Understanding Location Descriptions in the LEI System

David N. Chin
Dept. of Info. & Computer Sciences
University of Hawaii
2565 The Mall
Honolulu, HI 96822
Chin@Hawaii.Edu

Matthew McGranaghan
Dept. of Geography
University of Hawaii
2424 Maile Way, Porteus Hall 445
Honolulu, HI 96822
matt@uhunix.uhcc.hawaii.edu

Tung-Tse Chen
Dept. of Info. & Computer Sciences
University of Hawaii
2565 The Mall
Honolulu, HI 96822

Abstract

Biological specimens have historically been labeled with English descriptions of the location of collection. To perform spatial, statistical, or historic studies, these descriptions must be converted into geodetic coordinates. A study of the sublanguage used in the descriptions shows much less frequent than typical usage of observer-relative relations such as "left of," but shows problems with name ambiguity, finding the referents of generic terms like "the stream," ordinal numbering of river forks and valley branches, object-oriented prepositions ("behind"), fuzzy boundaries (how close is "at," how far is still "north of"), etc. The LEI system implements a semi-automated understanding of such location descriptions. Components of LEI include a language analyzer, a geographical reasoner, an object-oriented geographic knowledge base derived from US Geological Survey digital maps with user input, and a graphical user interface. LEI parses prepositional phrases into spatial relations, converts these into areas, then computes polygon overlays to find the intersection, and returns the minimum bounding rectangle. The user is consulted on unknown words/phrases and ambiguous descriptions.

1 Introduction

Many biological specimens collected in the past[1] are labeled with only an English description of their location of collection.[2] To perform any statistical or spatial analysis of this historical data, these descriptions must be converted into geodetic coordinates (latitude-longitude or UTM), a time-consuming process that requires eye-straining poring over maps to search for each location. To automate this process requires understanding the natural language descriptions, reasoning about the spatial relations described by the natural language, and mapping these into a geographical object base to derive the collection coordinates.

2 Previous Research

Talmy [1983], Herskovits [1986], and André *et al.* [1986] among others have documented the many problems in interpreting and using spatial prepositions. For example, *in* and *on* have similar but different meanings: "in the car" means within the car, while "on the car," means on top of the car. However, "on the bus/plane," means within the bus or plane. Also, each preposition typically has several different meanings or usages. For example, one says "at home," but "at *the* bank," and the meaning of "the plane is at Honolulu airport," is within the area of Honolulu airport, but the meaning of "the dog is at the telephone pole" is not within the telephone pole, but near it. These context dependent usages make interpretation and application of spatial prepositions problematic.

Kuipers [1985], Davis [1986], Peuquet and Ci-Xiang [1987], and Frank [1991] have investigated qualitative and/or quantitative reasoning techniques for dealing with spatial relations. Freeman [1975] and Mark and Frank [1991] have identified commonly used spatial relations.

3 Characteristics of the Domain

Although general purpose natural language processing (NLP) is beyond current state-of-the-art, limited domains have frequently been amenable to NLP using specific techniques because the domains use a "sublanguage," a fairly restricted subset of a general natural language, which may have its own syntax and peculiarities. In this case, an analysis of one thousand three hundred and forty sample location descriptions from the Bishop Museum's Herbarium Pacificum collection (accumulated by about two hundred different collectors over a period of 160 years) shows a highly restricted use of language that is amenable to understanding using specialized techniques.

Because these descriptions are meant to be read later by a reader who is not at the site, they contain very few observer relative descriptions (e.g., behind). Also, there

[1] Current collectors can use hand-held satellite-based geopositioning systems to record collection coordinates.

[2] There are an estimated several hundred millions of such labeled specimens.

are limits to the scale of the descriptions. For example, out of a thousand descriptions that were located manually, about half of the descriptions were judged to be accurate to within 1/3 of a mile and 73% accurate to within 1 mile. At the other end, there were no descriptions with accuracy in the meter range and the best descriptions were only accurate to within several tens of meters.

A typical location description is: "Punaluu Valley; Castle Trail from Punaluu to Kaluanui Valley + stream, on E. side of Northern Koolaus." Associated accession information typically includes the date, collector name(s), genus, elevation, the museum's collection number, and the collector's accession number. This sublanguage is made up mostly of named objects (e.g., "Punaluu Valley" and "Castle Trail") and prepositional phrases (e.g., from Punaluu to Kaluanui Valley + stream, on E. side of Northern Koolaus). The relation of the collection location to the unmodified named objects is almost always "within," that is, the collection location is within the area designated by the named geographic object. The interpretation of the prepositions is somewhat simpler than the general case because the sublanguage deals only with a two dimensional cartographic space supplemented by elevation markings. This sublanguage is relatively simple syntactically, but there are still many problems for automatic interpretation of the sublanguage.

One of the most common problems in interpreting this sublanguage is the inconsistent use of names. For example, Waikane-Schofield Trail also appears as Schofield-Waikane Trail, Schofield-Waikane Ditch Trail, Schofield Trail, W-S Ditch Trail, Schofield Waikane Trail, and Waikane Ditch Trail. A mountain like Kaala may be referred to as Mt Kaala, Mt. Kaala, Kaala Mts., Kaala Mountain(s), Mount Kaala, Kaala Puu (*puu* is the Hawaiian word for mountain), Kaala Summit, or Kaala Range. Another problem is that names are often not unique. For example, Manoa is the name for Manoa Falls, Manoa Valley, Manoa Valley Park, Manoa Triangle Park, Manoa Stream, Manoa Elementary School, Manoa Japanese Language School, Manoa Tunnel, and Manoa Falls. When Manoa appears by itself, which Manoa is meant is usually clear from the context. In many cases, the same name is even used for the same type of object (e.g., many cities have Elm Streets and Main Streets). Luckily, similar objects with shared names tend to be geographically separated, otherwise confusion would result. Another very frequent problem is missing names. Collectors often use generic terms like *stream* and *gulch* to refer to landmarks that do have names. As before, the context is usually enough to find the correct object reference even without knowing the name. This heavy reliance on context for disambiguation is also a frequent problem for general purpose NLP systems.

A difficult problem is the interpretation of ordinal numberings, which are used to differentiate forks of streams and branches of valleys. For example, "3rd branch S of S fork of Kahanaiki Stream," refers to the third branch after the main South fork of Kahanaiki Stream counting from the head of the stream. Unfortunately, this description could also refer to the 3rd branch following the path of the collector going up the stream. Similarly, "Honolulu Valley, 4th small gulch," refers to 4th small gulch counting from the open head of the valley, although this could easily be interpreted as the 4th small gulch along some trail that might enter the valley from some pass over the mountains at the tail end of the valley. Another problem is the occasional use of land coverage types such as "middle Metrosideros forest," "wooded gulch," and "Fern Forest." Not only do most geographical databases lack land coverage information, but such information changes frequently over time. Also, descriptions sometimes refer to rainfall frequency, sun exposure or other ephemeral attributes of the area: "in dryish forest," "wet valley," "deep shade in wet gulch," and "shady hillside."

Even after converting the location descriptions into the appropriate spatial relations, there are still many problems in the correct interpretation of the relations. For example, "along a stream" does not mean that the collection site was in the stream, but within some distance of the stream. The problem is what exactly is the value of that distance. Even cardinal directions like "north of" are fuzzy concepts. The region north of a point can be bounded by two vectors pointing NE and NW (the triangular model), but this model fails when computing north of an object that is elongated in the E-W direction. Some spatial relations like "in front of," "behind," and "beyond" are relative to the observer's direction. Although these are not very frequent (only 25 cases in the 1340 sample descriptions), they still appear. Other spatial relations like "above" are dependent on understanding the slope of elevation around the object.

To solve some of these problems, we have developed and implemented the LEI system to partially automate interpretation of this sublanguage. LEI is described in the following section.

4 The LEI System

4.1 Organization

The LEI[3] (Location and Elevation Interpreter) system is an implementation of our algorithms for interpreting the sublanguage of location description labels for biological specimens. LEI is composed of four main components: the language analyzer PPI, the geographical reasoner GR, the user interface LEIview, and the geographic knowledge base GKB. The geographic knowledge base contains an object-oriented description of geographical objects such as valleys, streams, and waterfalls with their associated locations and names. The user interface displays maps and allows users to add or modify object locations. The language analyzer parses the English location description and produces a collection of spatial relations that relate the actual collection point to geographical objects. It uses knowledge of geographical objects

[3] *Lei* is also the Hawaiian word for "garland," typically made out of flowers, leaves, or feathers.

and their associated names from the geographic knowledge base. The geographical reasoner translates spatial relations from the language analyzer into polygons and performs polygon intersection calculations to obtain the area specified by the spatial relations. Each component is described in detail below.

4.2 GKB, the Geographic Knowledge Base

LEI uses three U.S. Geological Survey (USGS) digital cartographic databases as the starting point for GKB, the Geographic Knowledge Base. These include the DLG (Digital Line Graph), GNIS (Geographic Name Information System), and DEM (Digital Elevation Model). Unfortunately, these databases are not object-oriented, that is, they do *not* link the names in GNIS to the object locations in DLG. The GNIS database contains only names, USGS quadrangles, a feature class, and the coordinates of the name as it appears on a USGS map. The DLG database contains a hierarchical organization of points, line segments (composed of points), and areas (composed of line segments) along with a two-level type hierarchy composed of major and minor codes. Unfortunately, the 60 plus GNIS feature classes do not correspond to the over 200 DLG major and minor codes. The DEM database consists of a raster style set of elevation values.

To convert the three USGS databases into a useful object-oriented format requires human intervention to associate the names with the line and area objects (point objects are already correctly located by the GNIS coordinates). This process uses the LEIview component to display the appropriate section of the map to the user, then displays names one by one along with the geographically closest object of the same type as the name. The user can confirm the match, ask for alternatives, or modify the set of line segments or areas to give the actual location of the named object. In many cases, there are no corresponding objects, so the user must draw the outline of the area from scratch. This is required for all valleys and mountains since these are missing from the DLG specifications.

Determining the closest object of the same type requires matching the GNIS feature class to appropriate DLG major and minor codes. This is done using LEI's own type hierarchy that includes type classes corresponding to each GNIS feature class and to each DLG major and minor code along with many bridging type classes and higher level types. Given a GNIS feature class, LEI first indexes into LEI's type hierarchy to find the corresponding LEI type. If this type has a corresponding DLG code, then that is the most likely match. Less likely matches consist of any subtypes that might have associated DLG codes. If there are no DLG codes at this type level nor at subtype levels, then LEI searches up the hierarchy for supertypes that have associated DLG codes. Using this algorithm, the matching process manages to find the correct match most of the time, so the user's time is freed to worry about the many missing entries and errors in USGS databases (e.g., rivers that extend into what should be coastlines, disconnected lines, etc.).

4.3 LEIview, the User Interface

The LEIview component provides a graphical interface that allows users to view maps; zoom; scroll; rearrange, add, and delete layers of the map (including DLG lines, GNIS names, and DEM elevations); search for named objects; enter points, line segments, or areas for new objects; modify existing objects; and view the results of interpreting location descriptions (both the English description and the area resulting from processing are displayed). LEIview is written in C under X windows with Motif widgets.

LEIview is used to associate names with object locations in building the GKB geographic knowledge base. It is also used to display the results of interpreting location descriptions. When there are sections of the description that are not comprehensible to the PPI language analyzer, LEI sends the description to LEIview, which displays the description with the incomprehensible parts highlighted and displays the regions corresponding to the understood portions of the description. The user can tell LEI to ignore the unknown parts of the description, delay processing this description until later, send the description back for reprocessing, or add new geographic objects by entering new points, line segments, or areas and selecting the corresponding words in the description. Any new objects are stored in GKB and the correspondence between the words and the new object are stored in the PPI language knowledge base.

4.4 PPI, the Language Analyzer

The PPI (Prepositional Phrase Interpreter) component is responsible for parsing the natural language location descriptions and converting them into spatial relations. PPI uses the PAU[4] parser and understander [Chin, 1992] to interpret the English descriptions and convert them into spatial relations represented in the MERA (Meta Entity Relation Attribute) semantic-network-style knowledge representation language [Takeda *et al.*, 1992]. PAU is an all-paths, chart-based, unification parser that completely integrates syntactic and semantic processing.

Figure 1 shows the MERA graph for the grammar rule, PP ← Prep NP (i.e., a Prepositional-Phrase is a Preposition followed by Noun-Phrase), along with its semantic interpretation. The node PP-pat represents the left-hand-side of the rule, and the relations Pca (pattern component A) and Pcb (pattern component B) point to the components on the right-hand-side of the rule. The Ref relation denotes the meaning of the rule: a Geographic-object that has a Spatial-relation to some other Geographic-object. The Unify relation between the Prep and the Spatial-relation indicates that the meaning of the Prep should be unified with the relation, Spatial-relation. Likewise, the Unify relation between the NP and the lower Geographic-object indicates that the meaning of the NP should be unified with the lower Geographic-object. Figure 2 specifies the meaning of the Prep, "along" as an instance of

[4] *Pau* is also the Hawaiian word for "finished."

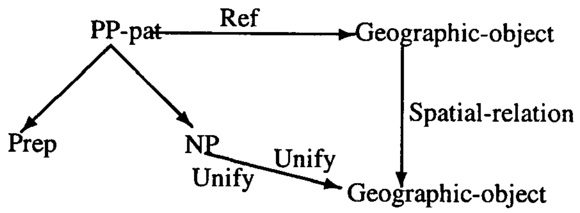

Figure 1: Rule for PP ← Prep NP.

the Spatial-relation *Near* relating a Geographic-object to a Linear-object (a subtype of Geographic-object).

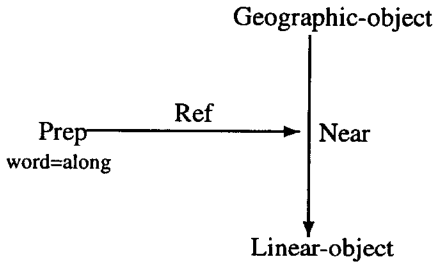

Figure 2: The meaning of the preposition "along."

The interpretation of the PP "along Ainapo" demonstrates how the integration of syntactic and semantic processing in PAU allows the early use of semantic constraints to reject semantically anomalous parses. "Along Ainapo" is ambiguous because Ainapo is both a trail and an area. However, since "along" only applies to linear objects such as trails, the Ainapo area interpretation is rejected by PAU. This happens when PAU is applying the grammar rule of Figure 1. When unifying the meaning of the Prep "along" (shown in Figure 2) with the Spatial-relation, the result is a Near relation. However, the sources and sinks of both relations must also be unified. This changes the lower Geographic-object into a Linear-object as seen in Figure 3, which shows the state of the "PP ← Prep NP" rule just before parsing "Ainapo." In PAU, both meanings of Ainapo are tried in parallel. The area meaning of Ainapo is rejected because an Area-object cannot unify with a Linear-object. This leaves only the Ainapo trail meaning to parse successfully.

Table 1 shows the spatial relations in PPI along with the corresponding prepositions.

4.5 GR, the Geographical Reasoner

The GR (Geographical Reasoner) component takes the output from the PPI component, which is a set of spatial relations, converts these into polygons, performs polygon overlay operations to find the common intersection of all the polygons, computes the center of the minimum bounding rectangle (mbr) of the polygon intersection, then returns the coordinates and centroid of the mbr. GR like PPI is written in Common LISP and converses with LEIview through UNIX sockets.

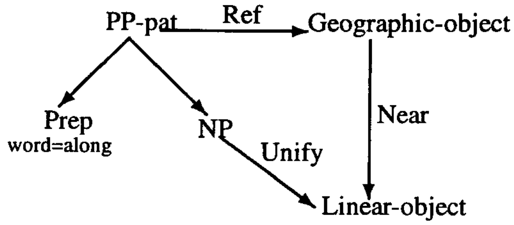

Figure 3: The PP-pat rule after parsing "along" and before parsing "Ainapo."

The first step is the most difficult since there are no generally accepted algorithms for converting spatial relations into areas. For the spatial relations based on cardinal directions such as East-of, Peuquet and Zhan (1987) give a complex algorithm for determining if one polygon is in a particular directional relationship with another. Their algorithm takes into account the shapes of the polygons (e.g., east of an elongated N-S polygon covers a different area than east of a small point polygon) and considers polygons that partially enclose or intertwine one another. Their algorithm is a refinement of the basic triangular model (in which North is the open-ended triangular region between two vectors pointing NE and NW), but it still does not give any limits concerning the distance between the polygons. Unfortunately there is no absolute distance that forms the edge of the region North-of some polygon. In a sense, the edge is given by the limit of human sight in that direction.

The algorithm currently used in GR for interpreting cardinal-direction relations around an geographic object starts by computing the minimum bounding rectangle (mbr) for the object. The area next to the mbr with the same size as the mbr is taken as the meaning of the spatial relation. Since the resultant area is the same size as the original object, this makes the meaning of cardinal directions relative to the size of the reference object, taking into account the fact that larger objects are visible from farther away. Cardinal directions relative to point objects are interpreted as a square, 500 meters on a side, lying in the appropriate direction.

The observer or object oriented relations such as Adjacent-to ("beside waterfall," "on Kona-Hilo Hwy"), Beyond ("1 1/2 mile beyond end of 20 Mile Road," "at back of Waihoi Valley"), Front-of (no examples in the sample data), Right-of ("right hand side of Kupu Kai Gap"), and Left-of ("Kulani Prison Road, toward Kulani Prison from intersection w/ Volcano Road, left road-

Spatial-relation	Prepositions
Adjacent-to	adjacent to, beside, next to, on
At-elevation	above, at, below, down, up
Between	between
Beyond	[in/at] back of, behind, beyond
East-of	east of
From	from
Front-of	before, [in] front of
Left-of	[to [the]] left [hand side] of
Near	adjacent to, along, around, at, by, near, outside [of]
North-of	north of
Right-of	[to [the]] right [hand side] of
South-of	south of
Toward	to, toward
West-of	west of
Within	among, at, in, inside, into, on, on top of

Table 1: Spatial Relations and Prepositions in PPI.

side") require understanding the orientation of the object or observer. Currently in GR, only object-oriented relations are processed. Given an object with a front, back, and sides (left and right), the corresponding relations are Front-of, Beyond, and Adjacent-to (Left-of and Right-of). These areas are calculated from the object's mbr in a similar fashion to the cardinal direction relations.

The At-elevation relation with respect to a given altitude requires computing the subregion of the common intersection area that is within 40 meters[5] of the given elevation. The At-elevation relation relative to an object (e.g., "above Schofield") requires computing the prevalent slope of the terrain around the object. GR takes a 200 meter square on the up/down side of the object.

Between, From, and Toward are handled by taking the mbr of the two objects, then computing the two corner points on each mbr that is furthest on either side from the line connecting the centers of the mbrs. These four points are then connected to form the area between the two objects.

The Near relation is converted into a buffer zone around the area. Currently GR uses a fixed distance of 200 meters for simplicity, however further study is needed to determine if this corresponds to most people's interpretation. There may be individual, cultural, or regional differences in interpretation. Also, the size of the buffer zone may depend on the size of the geographic object.

Currently GR does not handle references to terrain type, wetness, or typical sun exposure because this type of data is not available in the USGS databases. References to ordinal forks and branches are assumed to start from the head of the rivers or valleys. Generic terms are handled after processing all other spatial relations by exhaustively searching for any instances of the same type (or subtypes) that intersect with the intersection of the other known areas. In cases of multiple matches, the user is asked to help disambiguate through LEIview.

5 Future Directions

Because collectors often collect specimens on trips (either day hikes or multi-day camping expeditions), an analysis of the path of the collectors should yield valuable information about the location of collection. Specimens are typically labeled with the collection date and the collector's accession number, which provides the relative time of collection for specimens on that day. This information can be used to disambiguate location descriptions and to pinpoint vague locations. For example, in Hawaii, there are not only two Waihee Streams, but also a Waihee River. In a location description that mentions, "along Waihee Stream," there is ambiguity as to which of these three waterways is actually meant. In the current version of LEI, disambiguation is possible only if the description contains more information that specifies an area that intersects with only one of the three streams. By adding reasoning about time using accession dates and numbers and combining this with reasoning about paths, LEI could determine that it is unlikely that the collector stopped collecting specimens along one Waihee Stream, flew to another island to collect a specimen along another Waihee Stream, then flew back to continue collecting along the first Waihee Stream.

This type of reasoning can also help to pinpoint which part of Waihee Stream is meant by "along Waihee Stream." If LEI knows that the previous specimen was collected at point A and the following specimen was collected at point B, then LEI can make the reasonable assumption that this specimen was collected somewhere near the intersection of Waihee Stream and a region between points A and B. Using this type of reasoning, LEI can even make a reasonable guess about the collection

[5] USGS DEM data have a vertical resolution of one meter and a horizontal resolution of thirty meters.

location of specimens that have no location labels (provided only that they have an accession number and accession date given by the collector and the adjacent specimen numbers can be located). Adding such reasoning about time and paths would improve the accuracy of LEI's processing.

6 Conclusions

The LEI system demonstrates the feasibility of understanding the sublanguage used in location descriptions for biological specimens. Although this is an important and valuable task in and of itself, there is a much greater potential for application of the NLP and geographical reasoning techniques demonstrated in LEI to other areas such as natural language interfaces to general GISs (Geographic Information Systems). There is a need for validation of these techniques and a study is currently planned to compare the results of LEI with results obtained manually. Finally, the problems encountered in building LEI point to several new directions. First, the GKB component shows how object-oriented geographic databases should be organized in the future. Second, many new studies are required to determine the limits of fuzzy spatial relations like North-of, Front-of, and Near. Such studies should investigate task dependencies, context dependencies, individual variances, and cultural/regional variances. Such studies would lead to advances in understanding human cognition of spatial relations that would be directly applicable in GISs like LEI.

7 Acknowledgements

This research was sponsored in part by the National Science Foundation BSR&SES Grant No. BSR-9019041.

References

André, E., G. Bosch, G. Herzog, and T. Rist (1986). Coping with the Intrinsic and Deictic Uses of Spatial Prepositions. In Ph. Jorrand and V. Sgureg (Eds.), *Artificial Intelligence II, Proceedings of AIMSA-86,* pp. 375–382.

Chin, D. N. (1992). "PAU: Parsing and Understanding with Uniform Syntactic, Semantic, and Idiomatic Representations." In *Computational Intelligence,* **8**(3), pp. 456–476.

Davis, E. (1986). *Representing and Acquiring Geographic Knowledge.* Morgan Kaufman, Los Altos, CA.

Frank, A. (1991). Qualitative Reasoning about Cardinal Directions. In D. Mark and D. White (Eds), *Proceedings of Autocarto 10,* pp. 148–167.

Freeman, J. (1975). The Modeling of Spatial Relations. In *Computer Graphics and Image Processing,* **4**, pp. 156–171.

Herskovits, A. (1986). *Language and Spatial Cognition.* Cambridge University Press, Cambridge.

Kuipers, B. J. (1985). Modeling Human Knowledge of Routes: Partial Knowledge and Individual Variation. In the *Proceedings of the Third National Conference on Artificial Intelligence,* pp. 216–219.

Mark, D.M. and A.U. Frank (1991). (Eds.), *Cognitive and Linguistic Aspects of Geographic Space,* Klewer Academic Publishers, Boston.

Peuquet, D. and Z. Ci-Xiang (1987). An Algorithm to Determine the Directional Relationship between Arbitrarily-Shaped Polygons in the Plane. In *Pattern Recognition* **20**(1), pp. 65–74.

Takeda K., D. N. Chin, and I. Miyamoto (1992). MERA: Meta Language for Software Engineering. In the *Proceedings of the 4th International Conference on Software Engineering and Knowledge Engineering,* Capri, Italy, June, pp. 495–502.

Talmy, L. (1983). How Language Structures Space. In H. Pick and L. Acredolo, Eds., *Spatial Orientation: Theory, Research, and Application* Plenum Press, New York, pp. 225–282.