, it is customary to evaluate the performance of a system on development or test datasets with *gold* annotations—that is, annotations that have been manually created or approved and that are accepted as correct. Because we are working in domains without appropriately annotated corpora that we could use for evaluation, the ability to form a subjective characterization of the quality of the system's output quickly and accurately is essential to the development process.

4. Scaling up to Large Corpora

In many machine-learning techniques for natural language processing, the benefit of adding more training data levels off relatively quickly—after the first half-million words, for example, or 5 million. This does not seem to be the case in our initial experiments on TextGrounder. Training the topic model on the full 10 million words of the PCL Travel corpus produces recognizably better output than training it on half or three-quarters of the corpus, and using twice as much data should give a similar degree of improvement. Because the methods we are using are *unsupervised*—that is, they take unannotated text as input—adding 10 million more words, or 50 million, is relatively easy. While the PCL Travel corpus has been

thoroughly proofread by human editors, our methods do not even require corrected text, and resources such as Project Gutenberg and Google Books give us access to hundreds of millions of words of freely available English text, much of which has never been read (in its digital form) by a human.

To support computation of models at massive scale, we are adapting TextGrounder for use with the MapReduce programming paradigm pioneered by Google to support distributed computation over massive datasets. We are participating in an effort to support MapReduce on the Longhorn compute cluster at the Texas Advanced Computing Center (TACC) with support from a Longhorn Innovation Fund for Technology (LIFT) grant at The University of Texas at Austin. The cluster will run Hadoop, an open-source implementation of MapReduce created by Yahoo! and the Apache Software Foundation. The hardware is in two groups, with the first group comprised of 16 nodes, each with 144 gigabytes of memory, 1 terabyte of disk storage, and 8 processors; and the second of 64 nodes, each with 48 gigabytes of memory, 1.5 terabytes of disk storage, and 8 processors. This provides 16 terabytes and 96 terabytes, respectively, of total storage for each group. While more storage space could have been obtained, these are high-performance drives that will better support the scale we are interested in. Even accounting for the data replication required by Hadoop, the 16-node group will be able to work with large datasets of up to 5 terabytes, which is sufficient for many of the datasets we are interested in. For example, the PCL Travel corpus, which contains approximately 10 million words, can be stored in a fraction of a gigabyte, while a corpus of 1 million public-domain books requires 1.3 terabytes. The second 64-node group will be able to handle larger datasets than these, such as billions of words of nineteenth-century writing together with terabytes of more recent text from the Internet and other sources. These kinds of century-spanning historical corpora could lead to novel connections and visualizations; one can, for example, imagine an interface that would enable readers of a nineteenth-century novel to track two hundred years of news stories, travel narratives, or blog posts that are spatially or topically related to the passage they are reading.

5. Conclusion

Unsupervised (or semisupervised) methods such as topic models are part of a general trajectory in the history of natural language processing away from rule-based systems and toward more data-driven approaches. Instead of requiring humans to build large sets of syntactic production rules or manually annotate large bodies of text, these methods allow human effort and expertise to be redirected toward building smarter models, better algorithms, and more useful ways of navigating and visualizing the

output of the system. As the hardware necessary to work on extremely large datasets becomes cheaper and the amount of digitized text in any given domain becomes larger, we believe that these kinds of methods will become increasingly important for research in the humanities. Instead of a future where large-scale natural language processing is the domain of companies like Google and careful scholarly curation is a largely unrelated endeavor driven by a small group of humanities researchers, we foresee the development of a more collaborative model of research, where the datasets developed by companies like Google are processed by methods developed by scholars across disciplines.

The University of Texas at Austin
Austin, Texas

NOTES

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

1. TextGrounder is an open source system written in Java and Scala. Downloads, documentation, and news are accessible at https://github.com/utcompling/textgrounder.

2. For example, state of the art part-of-speech taggers for English can have a token accuracy for in-domain text of up to 97%. For sentence-boundary detection and tokenization the accuracy is generally even higher.

WORKS CITED

Andogah, Geoffrey. *Geographically Constrained Information Retrieval*. Diss. University of Groningen, 2010. Groningen: Groningen Dissertations in Linguistics, 2010. Print.

Block, Sharon. "Doing More with Digitization: An Introduction to Topic Modeling of Early American Sources." *Common-Place* 6.2 (January 2006). Web. Accessed 31 September 2010. http://www.common-place.org/vol-06/no-02/tales/.

Block, Sharon, and David J. Newman. "Probabilistic Topic Decomposition of an Eighteenth-Century Newspaper." *Journal of the American Society for Information Science and Technology* 57.5 (March 2006). Web. Accessed 10 October 2010. http://onlinelibrary.wiley.com/doi/10.1002/asi.20342/full.

Coyne, Bob, and Richard Sproat. "WordsEye: An Automatic Text-to-Scene Conversion System." *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. New York: ACM, 2001. Print.

Elliott, Tom, and Sean Gillies. "Digital Geography and Classics." *Digital Humanities Quarterly* 3.1 (Winter 2009). Web. Accessed 18 August 2010. http://www.digitalhumanities.org/dhq/vol/3/1/000031/000031.html.

Joshi, Mahesh, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. "Movie Reviews and Revenues: An Experiment in Text Regression." *Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference*. Los Angeles: ACL, 2010. Print.

Leidner, Jochen. 2008. "Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names." Diss. University of Edinburgh, 2007. Edinburgh: Edinburgh Research Archive, 2007. Print.

McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." *MALLET*. Web. Accessed 21 September 2010. http://mallet.cs.umass.edu/.

Moretti, Franco. *Atlas of the European Novel, 1800–1900*. London: Verso, 1999. Print.

———. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso, 2005. Print.

Nelson, Robert K. "Introduction." *Mining the* Dispatch. Web. Accessed 10 October 2010. http://americanpast.richmond.edu/dispatch/pages/intro.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." *Proceedings of the International AAAI Conference on Weblogs and Social Media*. Washington, DC: Advancement of Artificial Intelligence, 2010. Print.

Smith, David A., and Gregory Crane. "Disambiguating Geographic Names in a Historical Digital Library." *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*. New York: Springer-Verlag, 2001. 127–36. Print.

Speriosu, Mike, Travis Brown, Taesun Moon, Jason Baldridge, and Katrin Erk. "Connecting Language and Geography with Region-Topic Models." *Proceedings of the Workshop on Computational Models of Spatial Language Interpretation*. Portland: CoSLI, 2010. Print.

Steyvers, Mark, and Tom Griffiths. "Probabilistic Topic Models." In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch (eds.), *Latent Semantic Analysis: A Road to Meaning*. Hillsdale, NJ: Erlbaum, 2007. Print.

Wang, Xuerui, and Andrew McCallum. "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends." *Proceedings of the Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2001. Print.

Whitman, Walt. *Leaves of Grass*. Philadelphia: David McKay, 1891. *The Walt Whitman Archive*. Web. Accessed 12 September 2010. http://whitmanarchive.org/published/LG/1891/.

Wick, Marc. "GeoNames User Manual." *GeoNames*. Web. Accessed 10 October 2010. http://www.geonames.org/manual.html.