Mining Geographic Information in Text

Geoffrey Andogah (Gulu University)
G.Andogah@gu.ac.ug
The royal burgh of Annan is a well-built town, red sandstone being the material mainly used. Among its public buildings is Annan Academy of which the writer Thomas Carlyle was a pupil, a Georgian building now known as "Bridge House". The Town Hall, built in Victorian style in 1878, uses the local sandstone. Annan also features a Historic Resources Centre. In Port Street, some of the windows remain blocked up to avoid paying the window tax.

Kofi Atta Annan (born 8 April 1938) is a Ghanaian diplomat who served as the seventh Secretary-General of the United Nations from 1 January 1997 to 31 December 2006. Annan and the United Nations were the co-recipients of the 2001 Nobel Peace Prize for his founding the Global AIDS and Health Fund to support developing countries in their struggle to care for their people.

Amsterdam is a city located in Montgomery County, New York, USA. As of the 2000 census, the city had a total population of 18,355. The name is derived from the city of Amsterdam in the Netherlands. The city of Amsterdam is surrounded on the north, east, and west sides by the town of Amsterdam and is on the banks of the Mohawk River. The majority of the city lies on the north bank, but the Port Jackson area on the south side is also part of the city.

Amsterdam is the capital and largest city of the Netherlands, with an urban population of 747,290 and a metropolitan population of 2,158,592. The city is in the province of North Holland in the west of the country. It comprises the northern part of the Randstad, the sixth-largest metropolitan area in Europe, with a population of approximately 8.1 million according to larger estimates.

Definitions

- Geo/Non-Geo ambiguity resolution:
  - Toponym recognition: recognize and classify names spotted in documents as names of places.

- Geo/Geo ambiguity resolution:
  - Toponym resolution: assigns names of places to locations they referred to in documents.
  - Scope resolution: assigns geographic scopes to documents.
Toponym recognition

- Toponym recognition is a subset of named entity recognition and classification (NERC) task concerned with recognition and classification of place names.
- NERC employ both Machine Learning and Dictionary-based approaches.
Toponym recognition

NERC Tools and Services

– OpenNLP tools: http://incubator.apache.org/opennlp/
– GATE Annie: http://gate.ac.uk/ie/annie.html
– Alias-i Lingpipe: http://alias-i.com/lingpipe/
Toponym recognition

*Mahali PlacenameFinder*

- Mahali PlacenameFinder is derived from OpenNLP NameFinder.

```
outputME - OpenNLP NameFinderME output.
outputDT - OpenNLP DictionaryNameFinder output.

if pname in outputME AND outputDT AND NOT in stopWord
  set nerConf = 4

else if ( pname in outputME AND NOT ( outputDT AND stopWord ) )
  OR ( pname in outputDT AND NOT ( outputME AND stopWord ) )
  set nerConf = 2

else
  set nerConf = 1
```
Scope resolution

Geographic Distribution of Places

- Significant fraction of all locations mentioned in the document are either the scope itself or locations within the scope.
- Location references mentioned in the document are distributed smoothly across the scope.
- For example, a document with scope Groningen Province is expected to mention Groningen Province or locations within Groningen Province more frequently than places belonging to other provinces or countries.
Scope resolution

Toponym Frequency

- The most commonly occurring place in the document dictates the scope of the containing document.
- For example, if Groningen Province is mentioned more frequently than Utrecht Province, the scope of the document is most likely the Province of Groningen.
Scope resolution

*Geographic Feature Importance*

- The scope of a document is set to the country or region containing the most important unambiguous places (e.g., capital cities and other major cities) identified in the document.
- For example, spotting Rotterdam, Hague and Amsterdam in a document sets the scope of the document to the Netherlands.
Scope resolution
Tools and Services

– Yahoo! PlaceMaker http://developer.yahoo.com/geo/placemaker/
– Mahali ScopeFinder – demo soon coming online.
Mahali Scope
Data Model

parent-region

target-region

neighbors

smallest-cities

child-regions

primary-cities

child-child-regions

secondary-cities
Sample Scope Data Model
For Netherlands

Europe

Netherlands

Belgium

Germany

Boskop
Hillegom
Lisse
Vlist
etc.

North Holland
South Holland
Groningen
etc.

Amsterdam
Rotterdam
The Hague
etc.

Delft
Leiden
The Hague
Rotterdam
etc.

Delft
Gouda
Leiden
etc.
Yahoo! GeoPlanet

Mahali Scope data model fairly resembles Yahoo! GeoPlanet structure. GeoPlanet provides:

- WOEID or Where-On-Earth IDentifier: a number that uniquely identifies a place;
- Hierarchical containment of all places up to the Earth level;
- Zip codes are included as place names;
- Adjacencies: places neighbouring each WOEID;
- Aliases: synonyms for each WOEID.
Yahoo! GeoPlanet

- GeoPlanet's goal is to enable location-based and geographically-enabled web services.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Zip code</td>
<td>4,321,711</td>
</tr>
<tr>
<td>Towns</td>
<td>863,749</td>
</tr>
<tr>
<td>Adjacencies</td>
<td>9,643,734</td>
</tr>
<tr>
<td>Places</td>
<td>5,653,968</td>
</tr>
</tbody>
</table>
Toponym resolution

_default Sense Heuristics_

- Land surface area: select location with the largest land surface area.
- Hierarchy distance: select place at the top of the regions hierarchy, e.g., Holland (Europe) is preferred over Holland (Michigan).
- Place type: Select place in order of place type importance: country > city > town > village.
- Corpus popularity: select place that occurs more commonly in a collection.
- Population: select a place with the largest population.
Toponym resolution

Pattern Matching and Hierarchy Overlap

- Feature type qualifier: select location with the matching type.
- Country scope restriction: assigns a country scope to a document, and any ambiguous toponym is treated to belong to the county.
- Smallest polygon: resolves toponyms to the smallest polygon that completely grounds the whole set.
- Spatial distance: chooses a location closest to all the non-ambiguous locations in the document.
Toponym resolution

One Referent per Discourse

- Assumes one and only one meaning to toponyms mentioned in discourse. Subsequent mentions of the toponyms is assumed to convey the meaning as previous meanings.
Toponym resolution

Tools and Services

- Yahoo! PlaceFinder http://developer.yahoo.com/geo/placefinder/
- Google Geocoding API http://code.google.com/apis/maps/documentation/geocoding/
- OpenStreetMap's Nominatim http://wiki.openstreetmap.org/wiki/Nominatim
- Geonames Geocoder http://www.geonames.org/
- Mahali LocationFinder – demo soon coming online.
Mahali system architecture
Toponym resolution

Mahali LocationFinder

- Mahali LocationFinder exploits 26,820 geographic scopes automatically assigned to documents, type of place (e.g., city, airport, etc.), classification of place (e.g., populated place, administrative division, etc.), population of place, and frequency of non-ambiguous or resolved places.
Geo-meta

Application

- Enable local news and information services.
- Enable marketers target local content users will locally targeted advertisements.
- Enable search engines answer geography dependent user queries.
- Enable QA systems answer geography dependent question.
Mahali Tool Kit

What next?

- Working to bring Mahali demo online.
- Talking to businesses interested in adapting Mahali for their work.
- Create more scope reference models for individuals countries and regions.
- Integrate Mahali in Apache UIMA framework.
Reference


THANK YOU