Introducing Locations and Linking

Place names are ambiguous –

- **synonym:** Queensland and Sunshine State
- **multiple referents:** Cambridge, Massachusetts, USA or Cambridge, England, UK.
  - Smith and Mann (2003) analysed over 1 million entries in The Getty Thesaurus
  - Almost 33% of places in Asia had multiple names
  - Almost 30% in Oceania had multiple referents

**Named entity linking (NEL)** resolves the names of people, places and organisations to a knowledge base (KB) like Wikipedia (Figure 1).

**Toponym resolution (TR)** focuses on linking place names to their referent (Leidner, 2007). Many systems use gazetteers and machine learning using context for TR and geolocation.

**Better linking of locations may improve NEL**
- Motivation: NEL has not exploited meta-data
- Current state-of-art linking gets about 80%

Analysis of Data and Errors

- Motivated by error analysis of locations in linker
- Almost 40% entities are locations in TAC 11 data
- Locations account for almost 50% of errors – disproportionately difficult

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly linked named entities</td>
<td>1955</td>
<td>87.1</td>
</tr>
<tr>
<td>Locations in TAC 11 data</td>
<td>896</td>
<td>39.9</td>
</tr>
<tr>
<td>Correctly linked LOCs (%)</td>
<td>726</td>
<td>38.7</td>
</tr>
<tr>
<td>Location errors (out all errors)</td>
<td>137</td>
<td>47.2</td>
</tr>
</tbody>
</table>

Table 1: Analysis of TAC 11 location errors overall

**GeoNames** is a geographical database used in TR and in our method:
- Over nine million place names including aliases of cities, countries, mountains, rivers etc
- Each place has administrative level (e.g. country, or level 1 to 4 to represent state)
- Other information like population, capital city, continent included if available
- Users can edit database using web interface
- **Coverage is uneven** because combines multiple knowledge bases, sources

<table>
<thead>
<tr>
<th>Type</th>
<th>Total</th>
<th>% % Zeroes</th>
</tr>
</thead>
<tbody>
<tr>
<td>First order admin (state)</td>
<td>8</td>
<td>0.00</td>
</tr>
<tr>
<td>Second order admin</td>
<td>559</td>
<td>2.50</td>
</tr>
<tr>
<td>Populated place (PP)</td>
<td>12522</td>
<td>81.30</td>
</tr>
<tr>
<td>Section of PP</td>
<td>1977</td>
<td>95.35</td>
</tr>
</tbody>
</table>

Table 2: GeoNames population counts by administrative code for Australia - uneven by type, higher administrative regions have fewer zeros

Disambiguating Location Candidates

Intuition: more populous places are more likely to occur in news text and nearby places like likely to occur together. Features:

1. **Highest Population** – link most populous candidate
2. **Hierarchy filtering** – keep candidates in the same country/continent as selected pivot candidate, see Figure 1; experimented with different selections e.g. most frequent, highest ranked (top) concept or first concept in document

**Hypothesis:** improving location disambiguation and linking impacts NEL, including non-geographical entities.

**Experimental Results**

**Table 3: TAC 12 results**

**Table 4: CompNews test set results**

**Future Work**
- Use scaled, relative population or combine with other features in supervised linker
- Filter multiple concepts or LOC intersection
- Other methods to determine most salient location or don’t filter some e.g. when multiple locations or stories in the same article

**Locations are difficult**

Error analysis showed:
- **Missing population:** e.g. Africa, small towns
- Population or most frequent incorrect: e.g. New York State linked when actually New York City
- Local news biased towards smaller locations (American news or CompNews)
- Better filtering for regional? e.g. Springfield: 73 entries in GeoNames in US, 236 worldwide

Finding most salient place not easy - headlines and first paragraph can be misleading e.g. He told her he was a young Marine, recently back from Iraq is in US.

**Figure 1: NEL candidates for Washington and how GeoNames can filter by population and hierarchy**

**Figure 2: NEL for Sydney – ambiguity**

**Mention:** Washington

**Candidates:**
- Washington State
- Washington, D.C.
- Washington County, Oregon
- Washington (UK)
- George Washington

**GeoNames**
- Population: 6271775
- Country: USA
- Continent: North America
- Administrative region: Washington
- Population: 1930
- Country: United Kingdom
- Continent: Europe
- Administrative region: England

**Locations are difficult**

Error analysis showed:
- **Missing population:** e.g. Africa, small towns
- Population or most frequent incorrect: e.g. New York State linked when actually New York City
- Local news biased towards smaller locations (American news or CompNews)
- Better filtering for regional? e.g. Springfield: 73 entries in GeoNames in US, 236 worldwide

Finding most salient place not easy - headlines and first paragraph can be misleading e.g. He told her he was a young Marine, recently back from Iraq is in US.

**Future Work**
- Use scaled, relative population or combine with other features in supervised linker
- Filter multiple concepts or LOC intersection
- Other methods to determine most salient location or don’t filter some e.g. when multiple locations or stories in the same article