Extracting Geographic Features from the Internet to Automatically Generate High Resolution Gazetteers

Presented to: Conference / Class Name
Location, Date

Your Name Here
GIS Research Laboratory
Department of Geography
University of Southern California
http://uscgislab.net

Background Photo by NASA, date unknown
Outline

- Problem Description and Motivation
- Methods
  - Feature Generation
  - Type Association
  - Multiple Source Integration
  - Footprint Generation
- Results
  - Feature Types
  - Comparison to Existing Gazetteers
  - Ground Truth Evaluation
- Future Work
Problem

- There is a need for highly detailed gazetteers
  - Feature identification in imagery
  - Named entity recognition (NER)
  - Etc.
- But there are very few around
Identifying every feature in an image requires a gazetteer that has every feature in it.

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
Named Entity Recognition

- Successful NER requires the existence of the name in the underlying gazetteer
- Easy at the global level
  - Major City Names, Country Names
- Hard at the local level
  - Local newspaper police report:
    “The suspect entered *McDonalds* on *3rd Street* at 5:00 pm”
  - Highly Detailed Digital Geo-Library
    “All books, newspaper articles, and pictures relating to the *Sheraton Hotel* on *Grand Avenue*”
Motivation

• The information to make these things possible is out there
  – Online Phone books
  – Online Property Tax sites
  – Etc.

• We can extract and integrate information from multiple sources to build detailed regional gazetteers
Outline

- Problem Description and Motivation

- Methods
  - Feature Generation
  - Type Association
  - Multiple Source Integration
  - Footprint Generation

- Results
  - Feature Types
  - Comparison to Existing Gazetteers
  - Ground Truth Evaluation

- Future Work
Methods - Overview

Generate Features

Associate Types

Generate Footprints

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
Feature Generation

• Consider the following image
  – part of our test area
  • El Segundo, CA 90245
Feature Generation Cont.
Feature Generation Cont.

• What can we see?
  – For an urban area most of the ground is covered by buildings
  – Buildings have addresses
  – If we can generate the set of all addresses in an area, it would give us a pretty good representation of what is on the ground
Feature Generation Cont.
Data Source – Zip+4 Files

- Published by the USPS
  - List street segments
  - List valid addresses per segment

<table>
<thead>
<tr>
<th>ID, Pre, Name, Suffix, Post, Start, End, …</th>
</tr>
</thead>
<tbody>
<tr>
<td>100, E, Maple, Ave, N, 700, 798, Even, …</td>
</tr>
<tr>
<td>101, E, Maple, Ave, N, 701, 899, Odd, …</td>
</tr>
<tr>
<td>102, E, Maple, Ave, N, 800, 898, Even, …</td>
</tr>
<tr>
<td>103, E, Maple, Ave, N, 801, 899, Odd, …</td>
</tr>
<tr>
<td>104, E, Maple, Ave, N, 900, 950, Even, …</td>
</tr>
</tbody>
</table>

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
Data Source – Assessor

Provides:

- Information about features
- Verification of existence
- Type (Residential/Commercial)

- Interface is queryable by
  - Address

http://assessormap.co.la.ca.us/mapping/viewer.asp
Feature Generation

- USPS Zip+4 files
  - Can generate all addressable addresses
- Los Angeles County Assessor (LACA) Site
  - Can verify addresses which do exist

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
Method So Far

All Possible Addresses

Assessor

All Valid Addresses

Zip+4

id, 12231, Maple,Ave
id, ................
id, ............. ... 
id, ................
id, ....................

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
Feature Types

• Now that we have features
  – What are their names and types?

• All we know so far about a feature
  – It is a building
    • Commercial
    • Residential

• Can we get more information?
Feature Types Cont.

• Supplement with phone book information
  – Yellow Pages (Commercial)
  – White Pages (Residential)
Data Source – Superpages

Provides:

- Information about features
  - Name
  - Category
  - Address
  - Phone Number
- Interface is queryable via
  - Type Category
  - Location (Zip Codes)

http://www.superpages.com

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
Superpages – Extraction Method

Step 1: Model source

Step 2: Extract Category Hierarchy

Step 3: Generate Zip Specific URLs for Categories

http://Superpages.com?catid=XXX&zip=YYYYYY

Step 4: Extract Features

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
Feature Extraction

http://superpages.com?catid=XXX&zip=YYYYYY
### Feature Extraction Cont.

Below is a screenshot of a webpage with extracted data. The screenshot shows a list of businesses, each with detailed information such as name, address, and phone number. The data is extracted using a wrapper tool, which is used to process and interpret web pages for data extraction.

#### Example of Extracted Data:

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chicken Dips</td>
<td>234 Main Street, El Segundo, CA 90245</td>
<td>(310) 414-9000</td>
</tr>
<tr>
<td>Chino Verde</td>
<td>123 South Douglas Street Suite 114, El Segundo, CA 90245</td>
<td>(310) 444-7666</td>
</tr>
<tr>
<td>Contour Coffee Shop</td>
<td>831 South Douglas Street Suite 114, El Segundo, CA 90245</td>
<td>(310) 663-8228</td>
</tr>
<tr>
<td>Cazaynes Coastal Grill</td>
<td>2171 Rosecrans Avenue, El Segundo, CA 90245</td>
<td>(310) 444-8228</td>
</tr>
<tr>
<td>La Cabin De Lulan</td>
<td>130 North Sepulveda Boulevard Suite 2A, El Segundo, CA 90245</td>
<td>(310) 444-8228</td>
</tr>
<tr>
<td>Daily Grill</td>
<td>2121 Rosecrans Avenue, El Segundo, CA 90245</td>
<td>(310) 444-8228</td>
</tr>
</tbody>
</table>

This extracted data can be further processed and analyzed for various applications such as recommendation systems, customer analysis, and more.
Data Source – Switchboard

Provides:

- Information about features
  - Type (Residential)
  - Owner Name
  - Address
  - Phone Number
- Interface is queryable by
  - City
  - First Letter of Last Name (a, b, ..., z)

http://www.switchboard.com

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
**Switchboard – Extraction Method**

**Step 1: Model source**

**Step 2: Generate Name Specific URLs**

http://switchboard.com/bin/cgiQA.dll?city=el+segundo&LastN=XXXX

**Step 3: Extract Features**
Method So Far

Zip+4
id, 12231, Maple,Ave
id, id, All Possible id, id, Addresses

Assessor

Superpages

Category
Zip

All Possible Addresses

Name, Address, Type, Phone

City

A ... Z

Switchboard

Name, Address, Phone

Need a Single Integrated Set

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
Integration Process

• **Normalization**
  – Put all records into a consistent format

• **Record Linkage**
  – Eliminate duplicate records
  – Merge attributes from multiple records
Normalization

800 North Sepulveda Bl. West, apartment 15

800 Number
N Pre Directional
SEPULVEDA Name
BLVD Suffix
W Post Directional
APT Suite
15 Suite Number

800 N SEPULVEDA BLVD W APT 15

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
# Record Linkage

<table>
<thead>
<tr>
<th>Bob’s Pizza</th>
<th>123 Main St</th>
<th>Pizza</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob’s Pizza</td>
<td>123 Main St</td>
<td>Take Out</td>
</tr>
<tr>
<td>Bob’s Pizza</td>
<td>123 Main St</td>
<td>Italian</td>
</tr>
</tbody>
</table>

**Deduplication**

| Bob’s Pizza | 123 Main St | Pizza Takeout Italian |

**Object Consolidation**

Goldberg, Wilson, Knoblock  
CSCI 548 - 2009
Method So Far

Assessor

All Possible Addresses

Zip+4
id, 12231, Maple Ave
id, id
id, All Possible
id, id
id, Addresses

Category

Zip

Superpages

City

A ... Z

Switchboard

Name, Phone

Integration

Need a footprint

Name, Address, Type, Phone

Name, Address, Phone

All Valid Addresses

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
Footprint Generation

- Generate points as first approximation

Geocode

**Test**

To test the operation using the HTTP GET protocol, click the 'Invoke' button.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>streetaddress</td>
<td>2652 allendale place</td>
</tr>
<tr>
<td>city</td>
<td>los angeles</td>
</tr>
<tr>
<td>state</td>
<td>ca</td>
</tr>
<tr>
<td>zipstr</td>
<td>90007</td>
</tr>
</tbody>
</table>

`<geopoint>`

<lat>34.005206532674805</lat>
<lon>-118.47605547145031</lon>
<geoerror/>
<errorbounds>0.0015020408163265305</errorbounds>
<errorboundsunit>Decimal Degrees</errorboundsunit>
</geopoint>`
Complete Method

Zip+4
id, 12231, Maple Ave
id, --------, --------, --------, id,
All Possible Addresses
id,
--------,
--------,
--------,
--------,

Category
Superpages
Zip
City
Switchboard
A ... Z

Name, Address,
Type, Phone

Integration

All Valid Addresses

Geocode

Name, Type,
Latitude / Longitude

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
Outline

- Problem Description and Motivation

- Methods
  - Feature Generation
  - Type Association
  - Multiple Source Integration
  - Footprint Generation

- Results
  - Feature Types
  - Comparison to Existing Gazetteers
  - Ground Truth Evaluation

- Future Work
# Feature Types

<table>
<thead>
<tr>
<th>Path</th>
<th>SubCategory</th>
<th>#SubCategories</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Arts &amp; Entertainment</td>
<td>309</td>
<td>167</td>
</tr>
<tr>
<td>2</td>
<td>Automotive</td>
<td>560</td>
<td>314</td>
</tr>
<tr>
<td>3</td>
<td>Business &amp; Professional Services</td>
<td>825</td>
<td>821</td>
</tr>
<tr>
<td>4</td>
<td>Clothing &amp; Accessories</td>
<td>247</td>
<td>156</td>
</tr>
<tr>
<td>5</td>
<td>Community &amp; Government</td>
<td>427</td>
<td>244</td>
</tr>
<tr>
<td>6</td>
<td>Computers &amp; Electronics</td>
<td>283</td>
<td>413</td>
</tr>
<tr>
<td>7</td>
<td>Construction &amp; Contractors</td>
<td>1097</td>
<td>516</td>
</tr>
<tr>
<td>8</td>
<td>Education</td>
<td>89</td>
<td>127</td>
</tr>
<tr>
<td>9</td>
<td>Food &amp; Dining</td>
<td>460</td>
<td>265</td>
</tr>
<tr>
<td>10</td>
<td>Health &amp; Medicine</td>
<td>774</td>
<td>463</td>
</tr>
<tr>
<td>11</td>
<td>Home &amp; Garden</td>
<td>1092</td>
<td>607</td>
</tr>
<tr>
<td>12</td>
<td>Industry &amp; Agriculture</td>
<td>1476</td>
<td>416</td>
</tr>
<tr>
<td>13</td>
<td>Legal &amp; Financial</td>
<td>224</td>
<td>485</td>
</tr>
<tr>
<td>14</td>
<td>Media &amp; Communications</td>
<td>321</td>
<td>248</td>
</tr>
<tr>
<td>15</td>
<td>Personal Care &amp; Services</td>
<td>104</td>
<td>310</td>
</tr>
<tr>
<td>16</td>
<td>Real Estate</td>
<td>166</td>
<td>333</td>
</tr>
<tr>
<td>17</td>
<td>Shopping</td>
<td>668</td>
<td>439</td>
</tr>
<tr>
<td>18</td>
<td>Sports &amp; Recreation</td>
<td>521</td>
<td>149</td>
</tr>
<tr>
<td>19</td>
<td>Travel &amp; Transportation</td>
<td>596</td>
<td>529</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10239</td>
<td>7002</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Path</th>
<th>SubCategory</th>
<th>#SubCategories</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.1</td>
<td>Shopping &gt;&gt; Home &amp; Garden &gt;&gt; Appliances</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>17.1.1</td>
<td></td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>17.1.2</td>
<td></td>
<td>90</td>
<td>41</td>
</tr>
<tr>
<td>17.1.3</td>
<td></td>
<td>29</td>
<td>19</td>
</tr>
<tr>
<td>17.1.4</td>
<td></td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>17.1.5</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17.1.6</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17.1.7</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17.1.8</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17.1.9</td>
<td></td>
<td>175</td>
<td>94</td>
</tr>
</tbody>
</table>

**Direct Link**

- Garden Centers (has 1 names) 0
- Home Improvement Stores (has 1 names) 11
- Home Furnishings Stores (has 1 names) 12
- Plants & Trees (has 2 names) 0
- Seeds & Bulbs (has 1 names) 0
## Feature Types Cont.

### Table

<table>
<thead>
<tr>
<th>Name</th>
<th>Category Name(s)</th>
<th>City</th>
<th>State</th>
<th>Zip</th>
<th>Lat</th>
<th>Lon</th>
<th>Eon</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Mobile Services</td>
<td>Specialty Stores</td>
<td>El Segundo</td>
<td>CA</td>
<td>90245</td>
<td>33.89</td>
<td>118.40</td>
<td>2009</td>
</tr>
<tr>
<td>Body Workouts</td>
<td>Specialty Stores</td>
<td>El Segundo</td>
<td>CA</td>
<td>90245</td>
<td>33.89</td>
<td>118.40</td>
<td>2009</td>
</tr>
<tr>
<td>Body Massage Therapy</td>
<td>Specialty Stores</td>
<td>El Segundo</td>
<td>CA</td>
<td>90245</td>
<td>33.89</td>
<td>118.40</td>
<td>2009</td>
</tr>
<tr>
<td>Cosmetic Salons</td>
<td>Specialty Stores</td>
<td>El Segundo</td>
<td>CA</td>
<td>90245</td>
<td>33.89</td>
<td>118.40</td>
<td>2009</td>
</tr>
<tr>
<td>Corporate Consulting Services</td>
<td>Specialty Stores</td>
<td>El Segundo</td>
<td>CA</td>
<td>90245</td>
<td>33.89</td>
<td>118.40</td>
<td>2009</td>
</tr>
<tr>
<td>Face Place and Glow</td>
<td>Specialty Stores</td>
<td>El Segundo</td>
<td>CA</td>
<td>90245</td>
<td>33.89</td>
<td>118.40</td>
<td>2009</td>
</tr>
<tr>
<td>Hair Salons</td>
<td>Specialty Stores</td>
<td>El Segundo</td>
<td>CA</td>
<td>90245</td>
<td>33.89</td>
<td>118.40</td>
<td>2009</td>
</tr>
<tr>
<td>Health &amp; Beauty</td>
<td>Specialty Stores</td>
<td>El Segundo</td>
<td>CA</td>
<td>90245</td>
<td>33.89</td>
<td>118.40</td>
<td>2009</td>
</tr>
<tr>
<td>Health &amp; Beauty</td>
<td>Specialty Stores</td>
<td>El Segundo</td>
<td>CA</td>
<td>90245</td>
<td>33.89</td>
<td>118.40</td>
<td>2009</td>
</tr>
<tr>
<td>Natural Health Products</td>
<td>Specialty Stores</td>
<td>El Segundo</td>
<td>CA</td>
<td>90245</td>
<td>33.89</td>
<td>118.40</td>
<td>2009</td>
</tr>
<tr>
<td>Specialty Stores</td>
<td>Specialty Stores</td>
<td>El Segundo</td>
<td>CA</td>
<td>90245</td>
<td>33.89</td>
<td>118.40</td>
<td>2009</td>
</tr>
</tbody>
</table>

### Categories

- Specialty Stores (has 2 names)
- Hair Care & Treatment (has 1 name)
- Medical Equipment (has 1 name)
- Health & Beauty (has 1 name)
- Health & Beauty (has 1 name)
- Natural Health Products (has 1 name)
- Specialty Stores (has 1 name)
- Specialty Stores (has 1 name)
- Specialty Stores (has 1 name)
- Specialty Stores (has 1 name)
- Specialty Stores (has 1 name)

---

Goldberg, Wilson, Knoblock

CSCI 548 - 2009
Issues

• Definition of Feature Types
  – Depends on
    • The physical structure?
    • The occupants/uses?

  – Include different types of public structures
    • School, Police Station, Hospital
  – But not different types of businesses
    • Pizzeria, Brake Repair Shop, Dentist
Comparison – Existing Gazetteers

- ADL
- Los Angeles Comprehensive Bibliographical Database (LACBD)

ADL (22 Features)

LACBD (11 Features)

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
Comparison –
Automatically Generated Gazetteer

AGG (5,046 Features)

Goldberg, Wilson, Knoblock
CSCI 548 - 2009
### Feature Comparison

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Feature Name</th>
<th>ADL</th>
<th>LACBD</th>
<th>AGG</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>El Segundo City Hall</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>T</td>
<td>Old Town Music Hall</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Center Street Elementary</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>E</td>
<td>St Johns Lith Child Dev</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>E</td>
<td>St Anthony Elementary</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Richmond Street Elem</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Webster University</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Arena High</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>El Segundo High</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>E</td>
<td>El Segundo Mid</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>H</td>
<td>Chevron Refinery</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>H</td>
<td>Airport Towers Number 1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>El Segundo Public</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PK</td>
<td>Holly Valley</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>PK</td>
<td>Candy Cane</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>PK</td>
<td>El Segundo</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>PK</td>
<td>Dockweiler Beach St</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PK</td>
<td>Kansas</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>PK</td>
<td>El Segundo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Pacific Baptist</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>St Andrews Church</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>R</td>
<td>New Mt Calv Mis. Bap</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>R</td>
<td>Temple Rodeph Shalom</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>R</td>
<td>El Segundo Christian</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>R</td>
<td>El Segundo Foursquare</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>S</td>
<td>El Segundo Golf Course</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Ratio of total (superset): 21/26, 10/26, 15/26
Recall % (superset): 81%, 38%, 58.00%

Ratio of total (buildings): 14/18, 9/18, 14/18
Recall % (buildings): 78%, 50%, 78%

B building, T theater, E educational facility, H heliport, L library, PK park, PO post office, Religious site, and S sports facility
Ground Truth

• Surveyed 43 street segments of different types by walking the blocks
  • Residential (R)
  • Commercial (C)
  • Industrial (I)
  • R/C
  • I/C

• Compared
  – What the AGG said was there
  – What was actually there
Accuracy

• Location
  Precision:
    Of the number of features we extracted, how many actually existed?
  Recall:
    Of the number of features in existence, how many did we get?

• Name / Type
  Precision:
    Of the names/types we extracted, how many were correct?
  Recall:
    Of the possible names/types in existence, how many did we get?
**Superpages**

<table>
<thead>
<tr>
<th>Type</th>
<th>Name/Type Precision</th>
<th>Name/Type Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1.00</td>
<td>0.68</td>
</tr>
<tr>
<td>I/C</td>
<td>1.00</td>
<td>0.83</td>
</tr>
<tr>
<td>C</td>
<td>0.95</td>
<td>0.88</td>
</tr>
<tr>
<td>C/R</td>
<td>0.92</td>
<td>0.87</td>
</tr>
<tr>
<td>R</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Overall</td>
<td>0.96</td>
<td>0.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>Location Precision</th>
<th>Location Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.90</td>
<td>0.39</td>
</tr>
<tr>
<td>I/C</td>
<td>0.88</td>
<td>0.78</td>
</tr>
<tr>
<td>C</td>
<td>0.75</td>
<td>0.99</td>
</tr>
<tr>
<td>C/R</td>
<td>0.94</td>
<td>0.63</td>
</tr>
<tr>
<td>R</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Overall</td>
<td>0.72</td>
<td>0.58</td>
</tr>
</tbody>
</table>

- Very accurate name/type
- Overestimated features (shopping centers, downtown blocks, etc.)
Switchboard

<table>
<thead>
<tr>
<th>Type</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>I/C</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C</td>
<td>0.67</td>
<td>0.17</td>
</tr>
<tr>
<td>C/R</td>
<td>0.78</td>
<td>0.45</td>
</tr>
<tr>
<td>R</td>
<td>1.00</td>
<td>0.44</td>
</tr>
<tr>
<td>Overall</td>
<td>0.49</td>
<td>0.22</td>
</tr>
</tbody>
</table>

- Underestimated Features (cell phones, unlisted numbers, etc.)
Assessor

<table>
<thead>
<tr>
<th>Type</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.86</td>
<td>0.71</td>
</tr>
<tr>
<td>I/C</td>
<td>0.94</td>
<td>0.84</td>
</tr>
<tr>
<td>C</td>
<td>0.84</td>
<td>0.72</td>
</tr>
<tr>
<td>C/R</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td>R</td>
<td>1.00</td>
<td>0.91</td>
</tr>
<tr>
<td>Overall</td>
<td>0.89</td>
<td>0.81</td>
</tr>
</tbody>
</table>

- Very accurate features
- Not much type information
Automatically Generated Gazetteer

After integrating the three sources

<table>
<thead>
<tr>
<th>Type</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.94</td>
<td>0.80</td>
</tr>
<tr>
<td>I/C</td>
<td>0.92</td>
<td>1.00</td>
</tr>
<tr>
<td>C</td>
<td>0.64</td>
<td>1.00</td>
</tr>
<tr>
<td>C/R</td>
<td>0.79</td>
<td>0.99</td>
</tr>
<tr>
<td>R</td>
<td>0.99</td>
<td>0.91</td>
</tr>
<tr>
<td>Overall</td>
<td>0.85</td>
<td>0.94</td>
</tr>
</tbody>
</table>

- Very accurate features
- Very detailed type information
Outline

- Problem Description and Motivation
- Methods
  - Feature Generation
  - Type Association
  - Multiple Source Integration
  - Footprint Generation
- Results
  - Feature Types
  - Comparison to Existing Gazetteers
  - Ground Truth Evaluation
- Future Work
Future Work

• Automatically merge feature type hierarchies
  – Our types are very detailed and would not translate directly/automatically into existing hierarchies
  – How can we automatically augment existing FTT to accommodate highly detailed feature types?

• More feature types
  – New Sources
Future Work

• Better footprints
  – Extract parcel boundaries from LACA images
The End

• Questions?
• Comments?

• Thanks!