

WWW 2007 - I3 Workshop Talk - volz@fzi.de

Towards ontology-based disambiguation of geographical identifiers

8-May-07

R. Volz, J. Kleb, W. Müller



Content licenced under Creative Commons

<http://creativecommons.org/licenses/by-nc-sa/2.0/de/>



FZI

Geographical identifiers are highly ambiguous

Example Scenario

■ Tripreport

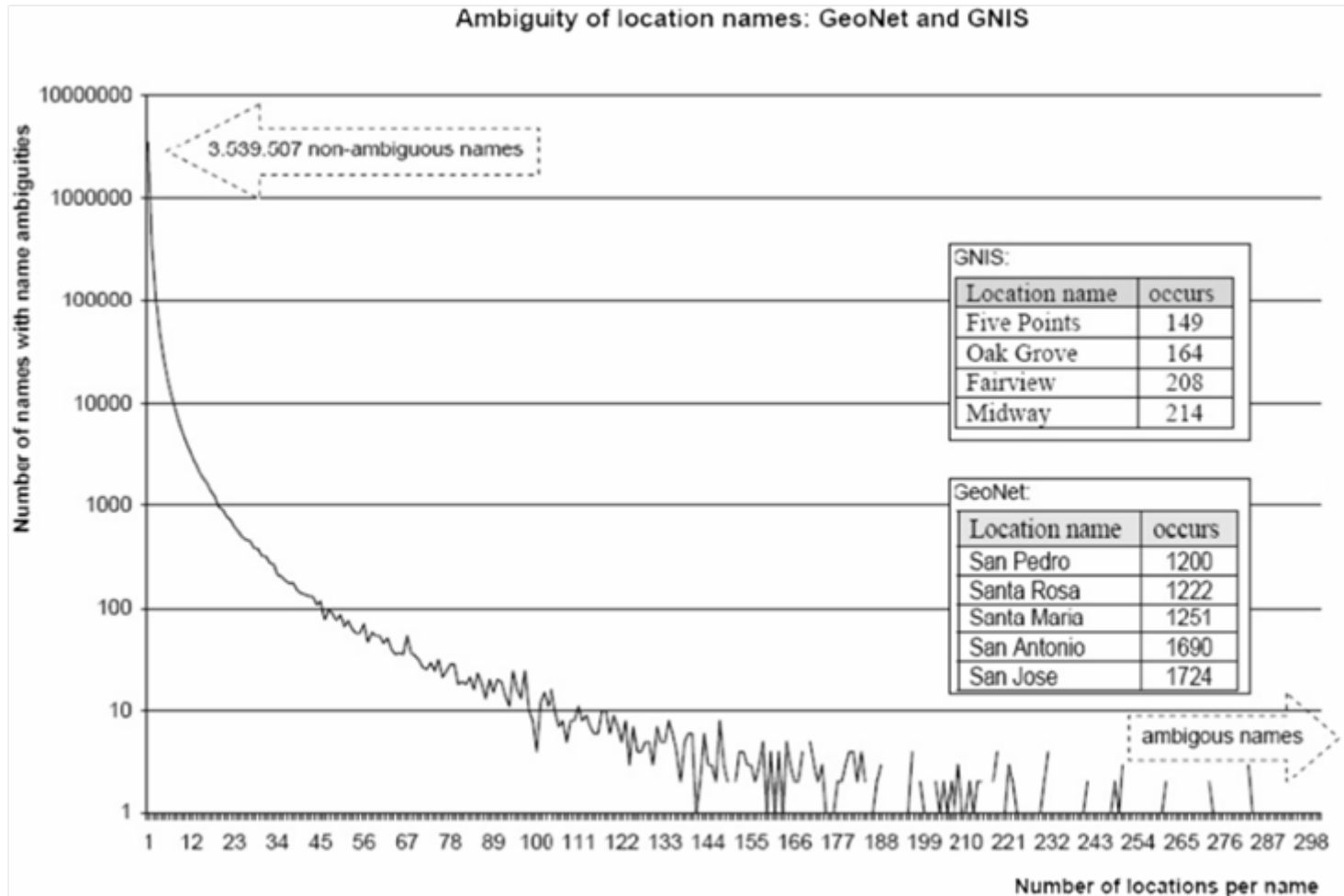
You will arrive at the Luis Muñoz Marín International Airport in San Juan where our guide will be waiting for you. From that moment on we will spend our time enjoying the splendid sites, food, and music of the small island of Puerto Rico. We will spend 5 nights in our base camp, Villas del Mar Hau Parador, and another night at the historic Spain's Consulate mansion, Hotel El Consulado, an elegant bed and breakfast at walking distance from casinos, excellent restaurants, and old San Juan. We will explore most of the sites from our base camp Villas del Mar Hau Beach Parador. During the trip, Lila will share her perspective on the arts, costumes, music, and politics of Puerto Rico.

Are
*geographic
entities*
in the text?

Which *identifier*
represents the
overall text
best ?

There are 1724 San Jose on this planet

Ambiguity Example



We distinguish three types of ambiguity

General Problem

- Types of ambiguity:
 - Multi-referent ambiguity
 - Different locations share the same name
 - Name variant ambiguity
 - Location is denoted by multiple names
 - geo / non-geo ambiguity
 - Word can refer to a geographic entity as well as to another entity type
 - E.g. a person name

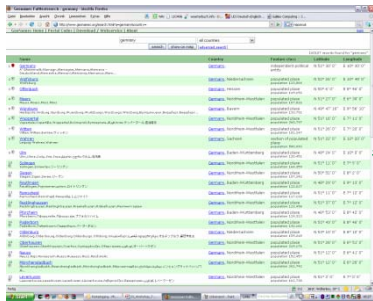
Two step approach:

- Identification of correct term within text
- Identification of correct geographic reference

We use two large geographic gazetteers as background knowledge

Geographic data sources

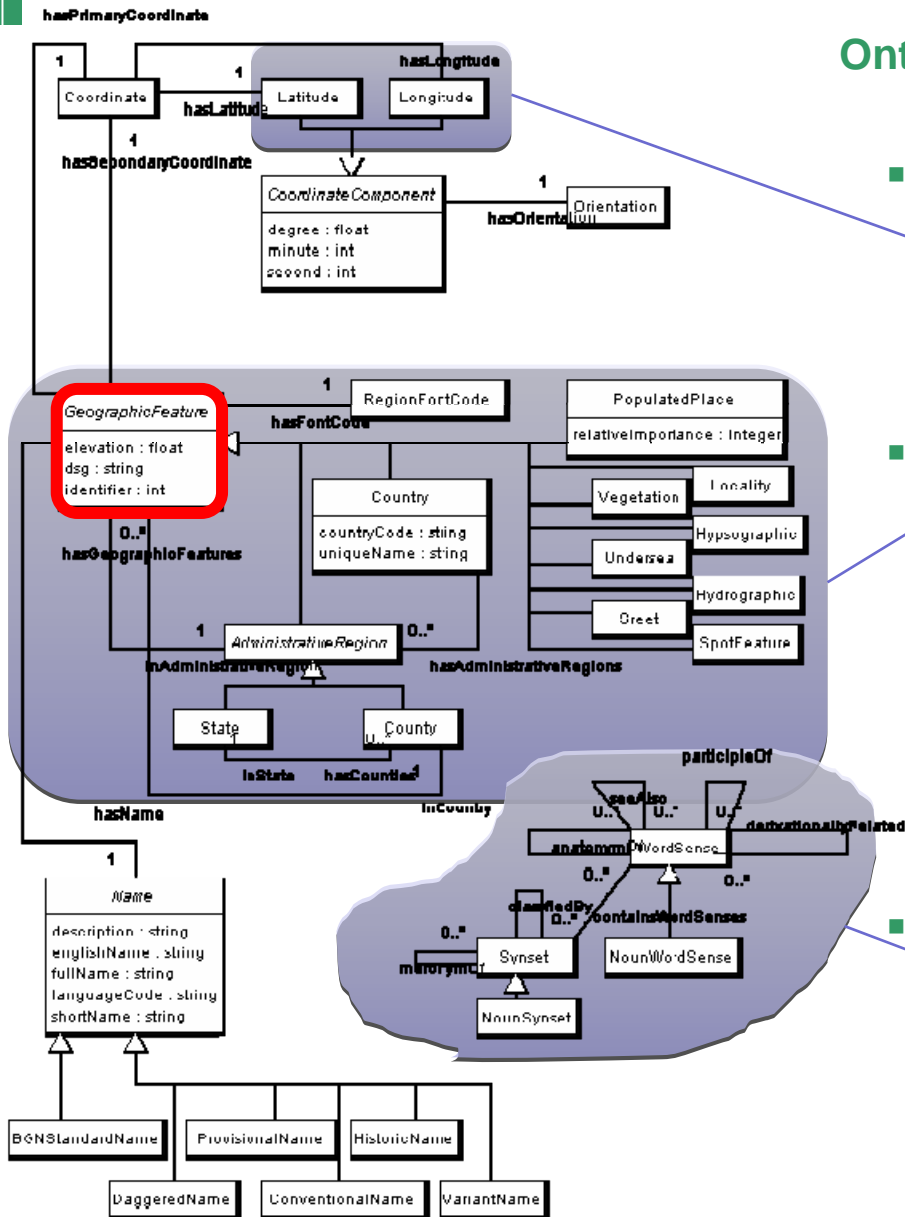
- GEONet Names Server (GNS)
 - 5.5 million names for roughly 4.0 million geographic features outside the USA
- Geographic Names Information System (GNIS)
 - contains information about almost 2 million features in the United States
- Next steps: More Geo Information
 - E.g. CIA World Fact Book
 - E.g. Parts of Wikipedia



A screenshot of a geographic data table, likely from a gazetteer. The table has several columns, including 'name', 'lat', 'lon', and 'feature'. The data is organized into a tree-like structure with expandable rows. The table contains various geographic features, such as 'GARDNER' and 'GARDNERVILLE', with their respective coordinates and feature types.

name	lat	lon	feature
GARDNER	37.500000	-122.500000	Populated place
GARDNERVILLE	37.500000	-122.500000	Populated place
GARDNERVILLE	37.500000	-122.500000	Populated place
GARDNERVILLE	37.500000	-122.500000	Populated place
GARDNERVILLE	37.500000	-122.500000	Populated place
GARDNERVILLE	37.500000	-122.500000	Populated place
GARDNERVILLE	37.500000	-122.500000	Populated place
GARDNERVILLE	37.500000	-122.500000	Populated place
GARDNERVILLE	37.500000	-122.500000	Populated place
GARDNERVILLE	37.500000	-122.500000	Populated place

An ontology for geography has been designed



Ontology

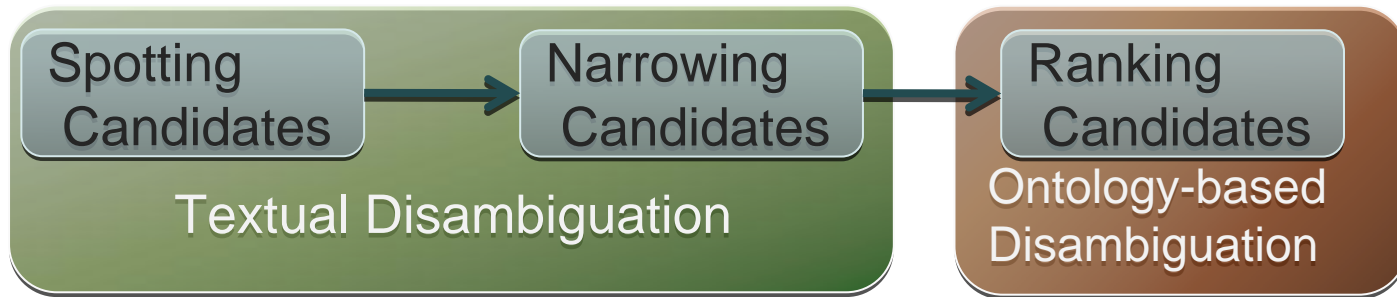
- Unique identifier
 - Latitude
 - Longitude

- Reflecting GONet & GNIS structure
 - Subclasses of “Geographic feature”
 - Names of classes
 - Names reference instances
 - Instances

- Parts from WordNet Ontology
 - Exception: no hyponyms of synset “city” and “country”

The disambiguation involves three steps in two disambiguation phases

Process



The ontology is used as lexicon in NLP processing

Spotting candidates

- Use of Natural Language Processing (NLP-Techniques)
 - Annotations based on gazetteer lists (without reduction of feature space)
 - Person names, organization names, stop words
 - Using the signs of concepts from knowledgebase
 - Annotations based on gazetteer (**with** reduction of feature space)
 - Candidate selection
 - Utilizing the instance lexicon of the ontology to obtain references

InsideTravel invites you to join Lila, a native of Puerto Rico, on this economical and culturally broadening trip to her homeland. The Taste of Puerto Rico trip takes you to the spirited capital of the island, San Juan; to its fortresses and lush rain forest; and to the warm waters of Villas del Mar Hau Beach Parador, "your little hidden, secret spot in paradise." On this trip this enticing island will reveal other treasures: the phosphorescent bay, Ponce, its main plaza with its historic museum and firehouse, and the splendid subterranean caves of Camuy. Swimming in the Caribbean or pool, or walking the beach at sunset invites you to "an adventurous and back-to-nature experience."



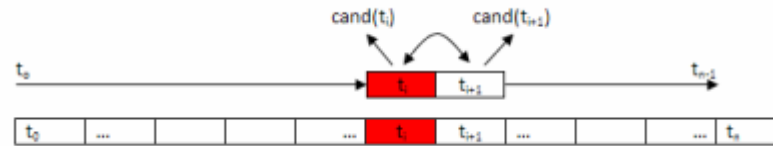
InsideTravel invites you to join Lila, a native of Puerto Rico, on this economical and culturally broadening trip to her homeland. The Taste of Puerto Rico trip takes you to the spirited capital of the island, San Juan; to its fortresses and lush rain forest; and to the warm waters of Villas del Mar Hau Beach Parador, "your little hidden, secret spot in paradise." On this trip this enticing island will reveal other treasures: the phosphorescent bay, Ponce, its main plaza with its historic museum and firehouse, and the splendid subterranean caves of Camuy. Swimming in the Caribbean or pool, or walking the beach at sunset invites you to "an adventurous and back-to-nature experience."

NLP analysis for secondary informationa
-Annotations
Results in matrix presentation

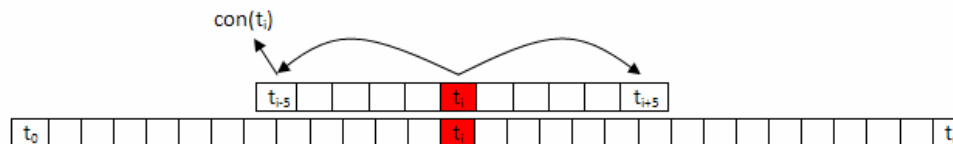
Candidates are narrowed using textual patterns

Narrowing candidates

- Use of textual patterns to retrieve further information
 - Relations between “candidates”
 - E.g. “Paris, France” or “Paris in France”



- Relations between concept labels and “candidates”
 - E.g. “the city of New York”



Candidates are determined by a ranking primarily based on weights attached to concepts

Approach

- Ranking the set of possible geographic references for a “candidate”
 - Use of weight values
 - Value associated to each ontology concept (Based on the character of human use)
 - Transitive propagation to subclasses
 - Negative weights for concepts without relation to geographic references
 - Differentiation within a concepts through use of attributes
 - Populated Place → population (population/1000=weight)
 - Each candidate is ranked according its concept relation
 - E.g. - Lancaster(CA) = 3129
 - - Lancaster(UK) = 3049
- Page focus uses geographic reference with highest weight value

Concept	Weight
Country	+3000
Populated Place	+3000
Administrative Region	+1000
Locality	+1000
Hydrographic	+10
Hypsographic	+10
Spot Feature	+10
Vegetation	+10
Street	+5
Undersea	-10
WordSense	-10000

Results need to be further improved

Evaluation Results

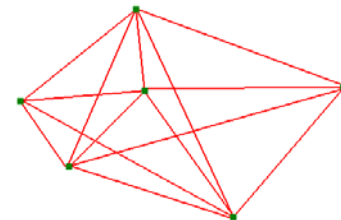
Test Run	I	II	III
Precision	40,1%	68,9%	67,9%
Recall	86,7%	86,7%	86,7%
Weights			
WordNet senses	0	-20000	-10000
Administrative Region	1000	1000	1000
Country	3000	3000	3000
Hypsographic	10	1	10
Locality	1000	500	1000
PopulatedPlace	3000	10000	3000
Road	5	10	5
Spotplace	1000	2500	1000
Hydrographic	10	10	10
Undersea	-10	1	-10
Vegetation	10	20	10

Recall not 100% due to misspellings and incorrect annotations by student annotators

We are currently improving the approach

Next Steps

- Algorithms
 - Additional patterns for textual disambiguation
 - Apply data mining algorithms (analysis of reference data)
 - Ontology based ranking:
 - Iterative Algorithm for dynamic weight value estimation
 - Learn algorithm (analysis of reference data)
 - Restricted candidate set for page focus
 - Use of spatial information (minimize distance between places)
- Ontology
 - Leverage additional sources :
 - “CIA World Fact Book”, Wikipedia, etc
- Change to reference corpus to enable comparison between approaches
- Generally, part of work on ranking search results



Ontology-based disambiguation of geographical identifiers is a promising approach

Summary

- Ontology used to map references to multiple ambiguous geographic feature candidates
- Ontology-based ranking among feature candidates recognized in text
- Weighting of concepts to prefer some concepts over other
- Use of attributes to prefer some instances of a concept over others (e.g. population size)
- Evaluation results promising but not fully satisfactory
- Improvements of the algorithm are work in progress