University of Hagen at GeoCLEF 2006: Experiments with metonymy recognition in documents

Johannes Leveling and Dirk Veiel
FernUniversität in Hagen (University of Hagen)
Intelligent Information and Communication Systems (IICS)
58084 Hagen, Germany
{johannes.leveling, dirk.veiel}@fernuni-hagen.de

Abstract

This paper describes the participation of the IICS group at the GeoCLEF task of the CLEF campaign 2006. We describe different retrieval experiments using a separate index for location names and identifying and indexing of metonymic location names differently. The setup of our GIR system is a modified variant of the setup for GeoCLEF 2005.

We apply a classifier for the identification of metonymic location names for preprocessing the documents. This classifier is based on shallow features only and was trained on manually annotated data from the German CoNLL-2003 Shared Task corpus for Language-Independent Named Entity Recognition and from a subset of the GeoCLEF newspaper corpus. After preprocessing, documents contain additional information for location names that are to be indexed separately, i.e. `LOC` (all location names identified), `LOCLIT` (location names in their literal sense), and `LOCMET` (location names in their metonymic sense).

To obtain an IR query from the topic title, description, and narrative, we employ two methods. In the first method, a semantic parser analyzes the query text and the resulting semantic net is transformed into database query. The second method uses a Boolean combination of a bag-of-words (consisting of topical search terms) with location names.

The results of our experiments can be summarized as follows: excluding metonymic senses of location names improves mean average precision (MAP) for most of our experiments. For experiments in which this was not the case, a more detailed analysis showed that for some topics the precision increased. Our experiments show that the additional use of topic narratives decreases MAP. For almost all experiments with the topic narrative, lower values for MAP and for the number of relevant and retrieved documents were observed. However, query expansion and the use of separate indexes improves the performance of our GIR application.

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

Measurement, Performance, Experimentation

Keywords

Geographic Information Retrieval, Metonymy Recognition, Indexing Location Names

1 Introduction

There are several essential tasks a geographic information retrieval (GIR) application has to perform, for example the identification and disambiguation of location names (see [3] for a general overview over tasks in GIR applications). Disambiguating location names includes differentiating between their literal (geographic) and metonymic senses. Metonymy is typically defined as a figure of speech in which a speaker uses *"one entity to refer to another that is related to it"* [4].

This paper presents an application of a classifier for literal and metonymic senses of location names. The classifier is trained on manually annotated data and uses shallow features from the textual context of location names only. We create different indexes corresponding to different senses of location names and investigate in a baseline experiment if utilizing a separate index containing location names will improve performance in GIR. The focus in our experiments lies on metonym identification in documents, because the GeoCLEF query topics did not contain any metonymic location names in topic titles, descriptions, or narrative.

2 System description

Our GIR system is based on the same system that was developed for GeoCLEF 2005 (see [8]). The WOCADI parser (Word Class Controlled Disambiguating (Parser), see [2]) analyzes the query topics and the GeoCLEF corpus of newspaper and newswire articles. From its parse results, concepts (or rather: lemmata) and compound constituents are extracted as index terms or search terms.

For the identification of metonymic location names, we employed a classifier trained on manually annotated data. The data consists of a subset of the German CoNLL-2003 Shared Task corpus for Language-Independent Named Entity Recognition [9] and a subset of the GeoCLEF newspaper corpus. The metonymy classifier [7] is based on shallow features only (e.g., part-of-speech information for closed word classes from list lookup, position of words in a sentence, word length, and base forms of verbs) and achieved a performance of 81.7% F_1-measure in differentiating between literal and metonymic senses of location names. In analyzing the annotated CoNLL data (1216 instances), we found that 16.95% of all location names were used metonymically, and 7.73% referenced both a literal and a metonymic sense at the same time (see [7] for a more detailed description of the metonymy classification). These numbers provide an upper bound for a performance increase for methods exploiting metonymy information. After preprocessing, the documents are structured with the following fields:

- `DOCID` – document ID
- `TEXT` – text of the document
- `LOC` – location names from the text
- `LOCLIT` – location names in their literal sense
- `LOCMET` – location names in their metonymic sense

All identified location names are indexed from the `LOC` field of a document. The result of the metonymy classifier determines how a given location name will be indexed, i.e. literal and metonymic senses of location names are indexed from the `LOCLIT` and `LOCMET` fields, respectively. Figure 1 shows an example document after preprocessing its text. The representations of 276,581 documents (after duplicate elimination) were indexed with the Zebra database management system [1], which supports a standard relevance ranking (*tf-idf* IR model).

Two methods were employed to obtain an IR query from a topic title, description, and narrative: In the first method, the WOCADI parser is applied to perform a deep linguistic analysis of the query text and the resulting semantic net is transformed into a database independent query representation (DIQR, see [6]). In the second method, a bag-of-words (topical search terms extracted from the query text) is combined with a subquery containing location names (identified by a name lookup) to obtain the DIQR. In both cases, a DIQR query consists of a Boolean combination of a subquery containing of topical search terms (or descriptors) and one containing location names.

<DOC>
<DOCID> ... </DOCID>
<TEXT>At the meeting of France and Germany
in Lisbon last year, Paris vetoed the decision.
... </TEXT>
<LOCALL>France Germany Lisbon Paris</LOCALL>
<LOCLIT>Lisbon</LOCLIT>
<LOCMET>France Germany Paris</LOCMET>
</DOC>

Figure 1: An example of a preprocessed GeoCLEF document. Note: In this sample document, the TEXT field contains the unprocessed text, before stemming and stopword removal.

In our baseline experiment (FUHddGNNNTD), the standard IR model (*tf-idf*) was utilized without additions or modifications. In experiments not accessing the separate name index, the location names were searched within the index for the TEXT field; in experiments using the separate index (e.g. FUHddGYYYTDN) location names were looked for in the index for the LOC field. For experiments with metonymy identification, the index for location names corresponds to the field LOCLIT, i.e. only location names in their literal sense were searched for.

3 Description of the runs submitted and results

The parameters used in our GeoCLEF experiments can be described as follows:

- BOP: Boolean operator for the combination of topic and location subquery
- LI: use separate location index (LOC or LOCLIT, as described in Section 2) vs. use TEXT index
- LA: apply deep linguistic analysis (WOCADI parser [2]) and transformation of resulting semantic net into a database query [6] vs. bag-of-words IR
- QEX: query expansion with semantically related terms and meronyms for locations [8] vs. no expansion
- MET: exploit metonymy information (as described in [7]) vs. no metonymy processing
- QF: query topic fields (TD: title and description, TDN: title, description, and narrative). A run with topic title, description, and narrative was mandatory to find out if extra narrative terms help improve performance.

Table 1 and Table 2 show the different parameter settings and results for our monolingual German and bilingual English-German GeoCLEF runs, respectively. The results shown consist of mean average precision (MAP) and the number of relevant and retrieved documents (rel_ret).

4 Analysis of the results

Using a separate index for location names leads to a better performance in our experiments. Additional experiments confirmed that the query expansion and the provided background knowledge have significant influence on results. As a preparation for the GeoCLEF 2006 experiments, we analyzed data from GeoCLEF 2005, in particular (but not exclusively) those topics for which we did not find any of the relevant documents. We modified the knowledge base containing our background knowledge to include some missing facts[1] and performed manual searching. For example, for topic GC019 (*"European golf tournaments"*),

[1] For the GeoCLEF 2006 experiments, the unmodified knowledge base from GeoCLEF 2005 was used.

Table 1: Parameter settings and results for monolingual German GeoCLEF experiments. 785 documents were assessed as relevant for the 25 queries. (Additional runs are set in italics.)

Run Identifier	Parameters						Results	
	BOP	LI	LA	QEX	MET	QF	MAP	rel_ret
FUHddGNNNTD	OR	N	N	N	N	TD	0.1694	439
FUHddGYYYTD	OR	Y	Y	Y	N	TD	**0.2229**	449
FUHddGYYNTD	OR	Y	Y	N	N	TD	0.1865	456
FUHddGNNNTDN	OR	N	N	N	N	TDN	0.1223	426
FUHddGYYYTDN	OR	Y	Y	Y	N	TDN	0.2141	**462**
FUHddGYYYMTDN	OR	Y	Y	Y	Y	TDN	0.1999	442
FUHddGYYNTD	AND	Y	Y	N	N	TD	0.1466	232
FUHddGYYNMTD	AND	Y	Y	N	Y	TD	0.1608	225
FUHddGYYYTD	AND	Y	Y	Y	N	TD	0.1718	267
FUHddGYYYMTD	AND	Y	Y	Y	Y	TD	0.1953	259

Table 2: Parameter settings for bilingual English-German GeoCLEF experiments. 785 documents were assessed as relevant for the 25 queries.

Run Identifier	Parameters						Results	
	BOP	LI	LA	QEX	MET	QF	MAP	rel_ret
FUHedGNNNTD	OR	N	N	N	N	TD	0.1211	**397**
FUHedGNNNTDN	OR	N	N	N	N	TDN	0.0548	333
FUHedGYYYTD	OR	Y	Y	Y	N	TD	**0.1280**	383
FUHedGYYNTD	OR	Y	Y	N	N	TD	0.1124	386
FUHedGYYYTDN	OR	Y	Y	Y	N	TDN	0.1234	375
FUHedGYYYMTDN	OR	Y	Y	Y	Y	TDN	0.1148	375

we found that with only slight differences in query formulation the MAP as well as the number of relevant documents increased. For topic GC019, 41 of 61 relevant documents were found after modifying the query. As there is now more data (topics and relevance assessments) available, we hope to identify regularities for search failures more easily.

As we observed from results of the GeoCLEF 2005 experiments, query expansion with meronyms leads to significantly better precision (0.1466 MAP vs. 0.1608 MAP; 0.1718 MAP vs. 0.1953 MAP) for most monolingual German runs, although recall is slightly worse. A detailed analysis of the runs FUHddGYYYTDN an FUHddGYYYMTDN shows that in this case, the metonymy identification task added this year improves the results for some topics.

Results for the bilingual experiments were found to be lower in general. There were a few errors in the translated topic titles, descriptions, and narratives. For some topic titles, there does not seem to be enough textual context to provide an adequate translation (topic titles and descriptions were translated separately).

One hypothesis to be tested was that the additional information in topic narratives (runs with QEX=TDN instead of TD) would improve results. We can not confirm this assumption because with our setup, MAP and relevant and retrieved documents are almost always lower for runs using the topic narrative than for runs with topic title and description.

5 Conclusion

We expected to find a significant increase in precision for all GIR experiments excluding metonymic senses of location names for a search. This assumption holds for most experiments, conforming results of earlier experiments [7]. However, the MAP for experiments using metonymy information is in one case lower. A

more detailed analysis of results for this experiment showed that at least for some topics, precision is in fact increased by metonymy identification.

A different explanation for a low performance might be that in our setup for GeoCLEF 2006, the location name index does not contain terms representing adjectives, language, or inhabitants for a given location name (e.g. the terms like *Dutch*, *Spanish*, or *Spaniard* do not occur in the location name index). Furthermore, instead of removing all metonymic senses for location names from the index, the corresponding terms should be indexed differently (with a lesser weight or with a different sense).

Additional information from the topic narratives did not improve precision or recall in our experiments, although one might have expected a similar effect as for query expansion with meronyms. We plan to rerun experiments with a state-of-the-art database management system (DBMS), such as Cheshire 3 [5], which offers a variety of different IR models. A more modern IR model will help to increase performance in general.

References

[1] Sebastian Hammer, Adam Dickmeiss, Heikki Levanto, and Mike Taylor. Zebra – user's guide and reference. Manual, IndexData, Copenhagen, Denmark, 2005.

[2] Sven Hartrumpf. *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag, Osnabrück, Germany, 2003.

[3] Christopher B. Jones, R. Purves, Anne Ruas, Mark Sanderson, Monika Sester, Marc J. van Kreveld, and Robert Weibel. Spatial information retrieval and geographical ontologies – an overview of the SPIRIT project. In *SIGIR 2002*, pages 387–388, 2002.

[4] George Lakoff and Mark Johnson. *Metaphors we live by*. Chicago University Press, 1980.

[5] Ray R. Larson and Robert Sanderson. Grid-based digital libraries: Cheshire 3 and distributed retrieval. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 112–113, New York, NY, USA, 2005. ACM Press.

[6] Johannes Leveling. *Formale Interpretation von Nutzeranfragen für natürlichsprachliche Interfaces zu Informationsangeboten im Internet*. Dissertation, Fachbereich Informatik, FernUniversität in Hagen, 2006. To appear in: Der andere Verlag, Tönning, Germany, 2006.

[7] Johannes Leveling and Sven Hartrumpf. On metonymy recognition for GIR. In *Proceedings of the 3rd ACM workshop on geographic information retrieval*, 2006.

[8] Johannes Leveling, Sven Hartrumpf, and Dirk Veiel. Using semantic networks for geographic information retrieval. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *CLEF 2005 Proceedings*, Lecture Notes in Computer Science (LNCS). Springer, Berlin, 2006.

[9] E. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language independent named entity recognition. In W. Daelemans and M. Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147, Edmonton, Canada, 2003.