# Integrating Methods from IR and QA for Geographic Information Retrieval

Johannes Leveling[1] and Sven Hartrumpf[2]

[1] Centre for Next Generation Localisation (CNGL),
Dublin City University, Dublin 9, Ireland
johannes.leveling@computing.dcu.ie
[2] Intelligent Information and Communication Systems (IICS),
University of Hagen, 58084 Hagen, Germany
sven.hartrumpf@fernuni-hagen.de

**Abstract.** This paper describes the participation of GIRSA at GeoCLEF 2008, the geographic information retrieval task at CLEF. GIRSA combines 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 separate indexes for content and geographic annotation, using complex term weighting, adding location names from the topic narrative, and merging results from IR and QA, which yields the highest mean average precision (0.2608 MAP).
For bilingual experiments, English and Portuguese topics were translated via the web services Applied Language Solutions, Google Translate, and Promt Online Translator. For both source languages, Google Translate seems to return the best translations. For English (Portuguese) topics, 60.2% (80.0%) of the maximum MAP for monolingual German experiments, or 0.1571 MAP (0.2085 MAP), is achieved.
As a post-official experiment, translations of English topics were analysed with a parser. The results were employed to select the best translation for topic titles and descriptions. The corresponding retrieval experiment achieved 69.7% of the MAP of the best monolingual experiment.

## 1 Introduction

GeoCLEF is the geographic information retrieval (GIR) task at CLEF. In recent years, we have developed GIRSA (GIR by Semantic Annotation), a system for exploring novel approaches at GIR. GIRSA supports methods to improve precision (e.g. annotation of metonymic uses of location names [1]) and methods to improve recall (e.g. normalisation of location names [2] and decompounding). For GeoCLEF 2008, the major improvement lies in the combination of results from information retrieval (IR) and question answering (QA).

## 2 System Description and Experimental Setup

GIRSA is a system for the evaluation of novel indexing and retrieval methods for GIR. Basically, the GIRSA setup introduced at GeoCLEF 2007 was used

**Table 1.** Selected results for 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 |
| FUHENVtd20 | EN | V | TD | B | N | 0.1817 | 808 | 0.32 | 0.32 | 0.29 |
| FUHENVtd20m | EN | V | TD | B | Y | 0.1857 | 877 | 0.33 | 0.31 | 0.29 |

for GeoCLEF 2008, too. This setup involves the identification and normalisation 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 details on the system's improvements, see [3].

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): the *topic source language* (lang.): German (DE) or English (EN) serves as topic source language; the *translation service* (transl.): Applied Language Solutions (A, http://www.appliedlanguage.com/free_translation.shtml), Google Translate (G, http://translate.google.com/), or Promt Online Translator (O, http://www.online-translator.com/), and – in post-official experiments – a combination of translations (V); the *content fields*: content words and location indicators are extracted from the topic title and description: with location names from the topic narrative (TDN) or without (TD); the *index type*: all words are stemmed and a single index is produced (A), content words are decompounded and stemmed, location names are identified, both are indexed separately (B); the *combination* (comb.): results from IR and QA are combined (Y) or not (N). Results are merged and the top 1000 documents are returned. Five metrics are employed to measure retrieval performance: mean average precision (MAP), the number of relevant and retrieved documents (rel_ret), and precision at $N$ documents (P@$N$).

## 3 Results and Discussion

The following four hypotheses were formulated before the experiments and investigated after the experiments as follows.

*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).

*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 (query decomposition, less strict matching) in InSicht, the QA subsystem. 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.

*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 60.2% of the best MAP for monolingual German); the MAP for the best bilingual Portuguese-German experiment is 0.2085 (about 80.0% compared to monolingual German). The experiments with topics translated by Google Translate returned the best results. 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.

*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, the initial MAP was quite high: GIRSA returned 69% MAP at 0% recall for monolingual German experiments (experiment FUHtd01m), other participants achieved 43% and 16%, respectively (see [4]).

A result analysis for the QA subsystem InSicht showed that query decomposition was vital: With decomposition, 1238 documents (232 assessed as relevant) were retrieved; only 125 documents (77 assessed as relevant) without decomposition. InSicht's orientation towards precision was confirmed because if documents were retrieved for a topic, also relevant documents were retrieved.

To find the causes of low performance for the bilingual experiments, we analysed the topic titles and descriptions translated by the MT web services into German. The topics show many types of errors: ending with a wrong translation (using a different word sense, e.g. *schießen*/*'shoot'* instead of *Feuer*/*'fire'* in topic GC88), using uncommon translations, using a wrong preposition, generating a translation with incongruence between words, using a wrong verb position, and untranslated words. Except for getting wrong prepositions, these errors do not seem to ultimately have much impact on the performance of a GIR system. Prepositions will become important in a GIR system which is capable of interpreting the prepositions as geographic semantic relations.

Analysing the translations per web service used, the following errors were observed for translations from English: Applied Language Solutions returns untranslated words or completely untranslated title and description fields for 4 topics. The translations also include uncommon words in 3 topics and wrong translations in 3 topics. Google Translate returns two untranslated words only, *'resons'* in GC99 and *'Douments'* in GC100, both spelling errors. The Promt Online Translator returns uncommon translations for 3 topics and wrong translations for 6 topics. The Promt translator added translation alternatives in brackets, which might have caused a topic shift if translations with different senses were added. The performance of these machine translation web services is reflected in the performance results for bilingual experiments: translations with Google Translate show the least number of errors and the corresponding experiments return the best performance.

As the three translators presented quite diverse translation mistakes, a virtual translator was implemented (after the official experiments) that picks one of the translations for a given sentence $t$ using the scoring function $q$:

$$q(t) := w_1 \cdot \text{parse\_quality}(t) - w_2 \cdot |\text{unknown\_words}(t)| \text{ with } w_1 = 1.0 \text{ and } w_2 = 0.1$$

The parse quality is a real number between 0 and 1 obtained from analysing the topics with InSicht's syntactico-semantic parser. The virtual English translator returned an acceptable translation for 92% of the topic titles and for 76% of the topic descriptions. Selection with the virtual translator gives much better results than using translations from the best single translator and allows better retrieval results: e.g., InSicht lost only 11 relevant documents compared to the monolingual run. GIRSA achieved 0.1817 MAP and 0.1857 MAP and a much higher initial precision when using the virtual translator (see FUHENVtd20 and FUHENVtd20m in Table 1).

Future work will continue in the field of integrating methods from information retrieval and question answering for geographic information retrieval, evaluating GIRSA in the GikiCLEF task planned for CLEF 2009.

## References

1. Leveling, J., Hartrumpf, S.: On metonymy recognition for geographic information retrieval. IJGIS **22**(3) (2008) 289–299
2. Leveling, J., Hartrumpf, S.: Inferring location names for geographic information retrieval. In Peters, C., et al., eds.: Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007. Volume 5152 of LNCS., Berlin, Springer (2008) 773–780
3. Leveling, J., Hartrumpf, S.: University of Hagen at GeoCLEF 2008: Combining IR and QA for geographic information retrieval. In: Results of the CLEF 2008 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark (2008)
4. Mandl, T., Carvalho, P., Gey, F., Larson, R., Santos, D., Womser-Hacker, C.: GeoCLEF 2008: the CLEF 2008 cross-language geographic information retrieval track overview. This volume