Chapter 5
Toponym resolution

Toponym resolution is a process of assigning a place name identified in text to a single non-ambiguous place on the surface of the earth by means of a reference coordinate system such as longitudes and latitudes.

In this chapter we return to one of the most central problems in geographical information retrieval (GIR), toponym resolution. Now that we have introduced the notion of document scope as well as the procedure designed to detect it, we are in a position to try to apply the notion to toponym resolution. In some sense, this chapter attempts to apply a top-down perspective (from document space to toponym) to improve toponym resolution. We shall conclude that it is successful. The research objectives of this chapter are:

1. How effective is the document’s geographical scope or coverage in aiding the resolution of toponyms contained in the document? [Question of heuristics.]

2. How best can the performance of toponym resolution systems be compared? [Question of binary vs. non-binary evaluation metric; and the question of reference evaluation corpora.]

The study of place names, their origins, meanings, use and typology is referred to as toponymy. A toponym is a named point of reference in both the physical and cultural landscape on the Earth’s surface. This includes natural features, such as streams (whose names are studied as hydronyms) and artificial ones (such as cities). Each toponym (or geographical name) is a
vital communication tool that reflect patterns of settlement, exploration, migration, and heritage that may otherwise be overlooked by residents, visitors, and future generations.\(^1\)

The names of places are ambiguous in many ways. They can reference other named entities (e.g., names of people, names of organizations, etc.), and may be used as common language vocabulary words (e.g., Split is a city in Croatia, Over a city in Germany, etc.). The use of place names outside their geographical or location context is sometimes referred to as Geo/Non-Geo ambiguity (Amitay et al., 2004). Besides reference or geo/non-geo ambiguity issues, names of places are referentially ambiguous. Referential ambiguity occurs when a single place name references multiple places. This is also termed as Geo/Geo ambiguity (Amitay et al., 2004). Many places are also referred to by more than one name, e.g., Netherlands vs. Holland.

The task of toponym resolution (Leidner, 2007) is to map a place name to a non-ambiguous location or a geographical point on the surface of the Earth. This mapping is normally done using a geographic reference coordinate system such as latitude and longitude. The terms geographical name, place name and toponym are synonymous, and they are used interchangeably to mean the same thing (i.e., a name of a place) throughout this chapter.

Figure 5.1 shows a schematic diagram depicting the toponym resolution process. In this work geo/non-geo ambiguity resolution is performed with the help of an off-the-shelf named entity recognition tool, the Alias-i LingPipe.\(^2\) The motivation is that place name recognition components of the state-of-the-art named entity recognizers have achieved near human performance (Tjong Kim Sang and De Meulder, 2003), therefore, existing off-the-shelf recognizers are sufficient to perform geo/non-geo resolution task. In this thesis we are concerned with geo/geo resolution task, i.e., mapping place

---

\(^1\)[08 October 2009]: http://en.wikipedia.org/wiki/Toponymy

\(^2\)[08 October 2009]: http://alias-i.com/lingpipe/
names to non-ambiguous geographical points on the surface of the Earth.

The toponym approach reported in this thesis exploits 26,820 geographical scopes automatically assigned to documents, the type of place (e.g., city), the classification of place (e.g., populated place, administrative division, etc.), the population of the place, and the frequency of non-ambiguous or resolved places.

5.1 Toponym resolution procedure

This section answers the research objective – How effective is the document’s geographical scope(s) in aiding the resolution of toponyms contained in the document. [Question of heuristics.]

This work proposes geographical scope-driven toponym resolution approach. The approach builds on previous works discussed in Chapter 2. The approach reported here applies the following techniques at various stages of resolution:

1. Single referent per discourse;
2. Scope restriction;
3. Population heuristics;
4. Place type restriction;
5. Default sense heuristics.

Figure 5.2 depicts geographic referent resolution routines, and the following is the explanation of how the various blocks work together to accomplish the resolution task. For illustration purposes, we shall consider a document containing the following place names:

(1) Sarajevo, Bosnia, Bihac, Tuzla, Britain, London

Our task is to resolve the place names to places they refer to on the surface of the earth.

Single referent per discourse and default sense heuristics

The toponym resolution algorithm starts by invoking the single referent per discourse and default sense heuristics. Place names mentioned more than once in a document are assumed to refer to the same place on the basis of the single referent per discourse heuristic. Next the default sense heuristic is
Chapter 5. Toponym resolution

Figure 5.2: Toponym resolution algorithm.
invoked to assign place names with continent sense to the continents. Candidate places for place names with senses other than continent senses are obtained from the geographical database (GeoDB). Place names with single candidate places are resolved to these places, and place names with multiple candidates are passed to lower processing modules starting with the scope restriction module. However, place names with no candidates in the GeoDB are ignored. This often happens when a geographic database lacks complete coverage.

**Scope restriction heuristic**

The scope restriction heuristic (see BK-A in Fig. 5.2) extends the country level restriction reported by Pouliquen et al. (2006). The heuristic exploits an elaborate list of ranked geographic scopes assigned to a document by the geographic scope resolver described in Chapter 4. A place name with multiple candidate referents is assigned to a single most highly ranked document geographic scope. The other candidates belonging to lower ranked document scopes are discarded. If a selected scope contains a single candidate, the candidate is marked as the place referred to by the name. However, if a selected scope contains multiple candidates with the same name, it is passed to the next processing block in the hierarchy, i.e., country and capitals resolution heuristic (see BK-B in Fig. 5.2). The accuracy of the scope restriction heuristic depends on how well the scope resolution performed. The greater the error in scope resolution the greater the error in referent resolution.

To illustrate the heuristics, refer to example (1) consisting of place names found in a document (see Sec. 5.1, page 97). The names are assigned to the following scopes:

(2) **Sarajevo:**

a. (Capital City, P; pop 696,731) Sarajevo → Federation of Bosnia and Herzegovina → Bosnia and Herzegovina → Europe

(3) **Bosnia:**

a. (Country, A) Bosnia and Herzegovina → Europe

(4) **Bihac:**

a. (City, P; pop 75,641) Bihac → Una-Sana County → Federation of Bosnia and Herzegovina → Bosnia and Herzegovina → Europe

b. (Municipality, A) Bihac → Una-Sana County → Federation of
Bosnia and Herzegovina $\mapsto$ Bosnia and Herzegovina $\mapsto$ Europe

(5) \textbf{Tuzla:}

\begin{itemize}
  \item a. (City, P; pop 142,486) Tuzla $\mapsto$ Tuzla County $\mapsto$ Federation of Bosnia and Herzegovina $\mapsto$ Bosnia and Herzegovina $\mapsto$ Europe
  \item b. (Municipality, A) Tuzla $\mapsto$ Tuzla County $\mapsto$ Federation of Bosnia and Herzegovina $\mapsto$ Bosnia and Herzegovina $\mapsto$ Europe
  \item c. (County, A) Tuzla $\mapsto$ Federation of Bosnia and Herzegovina $\mapsto$ Bosnia and Herzegovina $\mapsto$ Europe
\end{itemize}

(6) \textbf{Britain:}

\begin{itemize}
  \item a. (Country, A) United Kingdom $\mapsto$ Europe
  \item b. (Island, T) Britain $\mapsto$ United Kingdom $\mapsto$ Europe
\end{itemize}

(7) \textbf{London:}

\begin{itemize}
  \item a. (Capital City, P; pop 7,421,209) London $\mapsto$ United Kingdom $\mapsto$ Europe
  \item b. (Administrative Area, A) London $\mapsto$ United Kingdom $\mapsto$ Europe
\end{itemize}

The six place names reference places within Europe. Sarajevo, Bosnia, Bihac and Tuzla are in Bosnia and Herzegovina, and Britain and London in the United Kingdom. Within the Bosnia and Herzegovina scope, Sarajevo and Bosnia are non-ambiguous and are resolved to places in Bosnia and Herzegovina. Bihac, Tuzla, Britain and London remain ambiguous within the selected scopes. Note for instance that there are 55 places named London in the Geonames.org database used in this work. The scope-controlled heuristic eliminates the other candidates by selecting only candidates in the top ranked scopes as the most likely referred to locations.

\section*{Country and capitals heuristic}

The country and capitals heuristic (see BK-B in Fig. 5.2) is a kind of default sense heuristic, but restricted within the selected geographic scope for the name. A place name’s candidate place of type country or national capital or provincial capital is selected as the place being referred to within the selected scope. The order of preference is: \textit{country} $\mapsto$ \textit{country capital} $\mapsto$ \textit{provincial capital}. From the previous sub-section, we had resolved Sarajevo and Bosnia, but Bihac, Tuzla, Britain and London remained ambiguous within their respective scopes. Invoking the country and capitals heuristic, Britain and London are resolved the Britain the country and London the
5.1. Toponym resolution procedure

capital city. This leaves Bihac and Tuzla still ambiguous.

Type based heuristic

The type-based heuristic (see BK-C in Fig. 5.2) exploits the types of resolved places as a basis to resolve among competing candidate places. Commonly occurring types are preferred. The assumption is that places of the same type are more likely to be mentioned in a discourse. The candidate place of the type matching the most commonly occurring type among the resolved places is selected as the place being referenced in the text. With reference to our example, the following references are already resolved Sarajevo, Bosnia, Britain and London. Bihac and Tuzla remain ambiguous, and the type-based heuristics applied to see if either of them will be resolved at this stage. The resolved places are of two types: Sarajevo and London are of type capital city, and Bosnia and Britain are of type country. Note that the system does not exploit contextual clues from text segments surrounding the names in text for both scope and referent resolution. Therefore, we may be missing important clues such as place types that accompany place names in text. The name Bihac references a city and a municipality, and Tuzla references a city, a municipality and a county all in Bosnia and Herzegovina. None of the types of Bihac and Tuzla matches any of the types of the resolved places, and therefore, the type-based heuristic fails to resolve them at this stage of processing.

Class based heuristic

The class-based heuristic (see BK-D in Fig. 5.2) exploits geographic feature classifications in the Geonames.org database. As described in Chapter 3, the Geonames.org categorises geographic features into nine broad classes:

1. Administrative unit (i.e., country, state, region, etc.), A.
2. Hydrographic (i.e., stream, lake, bay, etc.), H.
3. Locality or area (i.e., parks, area, nature reserve, etc.), L.
4. Populated place (i.e., city, town, village, etc.), P.
5. Road or railroad (i.e., road, railroad, tunnel, etc.), R.
6. Spot (i.e., spot, building, farm, etc.), S.

7. Hypsographic (i.e., mountain, hill, island, etc.), T.
8. Undersea (i.e., basin, undersea, range, etc.), U.
9. Vegetation (i.e., forest, heath, pine grove, etc.), V.

The class-based heuristic procedure is similar to the type-based heuristic. Similar to the type-based heuristic assumption, the places of a similar class are more likely to be mentioned in a discourse. And therefore, this heuristic selects the candidate place of the class matching the most frequently occurring class among the resolved places as the referenced place. From the example we have been following to illustrate referent resolution procedure, already resolved places belong to two feature classification types: Britain and Bosnia are administrative units (A), and Sarajevo and London are populated places (P). Again this does not help the resolution of Bihac with classification P, A, and Tuzla with classifications P, A, T. However, at this stage the Tuzla candidate place with classification hypsographic, T is eliminated because none of the resolved places have the classification type hypsographic, T.

Population based heuristic

The population-based heuristic (see BK-E in Fig. 5.2) is straightforward in that the place with the largest population is selected as the place being referred to. However, this heuristic is applied to candidate places of the same type and class, e.g., the town of Groningen in Germany with population of 4,166, and the town of Groningen in Suriname with population of 3,216. The population-based heuristic can be effective only when the population information of population centers is complete in the geographic database. Turning to the example referent resolution task we have followed to illustrate referent resolution procedure, Bihac and Tuzla still remain unresolved. They both have two candidate places of class populated place and administration. Because the candidate places are of different type and class, the population-based heuristic can not handle the case and passes it on to be handled at next process level.

Manual resolution

The manual resolution (see BK-F in Fig. 5.2) is the last option when all the previous automated procedures fail to solve a given ambiguity problem. The task is passed over to the user to decide the meaning of the remaining ambiguous places by exploiting other sources of information at her or his
disposal. The unresolved Bihac and Tuzla from our example is handed over to the user to perform manual resolution.

**Geographic scope update**

Because the toponym resolution component is part of the place ambiguity resolution system (i.e., the Mahali system), its output is also used improve the accuracy of the scope resolution component. Therefore, upon completion of referent resolution (see BK-G in Fig. 5.2), the list of geographic scopes is updated by including scopes containing resolved places and their ancestor scopes. The other scopes remaining in the original ranked list are discarded.

### 5.2 Evaluation metric

An evaluation metric needs to consider the number of candidate places a given place name refers to. Place names with more candidate referents are more difficult to resolve than place names with fewer candidate referents. This diversity is in terms of both the number of locations and the number of feature types. For instance, the geographical reference Groningen refers to locations in the Netherlands, Germany, Suriname, South Africa, etc. These locations are of different types as well – province, city and farm. Therefore, a metric that incorporates some of this information in its calculation could be a better judge than a scheme that uses a binary metric such as the precision and recall metrics. Ignoring this in evaluation does not pass fair judgment on systems that perform well in tackling these kinds of problems.

This section describes a new referent resolution evaluation metric that exploits the number of candidate locations, geographical hierarchy and feature type to measure the nearness of the system resolved referent to the correct referent. The research objective addressed in this section is – How best can the performance of toponym resolution systems be compared? [Question of binary vs. non-binary evaluation metric]. Consider the schematic in Figure 5.3 showing three locations $L_1$, $L_2$, $L_3$ all referred to by the name $A$. From the metric distance perspective location $L_1$ is nearer to location $L_2$ than to location $L_3$. On the other hand, from hierarchical administrative distance perspective location $L_1$ is nearer to location $L_3$ than to location $L_2$ because they are under the same administrative jurisdiction (i.e., no passport is required to travel from $L_1$ to $L_3$, but one is required to have a relevant travelling document to travel from $L_1$ to $L_2$). The interpretation of the measurements is subject to topics being addressed. The argument falls short for instance when cases of disaster are considered – in case of spread of disease
location $L_1$ is nearer to location $L_2$ than to location $L_3$.

The proposed metric incorporates the following features in its calculations:

1. The number of candidate places sharing the ambiguous place name or reference.

2. The number of regions to traverse (or hop) from the system-resolved referent to the correct gold standard referent.

3. The number of feature classes to traverse (or hop) from the system-resolved referent type to the correct gold standard referent type.

The challenge of using hierarchical structure in a measurement metric is that the structure of administrative divisions varies from country to country and/or region to region. For the proposed metric to be effective a universally agreed administrative hierarchy level is needed, for instance, first or second administrative divisions of countries.

The hierarchical structure in Figure 5.4 is used to illustrate the application of the proposed metrics. The acronyms following the names of places in Figure 5.4 represent the types of the concerned geographical features – CONT stands for continent, PCLI stands for an independent political entity.
Figure 5.4: Example hierarchy structure.
Chapter 5. Toponym resolution

(e.g., country), ADM1 stands for first administrative division of a country (e.g., province, state, district, etc.), ADM2 stands for second administrative division of a country (e.g., municipality, county, etc.). The metric in Equation 5.1 uses hop counts or number of regions traversed with no consideration of the number of candidate locations sharing the name to be resolved.

\[
\text{Hop} = \frac{\sum_{i=1}^{K} D_i}{K}
\]

\[
D_i = \frac{\sum_{j=1}^{M} H_{ij}}{M}
\]

where, \(D_i\) is the score of the \(i^{th}\) document, \(M\) is the number of references with at least one candidate location, \(H_{ij}\) is the hop count (for the \(j^{th}\) reference in the \(i^{th}\) document) from the system resolved location to the correct gold standard location, \(K\) is the total number of documents in the collection.

The hop count is defined as the number of nodes traversed from the start node to the end node, inclusive of start and end nodes, e.g., the hop count from Groningen City, Netherlands to Groningen City, Germany is 7 (see Figure 5.4). With reference to Figure 5.4, SystemA and SystemB attempt to resolve references Netherlands and Groningen (assume the correct gold standard referents are Netherlands, Europe and Groningen City, Netherlands respectively) as follows:

(8) SystemA – Netherlands \(\rightarrow\) Netherlands, Europe and Groningen \(\rightarrow\) Groningen (City), Germany.

(9) SystemB – Netherlands \(\rightarrow\) Netherlands, Missouri and Groningen \(\rightarrow\) Groningen (City), Netherlands.

Using Equation 5.1 the scores of SystemA and SystemB are computed as

\[
\text{Hop(SystemA)} = \frac{1 + 1/7}{2} = 0.57
\]

and

\[
\text{Hop(SystemB)} = \frac{1/7 + 1}{2} = 0.57
\]

respectively. The second evaluation metric extends Equation 5.1 by incorporating candidate location counts as follows:

\[
e\text{Hop} = \frac{\sum_{i=1}^{K} D_i}{K}
\]
5.2. Evaluation metric

\[ D_i = \frac{\sum_{j=1}^{M} \sqrt{\frac{N_{ij}}{H_{ij}}}}{\sum_{j=1}^{M} \sqrt{N_{ij}}} \]

where, \( D_i \) is the score of the \( i \)th document, \( M \) is the number of references with at least one candidate location, \( N_{ij} \) is the number of candidates for the \( j \)th reference in the \( i \)th document, \( H_{ij} \) is the hop count (for the \( j \)th reference in the \( i \)th document) from the system resolved location to the correct location, \( K \) is the total number of documents in the collection.

The scores of SystemA and SystemB are computed using Equation 5.2 as follows

\[ eHop(SystemA) = \frac{\sqrt{\frac{2}{4}} + \sqrt{\frac{3}{7}}}{\sqrt{2} + \sqrt{5}} = 0.62 \]

and

\[ eHop(SystemB) = \frac{\sqrt{\frac{2}{7}} + \sqrt{\frac{3}{9}}}{\sqrt{2} + \sqrt{5}} = 0.76 \]

respectively. To smooth credit points to systems that handle more difficult resolution tasks (i.e., resolving references with many candidate locations), logarithmic or square root values of candidate counts can be used, e.g.,

\[ \text{candidate counts} = \langle 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, \ldots \rangle \]

\[ \text{SQRT(candidate counts)} = \langle 1.41, 1.73, 2, 2.24, 2.45, 2.65, 2.83, \ldots \rangle \]

However, using all the candidate locations may not be appropriate as some names are shared by hundreds of locations. Therefore, a scheme that groups references according to the number of candidate locations can be used, for example,

1. references with 2-4 candidate locations are assigned a candidate count of 2 (i.e., square root of 4).
2. references with 5-9 candidate locations are assigned a candidate count of 3 (i.e., square root of 9).
3. references with 10 and more candidate locations are assigned a candidate count of 4 (i.e., square root of 16).

The motivation is that candidates within these ranges pose roughly the same resolution challenge in terms of the number of candidate locations. Using the defined reference categories, the scores of SystemA and SystemB are:

\[ eHop(SystemA) = \frac{\sqrt{\frac{2}{7}} + \sqrt{\frac{3}{7}}}{\sqrt{2} + \sqrt{3}} = 0.66 \]
Chapter 5. Toponym resolution

![Figure 5.5: Sample Geonames.org feature code hierarchy.](image)

and

$$eHop(\text{SystemB}) = \frac{\sqrt{2} + \sqrt{3}}{\sqrt{2} + \sqrt{3}} = 0.72$$

respectively.

It is not uncommon for geographical features of different types to share names and belong to the same or different administrative divisions. For example, *Groningen* is a name of a *city* and *province* in the *Netherlands*. Geographical feature types can be used to measure the accuracy of referent resolution procedures. Assume that system SystemA and SystemB resolve *Groningen* to the *town* of Groningen in *Suriname* and Groningen farm in *South Africa* respectively. The correct location being referred to is the *city* of Groningen in the *Netherlands*. Feature types *city* and *town* are classified as *populated place* and a *farm* is classified as a *spot*. From the classification perspective, SystemA can be considered a better system because it resolved *Groningen* to a type closer to the correct type.

Figure 5.5 shows the structure of the Geonames.org feature type hierarchy, where, $A$ is administrative unit, $H$ is hydrographic, $L$ is locality or area, $P$ is populated place, $R$ is road or railroad, $S$ is spot, $T$ is hypsographic, $U$ is undersea and $V$ is vegetation. The type hierarchy structure as shown in Figure 5.5 is adopted for referent resolution evaluation metric in Equation 5.3.

$$tHop = \frac{\sum_{i=1}^{K} D_i}{K}$$  \hspace{1cm} (5.3)

[^4]: [http://www.geonames.org](http://www.geonames.org)
### 5.3 Evaluation

The toponym resolution scheme proposed in this work is evaluated on three news story datasets – (1) TR-CoNLL, (2) TR-CLEF and (3) TR-RNW. The research objective addressed in this section is – *How well can automated toponym resolution systems perform against humans? [Question of algorithms; and question of gold standard evaluation]*. Table 5.1 shows the overall characteristic of document collections used for the evaluation. More detailed information on the collections is found in Chapter 3.

The precision, recall and f-score measures are used in this evaluation. The metrics proposed in Section 5.2 are yet to be implemented. Following the straightforward instantiation of the standard definition from Leidner (2007),

<table>
<thead>
<tr>
<th></th>
<th>TR-CoNLL</th>
<th>TR-CLEF</th>
<th>TR-RNW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus size (in tokens)</td>
<td>204,566</td>
<td>360,559</td>
<td>68,010</td>
</tr>
<tr>
<td>Documents</td>
<td>946</td>
<td>321</td>
<td>556</td>
</tr>
<tr>
<td>Toponym instances</td>
<td>6259</td>
<td>5783</td>
<td>2338</td>
</tr>
<tr>
<td>Unique toponyms</td>
<td>763</td>
<td>802</td>
<td>432</td>
</tr>
<tr>
<td>Ambiguous toponyms</td>
<td>629</td>
<td>690</td>
<td>332</td>
</tr>
<tr>
<td>Non-ambiguous toponyms</td>
<td>134</td>
<td>112</td>
<td>102</td>
</tr>
<tr>
<td>kappa coefficient, K</td>
<td>0.9350</td>
<td>0.7135</td>
<td>–</td>
</tr>
<tr>
<td>Human annotators</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Percentage of ambiguity</td>
<td>0.8244</td>
<td>0.8603</td>
<td>0.7650</td>
</tr>
</tbody>
</table>

Table 5.1: Characteristics of TR-CoNLL, TR-CLEF & TR-RNW.

\[
D_i = \frac{\sum_{j=1}^{M} \sqrt{\frac{N_{ij}}{H_{ij} \times T_{ij}}}}{\sum_{j=1}^{M} \sqrt{N_{ij}}}
\]

where, \( D_i \) is the score of the \( i^{th} \) document, \( M \) is the number of references with at least one candidate location, \( N_{ij} \) is the number of candidates for the \( j^{th} \) reference in the \( i^{th} \) document, \( H_{ij} \) is the hop count (for the \( j^{th} \) reference in the \( i^{th} \) document) from the system resolved location to the correct location, \( K \) is the total number of documents in the collection, \( T_{ij} \) is the number of type nodes to traverse from the system resolved type to correct gold standard type. The scores of SystemA and SystemB in Examples (8) and (9) with Equation 5.3 are 0.62 and 0.68 respectively. The metric in Equation 5.3 penalized SystemB for wrong feature type, therefore, reducing its score from that scored by metric in Equation 5.2.
Chapter 5. Toponym resolution

**Precision**, $P$ is the ratio of the number of correctly resolved toponym instances, $T_C$ and the number of toponym instances that the system attempted to resolve (either correctly, $T_C$ or incorrectly, $T_I$)

$$P = \frac{T_C}{T_C + T_I} \quad (5.4)$$

**Recall**, $R$ is the ratio of the number of correctly resolved toponym instances, $T_C$ and the number of all toponym instances, $T_N$ (i.e., the number of resolvable toponyms in a text document or corpus)

$$R = \frac{T_C}{T_N} \quad (5.5)$$

Note that $T_N = T_C + T_I + T_U$ where $T_U$ is the number of toponym occurrences whose candidate referents are unresolved.

**F-Score**, $F_\beta$ for precision $P$ and recall $R$ is defined as

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad (5.6)$$

and the F-Score at $\beta = 1$ is

$$F_{\beta=1} = \frac{2PR}{P + R} \quad (5.7)$$

and all the F-Score values reported here are computed at $\beta = 1$.

**Evaluation on TR-CoNLL corpus**

The toponyms in TR-CoNLL corpus are annotated with a spatial footprint (i.e., latitude and longitude) information from the TextGIS (Leidner, 2007). The TextGIS gazetteer is created from the GNIS 5 gazetteer of the U.S. Geographic Survey and the GNS 6 gazetteers of the National Geospatial Intelligence Agency (NGA), and supplemented with the 267 CIA World Factbook 7 (WFB) country centroid. On the other hand, the scope resolution strategies and referent resolution routines proposed in this work use the Geonames.org 8 database as the source of their spatial information. The difference

5http://geonames.usgs.gov/domestic/index.html
8http://www.geonames.org
5.3. Evaluation

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leidner (2007)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAND</td>
<td>0.2973</td>
<td>0.2973</td>
<td>0.2973</td>
</tr>
<tr>
<td>MAXPOP</td>
<td>0.6506</td>
<td>0.1976</td>
<td>0.3032</td>
</tr>
<tr>
<td>LSW03</td>
<td>0.3650</td>
<td>0.3177</td>
<td>0.3397</td>
</tr>
<tr>
<td>POP</td>
<td>0.6829</td>
<td>0.2864</td>
<td>0.4035</td>
</tr>
<tr>
<td>CCC</td>
<td>0.6305</td>
<td>0.5326</td>
<td>0.5774</td>
</tr>
<tr>
<td>BK-A</td>
<td>0.5230</td>
<td>0.4409</td>
<td>0.4785</td>
</tr>
<tr>
<td>BK-AB</td>
<td>0.7529</td>
<td>0.6359</td>
<td>0.6895</td>
</tr>
<tr>
<td>BK-ABCD</td>
<td>0.7744</td>
<td>0.6541</td>
<td>0.7092</td>
</tr>
<tr>
<td>Mahali</td>
<td>0.7754</td>
<td>0.6549</td>
<td>0.7101</td>
</tr>
</tbody>
</table>

Table 5.2: Toponym resolution results on TR-CoNLL. POP and CCC are baselines, BK-A uses scope restriction, BK-B uses countries and capitals, and the rest combine heuristics. See text for details.

in geographical information results in spatial information mismatch at evaluation. For example, the coordinate of Seoul, the capital of South Korea is captured in the two databases as <37.5664;127.0> (in Geonames.org) and <37.5663889;126.9997222> (in TextGIS). In the evaluation with TR-CoNLL, any two referents (i.e., points in latitude/longitude representation) are assumed similar if and only if they are separated by not more than 0.04° (i.e., ≈ 4.48 Kilometers) in both latitude and longitude direction. Therefore, Seoul at <37.5664;127.0> and Seoul at <37.5663889;126.9997222> is assumed to be the same Seoul from the two databases.

The LSW03 system in Leidner (2007) is grounded on two minimality heuristics: (1) the one-referent-per-discourse heuristic that assumes that a place name mentioned in a discourse refers to the same location throughout the discourse, and (2) the spatial minimality heuristic that assumes that, in cases where there is more than one toponym mentioned in some span of text, the smallest region that is able to ground the whole set is the one that gives them their interpretation. The RAND and MAXPOP formed the baseline heuristics in Leidner (2007). The RAND heuristic selects a random referent if at least one referent was found in the gazetteer, and the MAXPOP picks the candidate referent with the largest population. Table 5.2 shows the performance of the various models of the new proposed toponym resolution strategy in comparison to RAND, MAXPOP and LSW03. The RAND, MAXPOP and LSW03 data are from Leidner (2007).

The top block in Table 5.2 shows the performance data from Leidner (2007) on the TR-CoNLL corpus, and the bottom block shows the perfor-
performance of the Mahali \(^9\) toponym resolution component on the TR-CoNLL corpus. The heuristics CCC and POP are treated as the baseline schemes. CCC selects the candidate location that refers to country, national capital or provincial capital as the place referred to. The POP heuristic picks the candidate location that has the largest population as the place referred to. For schemes BK-A (scope restriction), BK-AB (i.e., combination of BK-A and BK-B, countries and capitals), BK-ABCD (i.e., combination of BK-A, BK-B, BK-C and BK-D) and Mahali (i.e., when all models are activated) refer to Figure 5.2. The baseline heuristic CCC is very competitive on the TR-CoNLL corpus. This is the reflection of the types of toponyms used in stories with global scopes (i.e., countries and capital cities are commonly mentioned in stories with global scopes). The TR-CoNLL corpus consists of stories with scopes of the global, and therefore, a good performance is expected with CCC heuristic. The performance of the POP heuristics matches the performance of MAXPOP heuristic in Leidner (2007). The BK-A heuristic which selects candidate locations found in the top ranked scopes shows a good performance as well. And as more heuristics are combined the performance increases smoothly as seen in BK-AB, BK-ABCD and Mahali. Overall the scheme proposed in this work shows that it is very competitive.

Caution is needed in interpreting the performance of the toponym resolution strategy proposed in this thesis when it is compared to Leidner (2007) scheme on TR-CoNLL corpus because of the following disparities for which we have no explanation:

1. Leidner (2007) counted 6,980 toponym instances in TR-CoNLL while the proposed system counted 6,259 toponym instances.

2. Leidner (2007) counted 1,299 unique toponyms in TR-CoNLL while the proposed system counted 763 unique toponyms.

3. The toponyms in TR-CoNLL are annotated with TextGIS gazetteer, and the proposed system uses the Geonames.org database as its geographical knowledge.

### Evaluation on TR-RNW and TR-CLEF corpora

Table 5.3 shows the performances of the new toponym resolution models on TR-RNW and TR-CLEF corpora. As noted on the evaluation on TR-CoNLL, the CCC heuristic is competitive on both the TR-RNW and TR-CLEF. On TR-RNW, CCC heuristic is highly robust at precision, recall and

\(^9\)All the components developed in the course of this work form part of the system called Mahali system. Mahali means ‘place’ in Kiswahili.
5.4. The big picture

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Heuristic</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR-RNW</td>
<td>POP</td>
<td>0.7666</td>
<td>0.2642</td>
<td>0.3930</td>
</tr>
<tr>
<td></td>
<td>CCC</td>
<td>0.8279</td>
<td>0.6373</td>
<td>0.7202</td>
</tr>
<tr>
<td></td>
<td>BK-A</td>
<td>0.7973</td>
<td>0.6137</td>
<td>0.6936</td>
</tr>
<tr>
<td></td>
<td>BK-AB</td>
<td>0.9171</td>
<td>0.7060</td>
<td>0.7978</td>
</tr>
<tr>
<td></td>
<td>BK-ABCD</td>
<td>0.9306</td>
<td>0.7164</td>
<td>0.8096</td>
</tr>
<tr>
<td></td>
<td>LREC08</td>
<td>0.9461</td>
<td>0.7420</td>
<td>0.8317</td>
</tr>
<tr>
<td>TR-CLEF</td>
<td>POP</td>
<td>0.8044</td>
<td>0.3525</td>
<td>0.4902</td>
</tr>
<tr>
<td></td>
<td>CCC</td>
<td>0.5898</td>
<td>0.4915</td>
<td>0.5362</td>
</tr>
<tr>
<td></td>
<td>BK-A</td>
<td>0.7572</td>
<td>0.6310</td>
<td>0.6883</td>
</tr>
<tr>
<td></td>
<td>BK-AB</td>
<td>0.8645</td>
<td>0.7204</td>
<td>0.7859</td>
</tr>
<tr>
<td></td>
<td>BK-ABCD</td>
<td>0.8830</td>
<td>0.7358</td>
<td>0.8027</td>
</tr>
<tr>
<td></td>
<td>Mahali</td>
<td>0.8971</td>
<td>0.7726</td>
<td>0.8302</td>
</tr>
</tbody>
</table>

Table 5.3: Toponym resolution results on TR-RNW and TR-CLEF. The Mahali system combined heuristics proposed in this work.

*f-score*. The TR-RNW stories are of global scope, and being news summaries, they mostly mention countries and national capitals.

The TR-CLEF corpus is made-up of relevant stories to topics in GeoCLEF 2006 campaign, and therefore, consists of places that are more widespread in terms of type and population. This is reflected in the performance of the baseline heuristics CCC and POP. Though both CCC and POP still performs well, but BK-A out performs them on TR-CLEF. This proves that TR-CLEF corpus is a useful addition to existing resource to geographical information retrieval (GIR) community to evaluation toponym resolution systems.

5.4 The big picture

The toponym resolution forms the last piece of the Mahali system developed in the course of this work. The system consists of three components

1. *place reference recognition and classification* component which identifies and classifies name(s) as referring to place(s).

2. *geographical scope resolution* component which resolves the geographical scope(s) or coverage(s) of a document(s).

3. *toponym resolution* component which resolves a reference to a place to its exact location on the surface of the earth.
The toponym recognition component is an off-the-shelf system while the other two components are developed in the course of this work. The details of the toponym recognition component can be found in Chapter 3 Data and tools. Figure 5.6 shows the overall system configuration and how they interact to solve the problem of place name ambiguity.

5.5 Conclusion

This chapter described a new strategy to address the problem of toponym resolution. A novel evaluation metric is proposed to evaluate the toponym resolution task as well. The research objectives addressed in this chapter are

- How effective is the document’s geographical scope(s) in aiding the resolution of toponyms contained in the document. [Question of heuristics].

- How best can the performance of toponym resolution systems be compared? [Question of binary vs. non-binary evaluation metric].

- How well can automated toponym resolution systems perform when
compared to humans? [Question of algorithms; and question of gold standard evaluation].

The toponym resolution scheme proposed in this work exploits the geographical scopes assigned to documents, place types (e.g., city, mountain, etc.), classification of place (e.g., administrative unit, vegetation, etc.), population size, and frequency of non-ambiguous or resolved places to accomplish the task of resolving an ambiguous place name to the location it refers to on the surface of the earth.

The proposed scheme performed robustly on the three evaluation corpora – TR-CoNLL, TR-CLEF, TR-RNW. It is noticed that the selection of countries and national capitals as the location being referred to among competing candidate locations performs extremely well on stories with global scopes. However, on collections with multiple scopes such as the TR-CLEF collection, the performance of CCC and POP deteriorates. The new proposed scheme performed robustly on all the three corpora beating the baseline heuristics (CCC and POP) by a very significant margin.

The new proposed toponym resolution evaluation metrics integrates the following features in its calculations: (1) number of candidate places for a given reference, (2) number of regions to traverse from the system resolved referent to the correct gold standard referent, and (3) number of feature classes transversed from the system resolved referent type to the correct gold standard referent type.
Chapter 5. Toponym resolution