Disambiguating Geographic Names in a Historical Digital Library*

David A. Smith and Gregory Crane

Perseus Project
Tufts University
Medford, MA, USA
{dasmith,gcrane}@perseus.tufts.edu

Abstract. Geographic interfaces provide natural, scalable visualizations for many digital library collections, but the wide range of data in digital libraries presents some particular problems for identifying and disambiguating place names. We describe the toponym-disambiguation system in the Perseus digital library and evaluate its performance. Name categorization varies significantly among different types of documents, but toponym disambiguation performs at a high level of precision and recall with a gazetteer an order of magnitude larger than most other applications.

1 Introduction

Geographic interfaces provide natural, scalable visualizations for many digital library collections. Although domain-specific ontologies or automatic clusterings of documents may produce productive browsing tools in many cases, real world maps, along with timelines, can situate a wide range of information in a consistent, familiar space. When the contents of digital library documents are georeferenced, users can get a sense of the scope and focus points of a collection or a document, plot geographic places mentioned on any page of text, or find information about the places mentioned on a map or in a region [4,6]. At the Perseus Project, we have concentrated on representing historical data in the humanities from ancient Greece to nineteenth-century America [9]. With over one million identified toponym references, Perseus has built a rich digital library testbed and toolset that is available over the World Wide Web (`http://www.perseus.tufts.edu`; see Fig. 1, 2, and 3).

In order to reap the benefits of geographic interfaces, digital librarians must identify geographic names and link them to information about their location, in most cases their type (e.g. river, mountain, populated place), and other useful information such as dates of occupation, population at various times, and relation to other places. For documents of highly central importance to a scholar's work, it

* This research was supported by a grant from the Digital Libraries Initiative, Phase 2, with primary funding from the National Science Foundation and National Endowment for the Humanities.

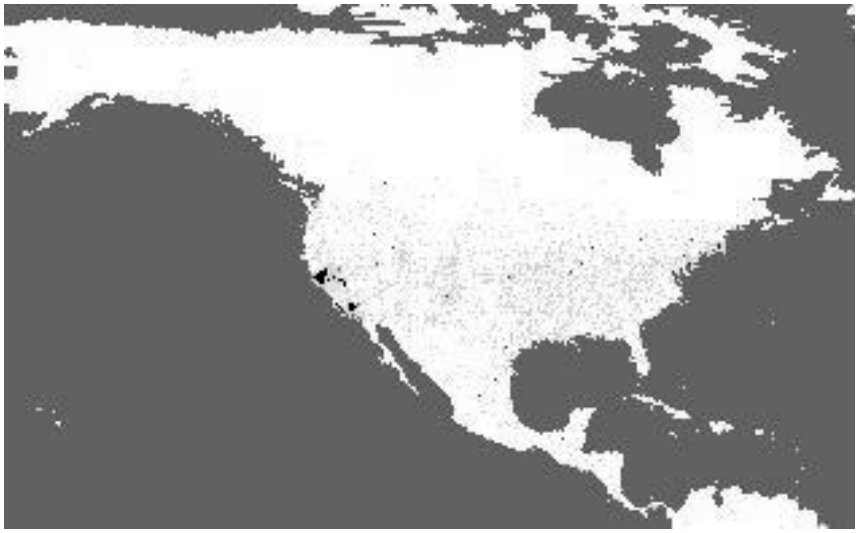

Fig. 1. The scope and focus of the collection on the settlement of California. Note the fainter spread of sites across the U.S. and the concentration in northern California.

Fig. 2. Sites mentioned in Herodotus. Note the strong concentration in present-day Greece and western Turkey.

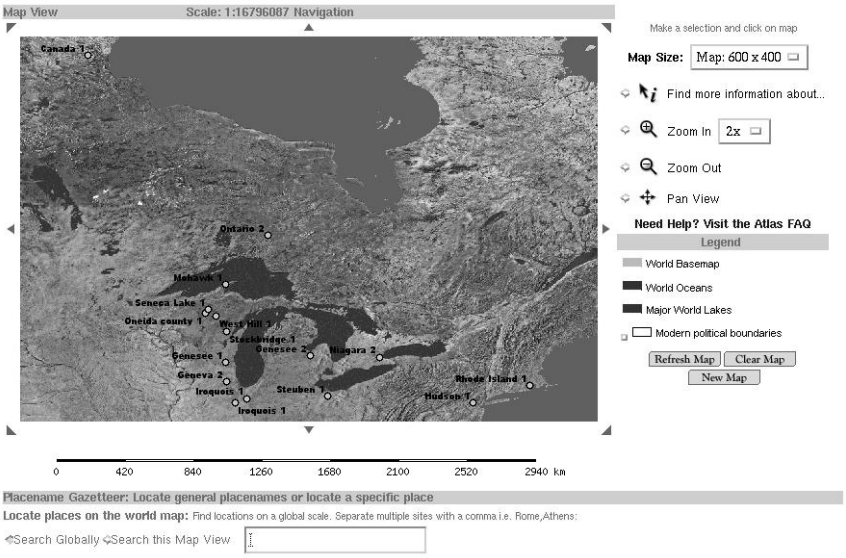

Fig. 3. Interactive map of the sites mentioned on one page of a diary of a voyage from Detroit to the source of the Mississippi, where the author is in Wisconsin but makes peripheral reference to the Naragansetts of Rhode Island. The user can zoom in on a particular region, such as the cluster in Oneida county, Wisconsin, or click on a site for more information.

might be worthwhile to spend the effort of manually tagging and disambiguating place names in a text, but manually tagging an entire corpus of any considerable size is impractical. Even at an optimistic ten seconds per toponym, it would take 28,000 person-hours to check the over one million toponyms in the Perseus DL of about 70 million words of English. We thus need automatic, or at least machine-assisted, methods for building georeferenced digital libraries.

2 Problem Description and Related Work

Linking strings in documents to locations on a map can be generally divided into two steps: name *identification and categorization* and *disambiguation* of those names classed as toponyms against a gazetteer with at least some coordinate information. In the past decade, many projects have devoted themselves to the first step and many fewer to the second. This concentration is not surprising; "named entity recognition", as the name categorization task is known, aims to classify entities as persons, organizations, dates, products, organisms, and so on, in addition to geographic entities, and has many applications in the fields of message understanding and information extraction as a whole. Two general strategies for named entity recognition can be represented by two widely known systems. BBN's Nymble [2] performs quite well with F-measures (see equation 1 below) at 90% or above, but requires at least 100,000 words of training data; IBM's Nominator [10] performs only slightly less well (F-measure approximately 88%) on Wall Street Journal documents with only simple heuristics, though the authors admit that the "heuristics are somewhat domain dependent".

The Geo-Referenced Information Processing System (GIPSY) described by [11] and [6] matches geographic names in text to spatial coordinates. Interestingly, this system also attempts to match such phrases as "south of Lake Tahoe" with fuzzy polygons. [8] describe a system to plot locations mentioned in transcripts of news broadcasts. Using a gazetteer of about 80,000 items, they report matching 269 out of 357 places (75%) in the test segments. Kanada reports 96% precision for geographic name disambiguation in Japanese text with a gazetteer of 55,000 Japanese and 41,000 foreign ones [5]. In an interesting parallel to our results below, he also reports significantly lower precision for Japanese toponyms than for foreign ones.

The documents in many digital libraries, however, present some particular problems for automatically identifying and disambiguating place names. Much work on proper names has dealt with news texts with useful discourse conventions for reducing ambiguity. A story mentioning Bill Clinton will use the full title "President Bill Clinton" on the first mention and "Mr. Clinton" or "Clinton" only afterwards. News stories also have relatively small scope, without long-distance anaphora. When a new story begins, President Bill Clinton is named in full all over again. Finally, place names themselves almost always have a disambiguating tag at their first mention, e.g. "London, Ontario" or "Clinton, New Jersey". Digital libraries, on the other hand, often contain documents of widely varying lengths written without benefit of journalistic style. Scholarly works of-

ten deal with several registers of time and place: a book on Shakespeare will talk not only about sixteenth-century Stratford (Warwickshire) but also about scholarship in nineteenth-century Cambridge (Massachusetts) and twentieth-century Berkeley (California). A wide historical purview can also make some pieces of knowledge in a gazetteer useless or misleading. Although the city of Samos is now in Greece and Miletus is in Turkey, they were both founded by Ionian Greeks and are only about 30 kilometers apart. Actual distance on the earth tells more than modern political categories.

Finally, a heterogeneous digital library can benefit from large knowledge bases to deal with its diverse materials but must deal with the cost of clashes among items in these authority lists. We can explore some of these ambiguities *a priori* by looking at the distributions in a gazetteer (table 1). Although the proportions are dependent on the names and places selected for inclusion in this gazetteer, the relative rankings are suggestive. In long-settled areas—such as Asia, Africa, and Europe—a place may be called by many names over time, but individual names are often distinct. With the increasing tempo of settlement in modern times, however, many places may be called by the same name, particularly by nostalgic colonists in the New World. Other ambiguities arise when people and places share names. Very few Greek and Latin place names are also personal names. This is less true of Britain, where surnames (and surnames used as given names) are often taken from place names; in America, the confusion grows as numerous towns are named after prominent or obscure people. In practice, we can express the scope of the disambiguation problem as follows: not counting the other names that could be mistaken for place names, some 92% of the toponyms in the Perseus digital library refer, potentially, to more than one place.

Continent	% places w/multiple names	% names w/multiple places
North & Central America	11.5	57.1
Oceania	6.9	29.2
South America	11.6	25.0
Asia	32.7	20.3
Africa	27.0	18.2
Europe	18.2	16.6

Table 1. Places with multiple names and names applied to more than one place in the Getty *Thesaurus of Geographic Names*

3 Disambiguation Procedure

As mentioned above, toponym disambiguation consists of name identification and categorization and disambiguation of those names. Our methods for performing these tasks rely on evidence that is internal or external to the text. (Note

the difference with the terminology and approach in [7], which uses evidence internal and external to the *name*.) Internal evidence includes the use of honorifics, generic geographic labels, or linguistic environment. External evidence includes gazetteers, biographical information, and general linguistic knowledge.

Before either identification or disambiguation could proceed, we gathered the knowledge sources used to make the categorization and disambiguation decisions. Perseus uses some knowledge sources, such as the Getty Thesaurus of Geographic Names or Cruchley's gazetteer of London, that were purpose-built for geocoding. We captured other information, such as lists of authors or the entries in the *Dictionary of National Biography*, as a by-product of constructing the digital library as a whole. In total, the gazetteer used for name identification and disambiguation contains over one million place names.

We then scan the documents in the digital library for possible proper names and assign the names, if possible, to broad categories such as person, place, or date. We chose to use simple heuristic methods like those used in Nominator [10] rather than learning systems, since we lacked training data for our types of documents, and since we were mostly interested in identifying geographic names and not in the broader task of named entity recognition and categorization. In English text, the Perseus system exploits generally used capitalization and punctuation conventions: initial candidate proper names are strings of capitalized words, and sentences are delimited with periods. Also at this stage, we exploit any markup that a document's editor has added, whether in tagging a string as a personal or place name, or in explicitly linking that name to an entry in a gazetteer. For initial categorization, the Perseus system uses language-specific honorifics (such as "Dr." or "Mrs.") as strong evidence that the following name is a personal name. In addition, once a "Col. Aldrich", for example, is seen in a document, further references to "Aldrich" are automatically classified a personal names. Generic topographic labels (such as "Rocky *Mountains*" or "Charles *River*") are taken as moderate evidence that the name is geographic. Standalone instances of the most common given names in Perseus' biographical dictionaries are labeled as personal names since a mere "John" is highly unlikely to refer to a town by that name in Louisiana or Virginia.

The system then attempts to match the names classed as geographic, as well as the uncertain names, against a gazetteer. As our aim is to allow for geographic browsing of a digital library, it is of little benefit if we identify a name that cannot be linked to spatial coordinates. As mentioned above, for the names in our corpus that have at least one match in the gazetteer, about 92% match more than one entity.

Disambiguating the possible place names then proceeds based on local context, document context, and general world knowledge. The simplest instances of local context are the explicit disambiguating tags that authors put after place names: e.g. "Lancaster, PA", "Vienna, Austria", or "Beverly Hills, 90210". More generally, a place will more likely than not be near to other places mentioned around it. If "Philadelphia" and "Harrisburg" occur in the same paragraph, a reference to "Lancaster" is more likely to be to the town in Pennsylvania than

to the one in England or Arizona. Document context can be characterized as the preponderance of geographic references in an entire document; for short documents, such as news articles, local and document context can be treated as the same. World knowledge may be captured from gazetteers or other reference works and comprises such facts about a place as its coordinates, political entities to which it belongs, and its size or relative importance.

The system begins by producing a simple characterization of the document context. All of the possible locations for all of the toponyms in the document are aggregated onto a one-by-one degree grid, with weights assigned for the number of mentions of each toponym. The system prunes some possibilities based on general world knowledge, so that only Spain the country, and not the town in Tennessee, will be counted. We compute the centroid of this weighted map and the standard deviation of the distance of the points from this centroid. We then discard points more than two standard deviations away from the centroid and calculate a new centroid from the remaining points, if any.

We then process the possible toponyms for final disambiguation. We represent the local context of a toponym's occurrence as a moving window of the four toponyms mentioned before it and the four after it. Only unambiguous or already disambiguated toponyms are taken into account, however, in constructing this context. Each possible location for a toponym is given a score based on (a) its proximity to other toponyms around it, (b) its proximity to the centroid for the document, and (c) its relative importance—e.g. all other things being equal, nations get a higher score than cities. Also at this stage, the system discards as probable false positives places that lack an explicit disambiguator, that receive a low importance score, and that are far away from the local and document centroids. If not thus eliminated, the candidate toponym identification with the highest score is declared the winner. Once the work of the disambiguation system is done, the resulting toponyms are loaded into a relational database for access by the runtime digital library system.

4 Evaluation

We evaluate the performance of the disambiguation system using standard precision and recall methods. Qualitatively, the system performs quite well at producing aggregate maps of the places mentioned in an entire document or corpus or in finding mentions of a particular place. For a more detailed look at the performance of the Perseus system on various texts, a human disambiguator worked through 20% of the output for a text from each of five representative corpora: ancient Greece, ancient Rome, the Bolles collection on the history and topography of London, and two Library of Congress collections on the settlement of California and the Upper Midwest. With a large gazetteer and conservative pruning rules, our system is biased towards more recall. In table 2, we show precision results for the system as a whole, which is what the end user actually experiences, and for the toponym disambiguation system independent of name categorization. We also show the F-measure for the whole system, a score that

combines recall (R) and precision (P) with the recall/precision weighting factor β^2 usually valued at 1:

$$F = \frac{(\beta^2 + 1)RP}{(\beta^2 R) + P} \tag{1}$$

Corpus	Precision	Perfect Categ.	Recall	F-measure
Greek	0.93	0.98	0.99	0.96
Roman	0.91	0.99	1.00	0.95
London	0.86	0.92	0.96	0.91
California	0.83	0.92	0.96	0.89
Upper Midwest	0.74	0.89	0.89	0.81

Table 2. Performance on five representative texts

From these figures, one can see that although our simple heuristic categorization algorithm was less adequate for certain tasks, the toponym disambiguator itself performed quite well. The evaluation of toponym disambiguation is, if anything, conservative since eliminating extraneous points from the local and document context should reduce the skew in the calculated centroids. Note also that the categorization performed better on the Greek and Roman history texts than on texts on the history of London, California, or the Upper Midwest. This reflects the degree to which toponyms are ambiguous with other names or non-names in the text (see table 1 above). This evaluation also turned up another linguistic issue: all of the mistaken toponym identifications in the Roman text—Caesar's *Gallic War*—were for the "Germans" whom Caesar is fighting. The ethnonym "German" is in the gazetteer in the record for Germany, but its plural is not. We could fill this deficiency by stemming the input, but proper names are not generally inflected in English, so on the whole stemming would do more harm than good. We can easily add these inflected geographic names to the gazetteer by hand. In general, the large gazetteer of over a million names probably depresses precision more than any other factor. In [5], for example, the gazetteer is an order of magnitude smaller (96,000) and precision reaches 96%.

5 Future Work

Although the Perseus toponym disambiguation system performs quite well, we will concentrate on improving the categorization system, especially for texts on North America. Many approaches to categorization require training data so that the system can learn context rules for the occurrence of various kinds of named entities. As noted above, important or canonical texts would in any case benefit from detailed hand markup, including name categorization and disambiguation,

and systems such as Alembic [3] have demonstrated computer-assisted methods to optimize the tagging task.

Restricting the available toponyms at any point in the text by time period would also improve the system's performance. In a heterogeneous digital library of historical information, however, a mix of temporal references may occur in close proximity. We could, however, use the preponderance of temporal references, as we now use the weighted map of spatial references, to rank the possibilities. "Ovid" in a discussion of Roman poetry is unlikely to refer to Ovid, Idaho. While the current system deduces this from the town's distance from Italy, where most of the other places in the document are located, the fact that the town was founded in the nineteenth century would also tend to exclude it from a document where most of the dates are in the first centuries B.C. and A.D.

As explained above, we characterize the document context or central "region of interest" of a document by the centroid of the most heavily referenced areas. There seems to be some lack of robustness in simply using the centroid, and we are experimenting with using a bounding rectangle or polygon to represent a document's region of interest.

Finally, we are compiling on a gazetteer of Greek and Latin toponyms to apply this work to the non-English texts in the Perseus digital library. Much of this information can be culled from digitized reference works such as the *Harper's Dictionary of Classical Antiquities* and Smith's *Dictionary of Greek and Roman Geography*. Although we note above that morphological stemming of the source text would be counterproductive for English, we will need to stem Greek and Latin texts with our existing tools in order to perform well with these highly inflected languages.

References

1. Association for Computational Linguistics. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, DC, April 1997.
2. Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: A high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* [1], pages 194–201.
3. David Day, John Aberdeen, Lynette Hirschman, Robyn Kozierok, Patricia Robinson, and Marc Vilain. Mixed-initiative development of language processing systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* [1], pages 348–355.
4. Linda L. Hill, James Frew, and Qi Zheng. Geographic names: The implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, 5(1), January 1999. See `http://www.dlib.org/dlib/january99/hill/01hill.html`.
5. Yasusi Kanada. A method of geographical name extraction from Japanese text for thematic geographical search. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, pages 46–54, Kansas City, Missouri, November 1999.
6. Ray R. Larson. Geographic information retrieval and spatial browsing. In Linda C. Smith and Myke Gluck, editors, *Geographic Information Systems and Libraries:*

Patrons, Maps, and Spatial Information, pages 81–123, April 1995. See `http://sherlock.berkeley.edu/geo_ir/PART1.html`.

7. David D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, pages 21–39. MIT Press, Cambridge, MA, 1996.
8. Andreas M. Olligschlaeger and Alexander G. Hauptmann. Multimodal information systems and GIS: The Informedia digital video library. In *Proceedings of the ESRI User Conference*, San Diego, California, July 1999.
9. David A. Smith, Jeffrey A. Rydberg-Cox, and Gregory R. Crane. The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25, 2000.
10. Nina Wacholder, Yael Ravin, and Misook Choi. Disambiguation of proper names in text. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* [1], pages 202–208.
11. Allison G. Woodruff and Christian Plaunt. GIPSY: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45(9):645–655, 1994.