Structured Toponym Resolution Using Combined Hierarchical Place Categories

Marco D. Adelfio   Hanan Samet

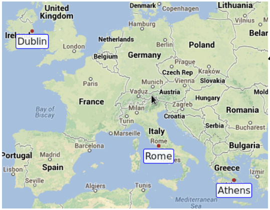
Department of Computer Science
Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA

GIR 2013

Toponym Resolution in Lists/Tables

- Many tables contain place names
  - spreadsheets, HTML tables, tables in PDF documents, etc.
- Often minimal external context for these tables
  - E.g., a company with a single "Paris" location does not need to specify which city named "Paris" is intended in intra-office spreadsheets
- Place lists also commonly found in plain-text comma groups

European Capitals



?

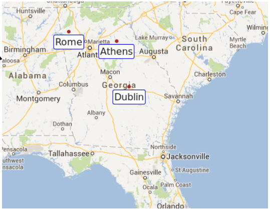
Location	Sales Data
Rome	...
Athens	...
Dublin	...

?

?

Other

Cities in Georgia, USA



Disambiguation Clues

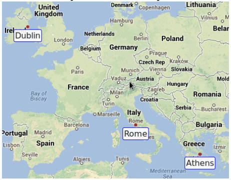
- Several attributes help determine geographic interpretations

Disambiguation Clues

- Several attributes help determine geographic interpretations
- Population, place type can be key indicators

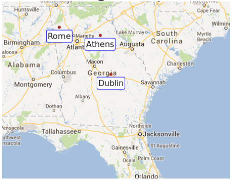
Toponyms
Rome
Athens
Dublin

European capitals



more likely than

Georgia cities

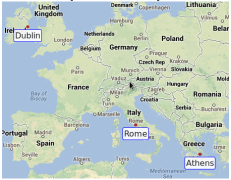


Disambiguation Clues

- Several attributes help determine geographic interpretations
- Population, place type can be key indicators

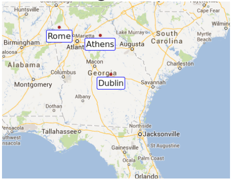
Toponyms
Rome Athens Dublin

European capitals



more likely than

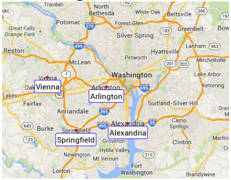
Georgia cities



- In other cases, a more constrained geographic container outweighs population

Toponyms
Alexandria Arlington Springfield Vienna

Virginia cities



more likely than

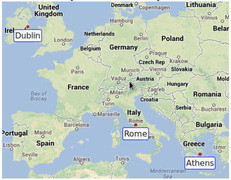
Larger cities around the world

Disambiguation Clues

- Several attributes help determine geographic interpretations
- Population, place type can be key indicators

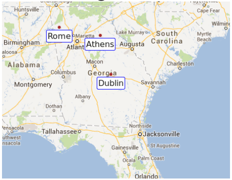
Toponyms
Rome Athens Dublin

European capitals



more likely than

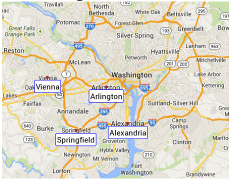
Georgia cities



- In other cases, a more constrained geographic container outweighs population

Toponyms
Alexandria Arlington Springfield Vienna

Virginia cities



more likely than

Larger cities around the world

- How to capture notion of “more likely” interpretations?

Table Geotagging Methods

- MapAList, BatchGeo, Google Fusion Tables, Wolfram Alpha
  - Provide different types of table geotagging services
  - Expect qualified place names
  - Geotag rows independently or based on simple geographic focus, so perform poorly when given single column lists of toponym
- Web-a-Where [Amitay et al. SIGIR'04] and other document geotagging methods reason about geographical hierarchies in order to identify geographic scope
  - We incorporate hierarchies for feature types and prominence
- STEWARD, NewsStand systems [Lieberman et al. GIS'07,'08 GIR'10] apply heuristic rules for either prominence, proximity, sibling place types
  - We use richer "category" descriptors, make decisions using machine learning

Outline of our approach

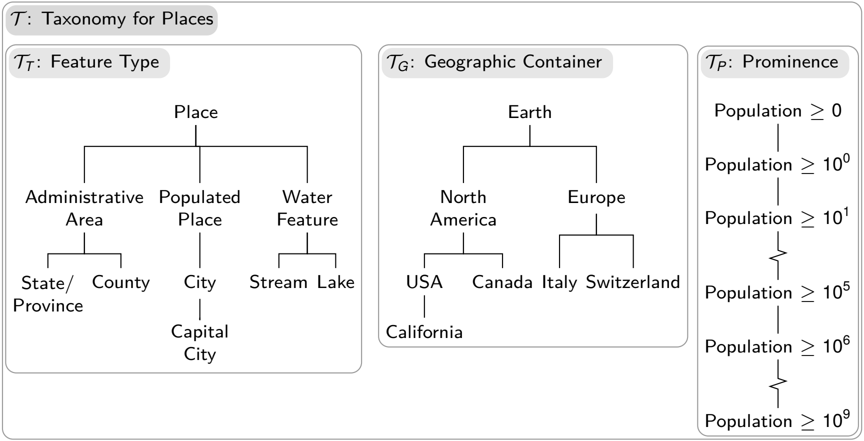
- Given set of toponyms D:
  1. Identify geographic **categories** that describe elements of D.
  2. Measure how well categories describe D using **coverage** and **ambiguity**.
  3. Apply **Bayesian classifier** to identify most likely category c_D.
  4. Return **geographic interpretations** of toponyms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

Combined Hierarchical Place Categories

- Attempt to identify coherent "category" for list
- Category components:
  - **Feature Type**. Ex: "capital cities," "parks," or "rivers."
  - **Geographic Container**. Ex: "in South Africa" or "in Shanghai, China."
  - **Prominence**. Ex: "with a population $\geq$ 10,000."
- Create strict containment hierarchy for each component using gazetteer
- Hierarchies constructed from raw GeoNames data
  - Feature Type hierarchy uses *feature class* attribute as first level, splits *feature code* attribute into two levels
  - Geographic Container hierarchy based on administrative regions plus continents
  - Prominence hierarchy nodes correspond to $log_{10}(pop)$ (multiple levels, no branches)

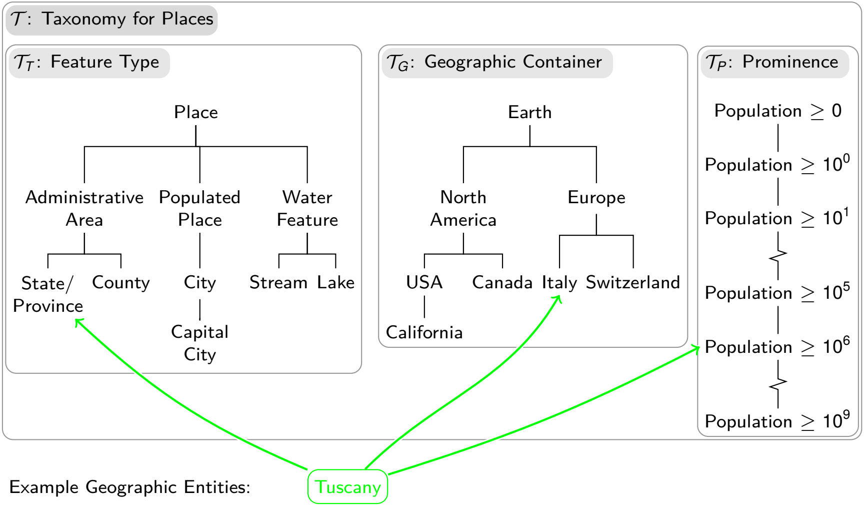
Combined Hierarchical Place Categories

- Hierarchies are combined to form Taxonomy $\mathcal{T}$.
- Simplified:



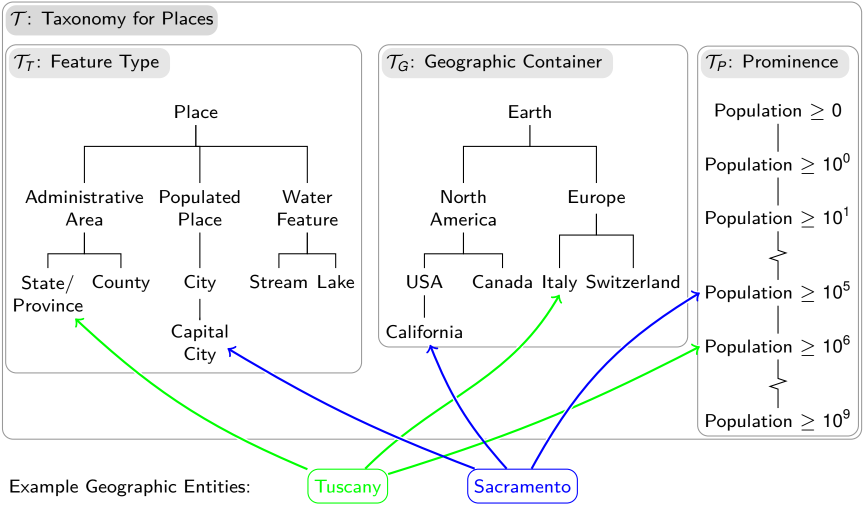
Combined Hierarchical Place Categories

- Hierarchies are combined to form Taxonomy $\mathcal{T}$.
- Simplified:



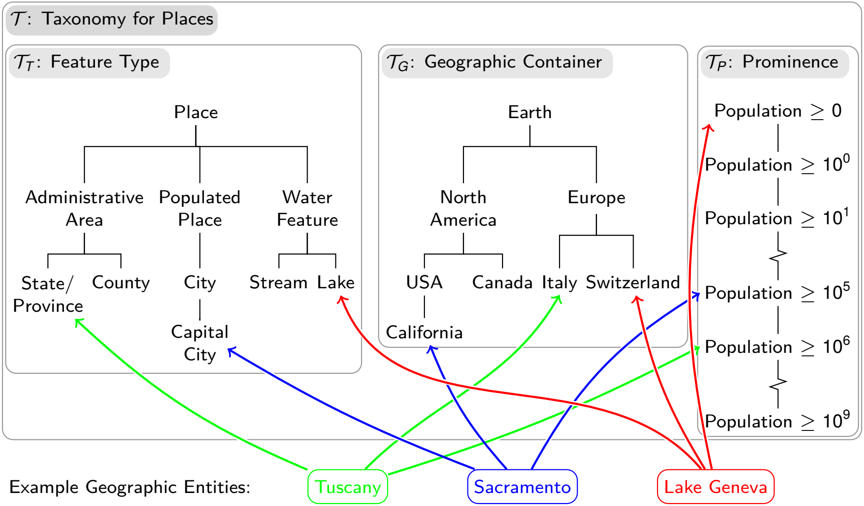
Combined Hierarchical Place Categories

- Hierarchies are combined to form Taxonomy $\mathcal{T}$.
- Simplified:



Combined Hierarchical Place Categories

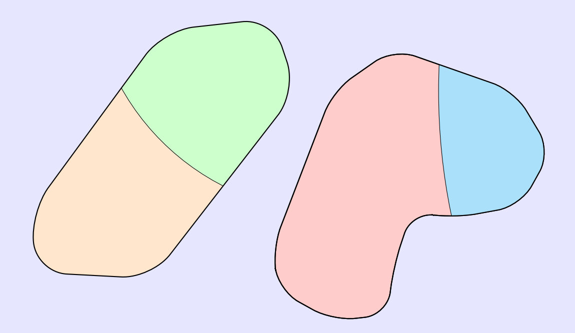
- Hierarchies are combined to form Taxonomy $\mathcal{T}$.
- Simplified:



Common Categories

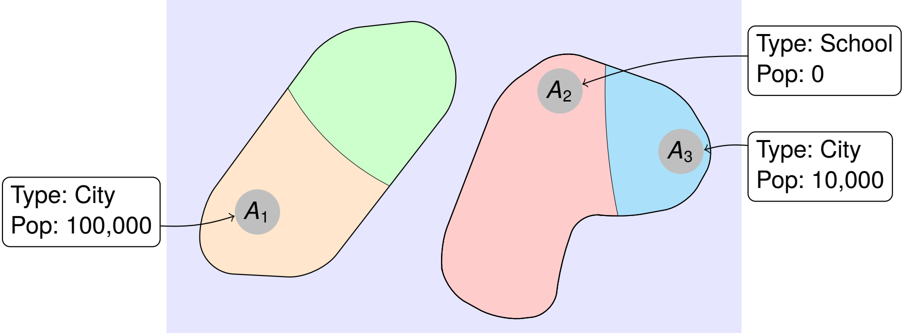
- Use 3-tuple to represent category: $\langle$Type, Container, Prominence$\rangle$
- Each geographic entity has one "specific" category and others that it "satisfies"
- Specific category determined by attributes in the gazetteer
  - Rome, Italy is most precisely described by category:
    $\langle$Capital 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$\rangle$
- Less specific categories also describe each entity
  - Geographic entity g **satisfies** category $c \in \mathcal{T}$ ($Sat(g, c)$) if and only if the nodes in the specific category of g are descendants of (or equal to) the nodes of c.
- All sets of entities satisfy at least one common category
  - Categories satisfied by *both* 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$\rangle$

Example: Geotagging toponyms [A, B]



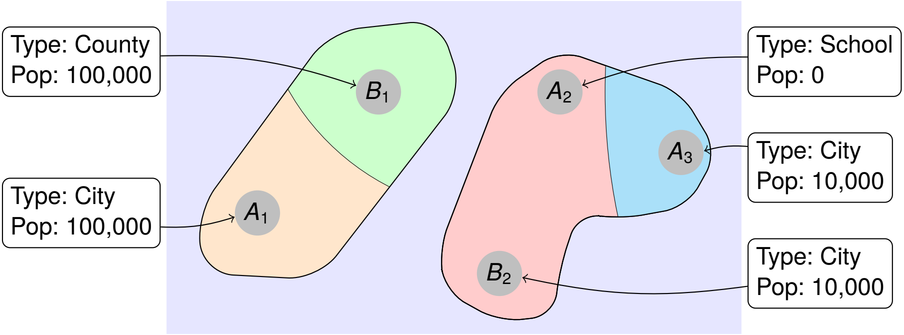
- Two continents: α and β
- Each contains two countries: α_1, α_2, β_1, β_2 (from left to right)
- Goal: Find interpretations for place names "*A*" and "*B*"

Example: Geotagging toponyms [A, B]



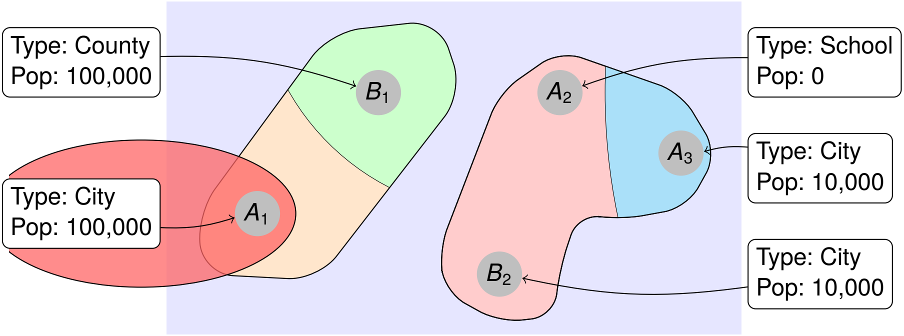
- Two continents: α and β
- Each contains two countries: α_1, α_2, β_1, β_2 (from left to right)
- Goal: Find interpretations for place names “*A*” and “*B*”

Example: Geotagging toponyms [A, B]



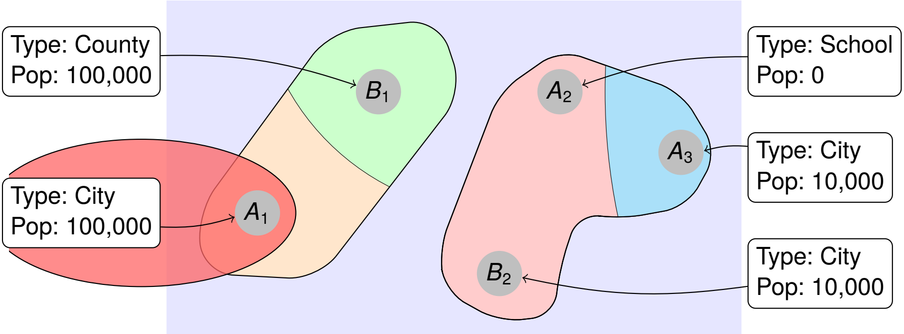
- Two continents: α and β
- Each contains two countries: α_1, α_2, β_1, β_2 (from left to right)
- Goal: Find interpretations for place names "*A*" and "*B*"

Example: Geotagging toponyms [A, B]



Place	Earth	Population $\geq$ 0
Populated Place	Continent-α	…
City	Country-α_1	Pop. $\geq$ 10,000
		Pop. $\geq$ 100,000

Example: Geotagging toponyms [A, B]



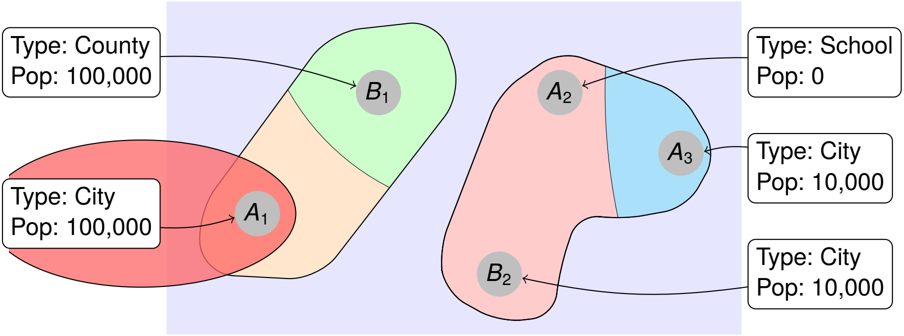
Place
|
Populated Place
|
City

Earth
|
Continent-α
|
Country-α_1

Population ≥ 0
...
Pop. $\geq$ 10,000
Pop. $\geq$ 100,000

Categories satisfied by A_1:
$\langle$CITY, COUNTRY-α_1, POPULATION $\geq$ 100,000$\rangle$

Example: Geotagging toponyms [A, B]



Place
|
Populated Place
|
City

Earth
|
Continent-α
|
Country-α_1

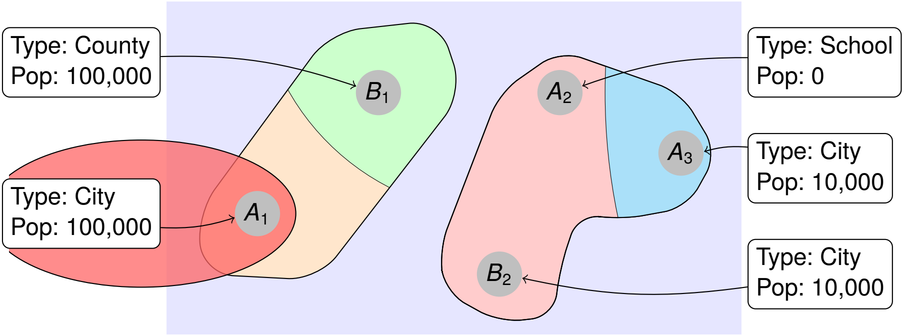
Population ≥ 0
...
Pop. $\geq$ 10,000
Pop. $\geq$ 100,000

Categories satisfied by A_1:

$\langle$City, Country-α_1, Population $\geq$ 100,000$\rangle$

$\langle$City, Country-α_1, Population $\geq$ 10,000$\rangle$

Example: Geotagging toponyms [A, B]



Place — Populated Place — City

Earth — Continent-α — Country-α_1

Population ≥ 0 — … — Pop. $\geq$ 10,000 — Pop. $\geq$ 100,000

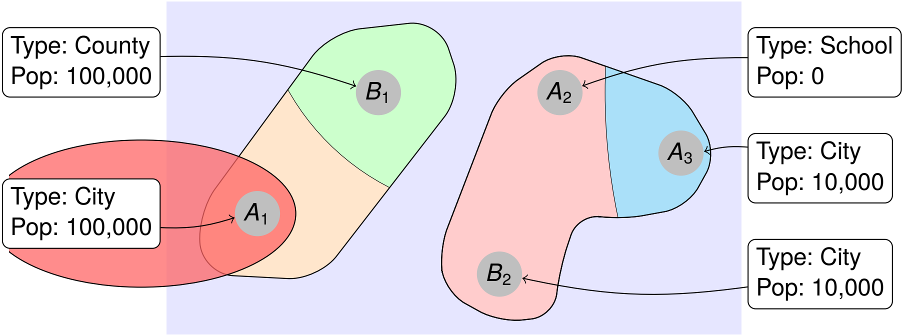
Categories satisfied by A_1:

$\langle$City, Country-α_1, Population $\geq$ 100,000$\rangle$

$\langle$City, Country-α_1, Population $\geq$ 10,000$\rangle$

$\langle$City, Continent-α, Population $\geq$ 100,000$\rangle$

Example: Geotagging toponyms [A, B]



Place
Populated Place
City

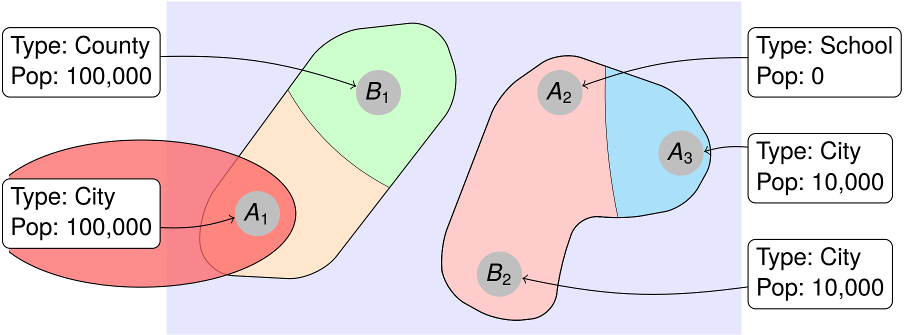
Earth
Continent-α
Country-α_1

Population ≥ 0
...
Pop. $\geq$ 10,000
Pop. $\geq$ 100,000

Categories satisfied by A_1:

⟨City, Country-α_1, Population $\geq$ 100,000⟩
⟨City, Country-α_1, Population $\geq$ 10,000⟩
⟨City, Continent-α, Population $\geq$ 100,000⟩
...
⟨Populated Place, Earth, Population $\geq$ 0⟩

Example: Geotagging toponyms [A, B]



Place
|
Populated Place
|
City

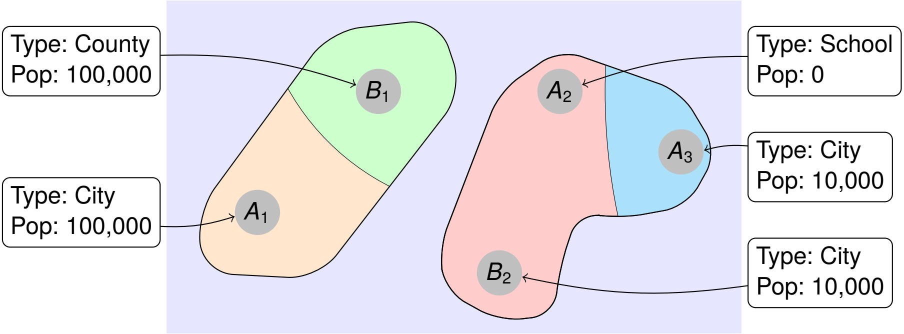
Earth
|
Continent-α
|
Country-α_1

Population ≥ 0
...
Pop. $\geq$ 10,000
Pop. $\geq$ 100,000

Categories satisfied by A_1:

⟨City, Country-α_1, Population $\geq$ 100,000⟩
⟨City, Country-α_1, Population $\geq$ 10,000⟩
⟨City, Continent-α, Population $\geq$ 100,000⟩
...
⟨Populated Place, Earth, Population $\geq$ 0⟩
⟨Place, Earth, Population $\geq$ 0⟩

Example: Geotagging toponyms [A, B]



Category	A	B
$\langle$Place, Earth, Pop $\geq 0\rangle$	A_1 A_2 A_3	B_1 B_2
$\langle$Place, Continent-β, Pop $\geq 0\rangle$	A_2 A_3	B_2
$\langle$County, Continent-α, Pop $\geq$ 100,000$\rangle$		B_1
$\langle$City, Continent-β, Pop $\geq$ 10,000$\rangle$	A_3	B_2
$\langle$Place, Continent-α, Pop $\geq$ 100,000$\rangle$	A_1	B_1
...		

Coverage and Ambiguity

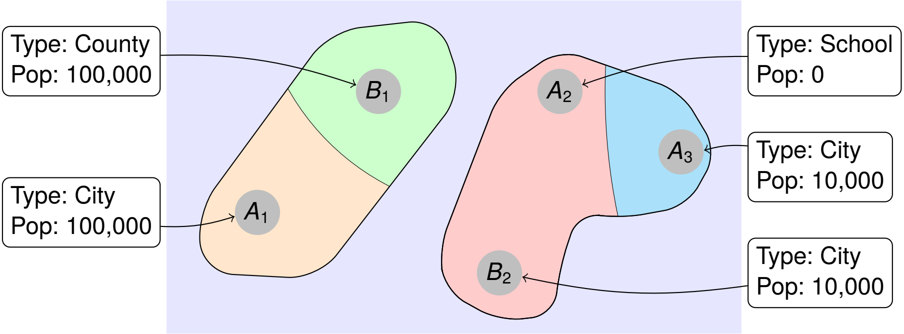
- We introduce two measures of how well a category *c* fits a toponym list *D*:
  1. **Coverage**
     - Fraction of toponyms with interpretations that satisfy the category

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

  2. **Ambiguity**
     - Number of interpretations per toponym that satisfy the category
     - Use product of interpretation counts to get total number of combinations, use geometric mean to normalize product
     - Lower value implies *specific* category
     - Higher value implies *vague* category

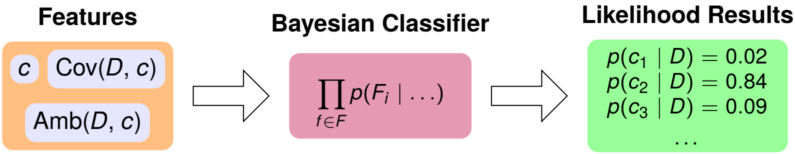
$$Amb(D, c) = \left(\prod_{d \in D} |\{g \mid g \in Geo(d), Sat(g, c)\}| \right)^{1/|D|}$$

Example: Geotagging toponyms [A, B]



Category	A	B	Coverage	Ambiguity
$\langle$PLACE, EARTH, POP $\geq$ 0$\rangle$	A_1 A_2 A_3	B_1 B_2	1.0	2.45
$\langle$PLACE, CONTINENT-β, POP $\geq$ 0$\rangle$	A_2 A_3	B_2	1.0	1.41
$\langle$COUNTY, CONTINENT-α, POP $\geq$ 100,000$\rangle$		B_1	0.5	1.0
$\langle$CITY, CONTINENT-β, POP $\geq$ 10,000$\rangle$	A_3	B_2	1.0	1.0
$\langle$PLACE, CONTINENT-α, POP $\geq$ 100,000$\rangle$	A_1	B_1	1.0	1.0

Calculating Category Likelihood



- Bayesian model computes category likelihood
- Model features are category nodes and coverage and ambiguity values
- Likelihood of features calculated independently – except coverage value
  - “Not-quite-Naive” Bayes
- Classifier setup
  - Train with 20 human categorized training samples (each sample has one true category and hundreds or thousands of false categories)
  - Use depth within $\mathcal{T}_G$ rather than node itself to avoid geographic bias
  - Discretize values of $Amb(D, c)$ to emphasize unambiguous categories (i.e., when $Amb(D, c) = 1.0$)
  - Model coverage values as truncated normal distribution based on mean and variance in training data

Location	Sales Data
Rome	...
Athens	...
Dublin	...

Category	Coverage	Ambiguity	Normalized Likelihood
country capitals with population $\geq$ 100,000 in Europe	1.00	1.00	70.13%
county seats with population $\geq$ 10,000 in Georgia, USA	1.00	1.00	15.07%
administrative regions with population $\geq$ 100,000 in Europe	1.00	1.26	13.88%
populated places with population $\geq$ 100 in Pennsylvania, USA	1.00	1.00	0.60%
populated places in Ohio, USA	1.00	2.15	0.05%
places in Missouri, USA	1.00	1.00	0.04%
farms in Limpopo, South Africa	1.00	2.47	0.04%
administrative regions with population $\geq$ 1,000,000 in Europe	0.67	1.41	0.03%
third-order administrative divisions with population $\geq$ 100,000 in Europe	0.67	1.00	0.03%
...	...	...	...

Dataset

- 20,000 spreadsheets and 20,000 HTML tables crawled from Web
- Tables preprocessed to discard non-relational tables [Adelfio PVLDB'13]
  - E.g., spreadsheets containing calendars and forms, or HTML layout tables
- Identify tables containing likely geographic columns
  - $\geq$ 3 strings matching GeoNames entities in first 100 values of a column
- Result: 12,861 geographic columns from 8,422 tables
- Categorized individual geographic columns using our method
- Place type characteristics
  - Most frequent column categories involved populated places and admin regions
  - Other common types: names of schools; airports; country, state/province, and region capitals; hospitals; rivers and streams
  - Root "place" type also common
    - American baseball team locations: Texas, Colorado, New York, Chicago (mix of states and cities)

Dataset (cont)

- Geographic container characteristics
  - 361 different geographic containers used as category component
  - Large geographic spread

39.7%	"Earth"
9.8%	continent level
41.6%	country level
7.4%	state/province level (admin level 1)
1.5%	county/region level (admin level 2+)

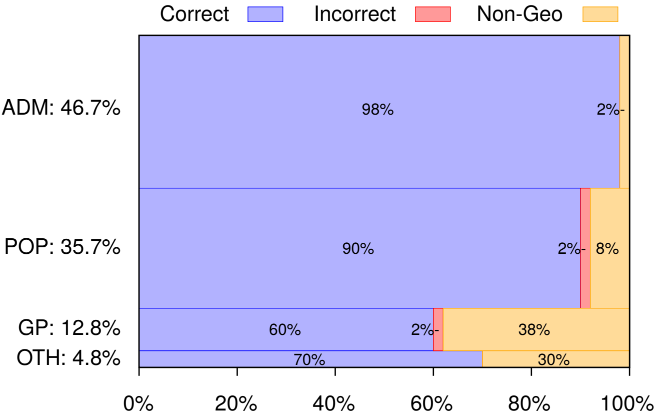
- Prominence characteristics
  - Large cities, states/provinces, and countries make up majority of place references
  - Non-populated places still referred to frequently, need to handle them

22.8%	high population ($\geq$ 1,000,000)
53.1%	medium population ($\geq$ 1,000 - 100,000)
8.5%	low population ($\geq$ 1 - 100)
15.6%	no population component ($\geq$ 0)

Experiment Setup

- Sampled 200 columns for category evaluation
- 50 from each group:
  - **ADM**: Administrative regions (or a descendant)
  - **POP**: Populated places (or a descendant)
  - **GP**: Generic places (i.e., the root of $\mathcal{T}_T$)
  - **OTH**: Other place types (e.g., schools, airports, etc.)
- For each selected column, manually specified if assigned category was:
  - Correct
  - Incorrect
  - Non-geo (mistakenly chosen as geographic column)

Experiment: Category Accuracy



- Bars scaled horizontally to reflect proportion of results within each group
- Bars scaled vertically to reflect the prevalence of each group within full dataset
- Overall accuracy rate (blue area) of 88.9%

Experiment: Toponym Resolution Accuracy

- Randomly select one toponym from each true geographic column found in previous experiment
- Use three methods for providing interpretation:
  - PROM considers only prominence of interpretations
  - 2D combines three classifiers that are each trained on only two of the hierarchies in $\mathcal{T}$
  - 3D uses full method (all three hierarchies)
- Manually evaluated each interpretation using full table context

Method	Accuracy
PROM	101/148 (0.682)
2D	130/148 (0.878)
3D	144/148 (0.973)

- Results show problem with prominence-only approach
- Demonstrate advantage of considering all three hierarchies together

Conclusions

- Introduced combined hierarchical place categories
- Devised coverage and ambiguity functions to measure how well category describes toponym list
- Used Bayesian model to select most likely categories and determine geographic interpretations
- Future Work
    - Augment prominence hierarchy using other gazetteers/databases
    - Improve method for disambiguating *within* categories
    - Examine usage for less coherent place lists (e.g., plain-text documents)
    - Handle multi-category columns

Acknowledgements

- Thanks to our sponsors:
    - Google Research
    - National Science Foundation

[The Numbers]

There are 33 places in the U.S. named Plymouth.

Call it Thanksgiving-induced nostalgia, but we sure are proud of our pilgrims—hence the nearly three dozen U.S. counties and towns named for one of our first settlements. The Plymouth that started it all, in Massachusetts, appeared in Money magazine's "Best Places to Live" in 2005. But it was not an easy place to inhabit in 1620, when colonists who set sail from England's port of Plymouth to escape the religious rule of King James I made landfall there. About half of them died during that first brutal winter. If only the pilgrims had come to shore in Plymouth, California, where fall temps hover around 62 degrees. You know what they say about hindsight...