Unnamed Locations, Underspecified Regions, and Other Linguistic Phenomena in Geographic Annotation of Water-based Locations

Johannes Leveling
Centre for Next Generation Localisation (CNGL)
Dublin City University, Dublin 9, Ireland
johannes.leveling@computing.dcu.ie

ABSTRACT

This short paper investigates how locations in or close to water masses in topics and documents (e.g. rivers, seas, oceans) are referred to. For this study, 13 topics from the GeoCLEF topics 2005-2008 aiming at documents on rivers, oceans, or sea names were selected and the corresponding relevant documents retrieved and manually annotated.

Results of the geographic annotation indicate that i) topics aiming at locations close to water contain a wide variety of spatial relations (indicated by different prepositions), ii) unnamed locations can be generated on-the-fly by referring to movable objects (e.g. ships, planes) travelling along a path, iii) underspecified regions are referenced by proximity or distance or directional relations. In addition, several generic expressions (e.g. *"in international waters"*) are frequently used, but refer to different underspecified regions.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; H.3.4 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search Process*

Keywords

GIR, Toponyms, Annotation

1. INTRODUCTION

Events and situations described in newspaper articles are referred to mostly by names of populated places on land. Topics aiming at events happening in or close to water bodies (e.g. rivers, seas, or oceans) are less frequent but more difficult to process in a geographic information retrieval (GIR) system. This short paper investigates how locations in or close to water masses in topics and documents are typically expressed in natural language.

.

Table 1: Selected GeoCLEF topics.

Topic	Topic title	# Rel.
01-GC	*"Shark Attacks off Australia and California"*	14
16-GC	*"Oil prospecting and ecological problems in Siberia and the Caspian Sea"*	15
21-GC	*"Sea rescue in North Sea"*	29
36-GC	*"Automotive industry around the Sea of Japan"*	0
41-GC	*"Shipwrecks in the Atlantic Ocean"*	4
47-GC	*"Champions Cup games near the Mediterranean"*	24
50-GC	*"Cities along the Danube and the Rhine"*	15
53-GC	*"Scientific research at east coast Scottish Universities"*	
63-GC	*"Water quality along coastlines of the Mediterranean Sea"*	0
66-GC	*"Economy at the Bosphorus"*	3
74-GC	*"Ship traffic around the Portuguese islands"*	3
85-GC	*"Nuclear tests in the South Pacific"*	43
91-GC	*"Forest fires on Spanish islands"*	2
95-GC	*"American troops in the Persian Gulf"*	53

In contrast to events on land, events happening in or close to water masses may be more difficult to describe. Most locations in water are unnamed (except for islands), because they are not populated, and there is no deep hierarchical structure of water bodies, i.e. no natural subdivisions exist and the typical *in*-relations cannot be exploited for GIR. Thus, these topics pose a challenge: if complex spatial relations are not correctly identified and interpreted, GIR effectiveness will decrease. For instance, a location name in a document may be part of a complex natural language expression to refer to events at a different, distant location. Furthermore, relevant documents may not be found because the same place is referred to by different natural language expressions.

2. TOPIC AND DOCUMENT ANNOTATION

Thirteen topics from the GeoCLEF evaluation campaign 2005-2008 were selected [1], where in the topic title or description a sea, river, or ocean is mentioned. The selected topic titles and the number of corresponding documents assessed as relevant (# Rel.) are shown in Table 1.

Note that for most topics, descriptions of events in or close to water will be relevant. For some topics however, events on land are also relevant (e.g. 16-GC, 47-GC, 50-GC). In

Table 2: Annotation frequencies.

Type	Frequency	
Location adjectives	1752	23.05%
Location nouns	3049	40.11%
Metonymic use	1142	15.02%
Demonyms (population name)	230	3.03%
Unnamed/underspecified locations	1428	18.79%
Sum	7601	100.00%

one case (GC-63), descriptions of both types may be relevant, depending on whether *"water quality"* refers to drinking water or bathing water. For topics GC-85 and GC-95, the highest number of relevant documents has been found (96 out of 205). For two other topics (36-GC and 63-GC), no relevant documents have been found, possibly because of complex spatial relations (i.e. *"around"* and *"along"*), which can not be easily interpreted by GIR systems.

For the thirteen topics, 205 relevant documents from the LA times and Glasgow Herald from 1994 and 1995 have been assessed as relevant. The average number of relevant documents per topic is 15.8. For comparison, the average number of relevant documents per topic for the set of all topics excluding the selected topics is 29.9 (2598 relevant documents for the remaining 87 GeoCLEF topics). There are several possible explanations for this difference: 1) There may be more relevant documents in the document collection, but these documents have not been found by the systems participating at GeoCLEF, have not been included in the pooled set of documents, and thus, have not been assessed for relevance. 2) There are no more relevant documents, which means that there could be a bias in articles towards events on land and/or events happening in maritime locations may be more difficult to describe in natural language.

In the relevant documents (about 940 KB text), geographic locations have been manually annotated. The following annotation guidelines were followed for the geographic annotation of documents: Location names (toponyms), location adjectives, and names for people inhabiting a place (demonyms), are annotated, using the longest possible geographic reference (typically a noun phrase, e.g. *"in Camp Doha, north of Kuwait City"*). Ambiguous names and location names which are part of organisation or institution names are not annotated as locations (e.g. *"the carrier George Washington"*). Coordinations of locations are annotated as a single instance, except if different annotations are required (e.g. *"Paris and Berlin"* is annotated as a single instance, but *"U.S. troops and forces from Kuwait"* is not). Metonymic use of location names is annotated and verbs or nouns indicating the metonymic use are linked to the location name (e.g. *"Russia (promised) Baghdad to campaign"*). Underspecified regions and unnamed locations, including anaphoric references to places are annotated as well (e.g. *"there"*, *"in the region"*). Table 2 shows the annotation frequency.

Confirming results from Leveling and Hartrumpf for annotated German newspaper texts [2], location names in the documents are often used in a non-geographic sense and locations can be referred to by means other than proper nouns. Examples of non-literal senses include metonymic use of location names (e.g. *"France said today it would provide emergency financial aid to victims"*, *"Australia and New Zealand want the issue put high on the agenda"*) and location names as part of other names (e.g. *"Ajax Amsterdam"*, *"European Union"*). Words other than location names implying a location include adjectives (e.g. *"Alsatian cuisine"*, *"Danish Foreign Minister"*) or demonyms (names for people inhabiting a place, e.g. *"Turkish officials"*).

Table 3 shows the most frequent geographic relations together with text excerpts from the annotated documents for the location nouns and unnamed locations (4477 instances). The geographic relations which have been observed typically correspond to prepositions and include proximity (e.g. *close*, *far*, *near*, *off*), containment or meronymy (e.g. *in*, *within*), paths (e.g. *across*, *along*, *around*, *between*, *through*), path start points (e.g. *from*, *of*, *out of*), path end points (e.g. *into*, *to*, *toward*), and other spatial relations (e.g. *above*, *below*, *beneath*, *over*). In addition, numeric distances and compass directions (e.g. *north*, *southeast*) have also been found in the documents. Prepositions which occur only rarely (such as *atop*) have not been included in Table 3. The document annotation shows that *in* is the most dominant relation, but a wide range and number of prepositions corresponding to spatial paths are also frequently used.

Similar to prepositions, verbs may also indicate a path (730 instances of the annotated location nouns and unnamed locations), including full paths (e.g. *"drive/fly from X to Y"*), path start points (e.g. *"leave X"*, *"depart from X"*), path end points (e.g. *"arrive at X"*, *"land in X"*), and stopping a movement on a path (e.g. *"crash/sink/stop at X"*). Some verbs are compatible with multiple prepositions (e.g. *"evacuate from/to X"*), other verbs do not require prepositions at all (e.g. *"reach X"*). In summary, geographic relations should be identified considering the preposition preceding the location together with the associated verb and identification should preferably be based on natural language processing techniques.

3. DISCUSSION

The following observations have been made with respect to natural language description of locations in water. Due to the nature of the selected topics, locations in or near the water bodies are referred to mostly in the document parts indicating relevance. These locations include oil platforms, oil fields, and moving vessels (e.g. ships, planes, trains). However, the annotated newspaper articles do not exclusively cover stories about situations and events at sea. Therefore, many mentioned locations are actually located on land. In particular, topic 85-GC has the most relevant documents. However, the events described in the documents (nuclear tests) incite responses from all over the world, so that most of the articles are about responses and reactions of countries (which is one reason for the high number of metonymically used location names), not mainly about nuclear tests in the South Pacific.

Locations in the water can be referenced by compass directions and distance from a known location (e.g. *"about 30 miles west of San Francisco's Golden Gate Bridge"*, *"22 miles from Brae Alpha"*, *"off the North-west of Scotland"*). Some proper nouns for seas are abbreviated if the geographic reference is clear from the document context (e.g. *"Caspian"* = *"Caspian Sea"*, *"in the Gulf"* = *"in the Gulf region"*).

Underspecified regions and unnamed locations are referenced as well. In the first case, a large, possibly indeterminate area or region is referenced (e.g. *"in international*

Table 3: Relation types and relations in the annotated documents.

Type	Relation	Document text	Frequency	Sum
Proximity	*close*	*"close to Gatwick Airport"*	14	165
	far	*"far from London"*	21	
	near	*"near Death Valley"*	70	
	off	*"off San Miguel Island"*	60	
Meronymy	*in*	*"in Alaska"*	1192	1909
	within	*"within 12 1/2 miles of the Kuwaiti border"*	21	
	at	*"at Clearwater Bay"*	332	
	on	*"on Christmas Island"*	364	
Path	*across*	*"across the Pacific"*	33	171
	along	*"along the coast of Northern Europe"*	34	
	around	*"around Loch Lomond"*	44	
	between	*"between the Black and Caspian seas"*	21	
	through	*"through the Maas (Meuse) valley"*	39	
Path start point	*from*	*"from Salzburg"*	327	875
	of	*"of Dereham, Norfolk"*	465	
	out of	*"out of Saudi Arabia"*	83	
Path end point	*into*	*"into the South Atlantic"*	65	655
	to	*"to Grimsby"*	560	
	toward	*"toward Las Vegas"*	30	
Distance	–	*"40 miles south of the Iraqi border"*	84	84
Direction	–	*"just north of the Kuwaiti border"*	125	125
Other	*above*	*"above the Indian Ocean"*	8	87
	below	*"below the 32nd Parallel"*	8	
	beneath	*"beneath the South Pacific"*	9	
	over	*"over the North Sea"*	62	

waters off Mururoa atoll", *"in the waters around San Miguel Island"*), in the latter case, a specific location is referenced, but remains underspecified (e.g. *"(on the way) from Aberdeen to Marathon's Brae Alpha platform"*, *"(en route) to the Red Sea"*). These locations typically are places along a path, where the path is either implicitly given (e.g. *"from X to Y"*) or defined by artifacts including pipelines, streets, canals, and railways. Unnamed locations can be referred to by paths, often without a start or end point (e.g. *"from Aberdeen to Marathon's Brae Alpha platform"*, *"to Kuwait"*). Rarely, locations are defined by intersections of paths or regions (e.g. *"where the Amsterdam-Rhine Canal meets the Waal"*, *"where the Rhine enters the Netherlands"*). In a few instances, temporal expressions relate to spatial distance (e.g. *"an hour northeast of Zurich"*, *"before (reaching) the Rhine Bridge"*).

4. CONCLUSION AND FUTURE WORK

The annotation illustrates that documents describing events located in or near water bodies, contain complex natural language descriptions. Due to the effort involved in manual annotation, a sample of documents relevant to land-based events has not yet been annotated and a comparison with results from corresponding documents on land-based events has not yet been performed.

In some cases, the document parts implying the relevance of a document for a topic contain references to unnamed locations, which are typically locations on a path. GIR systems will profit from a deeper analysis of the text which goes beyond identification and grounding of location names only. As part of the future work, the annotation effort will be used to produce rules or annotation patterns to improve the accuracy of location identification. The interpretation of spatial prepositions and verbs in natural language expressions referring to locations will be investigated in more detail.

Furthermore, the automatic resolution of anaphoric references to locations will be investigated. For example, if coreferences are resolved and pronouns referring to a location name are replaced by the location name (instead of removing them as stopwords), term frequencies are changed which will affect the result ranking of a GIR system.

5. ACKNOWLEDGMENTS

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/CE/I1142.

6. REFERENCES

[1] F. Gey, R. Larson, M. Sanderson, H. Joho, and P. Clough. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track. In C. Peters, F. Gey, J. Gonzalo, H. Müller, G. J. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Acessing Multilingual information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF'2005. R evised Selected papers*, volume 4022 of *LNCS*, pages 908–919. Springer, 2006.

[2] J. Leveling and S. Hartrumpf. On metonymy recognition for geographic information retrieval. *International Journal of Geographical Information Science*, 22(3):289–299, 2008.