CONTEXT-FREE RECURSIVE-DESCENT PARSING OF LOCATION-DESCRIPTIVE TEXT

Matthew McGranaghan

University of Hawaii
Department of Geography
2424 Maile Way
Honolulu, HI 96822

ABSTRACT

Databases in which locations are specified in (near) natural language text, rather than as coordinates or topological relations pose difficulties for current GIS and automated mapping systems. Such databases may include metes and bounds descriptions of properties, or textual descriptions of locations where biological specimens were collected. The difficulty of converting textual location descriptions to coordinate data was highlighted by recent efforts to map the collection sites of specimens in the Bishop Museum's Herbarium Pacificum.

Human interpretation of both the text-based locational information in the Herbarium records and a number of topographic maps was required to derive mappable coordinates from textual descriptions. An automated system which could interpret the textual descriptions, and return coordinates, is an attractive alternative.

This paper reports an effort to create such a system by modifying a context-free recursive-descent text parser. A model of the grammar used in location descriptions is presented. The model recognizes phrases which denote specific features, generic features and terms relating them. The parser will recognize words as elements of these phrases. Meaning is ascribed to the location descriptions by relating them to items in standard geographic data sets (USGS DEMs, DLGs and GNIS).

INTRODUCTION

McGranaghan and Wester (1988) reported deriving geographic coordinates from textual descriptions of sites where herbarium specimens were collected. The process was slow and tedious. Because current GIS technology relies on analytic geometry and cartesian coordinate systems, and because much existing spatial data is not referenced to such coordinate systems, it seems likely that many scientists will be faced with a similar task in trying to use GIS technology.

To make the problem more concrete, each of the 80,000 specimens in the Bishop herbarium is associated with a unique label. This label contains a museum accession number, the plant's identification, the name of its collector, the method used to preserve the specimen, and a description of where the specimen was collected, but it does not contain systematic spatial coordinates for the site. The locational information includes the name of the island from which the plant was collected, a locality (the name of a physical feature or land division where it was collected), the approximate elevation at which it was collected, and (usually) a more detailed narrative description of the collection site.

The narrative descriptions are constrained by several practical concerns. They tend to be terse, composed by field scientists, and able to fit on a few lines of a label form. They also tend to be nearly procedural, giving directions that one could take to reach the same site again. Information about the site which is relevant to plant habitat, such as ground cover, soil type and moisture, are often included in the description. Virtually none of the descriptions are written as complete, grammatically correct, sentences. All of the information from the labels has been entered (verbatim) into a database as part of an effort to automate herbarium management.

For the initial mapping project, interpreting these descriptions was the slowest part of the data conversion process (McGranaghan and Wester 1988). It required map reading skill, and judgement. The amount of detail provided in these descriptions varies. Consequently, the confidence in the derived map locations varied.

Converting the descriptions to mappable coordinates involved sorting the descriptions, interpreting them, plotting these points and then digitizing them. The location descriptions were sorted by island and locality, then the narratives were used to plot the sites on the correct USGS 7.5 minute series topographic maps. This was done manually, and usually involved visually scanning a number of possible maps for the name of an area or feature. It was possible that the name did not appear on any of the maps. Once the area had been found, the rest of the text was interpreted to fix the point with respect to, for instance, topography, elevation, ground cover, and what ever other information the collector had provided. All of the plotted points on each map sheet were digitized at once, and the table coordinates converted to latitude and longitude.

This paper describes an approach to automating this conversion process. The goal of this research is a computer program which can read text describing a location and produce the absolute position of the site where a specimen was collected and an estimate of the confidence associated with the position. To "understand" the textual description, the program must be able to parse the text, identify phrases and terms which locate the position with respect to the planet, as represented in standard USGS data sets.

NATURAL LANGUAGE UNDERSTANDING

Making computers understand natural language has occupied computer scientists for several decades. During this time, some general strategies have developed, much has been learned about the complexity of the task, and several programs capable of understanding simple English sentences about fairly restricted domains have been produced (Winston 1977).

Strategies for understanding natural language attempt to exploit regularities and constraints found in the language to break a sentence into meaningful units. This process is called parsing. The constraints which allow one to parse a sentence are related to both the meanings of words in the sentence, and to sentence structure.

Both the words and the structure contribute to the meaning of a natural language sentence. Some of the words guide in determining the sentence structure, while others identify what referents which are being related in the sentence. The sentence structure indicates how the things referred to by the words are related in the "meaning" of the sentence, and may guide expectations about where to find specific parts of the meaning. Sentence structure, or grammar, provides a great deal of information about the meaning of a sentence through the context it provides.

GRAMMAR AND PARSING

The Handbook of Artificial Intelligence (Barr and Feigenbaum 1981) provides an accessible introduction to formal languages and grammar. A grammar is a scheme for putting words together into the phrases and sentences allowed in a language. Grammars are generally defined as a tuple of elements and possible relations among them. Symbolically, a grammar can be represented as:

$$G(P,W,R,S)$$

Where the grammar (G) provides rules (R) relating a basic sentence (S) to a set of non-terminal phrases (P), which in turn are composed of members of a set of terminal units, words (W), available in the language. The intersection of P and W is the null set.

The rules are often represented as productions, in which the constituent parts of a non-terminal unit are indicated. The form of the rules can be used to classify the grammar. If the rules are such that a single non-terminal symbol is on the left-hand side of each production, the grammar is context-free. An example of such a grammar, drawn from The Handbook of Artificial Intelligence is:

```
   <SENTENCE> -> <NOUN PHRASE> <VERB PHRASE>
<NOUN PHRASE> -> <DETERMINER> <NOUN>
<NOUN PHRASE> -> <NOUN>
<VERB PHRASE> -> <VERB> <NOUN PHRASE>
 <DETERMINER> -> the
       <NOUN> -> boys
       <NOUN> -> apples
       <VERB> -> eat
```

This grammar could generate sentences such as: "boys eat apples", "the boys eat apples" or "the apples eat the boys". "Eat the apples" would not be a valid sentence in this grammar because there is no <NOUN PHRASE> preceding the <VERB PHRASE> (that is, this statement is only a <VERB PHRASE> and not a complete <SENTENCE>.

Algorithms exist for parsing sentences which can be characterized by such a grammar. NLP.C (Schildt 1987) is a simple parser written in C which can process slightly more complex sentences. Its rules are:

```
   <SENTENCE> -> <NOUN PHRASE> <VERB PHRASE>
<NOUN PHRASE> -> <NOUN>
<NOUN PHRASE> -> <DETERMINER> <NOUN>
<NOUN PHRASE> -> <DETERMINER> <ADJECTIVE> <NOUN>
<NOUN PHRASE> -> <PREPOSITION> <NOUN PHRASE>
<VERB PHRASE> -> <VERB> <NOUN PHRASE>
<VERB PHRASE> -> <VERB> <ADVERB> <NOUN PHRASE>
<VERB PHRASE> -> <VERB> <ADVERB>
<VERB PHRASE> -> <VERB>
```

```
         <VERB> -> { list }
         <NOUN> -> { list }
   <DETERMINER> -> { list }
       <ADVERB> -> { list }
    <ADJECTIVE> -> { list }
  <PREPOSITION> -> { list }
```

If the location descriptions followed these rules, NLP.C could parse them. By adding words to the data base in Schildt's NLP.C, it will parse, "the plant is in the valley." into a <NOUN PHRASE>, "the plant" and a <VERB PHRASE>, "is in the valley."

The NLP.C parser uses a recursive algorithm to parse an input sentence. Its routines find the parts of speech of words in the sentence, and use them to determin how to parse the sentence. The routines are mutually recursive, and the order in which they are called indicates the phrase structure of the sentence.

To determine the end of the <NOUN PHRASE> and the beginning of the <VERB PHRASE> NLP.C simply takes the first <NOUN> it encounters to be the end of the <NOUN PHRASE>. Similarly, the <VERB PHRASE> begins with a <VERB>, though it may end several ways.

PARSING LOCATION DESCRIPTIONS

The following examples of location description narratives are drawn from the herbarium database:

Niu

wet forest

Kaulani, on open hillside

In woods near base of pali directly back of Kaimi Farm, Koolau Mts.

Ko'olau Mts., along the Waikane-Schofield Trail

Ko'olau Range, Waikane-Schofield Trail, in woods along trail

South ridge of Kipapa Gulch, Waipio

Higher gulches

North Fork of Kipapa Gulch. Koolau Mts. Along stream and up the banks at elevation of 1100-1500 ft.

2nd Gulch E. of Pu'u Kaupakuhale, N.E. slope of Pu'u Ka'ala, moist bottom of gulch

Ridge North of Waimea Valley

The formal grammar in NLP.C does not adequately characterize these descriptions. The location descriptions are not proper sentences. Most do not contain a <VERB PHRASE>, and many contain several <NOUN PHRASE>s. In short, the plant location parser must deal with a somewhat less structured "sentence" than does NLP.C.

Still, there is some structure. The descriptions are composed of one or more "location-descriptive phrases". There is some regularity in the structure of these phrases. The order in which they appear is less regular than NLP.C would expect, and the phrase-types often are repeated in these "sentences". Inducing a grammar for the location descriptions requires identifying the forms of the structures used.

As an aside, the structure of these descriptions may reflect some feature of human spatial cognition. A location may be defined by intersecting constraints, to distinguish a location from all others. The organization of the constraints may not be important; rather, the meaning comes from the combination of them. There may be some advantage to listing the constraints from most general to most specific (closely related to procedural directions for finding the site). Details like "moist bottom of gulch" may only be useful if the location has already been limited to a particular gulch. However, even that pattern is not always used.

A model of this grammar must allow a description to be composed of one or more location description phrases. The location description phrases have a number of forms. They tend to be composed of nouns and modifiers. The nouns name either generic features (stream, ridge, gulch, etc.) or specific features such as Waikane-Schofield Trail, Kipapa Stream, or Kaimi Farm. The modifiers are usually prepositional phrases. A model of the grammar used in the descriptions might be represented as:

```
<LOCATION DESCRIPTION> -> <LOC_DES PHRASE>*
      <LOC_DES PHRASE> -> <NOUN PHRASE> | <NOUN> |
                          <PREPOSITIONAL PHRASE>
<PREPOSITIONAL PHRASE> -> <PREPOSITION> <NOUN PHRASE>
         <NOUN PHRASE> -> <DETERMINER> <NOUN> | <ADJECTIVE> <NOUN>
                          | <DETERMINER> <ADJECTIVE> <NOUN>
                <NOUN> -> <SPECIFIC FEATURE> | <GENERIC FEATURE>
    <SPECIFIC FEATURE> -> <UNKNOWN> <GENERIC FEATURE> | { gnis } |
                          <UNKNOWN>
          <DETERMINER> -> { list }
         <PREPOSITION> -> { list }
           <ADJECTIVE> -> { list }
     <GENERIC FEATURE> -> { list }
```

In this grammar, determiners, prepositions, and adjectives are considered closed-classes; each set contains a fairly small and fixed number of terms. As Leonard Talmy pointed out last June in Buffalo (Talmy 1988), such terms mark the grammatical structure of language. A parser can use them to track the structure of a sentence and, in turn, to identify the words that should refer to geographic features.

Another closed-class is being posited in the current version of the grammar. This is the "generic feature". Generic features are common landscape elements, such as "hill", "valley" or "stream". Two distinct sources for the members of this set were identified. The first was an exhaustive examination of the words in the herbarium database. The list produced this way included names for features of great local significance, such as "pali". Another, more standard, source of generic features is the set of "Feature Class Definitions" used in the USGS Geographic Names Information System (GNIS). This set of 63 terms for generic landscape features has the advantage of being documented and identical for the whole country. Terms with considerable local usage could be either added to this list or translated to the most appropriate term in the list.

Parser Strategy

In the NLP.C parser, the parts of speech of the words encountered by the parser signal which grammatical phrases the words belong to. Given the nature of the locational phrases, it seems that parsing these location descriptions amounts to recognizing sets of prepositional phrases, and interpretation will then be determining which data bases contain the nouns. An example of how the parser can recognize the parts of a prepositional phrase follows:

<PREPOSITIONAL PHRASE> -> <PREPOSITION> <NOUN PHRASE>

The <PREPOSITION> is the clue that a <PREPOSITIONAL PHRASE> is beginning. The function in the parser which recognizes <PREPOSITIONS> indicates that one has been encountered. A second <PREPOSITION>, when encountered, marks the close of this <PREPOSITIONAL PHRASE> and the beginning of another. Within a <PREPOSITIONAL PHRASE> there must be one or more words, which must be identified as parts of a <NOUN PHRASE>. These in turn, must be decomposed into some combination of determiners, adjectives, generic features and specific features. When the description has been broken down into its constituent parts, and each part's functions determined, the parsed description still needs to be interpreted.

Interpretation Strategy

Parsing the text is only part of the job. To assign meaning to the text, it must be interpreted. The information gained from parsing, will be used in the context provided by standard geographic databases (USGS Digital Elevation Models, GNIS, Digital Line Graphs and US SCS soil facet maps), to deduce the coordinates. Remember that in addition to the data derived through parsing, the Herbarium labels also provided a locality name and elevation data to use as a starting point in determining a location.

The parsed description produces a set of descriptive phrases. Each of these phrases can be thought of as a constraint on the described location. The role of each word in the description is known (or inferred) from its position a in phrase.

The prepositions indicate the spatial relations among the features identified in the descriptions. Containment and enclosure are indicated by "in" and "on". Position with respect to linear features might be indicated by "along". Proximity may be indicated by "near", "by" and others.

When a specific feature name can be found in the GNIS, coordinate determination becomes a "look-up" operation. Preliminary testing indicates that a high proportion of specific features named in the descriptions will be found in the GNIS database. This is a result of both the GNIS database and the collector's descriptions being derived from USGS topographic maps. The positions found this way may be modified by other parts of the description.

The generic features do not give a direct look-up key to geographic data sets but they may provide information to guide geographic pattern-matching search in several data sets. GNIS records generic feature names associated with mapped items. Even when a specific feature name is not available in the description, or does not match in GNIS, it is possible that the generic feature can be used to produce a list of possible sites of the right type. Together with other constraints this may prove sufficient to derive a location's coordinates.

Generic feature names might also indicate topographic configurations which might be recognizable in DEM or DLG data. For instance, a "hill" or a "draw" might be recognized as a particular configuration of elevations in a DEM. This type of pattern matching may not be exceedingly expensive if the rest of a description sufficiently limits the region to search. See O'Callaghan and Mark (1984), Band (1986) and Frank, Palmer and Robinson (1986) for discussions of techniques which might be employed and problems which must be overcome in this type of matching. Search in a DEM is further constrained by elevation data from the label data.

Adjectives and non-feature nouns also contain information that might be useful if other data sets are available. These might be especially useful if information about soil types, land use/land cover, and climate are available and spatially referenced.

The interpretation engine will need to resolve the set of constraints produced by the parser to a single location or set of possible locations. This involves spatial reasoning about the relations indicated by the description in light of information found in the standardized data sets. This is clearly a step beyond the parser.

FUTURE DIRECTIONS

The most pressing need in this project is to refine and generalize the grammar understood by the parser. In addition to improved utility of the parser, it is expected that this will aid understanding of how people conceive of, and describe, locations. Further evaluation of the value of conceptualizing specific and generic features as separate classes, and of considering generic features to be a closed set is needed.

A second objective is to make the parser more robust. One complication with joining diverse data sets is the need to match place names given spelling variations. The USGS data sets do not use diacritical marks in place or feature names. Considerable pride in the Hawaiian language, and a desire to maintain it, have resulted in many field scientists in Hawaii retaining the use of macrons, apostrophes (for glottal stops), and other diacritical marks. Diacritical marks are inconsistently used on the Herbarium labels and pose problems for word-matching software.

In the longer term, another goal is to develop a spatial reasoning system which can use knowledge from a wide range of domains, as does a human interpreter, it determining locations. Knowledge, such as the habitat normally associated with a species, or the time-space history of the collector, or even personal habits of collectors could be used.

REFERENCES

Band, L., 1986, "Topographic Partition of Watersheds with Digital Elevation Models", Water Resources Research, v. 22, n. 1, pp. 15-24.

Barr, A., and Feigenbaum, E. A., eds., 1981, The Handbook of Artificial Intelligence, vol. 1, William Kaufmann, Inc.

Frank, A., Palmer, B., and Robinson, V., 1986, "Formal Methods the for Accurate Definition of Some Fundamental Terms in Physical Geography", Proceedings: Second International Symposium on Spatial data Handling, July 5-10, Seattle, Washington, pp. 583-599.

McGranaghan, M., and Wester, L., 1988, "Prototyping an Herbarium Collection Mapping System", Technical Papers: 1988 ACSM-ASPRS Annual Convention: GIS, v.5, pp. 232-238.

O'Callaghan, J., Mark, D., 1986, "The Extraction of Drainage Networks from Digital Elevation data", Computer Vision, Graphics and Image Processing, v. 28, pp. 323-344.

Schildt, H., 1987, "Natural-Language Processing in C", Byte, December 1987, pp. 269-276.

Talmy, L., 1988, Presentation at a two day workshop, "Cognitive and Linguistic Aspects of Space", June 11-12, 1988, State University of New York at Buffalo.

Winston, P. H., 1977, Artificial Intelligence, Addison-Wesley.