

# Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language

Jochen L. Leidner  
Thomson Reuters Global Resources  
Catalyst Lab  
Neuhofstrasse 1  
CH-6340 Baar, Switzerland  
leidner@acm.org

Michael D. Lieberman  
University of Maryland  
Department of Computer Science  
Institute for Advanced Computer Studies  
College Park, MD 20742 USA  
codepoet@cs.umd.edu

## Introduction

Recognizing spatial language in text documents, termed *geoparsing*, is useful for many applications, because together with mapping such language to lat/long values, also known as *geocoding*, it enables the connection of the unstructured textual realm with the structured realm of *Geographic Information Systems (GIS)* [11]. For example, news stories about events happening in a particular location can be explored on a map for a spatial understanding of these events, as implemented by applications like the **European Media Monitor (EMM)** [18] and **NewsStand** [13, 20]. Web pages, blogs, encyclopedia articles, news stories, tweets and travel reports can all benefit from such interlinking with maps, which requires the recognition of spatial language. Note that geoparsing can be considered as a more specific application of the task of *Named Entity Recognition and Classification (NERC)*: NERC is concerned with automatically recognizing proper nouns of any kind, often meant to include monetary amounts, dates, and other types, while geoparsing is the NERC task applied to locations specifically. Geoparsing is also known by many names in the literature, including *geotagging*, *georecognition*, and *toponym recognition*, but for consistency, here we will refer only to geoparsing. In this paper, we provide an overview of the challenges related to geoparsing, several families of geoparsing methods, existing systems and data collections available for performing geoparsing, and open research questions related to geoparsing.

At the core of geoparsing’s difficulty are the many ambiguities present in natural language, including ambiguities related to toponyms. Indeed, these ambiguities form the focus of the next two notes by Overell and Buscaldi. The type of ambiguity most relevant for geoparsing is termed *geo/non-geo ambiguity*, i.e., that many non-locations share names with locations. For example, “Paris” can refer to “Paris, France”, but might also refer to the person “Paris Hilton”. Removing geo/non-geo ambiguity is crucial for successful geoparsing. Note that the amount of geo/non-geo ambiguity is affected by the level of granularity of toponyms considered in the procedure. A geoparsing process for country-level toponyms could be considered easier than one for city-level toponyms, due to the comparatively smaller number of country toponyms, which provides fewer opportunities for geo/non-geo ambiguity. Another challenge with geoparsing is how to deal with misspellings or errors in the documents themselves.

## Types of Geographic References

Many different types of entities can be considered geographic references. Perhaps the most obvious type of references are geopolitical entities, such as countries (e.g. “Spain”) and administrative divisions (“Brooklyn”; “Midlothian”), as well as populated places such as cities and towns (“Zürich”). Other types of region locations can include postal codes (“CB2 1RD”; “D-76887”) and municipal areas. At a smaller scale, various hyperlocal locations could be considered, such as streets (“Einstein Drive”), street addresses (“Sofienstraße 7”), street intersections (“51st St and Lexington Ave”), city centers (“downtown Seattle”), and buildings (“US Supreme Court”). In some applications, natural geographic features would be locations of interest, such as parks (“Hyde Park”), rivers (“Potomac”), and mountains (“Snowdon”). In contrast to formal place name gazetteers, which contain features named on maps, volunteered geographic information [7], which comprises an ever growing part of the Web, frequently also includes vernacular descriptions of locations, as well as references to imprecise areas (“east coast”; “southern France”; “downtown Washington”). Each of these location types affords different kinds of context that enable readers to understand that a location is being referred to. Some or all of these may be considered in the geoparsing task, depending on the application’s requirements or utility. Note that we can further distinguish between recognizing simple names referring to locations (e.g. “London”) and recognizing complex geographic phrases (e.g. “30 miles North of Austin”; “Washington, DC, USA”). The former refer to locations directly, whereas the latter can be analyzed compositionally, i.e., the meaning or reference of the expression is a function of the meaning of the parts and the way they are combined.

## Processing Textually-Encoded Spatial Data

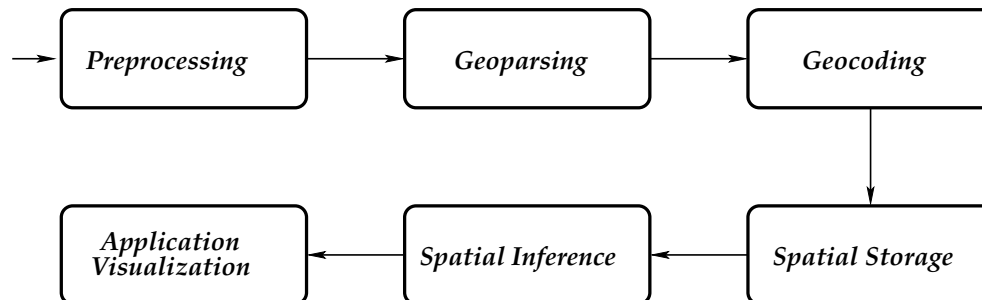


Figure 1: Reference model for processing textual geographic references.

Geoparsing is often integrated as one step of a multi-stage document processing pipeline. Figure 1 shows a typical processing pipeline. In a *preprocessing* step, the textual part of a document is separated from additional information such as metadata, formatting, layout, and the like. Depending on the nature of the document collection to be processed, this activity may range from a simple selection/projection of XML content elements containing digital textual prose, to automated layout analysis, Optical Character Recognition (OCR), automatic spelling correction, elimination of HTML markup in Web pages, or conversion from proprietary formats such as Microsoft Word. The *geoparsing* step comprises the detection of all ranges in the text that refer to place names (i.e. toponyms) or descriptions and, if more than one type are dealt with, the classification according to the type of geographic feature (e.g. anthropogenic artifact, natural place, human-inhabited dwelling). We describe geoparsing methods in the next section. After, a *geocoding* step (also

known as *toponym resolution*, a term coined in [11]) disambiguates toponyms with more than one candidate referent, and maps each toponym to its geographic representation, such as a polygon or the lat/long of the location's centroid. Once the geographic semantics of text documents are analyzed, a *spatial indexing* step may store the information in a data structure that permits fast retrieval using spatial operations, such as retrieval of all locations inside a bounding box or nearest to a given reference location, as an alternative or supplement to a textual inverted file index. Optionally, a *spatial inference* step may perform certain reasoning operations according to a *spatial logic*. For example, if a place *A* is south of *B*, and *B* is south of *C*, then *A* is also south of *C* (transitivity). Finally an *application* accesses and uses the extracted spatial knowledge to fulfill its purpose, e.g., crime mapping [12, 17]. A typical component of many geographic applications is the *visualization* of a single location or cluster of locations on a map. Note that the steps subsequent to geoparsing are aspects of the processing pipeline that are beyond the scope of this paper.

## Geoparsing Methods

There are three fundamental families of methods currently in use for the recognition of geographic language in text:

1. **Gazetteer Lookup Based.** The text is traversed either word by word or character by character, and searched for occurrences of a predefined set of toponyms. These toponyms are stored in a *gazetteer*, a database of place names and associated metadata [8]. Gazetteers are typically stored in tries (e.g. PATRICIA tries), hash tables [16], and/or in SQL databases on secondary storage. Note that special treatment of multi-word toponyms (e.g. "New York City") may be necessary, where a naive lookup-based approach can easily lead to inefficiencies. If the set of toponyms is not organized in a flat list but as a hierarchy, the term *ontology-based* geoparsing is used. Note that data quality is also an issue, as the incomplete and noisy nature of gazetteer data can lead to false positive and false negatives. Also, place names and administrative boundaries are constantly changing, and managing an update process for the gazetteer requires an integrated and automated workflow. Gazetteers often used in geoparsing systems include the NGA's GNS<sup>1</sup>, USGS's GNIS<sup>2</sup>, and GeoNames<sup>3</sup>.

*Example:* "Berlin" recognized by gazetteer lookup in GATE's ANNIE module [4].

2. **Rule Based.** A set of symbolic rules in a domain-specific language encodes a decision procedure that permits an interpreter to decide whether a word is a toponym or not. Typically, *Regular Expressions (REs)*, which correspond to *Finite State Automata (FSAs)*, or *Context-Free Grammars (CFGs)*, which correspond to *Push-Down Automata (PDAs)*, are used. *Definite Clause Grammars (DCGs)* are an extension of CFGs implemented by PROLOG. The former permit fast lookup and small storage at the same time [1], but permit only patterns with predefined maximum depth and limited forms of nesting [9], whereas the latter permit the formulation of less efficient, but more expressive grammars [2, 19].

---

<sup>1</sup><http://earth-info.nga.mil/gns/html>

<sup>2</sup><http://geonames.usgs.gov/domestic>

<sup>3</sup><http://www.geonames.org>

*Example:* city of ?  $\rightarrow$  <TOPO> in DIAL, the rule language of **OpenCalais**.  
*Example:* An automaton representing `[A-Z].+shire`, matching toponyms ending in “shire”.

3. **Machine Learning Based.** A sliding window is moved over the text, and at each position a set of properties known as *features* are computed. Features may comprise checks for particular strings, length computations, capitalization, and the like, and are frequently Boolean tests. Based on a *training corpus* containing *gold data*, feature configurations that are most highly correlated with toponyms are extracted. When run on a *test corpus* of unannotated text, the same features are computed and the most likely class for each word (i.e. toponym or non-toponym) is decided using, for instance, statistical inference [5, 6, 15].

*Example:* A feature  $F(W[*+1]=="Ave") == \text{true}$ ,  $P(LOC|F) = 0.9918$ . Here, for each word in the document, the subsequent word ( $W[*+1]$ ) is tested for equality with the string “Ave”, which comprises a Boolean test whether or not a sequence like “Madison Ave” has been found, in which case the value of this feature would be true, and otherwise false. Each feature is statistically correlated with the target outcome during training on a gold data corpus, which permits category prediction at runtime.

## Gold Data Collections

In order to extract geographic language from documents using supervised machine learning methods, such as those described above, and to aid in the evaluation of geoparsing methods, a *gold data corpus*, also known as a *reference corpus*, of documents is required, in which all occurrences of geographic names or phrases in the documents have been manually annotated. These annotations are compared with an automated system’s output to measure the system’s accuracy. A small number of such gold data corpora are available. The Message Understanding Conference (MUC) [3, 10], the ACE 2005 evaluation [14], and the CoNLL 2003 Shared Task [10, 21] are contests that provided participants with gold data corpora that are now commonly used to evaluate the quality of name taggers, including but not limited to location names. Note that despite the availability of these collections, there exists a general paucity of gold data corpora that are well-suited for use in geoparsing evaluation, due to several reasons. Gold data corpora are often derived from or make use of document collections that are under restrictive usage licenses, which hinders data sharing among researchers. Furthermore, creating gold data annotations is a large manual effort, especially if multiple human annotators are required. As a result, gold data corpora tend to be very limited in size, having at most a few hundreds of documents, which stands in stark contrast to, e.g., the ever-growing volume of data on the Web.

## Systems

There are several proprietary and open source systems available to recognize toponyms and other names, as well as geographic phrases. The **C&C** tagger [5] is a fast, open source, maximum entropy machine learning-based tagger implemented in C++. Apache’s **OpenNLP**<sup>4</sup> is a Java API for natural language processing components, including a module for named entity recognition.

---

<sup>4</sup><http://incubator.apache.org/opennlp>

Thomson Reuters's **OpenCalais**<sup>5</sup> and Yahoo!'s **Placemaker**<sup>6</sup> are popular, freely available Web services that tag names in documents, and provide additional metadata. The University of Sheffield's **GATE**<sup>7</sup> open source framework for text processing comprises APIs, a GUI and a component plugin architecture for natural language applications in Java under the LGPL license. Its distribution includes **ANNIE** [4], a rule-based name tagging component that also recognizes toponyms. **LingPipe**<sup>8</sup> is a commercial open source Java library that contains a name tagger based on statistical language models. Stanford University's open source machine-learning based tagger [6] uses linear chain Conditional Random Field (CRF) sequence models.

### Open Research Problems

1. **How can geoparsing be carried out in a semantically compositional way?** A phrase like "40 miles North of Kabul" describes a fuzzy area that might be computed from the respective meanings of "40 miles", "North" and the location related to Kabul.
2. **How can we recognize unknown toponyms?** In English, place names, as with other names, are capitalized, which is often the single most valuable feature in a machine learning-based system. However, methods must be devised for identifying and using alternative contextual clues, especially in languages where capitalization is not as valuable a cue.
3. **How can we perform joint geoparsing and geocoding?** It would be desirable to use geocoding (i.e. toponym resolution) knowledge to improve geoparsing (i.e. toponym recognition), and vice versa, by combining them in a single step. From a machine learning point of view, this corresponds to a joint optimization problem of two objective functions.
4. **How can we make geoparsing methods robust with respect to the geographic granularity of the input?** In other words, given that documents vary in geographic *scope* (i.e. distribution versus proximity of locations mentioned), how can we design methods that perform well on both global-scale and local-scale data?

### Conclusion

We have surveyed existing techniques for extracting geographic references, and have given an overview of the tools currently available to recognize spatial language in text documents. In doing so, we presented the techniques in context, using a proposed processing model that typical applications follow, and suggested several open problems for future research. Given the increasing importance of geography, and of understanding and leveraging the spatial aspects of data in Web and mobile applications, it is apparent that geoparsing methods will play a burgeoning role in comprehending this ever-expanding universe of data.

---

<sup>5</sup><http://www.opencalais.com>

<sup>6</sup><http://developer.yahoo.com/geo/placemaker>

<sup>7</sup><http://gate.ac.uk>

<sup>8</sup><http://alias-i.com/lingpipe>

## Acknowledgments

The second author was supported in part by the National Science Foundation under Grants IIS-10-18475, IIS-09-48548, IIS-08-12377, CCF-08-30618, and IIS-07-13501.

## References

- [1] K. R. Beesley and L. Karttunen. *Finite-State Morphology*. CSLI Publications, Chicago, 2003.
- [2] F. Bilhaut, T. Charnois, P. Enjalbert, and Y. Mathet. Geographic reference analysis for geographic document querying. In *Proceedings of the HLT/NAACL 2003 Workshop on Analysis of Geographic References*, pages 55–62, Edmonton, Canada, May 2003.
- [3] N. A. Chinchor. Overview of MUC-7/MET-2. In *MUC-7: Proceedings of the 7th Message Understanding Conference*, Fairfax, VA, 1998.
- [4] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *ACL'02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 168–175, Philadelphia, PA, July 2002.
- [5] J. Curran, S. Clark, and J. Bos. Linguistically motivated large-scale nlp with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic, June 2007.
- [6] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL'05: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370, Ann Arbor, MI, June 2005.
- [7] M. F. Goodchild. Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4):211–221, Aug. 2007.
- [8] L. L. Hill. *Georeferencing – The Geographic Associations of Information*. MIT Press, Cambridge, MA, 2006.
- [9] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, MA, 3rd edition, 2006.
- [10] J. L. Leidner. An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30(4):400–417, July 2006.
- [11] J. L. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, School of Informatics, University of Edinburgh, Edinburgh, Scotland, 2007.
- [12] J. L. Leidner, G. Sinclair, and B. Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT/NAACL 2003 Workshop on Analysis of Geographic References*, pages 31–38, Edmonton, Canada, May 2003.

- [13] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE'10: Proceedings of the 26th International Conference on Data Engineering*, pages 201–212, Long Beach, CA, Mar. 2010.
- [14] I. Mani, J. Hitzeman, J. Richer, and D. Harris. *ACE 2005 English SpatialML Annotations*. Linguistic Data Consortium, Philadelphia, PA, Jan. 2008. LDC Catalog Number LDC2008T03.
- [15] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.
- [16] D. Manov, A. Kiryakov, B. Popov, K. Bontcheva, D. Maynard, and H. Cunningham. Experiments with geographic knowledge for information extraction. In *Proceedings of the HLT/NAACL 2003 Workshop on Analysis of Geographic References*, pages 1–9, Edmonton, Canada, May 2003.
- [17] A. M. Olligschlaeger. *Spatial Analysis of Crime Using GIS-Based Data: Weighted Spatial Adaptive Filtering and Chaotic Cellular Forecasting with Applications to Street Level Drug Markets*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1997.
- [18] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A.-C. Forslund, and C. Best. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *LREC'06: Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 53–58, Genoa, Italy, May 2006.
- [19] F. Schilder, Y. Versley, and C. Habel. Extracting spatial information: grounding, classifying and linking spatial expressions. In *GIR'04: Proceedings of the SIGIR 2004 Workshop on Geographic Information Retrieval*, Sheffield, UK, July 2004.
- [20] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. NewsStand: A new view on news. In *GIS'08: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 144–153, Irvine, CA, Nov. 2008.
- [21] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *CoNLL'03: Proceedings of the 7th Conference on Natural Language Learning*, pages 142–147, Edmonton, Canada, 2003.