University of Hagen at GeoCLEF 2008: Combining IR and QA for Geographic Information Retrieval

Johannes Leveling and Sven Hartrumpf
Intelligent Information and Communication Systems (IICS)
University of Hagen (FernUniversität in Hagen)
58084 Hagen, Germany
firstname.lastname@fernuni-hagen.de

Abstract

This paper describes the participation of GIRSA at GeoCLEF 2008, the geographic information retrieval task at CLEF. GIRSA is a modified and improved variant of the system which participated at GeoCLEF 2007. It combines results retrieved with methods from information retrieval (IR) on geographically annotated data and question answering (QA) employing query decomposition.

For the monolingual German experiments, several parameter settings were varied: using a single index or a separate index for content and geographic annotation, using complex term weighting, adding location names from the narrative part of the topics, and merging results from IR and QA. The best mean average precision (MAP) was obtained by combining IR and QA results (0.2608 MAP).

For bilingual (English-German and Portuguese-German) experiments, topics were translated via various machine translation web services: Applied Language Solutions, Google Translate, and Promt Online Translator. Performance for these experiments is generally lower than for monolingual experiments. For both source languages, Google Translate seems to return the best translations. For English topics, 60% (0.1571 MAP) of the maximum MAP for monolingual German experiments is achieved. For bilingual Portuguese-German experiments, 80% (0.2085 MAP) of the maximum MAP for monolingual German experiments is achieved.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods; Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation; Search process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Experimentation, Measurement, Performance

Keywords

Geographic Information Retrieval, Question Answering, Cross-language Information Retrieval

1 Introduction

GeoCLEF is the geographic information retrieval (GIR) task at CLEF, the cross-language evaluation campaign. In recent years, we have developed GIRSA (Geographic Information Retrieval by Semantic Annotation), a system for exploring novel approaches at GIR. GIRSA supports methods to improve precision (e.g. annotation of metonymic location names [5]) and methods to improve recall (e.g. normalization of location name synsets [4]). For GeoCLEF 2008, the major improvement lies in the combination of results from information retrieval (IR) on geographically annotated documents with methods from question answering (QA).

2 System Description

GIRSA is a system for the evaluation of novel indexing and retrieval methods for GIR. Basically, the GIRSA setup introduced at GeoCLEF 2007 is used for the GIR experiments. This setup involves the identification and normalization of location indicators, i.e. text segments from which a geographic scope can be inferred. Location adjectives, names for inhabitants of a place, geographic codes, orthographic variants, acronyms, and abbreviations are mapped to location names. For its participation in GeoCLEF 2008, selected aspects of the IR subsystem have been improved:

- The IR indexing methods utilize an improved version of the German stemmer (in the old version, adjectives were often stemmed incorrectly due to an incorrectly implemented stemming rule).
- The resources for the identification of location indicators have been expanded. Additional lists of synonymous location names were extracted from Wikipedia articles and added to the geographic annotation data. For the normalization of multi-word names, missing inflectional variants of names were automatically generated and added. Furthermore, an automatic consistency check to find circular normalizations and other data inconsistencies was integrated and inconsistencies in the annotation data were removed (e.g. if the data contains entries to normalize *"Geneva"* to *"Genf"* and vice versa, this will be detected).
- The retrieval was modified to include a weighting scheme already used in our QA system [3]. The term weighting is meant to achieve a higher initial MAP by assigning weights according to the semantic contribution of words from the topic. Terms receive weights corresponding to their importance as follows (in order of increasing weights): lower case words (e.g. adjectives and adverbs), numeric expressions (e.g. temporal expressions), the answer subtype (similar to the expected answer type known from QA, typically the first noun from a question), nouns, and proper nouns.

The QA subsystem of GIRSA is InSicht, which also participates in QA@CLEF (see for example [2]). For the specific requirements in an IR setting, the QA system has been modified in the following ways:

- The normal processing of queries or questions stops after matching semantic representations of the query with semantic representations of documents. Answer generation is skipped because typical IR queries are not asking for answers, but for relevant documents.
- Semantic decomposition of queries, which was pioneered in the previous GeoCLEF [4], was extended by developing 6 decomposition methods aiming at improving recall for QA and/or IR (see [1] for details on the application of this approach to QA). For this year's experiments, only two decomposition methods were activated in order to reduce runtime and to avoid finding irrelevant documents. For the title of topic 91-GC (*"Waldbrände auf spanischen Inseln"*/*'Forest fires on Spanish islands'*), *description decomposition* produces the subquestion *"Nenne spanische Inseln."*/*'Name Spanish islands.'* The 14 subanswers found (e.g. *"Gran Canaria"*) are substituted on the level of semantic representations in the original question, leading to 14 revised queries, e.g. *"Waldbrände auf Gran Canaria"*. For the title of topic 96-GC (*"Wirtschaftsaufschwung in Südostasien"*/*'Economic boom in Southeast Asia'*), *meronymy decomposition* leads to subquestions like *"Welche Region/Welcher Staat/Welche Stadt liegt in Südostasien?"*/*'Which region/country/city is located in Southeast Asia?'*.

As these examples indicate, subquestions produce background knowledge (often of a geographic type) on the fly. Some pieces of knowledge are to be found in gazetteers, but there are many cases (*"Mittelmeeranrainerstaaten"*/*'Mediterranean countries'* in topic 81-GC, *"Nordafrika"*/*'Northern Africa'* in topic 83-GC, *"Südpazifik"*/*'South Pacific'* in topic 85-GC, etc.) where it is unlikely to find the relevant information in static, general-purpose gazetteers. To improve the answers for subquestions, these subquestions (in contrast to the original GeoCLEF queries) are answered also over the Wikipedia corpus used in QA@CLEF. With decomposition, 1238 documents (232 assessed as relevant) were retrieved; only 125 documents (77 assessed as relevant) without decomposition.

- The semantic network for a query can be split into two semantic networks at certain relations, e.g. splitting off temporal or local restrictions. In GeoCLEF 2007, these two parts had to be matched in the same document; this year, a NEAR operator (with 2000 characters) instead of the AND operator was applied in order to improve precision for these cases.

3 Experiments

We formulate our expectations regarding the MAP for different parameter settings in our experiments as hypotheses:

H1 *Experiments using additional location names from the narrative part of the topics will achieve a higher MAP than experiments that do not (to confirm results from GeoCLEF 2007).*

H2 *The MAP for experiments adding results from the QA subsystem will be somewhat higher than for experiments with pure GIR.*

H3 *Topic translations with the Promt Online Translator web service will be better (e.g. containing less untranslated words) than those from the other web services tested. The corresponding results will therefore have a higher MAP.*

H4 *Applying the weighting from QA (for all experiments), merging results from IR and QA, and combining indexes for location names and content words will result in a higher initial MAP.*

GIRSA was employed to produce results for a number of monolingual and bilingual experiments. The following parameter settings were varied in different retrieval experiments (see Table 1):

- language (lang.):
  German (DE), English (EN), or Portuguese (PT) serves as topic source language.
- translation (transl.):
  Applied Language Solutions[1] (A), Google Translate[2] (G), or Promt Online Translator[3] (O) was used to translate topics.
- fields:
  Content keywords and location indicators are extracted from the topic title and description: with location names from the topic narrative (TDN) or without (TD).
- index:
  - All words are stemmed; a single index is produced (A).
  - Content words are decompounded (if possible) and stemmed; location names are identified; both are indexed separately (B).
  - Content words are decompounded (if possible) and stemmed; location indicators are normalized; both are indexed separately (C).

[1] `http://www.appliedlanguage.com/free_translation.shtml`
[2] `http://translate.google.com/`
[3] `http://www.online-translator.com/`

Table 1: Results for monolingual and bilingual retrieval experiments on German GeoCLEF documents.

Run	Parameters					Results				
	lang.	transl.	fields	index	comb.	MAP	rel_ret	P@5	P@10	P@20
FUHtd01	DE	-	TD	A	N	0.2420	977	0.39	0.37	0.31
FUHtd01m	DE	-	TD	A	Y	0.2608	1028	0.38	0.37	0.35
FUHtd20	DE	-	TD	B	N	0.1719	914	0.20	0.29	0.27
FUHtd20m	DE	-	TD	B	Y	0.2211	998	0.36	0.35	0.34
FUHtdn20	DE	-	TDN	B	N	0.1478	834	0.17	0.24	0.20
FUHENAtd20	EN	A	TD	B	N	0.1076	644	0.18	0.17	0.17
FUHENAtdn20	EN	A	TDN	B	N	0.0962	610	0.14	0.15	0.13
FUHENGtdn20	EN	G	TDN	B	N	0.1571	800	0.21	0.21	0.21
FUHENOtd20	EN	O	TD	B	N	0.1179	703	0.23	0.23	0.21
FUHENOtdn20	EN	O	TDN	B	N	0.1146	699	0.21	0.21	0.19
FUHPTGtd01	PT	G	TD	A	N	0.2085	903	0.41	0.38	0.33
FUHPTGtd20	PT	G	TD	B	N	0.1776	907	0.29	0.30	0.27
FUHPTGtdn20	PT	G	TDN	B	N	0.1571	800	0.21	0.21	0.21
FUHPTGtd21	PT	G	TD	C	N	0.2002	913	0.34	0.34	0.31
FUHPTGtdn21	PT	G	TDN	C	N	0.1567	793	0.22	0.21	0.22

- combination (comb.):
  Results from IR and QA are combined (Y) or not (N).[4]

Three metrics are employed to measure retrieval performance (see Table 1):

- MAP: mean average precision,
- rel_ret: the number of relevant and retrieved documents (a total of 1417 documents was assessed as relevant for the GeoCLEF 2008 topics), and
- P@N: precision at N documents.

4 Results and Discussion

Let us revisit the hypotheses from Section 3.

H1 *Experiments using additional location names from the narrative part of the topics will achieve a higher MAP than experiments that do not (to confirm results from GeoCLEF 2007).* This turned out to be false. The MAP for experiments with additional location names from the topic narrative is lower than for the experiments using title and description only (e.g. FUHtd20 vs. FUHtdn20). Maybe additional location names from the topic narrative do not match the names in documents as exactly as in old topics; maybe too many additional location names are added, causing a topic shift. A solution would require a more elaborate weighting algorithm.

H2 *The MAP for experiments adding results from the QA subsystem will be somewhat higher than for experiments with pure GIR.* This is also not true: performance is considerably higher due to the improvements in the QA subsystem (query decomposition, less strict matching). The MAP for merged runs is higher in all cases. FUHtd01m shows a relative improvement of 7.8% in MAP compared to FUHtd01, FUHtd20m shows an improvement of 28.6% compared to FUHtd20; also, more relevant documents are retrieved in both cases. InSicht found documents for 13 (of the 25) topics, which is much better than last year. These results alone are not sufficient for GIR, but due to their high complementarity merging these results improves GIRSA significantly.

[4]To merge, the maximum score of results is chosen (for duplicate results), and the top-1000 documents are returned.

H3 *Topic translations with the Promt Online Translator web service will be better (e.g. containing less untranslated words) than those from the other web services tested. The corresponding results will therefore have a higher MAP.* The MAP for the best bilingual English-German experiment is 0.1571 (about 60% of the best MAP for monolingual German); the MAP for the best bilingual Portuguese-German experiment is 0.2085 (about 80% compared to monolingual German). The highest MAP was achieved with Google Translate. The experiments with topics translated by Google Translate returned the best results (FUHENGtdn20 vs. FUHENOtdn20 vs. FUHENAtdn20). Promt offers a web service (in beta status) different from previous years, which may be a reason why topics could not be translated well enough.

H4 *Applying the weighting from QA (for all experiments), merging results from IR and QA, and combining indexes for location names and content words will result in a higher initial MAP.* In comparison with results from the Berkeley group, the initial MAP was considerably higher: GIRSA returned 69% MAP at 0% recall for monolingual German experiments (experiment FUHtd01m), other participants achieved 43% and 16%, respectively (cf. the GeoCLEF overview paper in this volume); GIRSA achieved 63% MAP at 0% recall for bilingual experiments (experiment FUHPTGtd01), other participants achieved 47% and 16%, respectively.

To test GIRSA, experiments with the same parameter settings were conducted for the GeoCLEF 2007 topics before the 2008 campaign. The test experiments for topics from 2007 showed different results, e.g. the hypothesis H1 is true for the GeoCLEF 2007 topics, but not for the GeoCLEF 2008 topics (see also results for official experiments described in [4]). Future work will include a more thorough, per-topic analysis of errors.

References

[1] Sven Hartrumpf. Semantic decomposition for question answering. In Malik Ghallab, Constantine D. Spyropoulos, Nikos Fakotakis, and Nikos Avouris, editors, *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI)*, pages 313–317, Patras, Greece, July 2008.

[2] Sven Hartrumpf, Ingo Glöckner, and Johannes Leveling. University of Hagen at QA@CLEF 2008: Efficient question answering with question decomposition and multiple answer streams. In *Results of the CLEF 2008 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, September 2008.

[3] Johannes Leveling. On the role of information retrieval in the question answering system IRSAW. In *Proceedings of the LWA 2006 (Learning, Knowledge, and Adaptability), Workshop Information Retrieval*, pages 119–125. Universität Hildesheim, Hildesheim, Germany, 2006.

[4] Johannes Leveling and Sven Hartrumpf. Inferring location names for geographic information retrieval. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science (LNCS)*, pages 773–780, Berlin, 2008. Springer.

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