Geographical Scope Resolution

Geoffrey Andogah, Gosse Bouma, John Nerbonne, Elwin Koster

Alfa Informatica, University of Groningen
Broerstraat 5, P.O. Box 72, 9700 AB Groningen, The Netherlands
{g.andogah,g.bouma,j.nerbonne,e.a.koster}@rug.nl

Abstract

It is common for placenames to reference other named entities (e.g., names of people, names of organizations, etc.) and to be used as vocabulary words (e.g., city of Split). Apart from reference ambiguity, placenames are faced with the problem of referent ambiguity (i.e., a placename referring to multiple places). Many places are also referred to by multiple names (e.g., Netherlands vs. Holland). In this paper we describe an approach to place ambiguity resolution in text, i.e., place reference resolution, resolution of a document's geographical scope and placename referent resolution. The approach is composed of three components: (1) geographical tagger, (2) geographical scope resolver and (3) placename referent resolver.

1. Introduction

Placenames are highly ambiguous as they reference other named entities (e.g., names of people, names of organizations, etc.) and are commonly used as language vocabulary words (e.g., city of Split). Apart from reference ambiguity, placenames are faced with the problem of referent ambiguity (i.e., a placename referring to multiple places). Many places are also referenced by multiple names (e.g., Netherlands vs. Holland).
Before proceeding further, a brief definition of some terminology is necessary:

Place reference recognition and classification (PRRC): The process of recognizing names in text and classifying them as place names as opposed to names of other entities.

Place referent ambiguity resolution (PARR): The 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.

Geographic scope resolution (GSR): The process of assigning a geographical region or area to a document for which the document is geographically relevant.

We describe an approach to place ambiguity resolution in text consisting of three components: (1) a geographical tagger, (2) a geographical scope resolver, and (3) a placename referent resolver. The last two components were built in-house while the first component is off-the-shelf software. Figure 1 shows the overall system architecture where the slanted boxes with dashed line boundaries are system outputs at various stages of processing.
Non-ambiguous geographical information (e.g., geographical scopes and placename referents) could improve the performance of standard information retrieval (IR) systems where the answer to the user's information need is geographically restricted (e.g., retrieving documents about "cities along river Nile") (Mandl et al., 2007). Placenames, geographic scopes (geo-scopes) and placename referents are used in query processing, document retrieval, document ranking and document visualization (Martins et al., 2006; Andogah and Bouma, 2007; Cardoso et al., 2007; Graupmann and Schenkel, 2006; Fu et al., 2005; Larson et al., 2006). The GSR approach reported in this paper exploits placename frequency of occurrence, geographical adjectives, place type (e.g., city), place importance (e.g., based-on population size and place type), and vertical (transitive parent/child) and horizontal (adjacency) relationships among places. On the other hand PRAR exploits geo-scopes assigned to documents, place type, place classification, place population and frequency information (e.g., counts of types of non-ambiguous places). Our GSR is implemented using a standard information retrieval (IR) library whilst our PRAR component is composed of simple heuristics. As mentioned before, the geographical tagger used is an off-the-shelf software[1] component pre-trained to mark place names, organization names and person names in text.
Our system is innovative in a few ways: (1) the GSR uses unresolved place names to resolve geographical scopes of

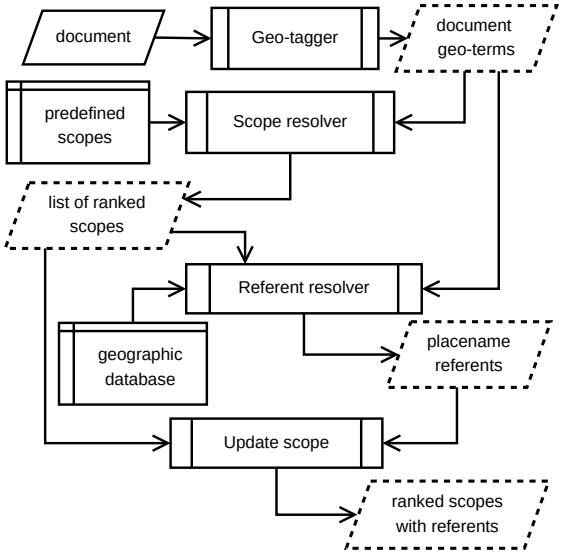

Figure 1: Placename ambiguity resolution system architecture.

[1] http://alias-i.com/lingpipe/

documents, (2) the GSR is implemented using a standard IR library, (3) the PRAR uses an elaborate range of geographical scopes assigned to a document as a basis to perform referent resolution and (4) the PRAR also makes extensive use of place types and classification to resolve among competing candidate places.

2. Geographic Scope Resolver

The geo-scope resolution approach discussed in this paper is based on Assumption 1.

Assumption 1 *Places of the same type or under the same administrative jurisdiction or near/adjacent to each other are more likely to be mentioned in a given discourse. For example, a discourse mentioning The Netherlands is most likely to mention places of the type country (e.g., Spain, Uganda) or places under the jurisdiction of The Netherlands (e.g., Amsterdam, Rotterdam) or places adjacent to The Netherlands (e.g., Belgium, Germany).*

To implement the assumption, six groups of geo-scope are pre-defined at administrative (i.e., continent, country, province) and directional (i.e., at continent, country, province) levels. Province is used in a broader sense to mean first order administrative division of a country. The pre-defined geo-scopes are indexed and searched using the Apache Lucene IR library.

2.1. Apache Lucene

Lucene's default similarity measure is derived from the vector space model (VSM). The VSM is a classic document and query modeling technique in IR systems. In VSM both the document and query are viewed as vectors (i.e., terms obtained from document and query texts with associated weights) in a multi-dimensional space (Lee et al., 1997).
The Lucene similarity score formula combines several factors to determine the document score for a query (Gospodnetic and Hatcher, 2005):

$$Sim(q,d) = \sum_{t\,in\,q} tf(t\,in\,d)\,.\,idf(t)\,.\,bst\,.\,lN(t.field\,in\,d) \quad (1)$$

where, $tf(t\,in\,d)$ is the term frequency factor for term t in document d, $idf(t)$ is the inverse document frequency of term t, bst is the field boost set during indexing and $lN(t.field\,in\,d)$ is the normalization value of a field given the number of terms in the field. In our implementation we leverage Lucene's capability to query on multiple fields and query term boosting.

2.2. Geographical knowledge

The Geonames.org[2] database is used as the basis of our geographical knowledge. It contains over eight million geographical names and consists of 6.5 million unique features including 2.2 million populated places and 1.8 million alternate names. All the features are categorized into one of nine feature classes and further subcategorized into one of 645 feature codes. We used features of the class administrative division (A) and populated place (P) to define geo-scopes.

Feature class	No. features	Unique names
All classes	6,603,579	4,230,969
Class A & P	2,564,814	1,640,422
Class P	2,393,808	1,565,458
Class A	171,006	144,684

Table 1: Geonames.org feature class A & P statistics.

Name type	No. features	Unique names
Standard	6,603,579	4,230,969
Alternative (EN)	1,237,759	1,735,528

Table 2: Geonames.org standard and alternate names statistics.

Tables 1 & 2 respectively show feature class and name statistics. Standard names are the feature names in the main Geonames.org database whilst alternative names consists of English name alternatives. Standard names have a one-to-many relationship with geographical features whilst alternative names stand in a many-to-one relationship with geographical features. Alternative names provide many surface forms of the name (e.g., Netherlands, the Netherlands, etc.). On the other hand standard names are more broad and may include feature specific qualifiers (e.g., Kingdom of the Netherlands, etc.). It is easier to find document placenames matching alternative names than standard names since people commonly use the shorter forms of placenames in documents.

2.3. Defining Geo-scopes

In this paper geo-scopes are limited to: (1) continent (CT) e.g., Europe, (2) continent directional (CD) as defined by the UN-statistics division[3] e.g., Western Europe, (3) country (PC) e.g., Netherlands, (4) country directional (PD) e.g., north-east-of Netherlands, (5) province (AM) e.g., Groningen and (6) province directional (AD) e.g., north-of Groningen. For directionally oriented scopes at country and province levels, the regions are divided into nine sections: north, north-east, east, south-east, south, south-west, west, north-west, and central.

2.3.1. Continent and continent-directional scopes

Continent and continent-directional scopes consists of the following constituents: continent, countries, country-capitals (LC), provinces, provincial-capitals (LA) and cities with over 49,999 inhabitants. Table 3 shows the distribution of scopes, locations and names at continent and continent-directional level. The average ratio of name-to-location within the scopes is 4.68. There are 7 continent scopes compared to 24 continent-directional scopes.

2.3.2. Country and country-directional scopes

Each country scope is defined by its child constituents, parent continent and adjacent countries. And each country-directional scope is defined by its child constituents and parent country. The following make up country and

[2] http://www.geonames.org

[3] http://unstats.un.org/unsd/default.htm

country-directional child constituents: country, country-capital, provinces, provincial-capitals, counties and cities with over 9,999 inhabitants. Distribution of scopes, locations and names at country and country-directional level is depicted in Table 3. The average ratio of name-to-location within the scopes is 1.73. There are 190 country scopes compared to 1089 country-directional scopes.

2.3.3. Province and province-directional scopes

Each province scope is defined by its child constituents, parent country, and adjacent provinces. And each province-directional scope is defined by its child constituents and parent province. Province and province-directional consist of the following child constituents: province, provincial capitals, country-capitals, counties and all populated places. Table 3 shows the distribution of scopes, locations and names at province and province-directional level. The average ratio of name-to-location within the scopes is 1.02. There are 4,749 province scopes compared to 20,761 province-directional scopes.

Scope	No. scopes	No. places	No. names
CT	7	13,226	61,939
CD	24	13,226	61,990
PC	190	105,576	182,442
PD	1,089	105,569	182,442
AM	4,749	2,311,244	2,354,716
AD	20,761	2,005,682	2,068,732

Table 3: Geographic scope statistics. [see Section 2.3. for scope abbreviations.]

2.4. Storing Geo-scopes in Lucene Index

Each geo-scope group (e.g., continent scope) is stored in a separate index. Lucene provides the capability to query across multiple indexes. Ten Lucene fields are defined to store geo-scope data in the index: (1) scope-id (ID), (2) names of the scope (SNM), (3) names of capitals and populated places (i.e., cities, towns & villages) with large population (CNM), (4) names of primary administrative units (PAN), (5) names of secondary administrative units (SAN), (6) names of primary cities, towns and villages (PCN), (7) names of secondary cities, towns and villages (SCN), (8) names of adjacent regions of the same type (ASN), (9) names of parent regions (PRN) and (10) names of relatively smaller child places (CPN). The type of a place (e.g., capital city, provincial capital) and population size is used to group places within a scope category. For example to populate CNM field; cities, towns and villages with over 500.000 inhabitants are considered in country scope while the threshold is lowered to 100.000 inhabitants in province scope. Table 4 shows an example Lucene index data for the scope Europe. A complete geo-scope data storage layout inside the Lucene index is shown in Table 5.

Field	Data
ID	EU
SNM	Europe, EU, Europa, etc.
CNM	–
PAN	Netherlands, Germany, Belgium, etc.
SAN	Groningen, Sachsen, Antwerp, etc.
PCN	Amsterdam, Berlin, Brussels, etc.
SCN	Utrecht, Hamburg, Antwerp, etc.
ASN	Africa, Asia, North America
PRN	Earth
CPN	Delft, Tournai, Unna, etc.

Table 4: Example Lucene Index for scope Europe. [see Section 2.4. for acronym explanation.]

2.5. Resolving document scopes

The general idea is to assign each document to geo-scopes in the Lucene index. This basically involves three steps: (1) extracting place names, place types and geographical adjectives from the document using the geographical tagger, (2) submitting extracted geographical information to query the Lucene index of pre-defined geo-scopes, and (3) returning a ranked list of geo-scopes for the document. To effectly resolve a document's geo-scope with the approach reported in this paper, query formulation is crucial. The following features are considered in our query formulation strategy: (1) perceived importance of Lucene field (2) type of place, (3) importance of place determined by population and (4) the number of occurrences of place name in a document. The importance of assigning different weights to fields comes into play when the same place takes different roles in different scopes e.g., in the hierarchy `Groningen` $\mapsto$ `Netherlands` $\mapsto$ `Europe` $\mapsto$ `Earth`, `Groningen` is a primary administrative unit in `Netherlands` while a secondary administrative unit within `Europe`. That is, `Groningen` carries more importance within the scope `the Netherlands` in comparison to the scope within `Europe`. Importance is assigned to Lucene fields in the following order (i.e., descending order of importance): SNM $\mapsto$ CNM $\mapsto$ PCN $\mapsto$ PAN $\mapsto$ SCN $\mapsto$ SAN $\mapsto$ PRN $\mapsto$ CPN $\mapsto$ ASN. And weights are assigned to types of places according to the following order (i.e., descending order of importance): CT $\mapsto$ PC $\mapsto$ LC $\mapsto$ LA $\mapsto$ AM $\mapsto$ A2. Other cities are assigned weights according to their population size.

The aforementioned features are factored into our query formulation strategy as query term boost factor using Equation 2:

$$QueryGeoTermBoostFactor = tf * FWT * GWT \quad (2)$$

where tf is the place name frequency count in the document, FWT is the weight of the Lucene field being queried against and GWT is place type or importance weight. Besides query formulation we pay attention to how the index is searched. Each geographical term in the query is analyzed to determine which field or fields to query against (e.g., Netherlands is submitted to search the field values of SNM and PAN as the Netherlands can be the name of scope Netherlands or the name of a primary administrative unit in scope Europe). Table 6 depicts feature weights implemented in our query formulation strategy. Geographical

Scopes ↦	CT	CD	PC	PD	AM	AD
ID	CT-ID	CD-ID	PC-ID	PD-ID	AM-ID	AD-ID
SNM	CT		PC		AM	
CNM			LC,P500	LC,P500	LA,LC,P150[a]	LA,P150
PAN	PC	PC	AM	AM	A2	A2
SAN	AM	AM	A2[b]	A2		
PCN	LC,P500[c]	LC,P500	LA,P100	LA,P100	P50	P50
SCN	LA,P100[d]	LA,P100	P50	P50	P5[e],P10	P5,P10
ASN	CT		PC		AM	
PRN	EH[f]	CT	CT	PC	PC	AM
CPN	P50[g]	P50	P10[h]	P10	P0[i]	P0

[a] P150: Population centers (population $\geq$ 100,000).
[b] A2: Second order administrative division of a country.
[c] P500: Population centers (population $\geq$ 500,000).
[d] P100: Population centers (100,000 $\leq$ population $<$ 500,000).
[e] P5: Population centers (5,000 $\leq$ population $<$ 10,000).
[f] EH: Earth.
[g] P50: Population centers (50,000 $\leq$ population $<$ 100,000).
[h] P10: Population centers (10,000 $\leq$ population $<$ 50,000).
[i] P0: Population centers (population $<$ 5,000).

Table 5: Geo-scope data layout in Lucene index. [see Section 2.4. for explanations of acronyms.]

Field	FWT	Type/Population	GWT
ID	-	CT	10.0
SNM	10.0	Country	9.0
CNM	9.0	Province	2.5
PAN	5.0	County	1.5
SAN	3.0	CountryCapital	9.0
PCN	8.0	ProvinceCapital	7.0
SCN	5.0	people $\geq$ 1M	9.0
ASN	1.5	0.5M $\leq$ people $<$ 1M	8.0
PRN	2.0	0.1M $\leq$ people $<$ 0.5M	7.0
CPN	2.0	50K $\leq$ people $<$ 100K	6.0
		10K $\leq$ people $<$ 50K	5.0
		5K $\leq$ people $<$ 10K	2.0
		people $<$ 5K	1.0

Table 6: Field and place type weights. [see Section 2.4. for explanations of acronyms.]

adjectives, like placenames are highly ambiguous – seeing the geographical adjective `French` in a document does not necessarily refer to things explicitly connected to the nation of France (e.g., `French` in a document may refer to a subject in school or a type of cooking). Nevertheless, if used judicially, geographical adjectives can provide useful information to geographically resolve document scopes. We map query geographical adjectives (e.g., Dutch) and placename abbreviations (e.g., UK) to their corresponding country names (e.g., Dutch mapped-to Netherlands) and assign lower weights to them. We did not try to resolve geographical adjective ambiguities, instead we assume that the places the adjective is referring to are mentioned in the document and therefore, the geo-scope resolver will use the adjective to further reinforce scope resolution.

To illustrate our geo-scope resolution approach, consider a sample document containing the following placenames with their respective term frequency in brackets: `New York (1), Rwanda (4), France (1), Kigali (1)`[4]. Table 7 depicts how query geographical terms are analyzed per field at querying processing. Each geographical term is assigned a weight (in square brackets) according to Equation 2. The document is geographically resolved to ranked geo-scopes as: `Rwanda (0.082667), Eastern Africa (0.007700), Africa (0.004359), France (0.003444), United States (0.001750).`

Field	Query Formulation
ID	-
SNM	new york[25.0] rwanda[360.0] france[90.0]
CNM	kigali[81.0]
PAN	new york[12.5] rwanda[180.0] france[45.0]
SAN	new york[7.5]
PCN	new york[56.0] kigali[72.0]
SCN	new york[35.0]
ASN	new york[3.75] rwanda[54.0] france[13.5]
PRN	new york[5.0] rwanda[72.0] france[18.0]
CPN	new york[12.0]

Table 7: Example query formulation for per field querying. [see Section 2.4. for explanations of acronyms.]

3. Placename Referent Resolver (PRR)

The placename referent resolver is a component that performs the PRAR task. PRR is fed the output of the (GeoSR) geographical scope resolver (i.e., a list of ranked document geo-scopes) and the output of the geographical tagger (i.e.,

[4] New York (State or City), Rwanda (Country), France (Country), Kigali (Country capital)

a list of place names extracted from the document) (see Figure 1). Figure 2 shows the algorithm to realize PRAR.

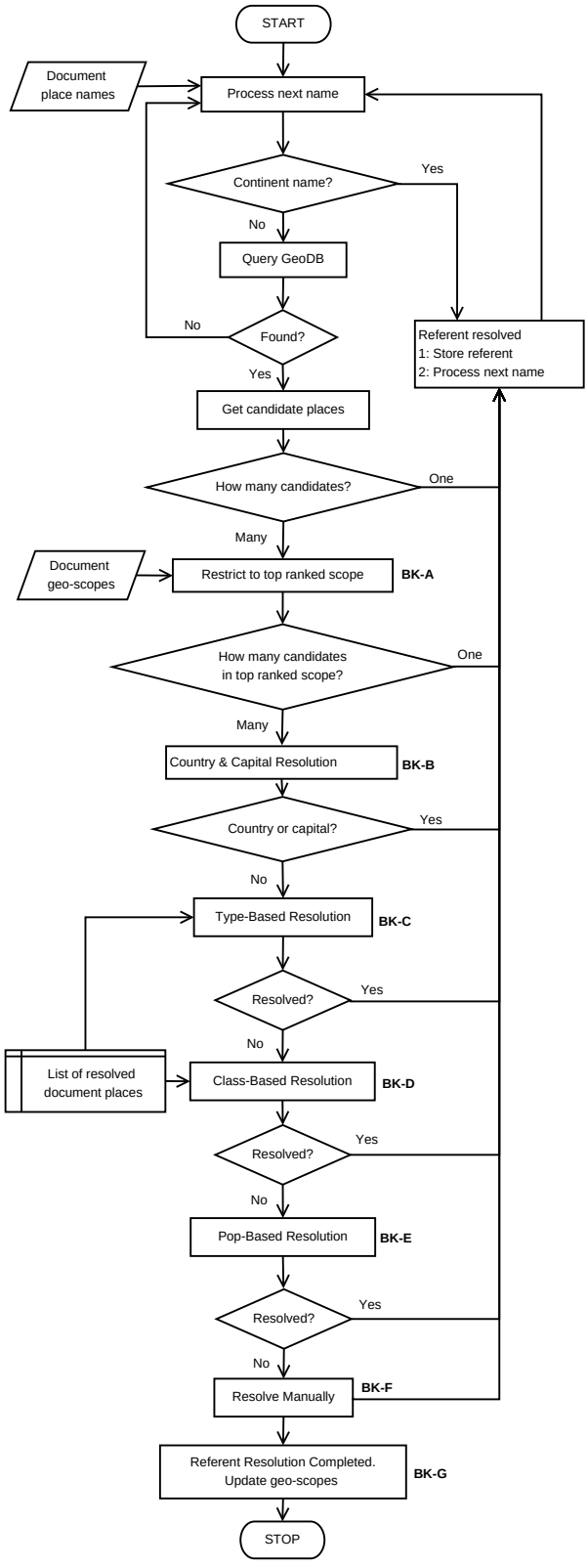

Figure 2: Referent ambiguity resolution algorithm.

Here we describe the functionality of the main processing blocks shown in Figure 2. The algorithm starts by assigning continent place names to continents. It then extracts candidate places for place names other than continent names from the geographical database (GeoDB). Place names with a single candidate place are resolved to these places while place names with multiple candidate places are passed to lower processing blocks starting with the scope restriction block. For illustration purposes, we use a sample document containing the following place names: `Sarajevo, Bosnia, Bihac, Tuzla, Britain, London`.

Scope restriction block (BK-A): This module extends the country-level restriction as reported in Pauliquen et al (2006). It exploits an elaborate list of ranked geographical scopes assigned to a document. A place name with multiple candidate referents is assigned to a single top ranked document geo-scope. The other candidates belonging to lower ranked document geo-scopes are discarded. If a selected scope contains a single candidate, the candidate is marked as the place being referred to by the name. The main source of error when using scope restriction arises from an inherited GeoSR error. However, if a selected scope contains multiple candidates, it is passed to the next processing block i.e., country & capitals resolution (BK-B). Back to our example above, the place names are restricted to the following scopes (scopes are presented as `NAME:COUNTRY@PROVINCE[CANDIDATE IDs]`): `Sarajevo:BA@01[1], Bosnia:BA@00[2], Bihac:BA@01[3], Tuzla:BA@01[4,5,6], Britain:GB@00[7,8], London:GB@H9[9,10].` `Sarajevo, Bosnia and Bihac` are non-ambiguously resolved through scope restriction because the assigned scopes contain one candidate place each. On the other hand, `Tuzla, Britain and London` remain ambiguous within selected scopes because they contain multiple candidate places.

Country & capitals resolution (BK-B): A place name's candidate place of type country (PC) or country-capital (LC) or provincial capital (LA) is selected as the place being referred to by the name. The order of preference is PC $\mapsto$ LC $\mapsto$ LA. If the ambiguity is not resolved at this stage, it is passed to the next processing block i.e., Type-based resolution (BK-C). Back to our example above; we select any candidates for `Tuzla, Britain and London` which are of type PC or LC or LA as the referent. This routine resolves `Britain and London` to places of type PC and LC respectively. `Tuzla` remains ambiguous within the selected scope.

Type-based resolution (BK-C): Type-based resolution exploits types of resolved places as the basis to resolve among competing candidate places. The commonly occurring types are preferred. The assumption is that places of a similar type are more likely to be mentioned in a discourse. The candidate place of type matching the commonly occurring type among the resolved places is selected as the place being referenced. Back to our example above; here is the list of already resolved referents with their types in curly brackets: `Sarajevo{PPLC}, Bosnia{PCLI}, Bihac{PPL}, Britain{PCLI}, London{PPLC}`. From this list there are two places of type `PPLC`, two places of type `PCLI` and one place of type `PPL`. The ambiguous `Tuzla:BA@01[4,5,6]` has

three candidate places in scope BA@01. The types of these candidate places are (candidate ID in square bracket and type in curly bracket): [4]{PPL}, [5]{ADM2} and [6]{ADM3}. Candidate [4]'s type matches one of the types of resolved referents and therefore, is selected as the place referred to by name Tuzla.

Class-based resolution (BK-D): The class-based resolution procedure is similar to the type-based resolution routine. The class-based procedure exploits feature classification of resolved places as the basis to resolve among competing candidate places (see Sec. 2.2. for feature classification detail). Again the assumption is that places of a similar class are more likely to be mentioned in a discourse. The candidate place of class matching the most frequently occurring class among the resolved places is selected as the place referred to. Back to our example above; we will try to resolve among the three candidates of Tuzla in scope BA@01 employing the class-based procedure. Here is a list of resolved places with their corresponding class in curly brackets: Sarajevo{P}, Bosnia{A}, Bihac{P}, Britain{A}, London{P}. There are two places classified as A and three places classified as P. The three candidates of reference Tuzla are classified as (candidate ID in square bracket and classification in curly bracket): [4]{P}, [5]{A} & [6]{A}. Candidate [4]'s class matches the most frequently occurring class among the resolved places and therefore, is selected as the place referred to by name Tuzla.

Pop-based resolution (BK-E) & manual resolution (BK-F): Population based resolution (BK-E) selects the place with the largest population as the place being referred to. Manual resolution (BK-F) passes the task of resolving among competing places to the user. Manual resolution is called when the preceding automated procedures fail to resolve the ambiguity.

Update geo-scopes (BK-G): Here the list of a document's ranked geographical scopes is updated by including only the scopes containing resolved places and their ancestor geo-scopes. The remaining geo-scopes in the ranked list are discarded. From our example above, scope list update with respect to London and Britain will include: Europe, GB@00, GB@H9, GB@S.East, Northern Europe, GB@H9@S.East. The following scopes in the original ranked scope list are discarded: CA@East, CA@08, CA@08@S.East, CA@00 where GB and CA stand for Great Britain and Canada respectively. The scope Canada featured in the original scope list because of a place named London in Ontario, Canada.

4. Evaluation

Here we report on geographical scope resolver (GSR) evaluation. Because of time constraints and lack of test dataset, we were unable to fully evaluate placename referent resolver (PRR) for this paper. However, a preliminary test on 102 documents containing 195 ambiguous place names, our PRR resolved 181 (92.8%) of the place names correctly[5].

[5]A comprehensive evaluation of our placename referent resolver (PRR) will be reported in the PhD thesis in preparation.

4.1. GSR Evaluation

4.1.1. Dataset

We evaluated our implementation using the CoNLL-2003 Shared Task (Sang and Meulder, 2003) training and development set of 1162 documents for English. The CoNLL-2003 English dataset is derived from the Reuters English corpus (RCV1) (Rose et al., 2002). Of the 1162 documents, 1124 documents contain geographical terms (place names and geographical adjectives). These documents have geographical scopes at country levels assigned to them. Of 1124 documents 686 were assigned single scopes, 313 double, 90 triple and 35 four or more.

4.1.2. Results

Our system can assign geographical scopes up to six levels: continent, continent-directional, country, country-directional, province and province-directional. For this evaluation, we turned on the country level scope resolver for that is the scope level assigned to our test document collection. Our system resolves documents geographically to multiple scopes ranking them from the most significant to the least significant scope.

Single Scoped Documents. Of the 686 documents with single scope, our system assigned scopes correctly to 645 (94%) documents (that is, the scopes assigned to the 645 documents were ranked at position one).

Two Scoped Documents. Of the 313 documents with two scopes, our system assigned scopes correctly to 197 (62.94%) documents (that is, the scopes assigned to the 197 documents were ranked at the top two positions). The remaining 116 (37.06%) documents had one scope correctly assigned to them in the top two rank postions.

Three Scope Documents. Of the 90 documents with three scopes, our system assigned scopes correctly to 18 (20%) documents (that is, the scopes assigned to the 18 document were ranked at the top three positions). Of the remaining 72 documents, 48 (53.33%) documents were correctly assigned two scopes in the top three rank postions. The remaining 24 (26.67%) documents had one scope correctly assigned to them in the top three rank positions.

5. Conclusion

We described a complete placename ambiguity resolution system consisting of three components: a geographical tagger, a geographical scope resolver (GeoSR) and a placename referent resolver (PRR). The last two components are built in-house while the geographical tagger is an off-the-shelf software component.

The novelty in GeoSR is that it uses unresolved place names as opposed to resolved place names used in previous works (Amitay et al., 2004; Martins and Silva, 2005). This means that geographical scopes can be computed independent of geographic name resolution, and thus does not suffer from mistakes in placename resolution. Also the GeoSR is implemented using a standard IR library exploiting a number of features, namely, placename frequency of occurrence, geographical adjectives, place type, population, vertical (transitive parent/child relation) and horizontal (adjacency relation) relationship among places. The GeoSR

achieved a promising result on a subset of the Reuters English corpus (RCV1) dataset comparable with (Amitay et al., 2004; Martins and Silva, 2005): single scoped documents (96%) and two scoped documents (62.94%). However, the system performance for a three or more scoped documents is very poor (20%).

The novelty in PRR is that it uses an elaborate list of ranked geographical scope as the basis to resolve place ambiguity. The PRR also makes extensive use of place types and classification to resolve among competing candidate places. However, we are unable to evaluate PRR because of time constraints and lack of test dataset.

Lastly, there is an urgent need for freely available datasets to evaluate referent and scope resolution approaches. The datasets should consist of various genres, e.g., news articles and webpages. Leidner's work on toponomy resolution is a step in the right direction (Leidner, 2007).

6. Acknowledgements

This work is supported by NUFFIC within the framework of The Netherlands Programme for the Institutional Strengthening of Post-secondary Training Education and Capacity (NPT) under the project titled "Building a sustainable ICT training capacity in the public universities in Uganda".

7. References

Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. 2004. Web-a-Where: Geotagging Web content. In *Proceedings of SIGIR-04, the 27th conference on research and development in information retrieval*, pages 273–280. ACM Press.

Geoffrey Andogah and Gosse Bouma. 2007. University of Groningen at GeoCLEF 2007. In *Working Notes for CLEF 2007*, Budapest, Hungary.

Nuno Cardoso, David Cruz, Marcirio Chaves, and Mário J. Silva. 2007. The University of Lisbon at GeoCLEF 2007. In *Working Notes for CLEF 2007*, Budapest, Hungary.

Gaihua Fu, Christopher B. Jones, and Alia I. Abdelmonty. 2005. Ontology-based Spatial Query Expansion in Information Retrieval. In *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*, volume 3761/2005, pages 1466–1482. Springer (Lecture Notes in Computer Science LNCS).

Otis Gospodnetic and Eric Hatcher. 2005. *Lucene in Action*. Manning Publications Co., 206 Bruce Park Avenue, Greenwich, CT 06830.

Jens Graupmann and Ralf Schenkel. 2006. GeoSphereSearch: Context-Aware Geographic Web Search. In *Proceedings of the 3rd Workshop on Geographic Information Retrieval held at The 29th Annual International ACM SIGIR Conference*, Seattle, WA, USA.

Ray R. Larson, Fredric C. Gey, and Vivien Petras. 2006. Berkeley at GeoCLEF: Logistic Regression and Fusion for Geographic Information Retrieval. In *Accessing Multilingual Information Repositories*, volume 4022/2006, pages 963–976. Springer (Lecture Notes in Computer Science LNCS).

Dik L. Lee, Huei Chuang, and Kent Seamons. 1997. Document ranking and the vector space model. *IEEE Software*, pages 67–75.

Jochen Lothar Leidner. 2007. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie. 2007. GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview . In *Working Notes for CLEF 2007*.

Bruno Martins and Mario J. Silva. 2005. A graph-ranking algorithm for geo-referencing documents. In *Proceedings of ICDM-05, the 5Th IEEE International Conference on Data Mining*.

Bruno Martins, Mario J. Silva, Sergio Freitas, and Ana Paula Afonso. 2006. Handling locations in search engine queries. In *Workshop on Geographical Information Retrieval, SIGIR'06*, August.

Bruno Pouliquen, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fluart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, and Clive Best. 2006. Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation. In *Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 53–58.

Tony Rose, Mark Stevenson, and Miles Whitehead. 2002. The Reuters Corpus Volume 1 - from yesterday's news to tomorrow's language resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, volume 3, pages 827–833.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language independent named entity recognition. In *Walter Daelemans and Miles Osborne, Editors, Proceedings of CoNLL-2003*, pages 142–147.