The Domain-Specific Task of CLEF - Specific Evaluation Strategies in Cross-Language Information Retrieval

Michael Kluck[1] and Fredric C. Gey[2]

[1] InformationsZentrum Sozialwissenschaften Bonn
[2] UC Data Archive & Technical Assistance
University of California, Berkeley, CA 94720 USA
e-mail: kluck@bonn.iz-soz.de, gey@ucdata.berkeley.edu

Abstract. This paper describes the domain-specific cross-language information retrieval (CLIR) task of CLEF, why and how it is important and how it differs from general cross-language retrieval problem associated with the general CLEF collections. The inclusion of a domain-specific document collection and topics has both advantages and disadvantages

1 Introduction

For the past decade, the trend in information retrieval test-collection development and evaluation has been toward general, domain-independent text such as newswire information. This trend has been fostered by the needs of intelligence agencies and the non-specific nature of the World Wide Web and its indexing challenges. The documents in these collections (and in the general CLEF collections) contain information of non-specific nature and therefore could potentially be judged by anyone with good general knowledge.

Critics of this strategy believe that the tests are not sufficient to solve the problems of more domain-oriented data collections and topics. Particularly for cross-language information retrieval, we may have a vocabulary disconnect problem since the vocabulary for a specific area may not exist in a Machine Translation (MT) system used to translate queries or documents. Indeed, the vocabulary may have been redefined in a specific domain to mean something quite different from its general meaning. The rationale of the inclusion of domain specific collections into the tests is to test retrieval systems on another type of document collection, serving a different kind of information need. The information provided by these domain specific documents is far more targeted than news stories. Moreover, the documents contain quite specific terminology related to the respective domain. The hypothesis to be tested is whether domain-specific enhancements to information retrieval provide (statistically significant) improvement in performance over general information retrieval approaches.

Information retrieval has a rich history of test collections, beginning with Cranfield, which arose out of the desire to improve and enhance search of scientific and technical literature. The GIRT collections (defined below) of the

TREC-8 evaluation and of this first CLEF campaign provide an opportunity for IR to return to its roots and to illuminate those particular research problems and specific approaches associated with domain-specific retrieval. Other recent examples of domain specific collections are the OHSUMED collection[10] for the medical domain and the NTCIR collection[12] for science and engineering. The OHSUMED collection has been explored for its potential in query expansion[13, 11] and was utilized in the filtering track of the TREC-9 conference (see http://trec.nist.gov). The NTCIR collection is the first major test collection in Japanese and the NTCIR evaluations have provided the first large-scale test of Japanese-English crosss language information retrieval.

2 Advantages and disadvantages of domain specific CLIR

A domain-specific language requires appropriate indexing and retrieval systems. Recent results clearly show this difficulty of differentiating between domain-specific (in this case: sociological) terms and common language terms: "words [used in sociology] are common words that are [also] in general use, such as community or immigrant"[9]. In many cases there exists a clear difference between the scientific meaning and the common meaning. Furthermore, there are often considerable difference between scientific terms when used in different domains, owing to different connotations, theories, political implications, ethical convictions, and so on. This means that it can be more difficult to use automatically generated terms and queries for retrieval. For example, Ballesteros and Croft [1] have noted, for a dictionary-based cross-language query system: "queries containing domain-specific terminology which is not found in general dictionaries were shown to suffer an additional loss in performance". In some discipline (for instance in biology) different terminologies have evolved in quite narrow subfields as Chen at al.[3] have shown for the research dealing with the species of worms and flies and their diverging terminology.

For several domains Haas [9] has carried out in-depth-research and stated: "T tests between discipline pairs showed that physics, electrical engineering, and biology had significantly more domain terms in sequences than history, psychology, and sociology (...) the domains with more term sequences are those which may be considered the hard sciences, while those with more isolated domain terms tend to be the social sciences and humanities."

Nevertheless, domain specific test collections offer new possibilities for the testing of retrieval systems as they allow the domain specific adjustment of the system design and the test of general solutions for specific areas of usage. Developers of domain specific CLIR systems need to be able to tune their systems to meet the specific needs of a more targeted user group.

The users of domain specific collections are typically interested in the completeness of coverage. They may not be satisfied with finding just some relevant documents from a collection. For these users the situation of too much overlap between the relevant documents within the result sets of the different evaluated systems is much more important and has to be solved.

3 Domain specific evaluation procedures

Domain-specificity has consequences not only for the data but also for the topic creation and assessment processes. Separate specific topics have to be created because the data are very different from that found in newspapers or newswires. The GIRT documents treat more long-term societal or scientific problems in an in-depth manner; current problems or popular events (as they are represented in news articles) are dealt with after some time lag. Nevertheless, the TREC/CLEF domain-specific task attempted to cover German newswire and newspaper articles as well as the GIRT collection. Thus topics were developed which combined both general and domain specific characteristics. It proved to be challenging to discover topics which would retrieve news stories as well as scientific articles.

The topic developers must be familiar with the specific domain as well as the respective language in which the topic has been created or into which the topic is to be translated. The same is true for the assessors – they must have domain related qualifications and sufficient language skills to develop the relevance judgements.

Therefore each domain specific sub-task needs its own group of topic developers and relevance assessors in all languages used for the sub-task. Finally the systems being tested must be able to adjust general principles for retrieval systems to the domain-specific area.

4 The GIRT domain-specific social science test collection

The TREC-7, TREC-8 and CLEF 2000 evaluations have offered a domain specific subtask and collection for CLIR in addition to the generally used collections. The test collection for this domain specific subtask is called GIRT (German Information Retrieval Test database) and comes from the social sciences. It has been used in several German tests of retrieval systems [6, 14, 2] The GIRT collection was made available for research purposes by the InformationsZentrum Sozialwissenschaften (IZ; = German Social Sciences Information Centre), Bonn. For pre-test research by the IZ and the University of Konstanz a first version, the GIRT1 collection contained about 13,000 documents. For the TREC7 and TREC8 evaluations, the GIRT2 collection was offered which included GIRT1 supplemented with additional documents and contained about 38,000 documents. In the CLEF2000 campaign the GIRT3 collection was used which included the GIRT2 data and additional sampled documents for a total of about 76,000 documents. Figure 1 presents a sample document from the GIRT3 collection.

The GIRT data have been collected from two German databases offered commercially by the IZ via traditional information providers (STN International, GBI, DIMDI) and on CD-ROM (WISO III): FORIS (descriptions of social sciences current research projects in the German speaking countries), and SOLIS (references of social sciences literature originated in German speaking countries, containing journal articles, monographs, articles in collections, scientific reports,

<DOCNO>19940100925</DOCNO>
<TITLE>Psychisch kranke Mitarbeiter in Betrieben : die Sichtweise der betrieblichen Helfer</TITLE>
<TITLE-ENG>Mentally ill employees in companies : the viewpoint of company assistants</TITLE-ENG>
<AUTHOR>Schubert, Andreas</AUTHOR>
<PUBLICATION-YEAR>1988</PUBLICATION-YEAR>
<LANGUAGE>DE</LANGUAGE>
<CONTROLLED-TERM>psychische Krankheit,Mitarbeiter,Betrieb,Helfer,soziales Netzwerk,Bezugsperson,Integration</CONTROLLED-TERM>
<CLASSIFICATION>Industriesoziologie, Betriebssoziologie, Arbeitssoziologie, industrielle Beziehungen,soziale Probleme,Sozialpolitik</CLASSIFICATION>
<TEXT>"Ausgehend von der äußerst problematischen Situation psychisch kranker und behinderter Menschen auf dem allgemeinen Arbeitsmarkt wird die besondere Bedeutung innerbetrieblicher Hilfen dargestellt. Dazu wird modellhaft die Situation eines Mitarbeiters mit 'seelischen Problemen' in einem Betrieb skizziert, um somit die potentiellen Bezugspersonen und damit ein mögliches innerbetriebliches soziales Netzwerk zu kennzeichnen. Die Fragestellung der dargestellten Untersuchung ist, inwieweit die per Gesetz zur Unterstützung Behinderter und damit auch psychisch behinderter Mitarbeiter verpflichteten 'betrieblicher Helfer', diese Funktion tatsächlich wahrehmen, d.h. inwieweit das Hilfspotential dieser Gruppe sich umsetzt in ein für den Betroffenen erfahrbares innerbetriebliches soziales Netzwerk. Dazu werden die Ergebnisse einer schriftlichen Befragung von 144 betrieblichen Helfern referiert. Als Fazit der Untersuchung muß von einem relativ geringen Kenntnisstand betrieblicher Helfer bzgl. der Auswirkungen psychischer Krankheit ausgegangen werden, von negativen Einschätzungen der Leistungs- und Integrationsmöglichkeiten psychisch behinderter Mitarbeiter und von einer starken Tendenz dieser Gruppe, die Problematik und damit die Betroffenen auszugrenzen oder, bei betriebsinternen Vorfällen, an betriebliche Entscheidungsträger wie direkte Vorgesetzte, Personal- und Betriebsleitung 'abzuschieben'. Da häufig weder interne noch externe Fachleute hinzugezogen werden, ist der Aufbau eines innerbetrieblichen Netzwerkes als sehr schwierig einzuschätzen. Positive Beispiele belegen allerdings die Integrationsmöglichkeiten für psychisch Behinderte auch in 'normalen' Betrieben." (Autorenreferat)</TEXT>
<TEXT-ENG>"Because of the extremly problematical situation of psychologically disturbed people so far as the job market is concerned this paper stresses the importance of help inside the concerns. In order to show potential sources of help and thus a possible supportive network inside a firm a model case of a worker with 'psychological problems' is sketched. This investigation was aimed at discovering how far the legal obligation to assist handicapped people inside industrial concerns, and thus also psychologically handicapped workers, is actually fulfilled by the 'industrial helpers', i.e. how far the potential help offered by these

Fig. 1. GIRT Sample document(English text truncated)

dissertations). The FORIS database contains about 35,000 documents on current and finished research projects of the last ten years. As projects are living objects the documents are often changed; thus, about 6,000 documents are changed or newly entered each year. SOLIS contains more than 250,000 documents with a yearly addition of about 10,000 documents.

The GIRT3 data contain selected bibliographical information (author, language of the document, publication year), as well as additional information elements describing the content of the documents: controlled indexing terms, free terms, classification texts, and abstracts (TEXT) - all in German (GIRT1 and GIRT2 data contained some other fields). Besides the German information there are English translations of the titles (for 71% of the documents) available. For some documents (about 8%) there are also English translations of the abstracts (TEXT-ENG). One exception is the TITLE field where the original title of the document is stored: in some cases the original title has already been English, thus, no English translation has been necessary and the field TITLE-ENG is missing, although the title is in fact English. The information elements of the GIRT collection are quite similar to those of the OHSUMED collection which has been developed by William Hersh [10] for the medical domain, but that test collection is bigger (348,566 documents). The OHSUMED fields are: title, abstract, controlled indexing term (MeSH), author, source, publication type.

Most of the GIRT3 documents have German abstracts (96% of the documents), some have English abstracts (8%). For the 76,128 documents 755,333 controlled terms have been assigned, meaning, on average, each document has nearly 10 indexing terms. Some documents (nearly 9%) have free terms assigned which are only given by the indexing staff of the IZ to make proposals for new terms to be included in the thesaurus. The documents have on average two classifications assigned to each of them. The indexing rules allow assignment of one main classification, as well as one or more additional classifications if other (sub-)areas are treated in the document. The average number of authors for each document is nearly two. The average document size of the GIRT documents is about 2 KB.

Field label	# Occurrences of field	percent in GIRT3 docs	Avg. # of entries per doc
DOC	76,128	100.00	1.00
DOCNO	76,128	100.00	1.00
LANGUAGE	76,128	100.00	1.00
PUBLICATION YEAR	76,128	100.00	1.00
TITLE	76,128	100.00	1.00
TITLE-ENG	54,275	71.29	-
TEXT	73,291	96.27	-
TEXT-ENG	6,063	7.96	-
CONTROLLED-TERM	755,333	-	9.92
FREE-TERM	6,588	-	0.09
CLASSIFICATION	169,064	-	2.22
AUTHOR	126,322	-	1.66

Table 1. Statistics of the GIRT3 data collection

The GIRT multilingual thesaurus (German-English), based on the Thesaurus for the Social Sciences [4] provides the vocabulary source for the indexing terms within CLEF (see Figure 2). A Russian translation of the German thesaurus is also available. The German-English thesaurus has about 10,800 entries, of which 7,150 are descriptors and 3,650 non-descriptors. For each German descriptor there is an English or Russian equivalent. The German non-descriptors have been translated into English in nearly every case, but this is not true for the Russian word list. There are smaller differences to the trilingual German-English-Russian word list, because it was completed earlier (1996) than the latest version of the Thesaurus (1999). Thus, English or Russian indexing terms could be used for retrieval purposes by matching to the equivalent German terms from the respective version of the thesaurus.

The first GIRT collection (GIRT1), which was utilized for the pre-tests, contained a subset of the databases FORIS and SOLIS with about 13,000 documents

```
<entry>
  <german>Absatzpolitik</german>
  <related-concept>ABSATZPOLITIK</related-concept>
  <broader-term>Unternehmenspolitik</broader-term>
  <narrower-term>Werbung</narrower-term>
  <narrower-term>Produktgestaltung</narrower-term>
  <narrower-term>Preispolitik</narrower-term>
  <english>sales policy</english>
</entry>
```

Fig. 2. GIRT Thesaurus Entry

which were restricted to the publication years 1987-1996 and to the topical areas of "sociology of work", "women studies" and "migration and ethnical minorities" (with some additional articles without topical restrictions from two German top journals on sociology being published in this time-span). This topical restriction was obtained by choosing the appropriate classification codes as search criteria. The GIRT2 collection - offered in TREC7 and TREC8 - contained a subset of the databases FORIS and SOLIS, which included the GIRT1 data, followed the same topical restrictions, but was enlarged to the publication years 1978-1996. This led to a specific topicality of the data, which had to be considered during the topic development process and restricted the possibilities of selecting topics. The distribution of descriptors and even of the words within the documents was also affected by these topical restrictions. The GIRT3 collection - offered in the CLEF2000 campaign - has been broadened to all documents in this time-span regardless of their topics. Thus, this collection is an unbiased representative sample of documents in German social sciences between 1978 and 1996.

```
- <top>
    <num>girt002</num>
    <E-title>Kids and Computer Games</E-title>
    <E-desc>How are computer games used by children?</E-desc>
    <E-narr>Find information on how children use computer games and on the consequences of such
      use.</E-narr>
  </top>
- <top>
    <num>girt002</num>
    <G-title>Kinder und Computerspiele</G-title>
    <G-desc>Was gibt es über die Nutzung von Computerspielen durch Kinder?</G-desc>
    <G-narr>Alle Informationen über die Benutzung und Auswirkung der Nutzung von
      Computerspielen durch Kinder sind von Interesse. Ebenso sind Untersuchungen über die
      Gründe von Gewalt sowie Programme und Maßnahmen gegen Gewalt relevant.</G-narr>
  </top>
```

Fig. 3. GIRT Topic 002 – Children and computer games

5 Experiences and opportunities in TREC/CLEF with domain specific CLIR

Although specific terminology and vocabularies must be changed for each new domain, this is more than compensated for by features which can be exploited in domain-specific cross-language information retrieval. Existing domain-related vocabularies or thesauri can be utilized to reduce ambiguity of search and increase precision of the results. For multilingual thesauri an additional benefit accrues from using them as translation tools because the related term pairs of languages are available. Use of the MESH multilingual thesaurus for CLIR was explored by Eichmann Ruiz and Srinivasan[5] for the OHSUMED collection.

Additional aids are given if there exist translated parts of the documents (often the case for scientific literature, where English titles are frequently available for documents in other languages). This can allow a direct search against the translated document parts. The same advantage arises within existing document structures where the use of the specific meaning of different information elements allows a targeted search (i.e. if an author field exists, it possible to distinguish between a person as subject of an article or as the author of it).

Thus far the GIRT collections have received limited attention by groups engaged in cross-language information retrieval. At TREC-8 there were two groups participating and at CLEF three groups participated and one of those submitted only a monolingual entry. The best monolingual entry was submitted by the Xerox European Research Centre, while the cross-language entries came from the Berkeley Group[7] and the Dortmund Group[8].

6 Conclusion

This paper has discussed the domain-specific retrieval task at CLEF. The GIRT collection, oriented toward the social science domain, offers new opportunities in exploring cross-language information retrieval for specialized domains. The specific enhancements available with the GIRT collection are:

- a collection indexed manually to a controlled vocabulary
- bi-lingual titles (German and English) for almost all documents
- a hierarchical thesaurus of the controlled vocabulary
- multilingual translations of the thesaurus (German, English, Russian)

The multilingual thesaurus can be utilized as a vocabulary source for query translation and as a starting point for query expansion to enhance cross-language retrieval. Because each document is manually assigned, on average, by ten controlled vocabulary terms, the collection also offers the opportunity for research into multi-class text categorization.

References

1. Lisa Ballesteros and W. Bruce Croft. Statistical methods for cross-language information retrieval. In Gregory Greffenstette, editor, *Cross Language Information Retrieval*, pages 21–40. Kluwer, 1998.
2. Gisbert Binder, Matthias Stahl, and Lothar Faulborn. Vergleichsuntersuchung MESSENGER - FULCRUM projektbericht available at http://www.bonn.iz-soz.de/publications/series/working-papers/ab18.pdf. In *IZ-Arbeitsbericht Nr. 18, Bonn*, 2000.
3. Hsinchun Chen, Joanne Martinez, Tobun Ng, and Bruce Schatz. A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the worm community system. In *Journal of the American Society for Information Science*, volume 48 (1), pages 17–31, 1997.
4. Hannelore Schott (ed.). *Thesaurus for the Social Sciences. [Vol. 1:] German-English. [Vol. 2:] English-German. [Edition] 1999.* InformationsZentrum Sozialwissenschaften Bonn, 2000.
5. David Eichmann, Miguel Ruiz, and Padmini Srinivasan. Cross-language information retrieval with the UMLS metathesaurus. In W B Croft A Moffat C J van Rijsbergen R Wilkinson and J Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia*, pages 72–80, August 1998.
6. Elisabeth Frisch and Michael Kluck. Pretest zum projekt german indexing and retrieval testdatabase (GIRT) unter anwendung der retrievalsysteme messenger und freewaissf. In *IZ-Arbeitsbericht Nr. 10, Bonn*, 1997.
7. Fredric Gey, Hailing Jiang, Vivien Petras, and Aitao Chen. Cross-language retrieval for the CLEF collections - comparing multiple methods of retrieval. In *this volume.*
8. Norbert Govert. Bilingual information retrieval with HyREX and internet translation services. In *this volume.*
9. Stephanie W. Haas. Disciplinary variation in automatic sublanguage term identification. In *Journal of American Society for Information Science*, volume 48, pages 67–79, 1997.
10. William Hersh, Chris Buckley, TJ Leone, and David Hickman. OHSUMED: An interactive retrieval eavaluation and new large test collection for research. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of SIGIR94, the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 192–201, 1994.
11. William Hersh, Susan Price, and Larry Donohoe. Assessing thesaurus-based query expansion using the umls thesaurus. In *Proceedings of the 2000 Annual AMIA Fall Symposium*, pages 344–348, 2000.
12. Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. Overview of IR tasks at the first NTCIR workshop. In Noriko Kando and Toshihiko Nozue, editors, *The First NTCIR Workshop on Japanese Text Retrieval and Term Recognition , Tokoyo Japan*, pages 11–22, September 1999.
13. Padmini Srinivasan. Query expansion and MEDLINE. In *Information Processing and Management*, volume 32(4), pages 431–443, 1996.
14. Christa Womser-Hacker(ed.) et al. *Projektkurs Informationsmanagement: Durchführung einer Evaluierungsstudie, Vergleich der Information-Retrieval-Systeme (IRS) DOMESTIC, LARS II, TextExtender. University of Konstanz.* 1998.