GeoWhiz: Toponym Resolution Using Common Categories

Marco D. Adelfio **Hanan Samet**
University of Maryland – {marco, hjs}@cs.umd.edu

Introduction

GeoWhiz converts a spreadsheet column or list of place names (*toponyms*) into a list of geographic references. Toponym ambiguity is resolved using the other place names within the column as context.

Challenge: Fully Utilize Context Clues

Surrounding toponyms provide context that should be used to improve geotagging results. Instead, existing systems often resolve each geographic reference independently or make limited use of available context.

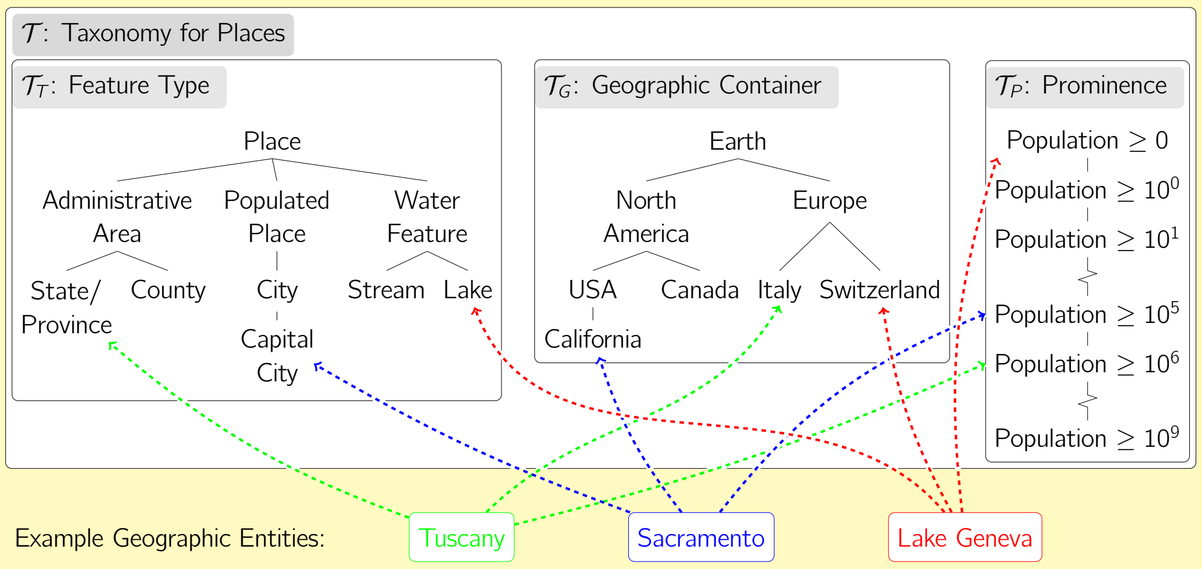
Approach

Given set of toponyms D:

1. Identify geographic entity **categories** that describe elements of D.
2. Measure how well categories describe D using **coverage** and **ambiguity**.
3. Apply **Bayesian classifier** to identify most likely category c_D.
4. Return **geographic interpretations** of toponyms that fall into c_D.

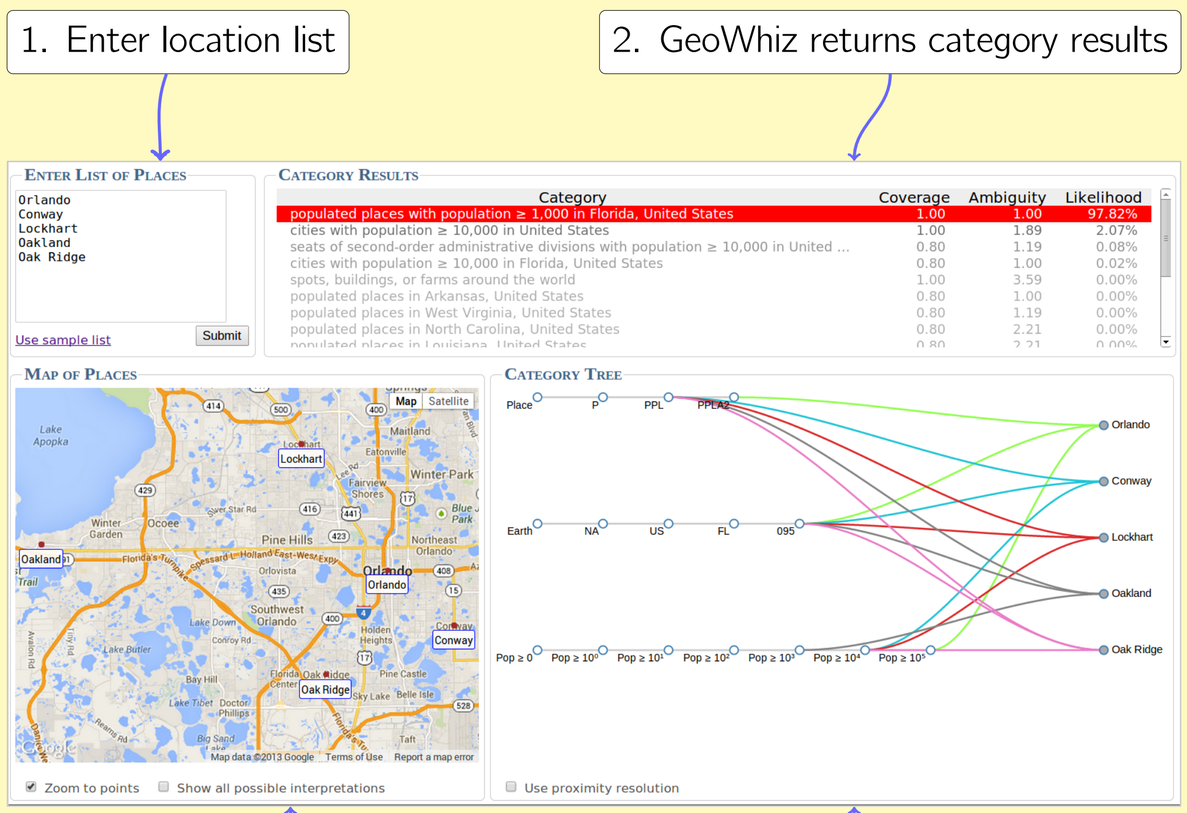
Place Taxonomy

- Geographic entities are described using taxonomy $\mathcal{T}$.
- Three components ($\mathcal{T}_T$, $\mathcal{T}_G$, $\mathcal{T}_P$) for describing each entity's feature type, geographic container, and prominence.
- $\mathcal{T}$ (simplified):



Application

- Browser-based DHTML implementation for resolving lists of toponyms

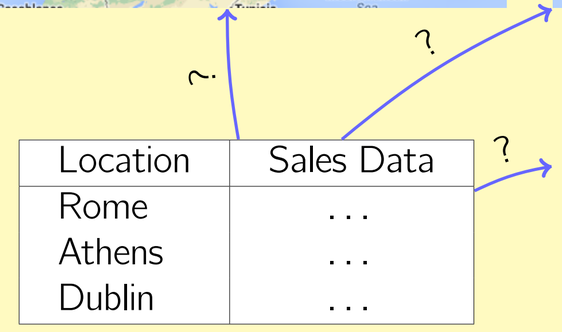


Example

Given a table column containing toponyms [Rome, Athens, Dublin], which geographic entity was most likely intended by each toponym?

European Capitals

Cities in Georgia, USA



Other Geographic Interpretations

- Administrative Regions in Europe
- Places in Pennsylvania, USA
- Places in Missouri, USA
- Places in New York, USA
- Places in Kentucky, USA
- Places in Ohio, USA

Common Categories

Geographic entity g *satisfies* (Sat) a category $c \in \mathcal{T}$ iff the attributes of g are descended from (or equal to) the attributes of c in each component of $\mathcal{T}$. Example:

- Rome, Italy is most precisely described by category:
  $\langle$Capital City, Region of Lazio (Italy), Population $\geq$ 1,000,000$\rangle$
- Athens, Greece is most precisely described by category:
  $\langle$Capital City, Region of Attica (Greece), Population $\geq$ 100,000$\rangle$
- Common categories of Rome, Italy and Athens, Greece include:
  $\langle$Capital City, Europe, Population $\geq$ 100,000$\rangle$
  $\langle$Populated Place, Europe, Population $\geq$ 100,000$\rangle$
  $\langle$Capital City, Earth, Population $\geq$ 10,000$\rangle$
  $\langle$Place, Earth, Population $\geq$ 0$\rangle$

Estimate Category Likelihood

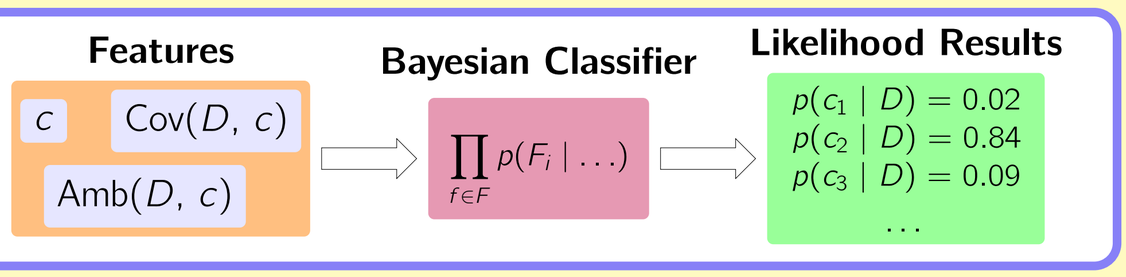
Input: toponym list D, category c.
Coverage: fraction of toponyms in D with interpretations satisfying c.

$$Cov(D, c) = |\{d \in D \mid \exists g \in Geo(d) : Sat(g, c)\}| \div |D|$$

Ambiguity: total number of combinations of interpretations that satisfy c, normalized over $|D|$. Represents category's specificity.

$$Amb(D, c) = \left(\prod_{d \in D} |\{g \mid g \in Geo(d), Sat(g, c)\}| \right)^{1/|D|}$$

Estimated Likelihood: output of Bayesian classifier.



References

[1] M. D. Adelfio and H. Samet. Structured Toponym Resolution Using Combined Hierarchical Place Categories. In *GIR 2013*.

[2] M. D. Adelfio and H. Samet. GeoWhiz: Toponym Resolution Using Common Categories. In *ACM SIGSPATIAL GIS 2013*.